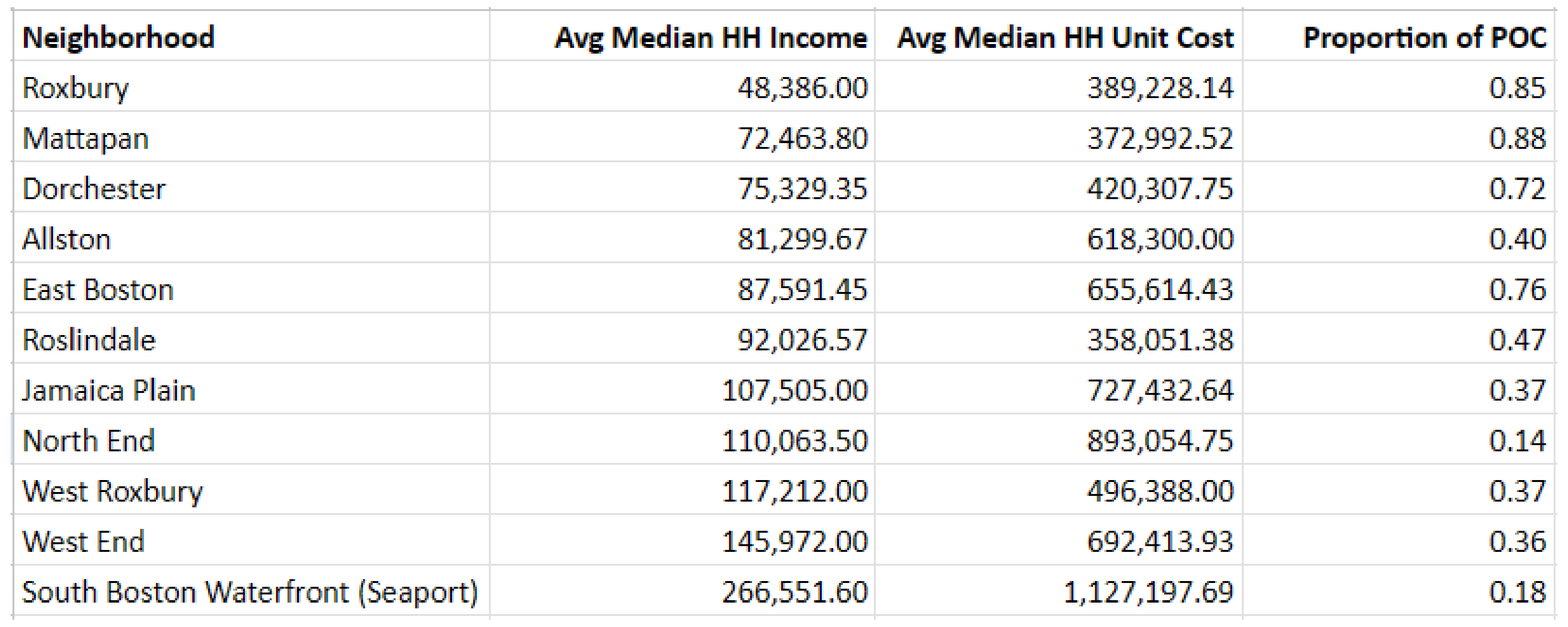
BOSTON: THE MOST SEGREGATED CITY IN AMERICA?

Across Boston neighborhoods, is race (the proportion of people of color) associated with housing cost (the average median household unit cost) adjusting for income (average median household income)?

HYPOTHESIS

We expected for the median household cost to increase if the average median household income in an area is higher. Additionally, we expected that the demographics of the area would be proportional to the financial properties of the neighborhoods we observe.



Neighborhood	Avg Median HH Income	Avg Median HH Unit Cost	Proportion of POC
Roxbury	48,386.00	389,228.14	0.85
Mattapan	72,463.80	372,992.52	0.88
Dorchester	75,329.35	420,307.75	0.72
Allston	81,299.67	618,300.00	0.40
East Boston	87,591.45	655,614.43	0.76
Roslindale	92,026.57	358,051.38	0.47
Jamaica Plain	107,505.00	727,432.64	0.37
North End	110,063.50	893,054.75	0.14
West Roxbury	117,212.00	496,388.00	0.37
West End	145,972.00	692,413.93	0.36
South Boston Waterfront (Seaport)	266,551.60	1,127,197.69	0.18

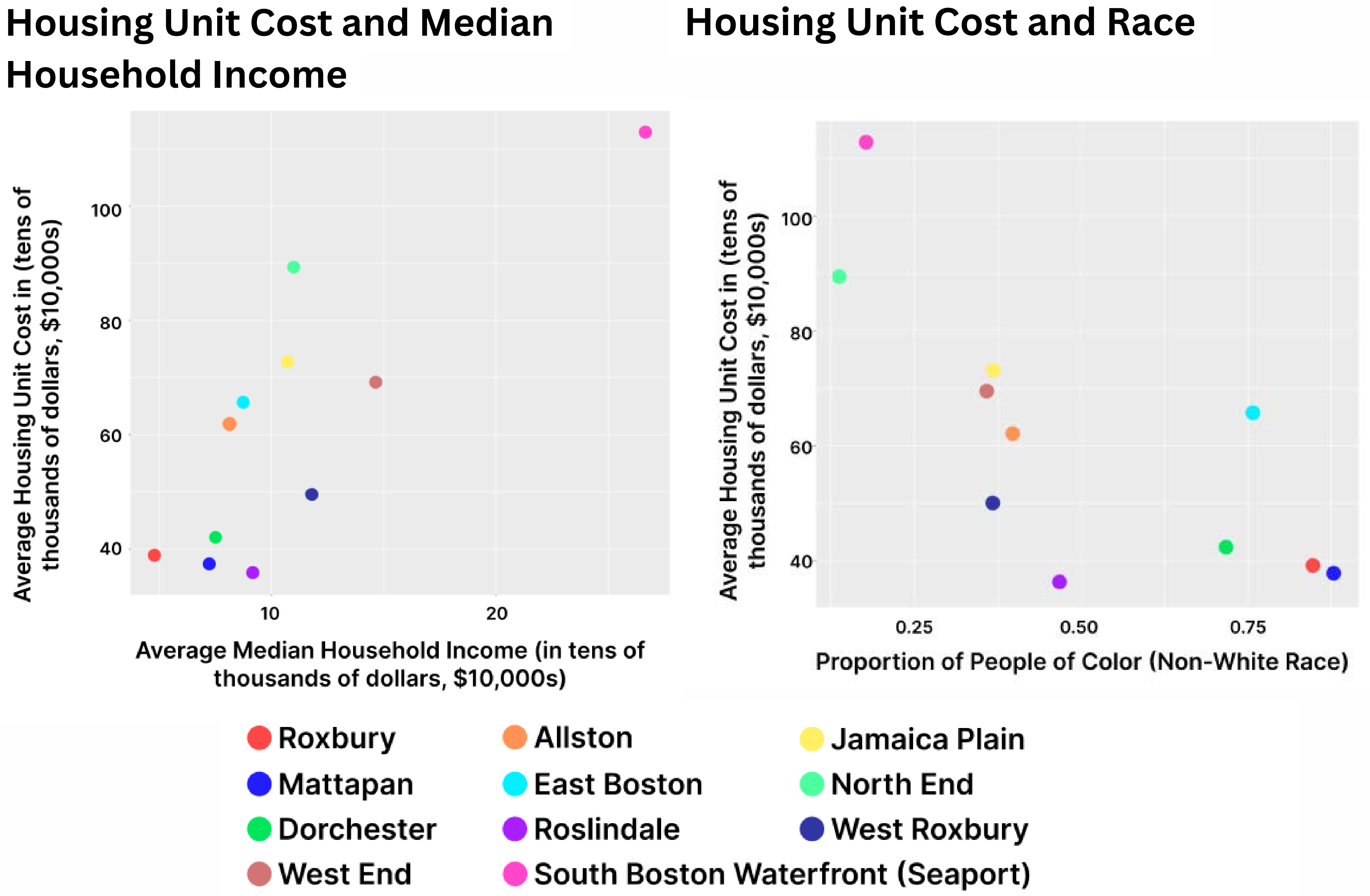
WHY IS THIS IMPORTANT?

Redlining as a practice has a dark history in Boston and many other cities in America, and we wanted to explore the potential effects of that on housing demographics and economic status today.

METHODOLOGY

We downloaded and imported census tract data from the US Census American Community Survey into Google Sheets. From there, we found the housing unit cost and number of housing units by census tract; we used this to calculate the weighted average housing unit cost and average median household income for each neighborhood from their census tracts.

To find the proportion of minorities in each neighborhood, we compiled US Census demographic data for each census tract in each neighborhood.



Housing Unit Cost and Median Household Income

Housing Unit Cost and Race

CHALLENGES/LIMITATIONS

Some of the neighborhoods we were researching had incomplete data, but we decided to keep these neighborhoods' data as excluding these neighborhoods would compromise the validity of our research. This was also divided race into Whites and people of color as one group (Asians, Blacks, Hispanics), rather than specific races. We did not take account of the size of each neighborhood for the average housing units.

CONCLUSION

Based on the data from the graph, there is a direct correlation between average median household income and average housing unit cost in the key neighborhoods of Boston. This highlights the significant impact of socioeconomic factors like race and income on housing dynamics in urban areas.

Boston Latin Academy

Betty Nguyen, Cynthia Jonah, Florie Donna, Phaedra Sanon, Jenna Zick

WHAT ACTIONS CAN STUDENTS TAKE TO RAISE THEIR GPA AT BROOKE HIGH SCHOOL?

BROOKE HIGH SCHOOL'S DATA SCIENCE CLUB
Jenelle Joeseph-Allen, Rodriguez Guerrier, Jerchel Porter, Xai Castilla
Advisors: Mr. Hempleman and Mr. Fleming

INTRODUCTION

Research Question and Hypothesis

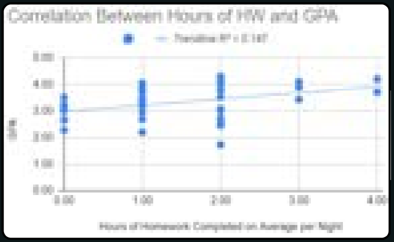
What replicable aspects of a student's daily life have the greatest impact on their GPA?

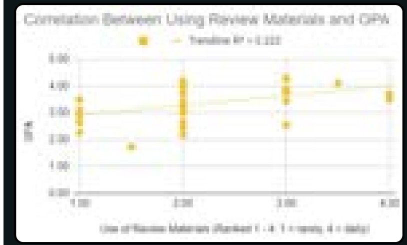
While any one behavior will not likely lead to meaningful increases in GPA, we believe that several academic practices in aggregate can incrementally contribute to significant change to your GPA. We hypothesize these factors will be the highest factors that contribute to GPA: Reviewing classwork outside of class, Reducing cellphone use, and spending more time on homework and studying each day.

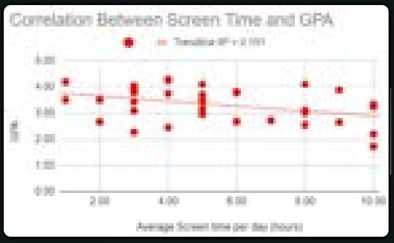
High school is hard. Especially when you're taking multiple AP Classes which is extremely common in our school. To improve our schools performance, the Brooke High School Data Science Club wanted to find ways to improve Grade Point Average (GPA) to provide guidance to our fellow students how to optimize their GPAs and thereby gain access to a wider array of opportunities beyond high school. At first, we hypothesized that GPA would be affected by the amount of schoolwork done, physical and emotional health, sleep, and many other factors. Our analysis showed that some of these factors affected GPA but others were unexpectedly uneffective.

RESULTS

After collecting our data from our sample, we identified the explanatory variables with the highest correlation of determination (R^2). We then created 3 scatter-plots and added trendlines to visualize the correlation between those variables and our response variable (GPA).







CONCLUSION

In conclusion, we have analyzed the correlation between many GPA-related factors. We were able to decide that by looking at our scatter plots and phone usage, spending time on homework, and studying had the highest r^2 values. Since all of these correlation values are fairly low we feel this supports our original hypothesis, that no one factor on it's own has a large effect on GPA but instead a combination of these factors lead to higher grades.

METHODS

Every two weeks, our club sent out surveys that asked certain questions to volunteers in our school to figure out their daily activities. At the end of the survey, we asked what their GPA was to see if students with certain habits had a higher or lower GPA. By using R^2 analysis we checked to see if there was any significant correlation between habits and GPA.

How we visualized our data

We visualized the data by using scatterplots. These scatterplots allowed us to see further if there was a trend line that showed any correlation and see if different things students did, in or outside of school, had any impact on GPA.

ANALYSIS

One of the values that we found was that people with a higher GPA were more satisfied with their academics (R^2=0.30). However, that value wasn't helpful, being satisfied with your GPA doesn't show anything actionable that students can use to increase their grades. Studying was not the highest correlation we found (R^2=0.32) but it helped prove that studying does play a factor in helping increase a student's GPA. Another factor that we saw was a decrease in phone usage caused an increase in GPA (R^2=0.20). We decided that this was an important factor because students often spend the majority of their time when getting home on there, so decreasing phone usage would hopefully lead to an increase in GPA.

CHALLENGES

Our main challenge was recruiting people to do the survey and experimental portions of our study. Our sample sizes were smaller than we would have liked (~40 total students from across the 3 surveys) Additionally, it was a big challenge to plan our experiment when we started because it was a big project that, in size, hadn't been done in our school

NEXT STEPS

After conducting three surveys and getting over 120 responses, we identified the three explanatory with the highest r^2s. The drawback of this analysis was that we could only show correlation and not cause and effect. In order to prove cause and effect, we designed an experiment using the three explanatory variables as individual treatments as well as a control. We are currently in week 2 of an 8 week experiment and are extremely excited to start getting our first round of comparative results soon!

Registered Drinkers

Avonworth High School

Team Members: Brayden Simmons, Amelia Hardiman, Madison Hollywood, Jackson Shields, Jagger Boyd, and Samuel Cavanaugh

Mentor: Sarah Sirakos, University of Pittsburgh

Central Question:

How do alcohol sales per capita in a county compare to the political ideology in that county?

Background:

With the rise of extreme politics and political party ideas, it is important to understand potential trends in politics. Addressing a potential correlation between the buying of alcohol in relation to political party voter registration may bring forth unexpected trends. Additionally, Pennsylvania has been a swing state in past elections, meaning that slight changes and trends in the percentage of voters in each political party could have dramatic effects on future state and county elections. These questions will coincide with a great majority of the voting population as well as teenagers who are politically active and nearing the voting registration age.

Hypothesis

We predict that alcohol sales will have a strong positive correlation with republican voter registration. Areas with higher alcohol sales per number of citizens will have a higher percentage with republican voters.

Challenges

Confounding Variables

- We had to ensure that we factored in possible confounding variables. Political Ideology is incredibly complex and affected by factors including socioeconomic status, age, gender, etc.
- We had to request specific data from various sources and companies unwilling to disclose data
  - We ultimately had to abandon looking at restaurants, an integral part of our initial proposal
- The data sets were organized by zip code, so we had to geocode the data

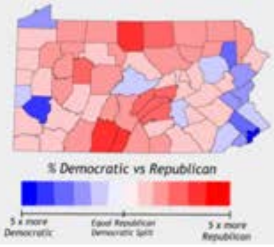
Datasets:

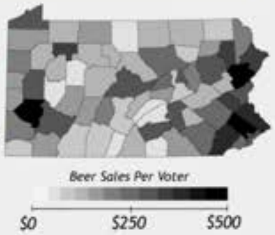
Alcohol Sales: Pennsylvania Liquor Control Board

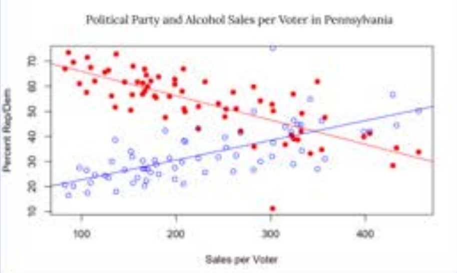
https://docs.google.com/spreadsheets/d/1snrXy2I5So5SvBQEQNVvJcZsxGUUV5PX/edit#gid=404422276

Political Ideology Statistics: Pennsylvania Center for Workforce and Information & Analytics

https://www.dos.pa.gov/VotingElections/OtherServicesEvents/VotingElectionStatistics/Documents/currentvotestats.xls

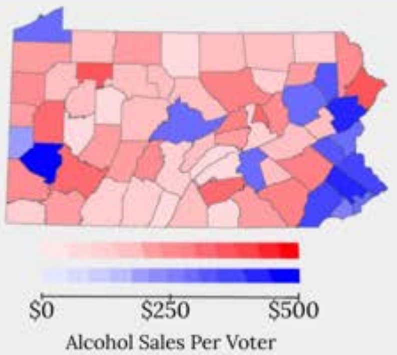





Political Party and Alcohol Sales per Voter in Pennsylvania

```
Residuals:
   Min      1Q  Median      3Q     Max
-35.082  -4.720   0.947   5.467  20.121
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 75.61002    2.52525   29.942   < 2e-16 ***
x           -0.09694    0.01020   -9.501   6.54e-14 ***
---
Residual standard error: 7.907 on 65 degrees of freedom
Multiple R-squared:  0.5814,    Adjusted R-squared:  0.575
F-statistic: 90.28 on 1 and 65 DF,  p-value: 6.543e-14
```

```
Residuals:
   Min      1Q  Median      3Q     Max
-15.451  -4.633  -1.007   4.035  36.605
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 15.04590   2.52537   5.958   1.20e-07 ***
x            0.07825   0.01018   7.689   1.14e-10 ***
Multiple R-squared:  0.4802,    Adjusted R-squared:  0.4721
F-statistic: 59.12 on 1 and 64 DF,  p-value: 1.139e-10
```



Methodology

Data Collection: We reached out to the PA Liquor Control Board for alcohol sales by county. And found current voter status for each county on the Pennsylvania Department of State Website.

Data Sorting: We imported the two data sets into excel spreadsheets and geocoded them. We reviewed each set to ensure that data was accurate and in a usable format.

Linear Regression Analysis: We imported the data sets into RStudio and created linear regression graphs. We then created lines of best fit to test the correlation values.

Analysis & Conclusion

Republican: Based on the Republican Voter percentage vs. alcohol sales per voter graph, with a corresponding P value of 6.543E-14 and r^2 value of 0.5814, we've found a moderate negative correlation between alcohol sales per voter and Republican voter registration. **Thus, we found that counties with a lower alcohol sale per voter had a larger percentage of registered republican voters.**

Democrat: Based on the Democrat voter percentage vs. alcohol sales per voter graph, with a corresponding P value of 1.139E-10 and r^2 value of 0.4802, we've found a moderate positive correlation between alcohol sales per voter and Democrat voter registration. **Therefore, we found that counties that had a higher percentage of voters registered democrat consisted of higher alcohol sales per voter.**

General: Since both trends are significant, it suggests that **the more democratic an area, the more alcohol sales per voter.**

Diving Deeper:

We were extremely surprised to find a significant p-value that further proved the correlation between political affiliation and drinking habits among voters. While we cannot conclude that drink sales result in political party registration, we believe this is an important topic that should be analyzed among other variables.

The Reality of Real Estate

What influences property values in Pittsburgh municipalities?

Avonworth High School Team 2: Vanessa Amayo, Kai Carlson, Beckham Chekan,
Kalea Wilson, Elena Zimmerman, Nickolas Veleke

Background

As the market for homes has fluctuated and become increasingly unstable, it is increasingly useful to see what goes into pricing a house. Many factors influence housing costs, like the resources used (both labor and materials), square footage, and the amount of land. However, we were most interested in the surrounding properties and their effects on the pricing of a house. We researched whether there is a correlation between a property's value and its locational assets.

Hypothesis: The correlation between residential property value and locational assets will be relatively low, as the cost to build it and the amount of land will outweigh the influence of locational assets.

Process

Once we selected our idea, our process was divided into 3 stages.

- **Stage 1:** Randomly select municipalities in Allegheny County and find the correct price values of homes within that area. This process came with the most challenges.
- **Stage 2:** Format the data into .csv files to process in Python. We created a program to clean and process the datasets. The program assigned subjective point values, which relate to the locational assets and their distances from the property and distance from the city. The closer to the home and further from the city, the higher the point value.
- **Stage 3:** Find the correlation and create heatmaps to represent the relationships.

Our Data Sets

- Parcel Centroids 2022 August
- Property Assessment Appeals 2015-2023
- Allegheny County Property Assessments
- Pittsburgh Public School Locations
- Supermarkets and Convenience Stores 2016 Data
- Allegheny County Restaurant/Food Facility Inspection and Locations

Definitions

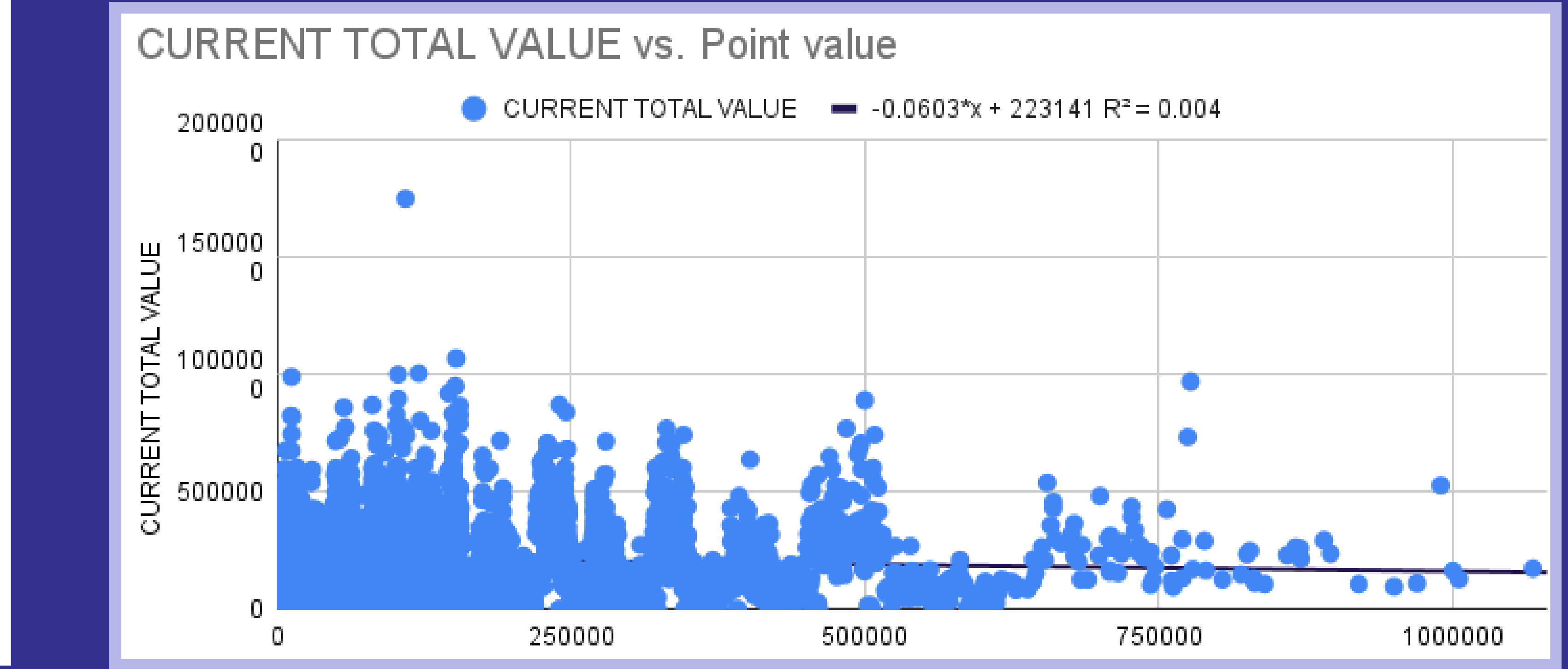
Locational Assets and Radii: Supermarkets (3mi) Convenience stores (2mi), Restaurants (4mi), Public schools (3mi), and Distance from the city (5mi)

Radii: Each distance above is the minimum distance for a property to get 1 point.

Municipalities: Ben Avon Borough, Castle Shannon Borough, Coraopolis Borough, East Pittsburgh Borough, Elizabeth Borough, Franklin Park Borough, Jefferson Hills Borough, Munhall Borough, Port Vue Borough, White Oak Borough

Challenges

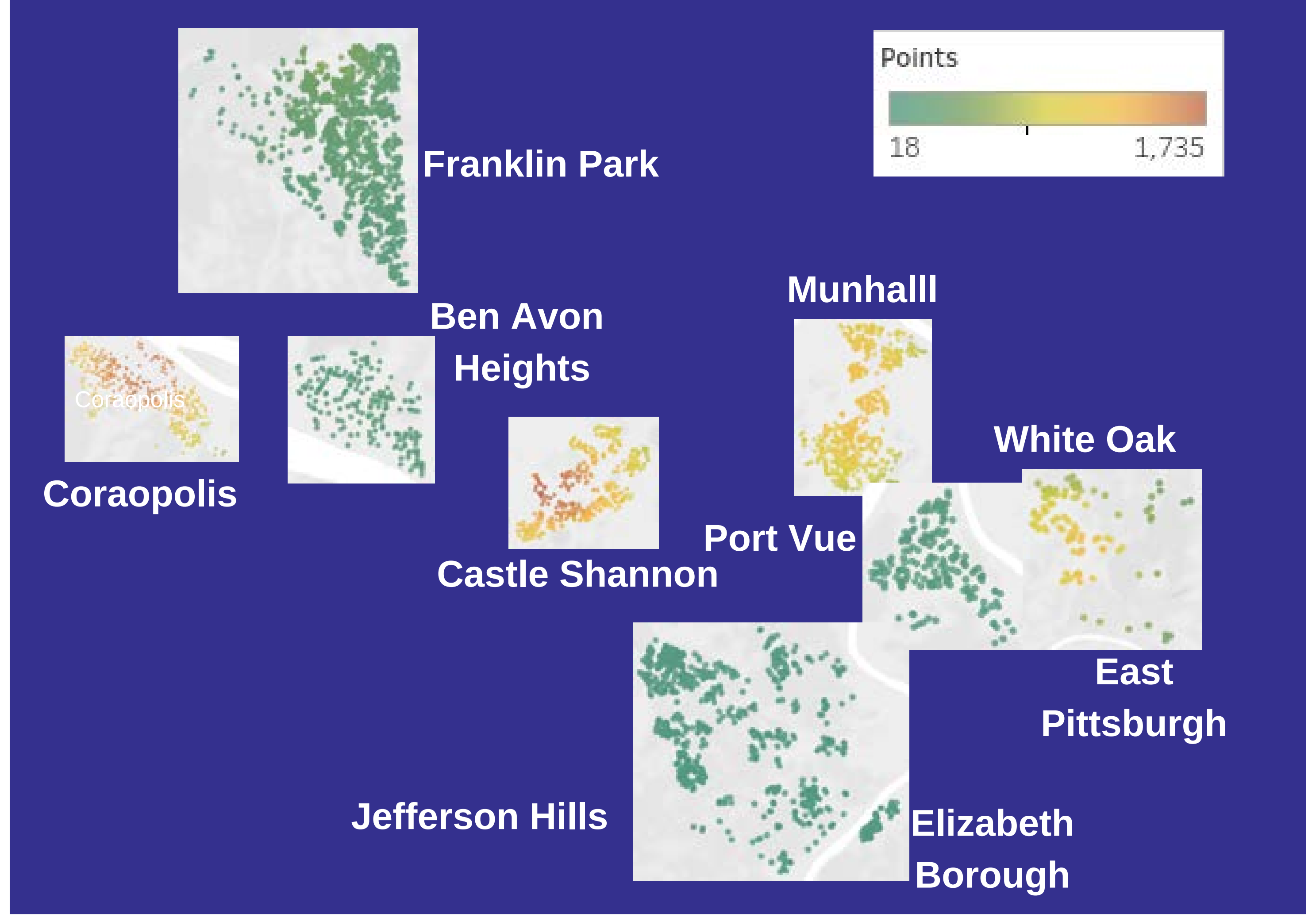
Our main challenge was **finding data sets with applicable** data for our hypothesis. We had to adapt the variables (eq. the type of area - municipalities, neighborhoods, or area codes) to fit our data sets. Another challenge was **developing the proper scales for the point values of our location assets.** We had to subjectively assign values to locational assets, which was difficult to decide.



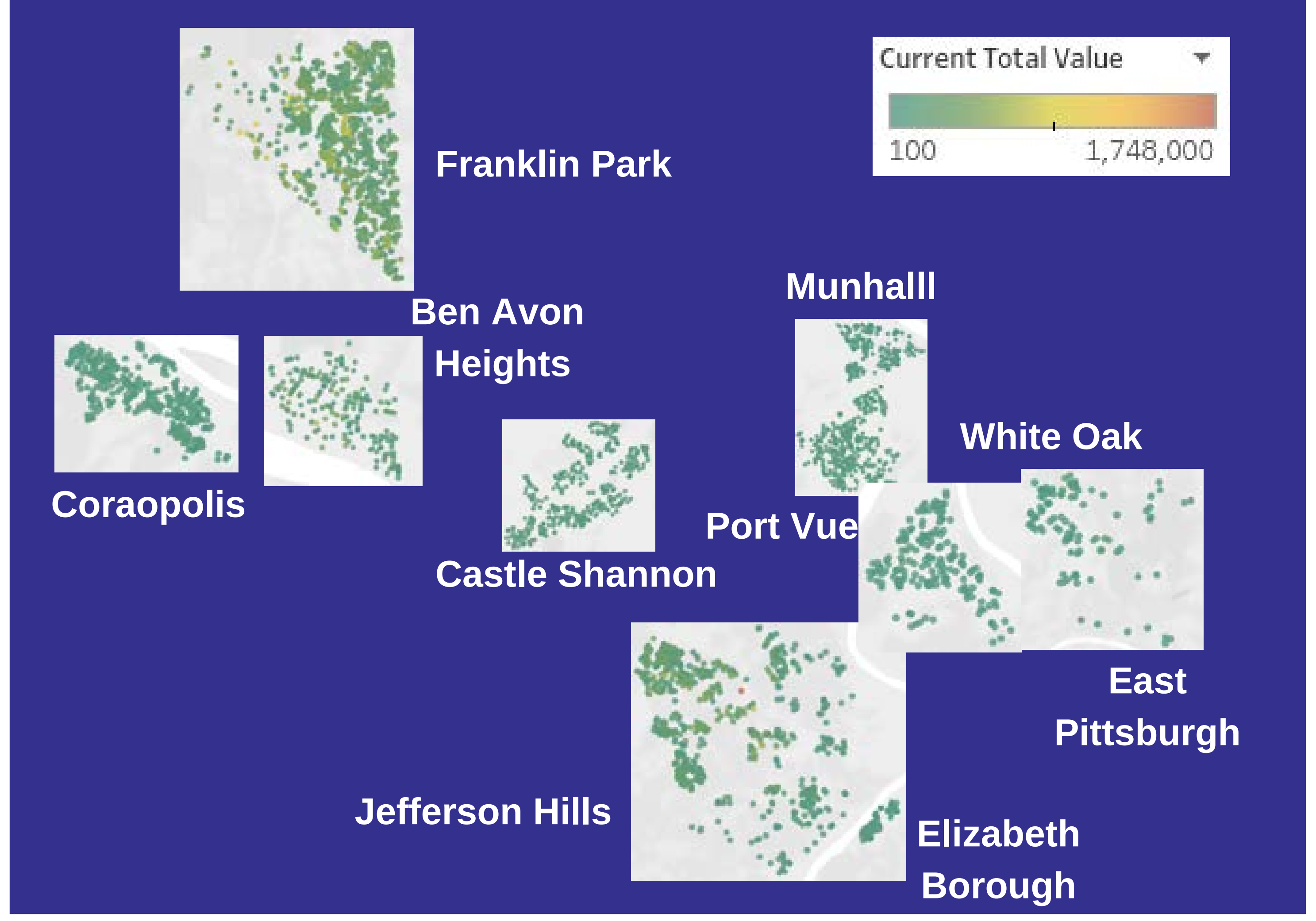
Conclusions

Based on our results our hypothesis was correct on how the correlation between residential property value and locational assets will be relatively low. The R^2 value indicates that .4% of the variability in property value can be attributed to the variability in point value. **<u>Location assets hardly affect residential properties' land property value.</u>** Our results can be applied to the Municipalities selected, as well as neighboring ones, however, as the distance from the Municipalities grows, the less confident our results can be applied.

Point Value



Actual Value



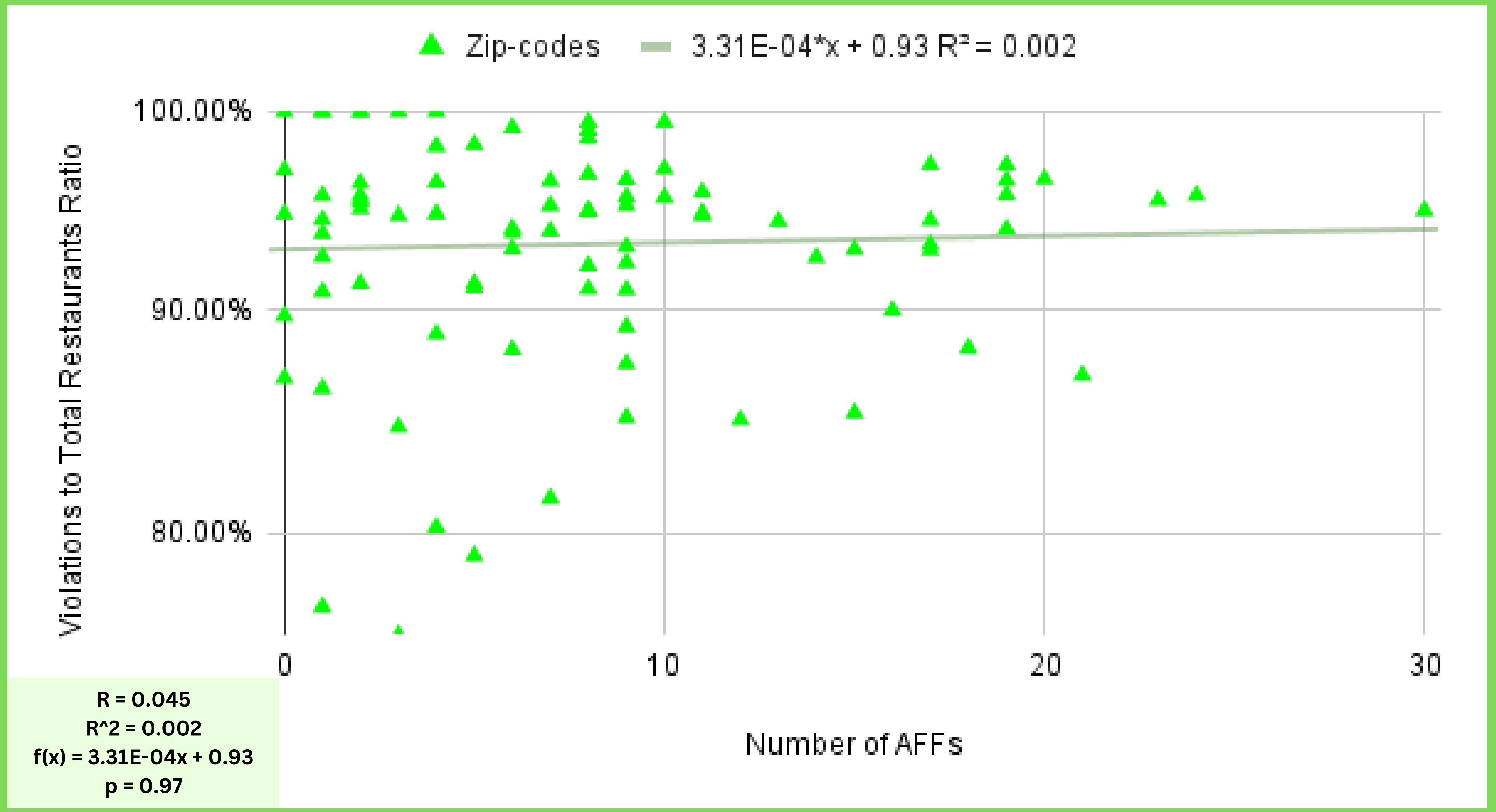
Let's Grab a Bite to Eat

How does where you live affect how well you can eat?

Background:

Access to safe food is access to healthy living. In the United States, access is by no means equal. By investigating the relationship between the places where ingredients can be purchased (AFFs) and and up-to-date health inspections of restaurants, we can determine if zip code directly correlates to healthy diets. The goal of this investigation is to become a stepping stone in further action towards correcting systematic inequality.

Number of AFFs against the Ratio of Violating Restaurants to Total Restaurants 2016-2022



Because the correlation coefficient is 0.002, and the p-score is 0.97, the violations ratio does not significantly correlate with the AFFs of a zip code.

Avonworth Data Jam Team 3 (Gbemi Odebode, Airah Shafiq, Peri Swiatkowski, Reese Theobald)

Research Question(s):

Q1: Does access to fresh food (farmers' markets, grocery, and convenience stores) impact the safety of restaurants in a given zip code?

Q2: Are farmers' markets, grocery, and convenience stores evenly distributed throughout the zip codes of Allegheny County?

Definitions:

We define "access to fresh food" (AFF) locations as the number of places were ingredients can be purchases (ie. farmers' markets, grocery, and convenience stores.)

Data Sets:

- Allegheny County Supermarkets & Convenience Stores-CKAN
- Allegheny County Farmers Markets Locations (2017)-CKAN
- Allegheny County Restaurant/Food Facility Inspections and Locations - Dataset - CKAN

Hypothesis:

The number of health inspection-violating restaurants will be inversely proportional to the number of AFFs in a given zip code

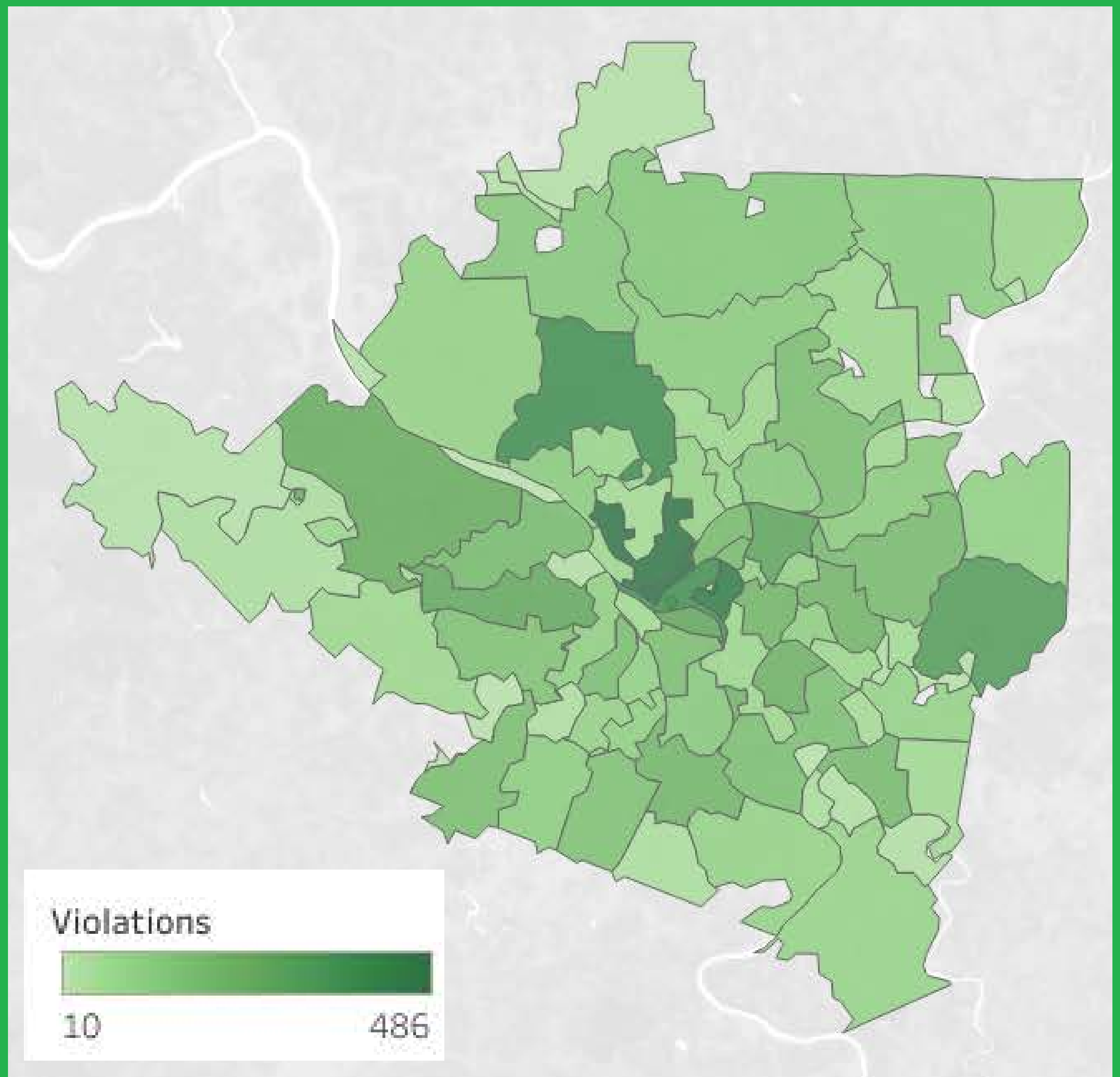
Methodology:

1. First, we examined the breadth of our data. Since two of our sets were restricted to 2016-17, we chose to limit our model to 2016-2022.
2. Using Excel, we extracted the number of violating restaurants in the period, their zip codes, and the total violations per zip code (VPZ). We ran the same analysis for the AFFs.
   a. In our investigation, the number of violations per restaurant is not of concern, one violation is enough.
3. We then compared the VPZ to the total number of restaurants per zip code to get a violation ratio.
   a. We chose to restrain our regression to only zip codes with 10 or more violations.
4. Using Google Sheets, we created a scatter plot and regression to model our data, where each point represents a zip code, the x-axis is the number of AFFs, and the y-axis is the violations ratio.
5. To investigate our second research question, we used Tableau to create a heat map of the AFFs and VPZ.

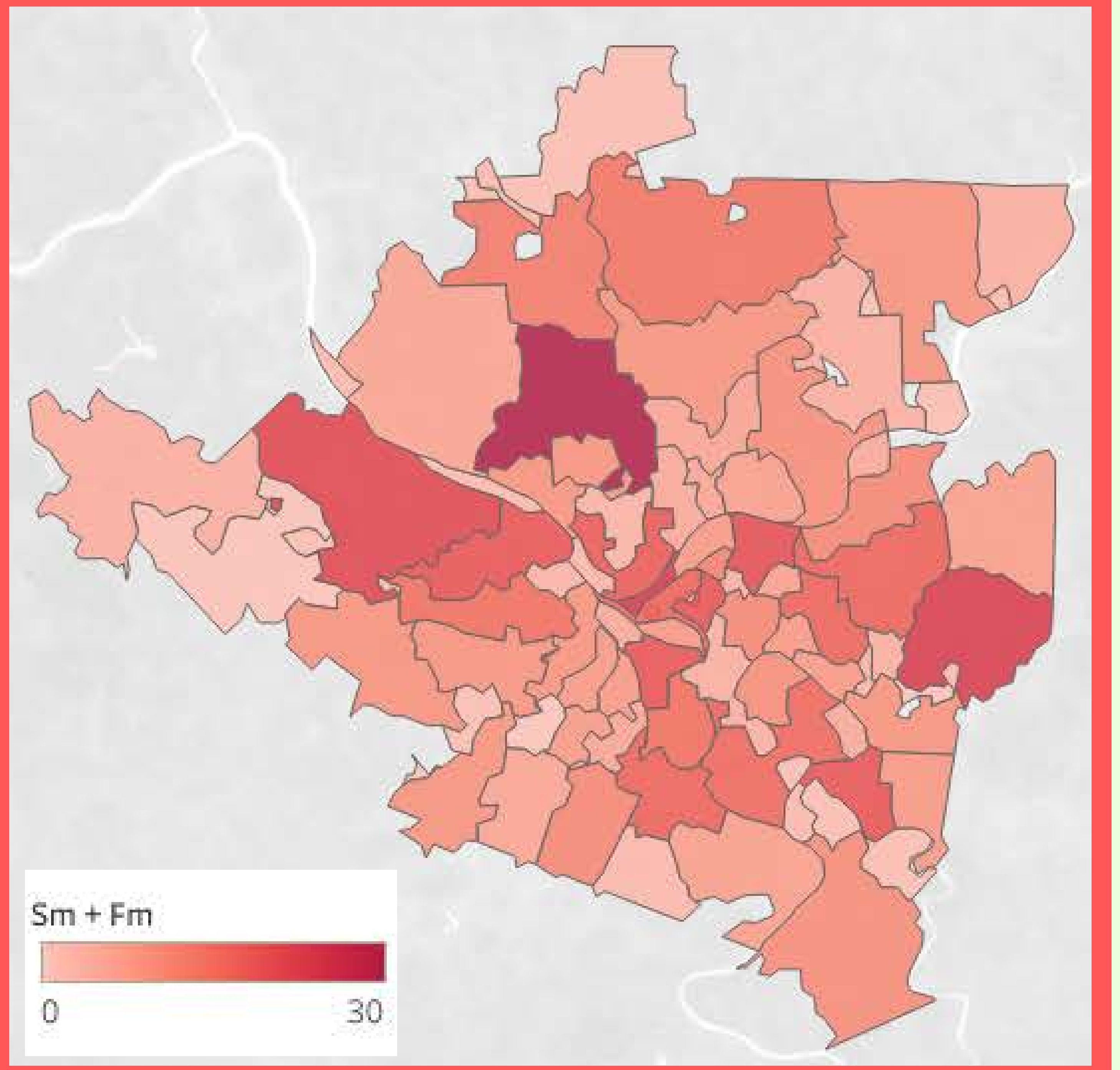
Challenges:

- we intended to compare our access parameters to the wealth of each zip code, as defined by housing value, however, but the data was not in a useful format
- our project works on too many assumptions to draw definitive conclusions

Number of Violations Per Zipcode



Number of AFF's Per Zipcode



Conclusion:

Although our conclusion was not statistically significant, it still provides great actionable value. This investigation has shown that access in Allegheny County is more nuanced than just AFFs per zip code. Due to restraints in our analysis capabilities and data access, we can not draw any definitive conclusions from this experiment. The trends in AFF and VPZ exhibited in the heat maps may be simply explained by local zoning laws. We recommend a deeper investigation with more extensive data sets.

Battle of the LSATS:
Are Public or Private Universities Better?

By: Lexi Dorfner and Sara Impellicceiri
Bethel Park High School

Problem

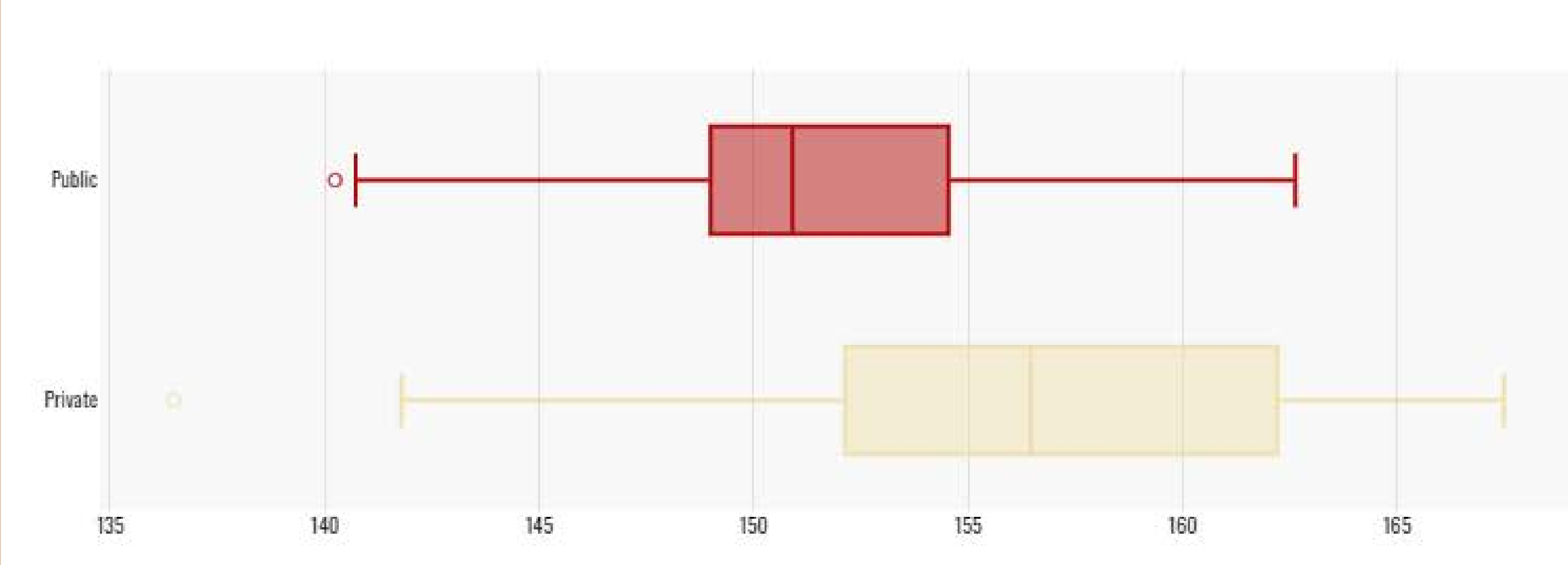
Do private or public schools produce better LSAT scores?

Importance

This is important to many students when selecting an undergraduate school. If a student knows they want to go to law school they would want to have a better understanding
of if the type of undergraduate school they pick matters. Not only do the amount of resources for law school matter, but also knowing how their outlook has been will be helpful.

Average LSAT Scores

Avg. of public school scores: 151.65
(44th percentile)
Avg. of private school scores: 156.2
(63rd percentile)



Conclusion/Policy

When looking at LSAT scores alone, we saw a positive correlation in the concept that private schools produce higher LSAT scores than public schools; however, when we looked beyond the LSAT scores alone, we realized there are more reasons why students score higher or lower than just the type of university they attend. We found a higher correlation with the following equation:

LSAT = 100.3 + 17.32 * Median GPA - 7.46 * Acceptance Rate

Policy: Public schools must offer study recourses such as law libraries for law students.

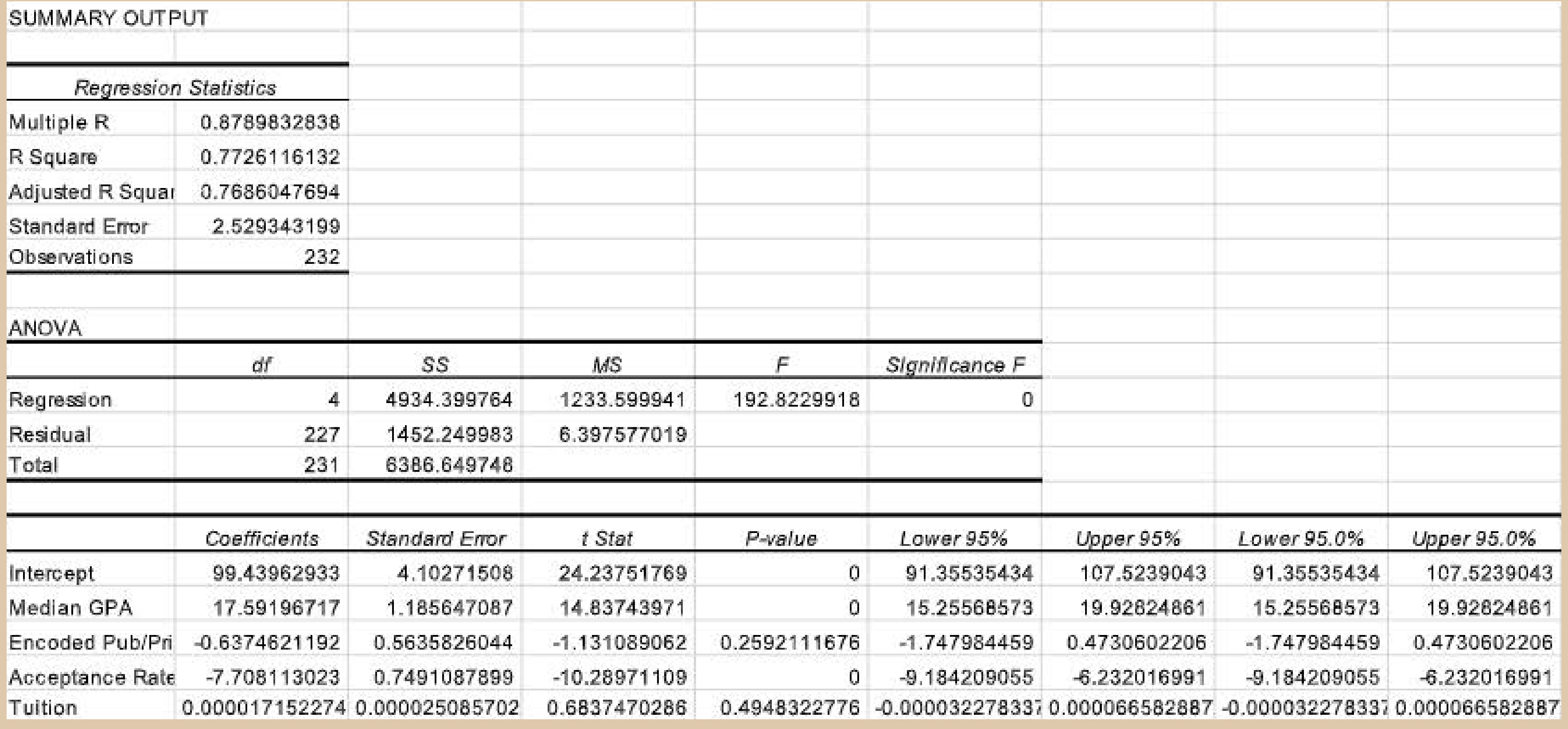
Hypothesis

If the student attends a private school, then they will have higher LSAT scores.

Analysis

We took data about the top 240 Feeder Schools for ABA (American Bar Association) Applicants, and we aggregate the data by whether the schools were public or private. We then conducted a 2 sample T test to find whether the type of undergraduate school was significant. Then we performed a multiple linear regression to look at how other factors such as GPA, Tuition, and acceptance rate affected the LSAT scores.

Multiple Linear Regression



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8789832838
R Square	0.7726116132
Adjusted R Squar	0.7686047694
Standard Error	2.529343199
Observations	232

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	4934.399764	1233.599941	192.8229918	0
Residual	227	1452.249983	6.397577019		
Total	231	6386.649748			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	99.43962933	4.10271508	24.23751769	0	91.35535434	107.5239043	91.35535434	107.5239043
Median GPA	17.59196717	1.185647087	14.83743971	0	15.25568573	19.92824861	15.25568573	19.92824861
Encoded Pub/Pri	-0.6374621192	0.5635826044	-1.131089062	0.2592111676	-1.747984459	0.4730602206	-1.747984459	0.4730602206
Acceptance Rate	-7.708113023	0.7491087899	-10.28971109	0	-9.184209055	-6.232016991	-9.184209055	-6.232016991
Tuition	0.000017152274	0.000025085702	0.6837470286	0.4948322776	-0.000032278337	0.000066582887	-0.000032278337	0.000066582887

Challenges

Calculating out T statistic
Entering data for all 240 schools
Lack of prerequisite Statistics knowledge

Sources

Niche.com
Top 240 feeder schools for ABA applicants: 2015–2017.

Clearing The Air: Identifying Key Factors In Pollution Levels

Alison Leung, Salaha Suleyman, Madeline McDine
Keystone Oaks High School Team #3

Research Question: What factors best predict air pollution?

Hypothesis: Elevation is the best predictor of pollution.

Background:

Air pollution poses significant threats to public health, particularly in densely populated areas where individuals face daily exposure to harmful pollutants.Particulate matter (PM) is a significant source of air pollution and a major factor in the negative effects on human health. PM is made up of solid particles and liquid droplets that are suspended in the air. PM is classified based on its size, with PM1, PM2.5, and PM10 representing particles smaller than 1 micrometer, 2.5 micrometers, and 10 micrometers. These particles originate from various sources, including vehicle emissions, industrial activities, and natural disasters such as dust storms and wildfires. Exposure to PM is associated with a range of health issues, including respiratory and cardiovascular problems, as well as death. Particles like PM2.5 can penetrate deep into the respiratory system and enter the bloodstream, allowing significant risks to human health.

Methods:

We started by choosing cities throughout the US. We used the data collected from our school roof(Dormont, PA) as a base metric. South Side(PA) was chosen due to its proximity to a river and its placement within the Monongahela river valley. Chicago and Denver were chosen based on their elevations, with Chicago having a low elevation and Denver having an extremely high elevation. We imported data from Purple Air to get the particulate matter measurements from each city and used the USGS topographic map to get the elevations.

After we collected all of the data, we used RStudio to conduct a univariate analysis on the PM levels, temperature, and relative humidity. Then we created scatter plots and linear regression models. We analyzed the correlations, r-squared values, and confidence intervals of the models to determine which factor was the best predictor of pollution.

Challenges:

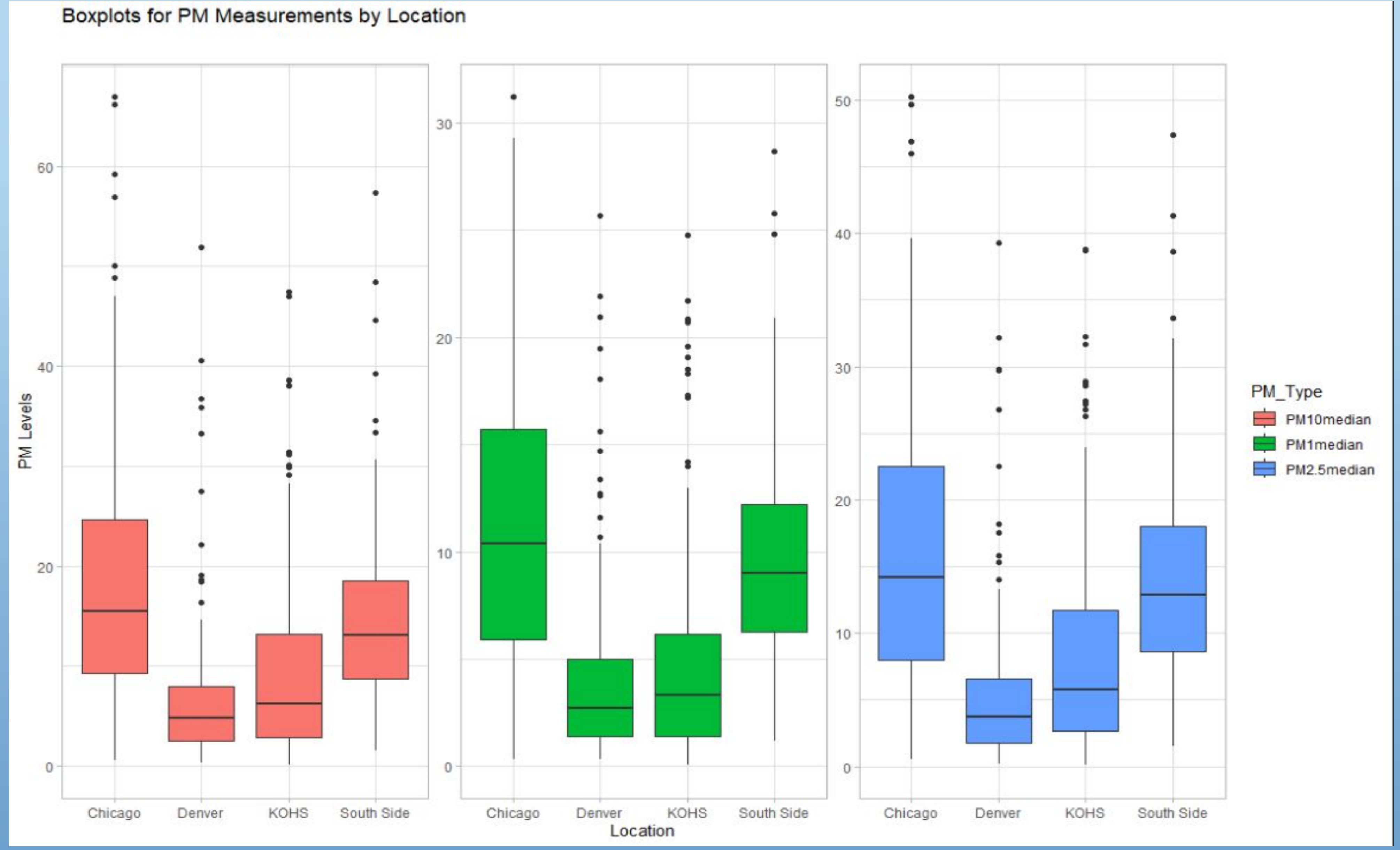
While the models predicting the levels of particulate matter are accurate in their predictions, since the PM levels are multicollinear, with all of the PM correlations being above 0.95, PM levels cannot be accurately predicted without each other. For example, trying to predict PM1 without using PM2.5 and PM10 produces low R-squared values, meaning that the predictions have a lot of unexplained variance.

Some challenges we faced included difficulty in collecting and cleaning the data. While it was helpful for the data to be aggregated daily, each city had a separate csv file, which took a considerable amount of time to clean and organize in a way that was usable for our analysis. We were unable to use the streamed data that was provided by Purple Air since it would take more time and computing power than what we had available. We were also unable to compare safe PM1 levels because we couldn't find air quality guidelines.

Recommendation:

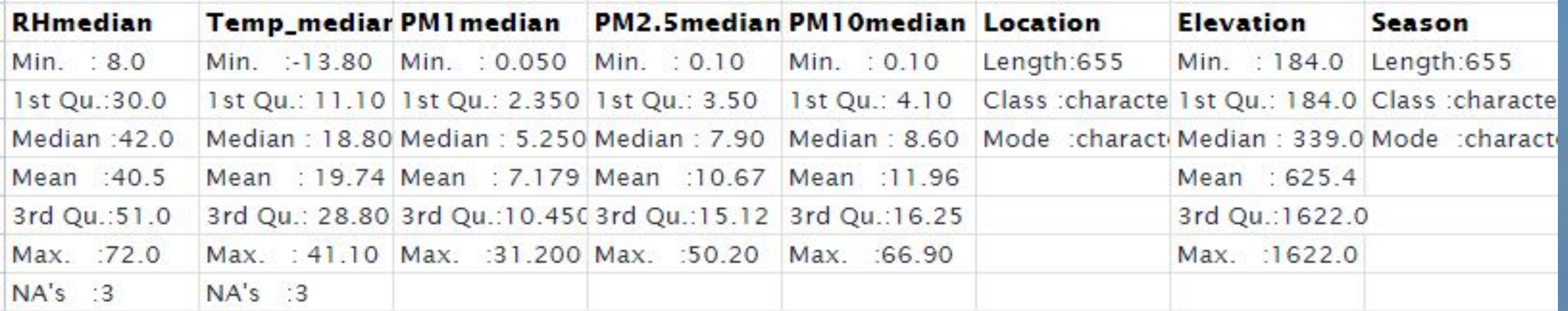
Temperature, relative humidity, and elevation do not seem to have a significant effect on pollution levels; however, it is likely that other factors such as population density, car usage, and natural disasters have a greater effect on pollution. Using public transit, recycling, and using reusable items are a few things that people can do to reduce pollution in their communities.

If we were to continue this analysis, we would look to include more factors such as population, population density, average carbon emissions, etc. We would also look to other countries with big, polluted cities, such as China, India, or South Africa.

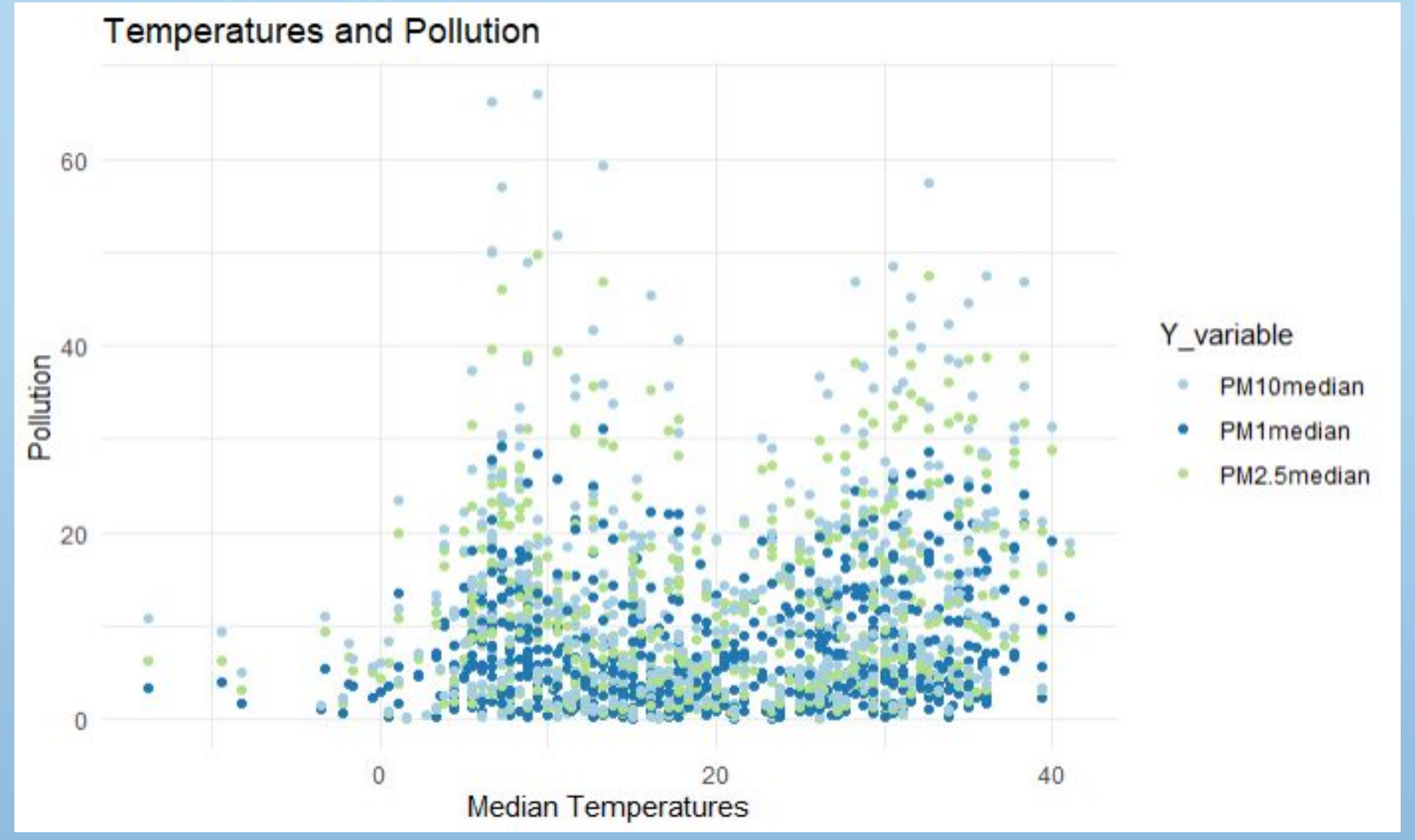


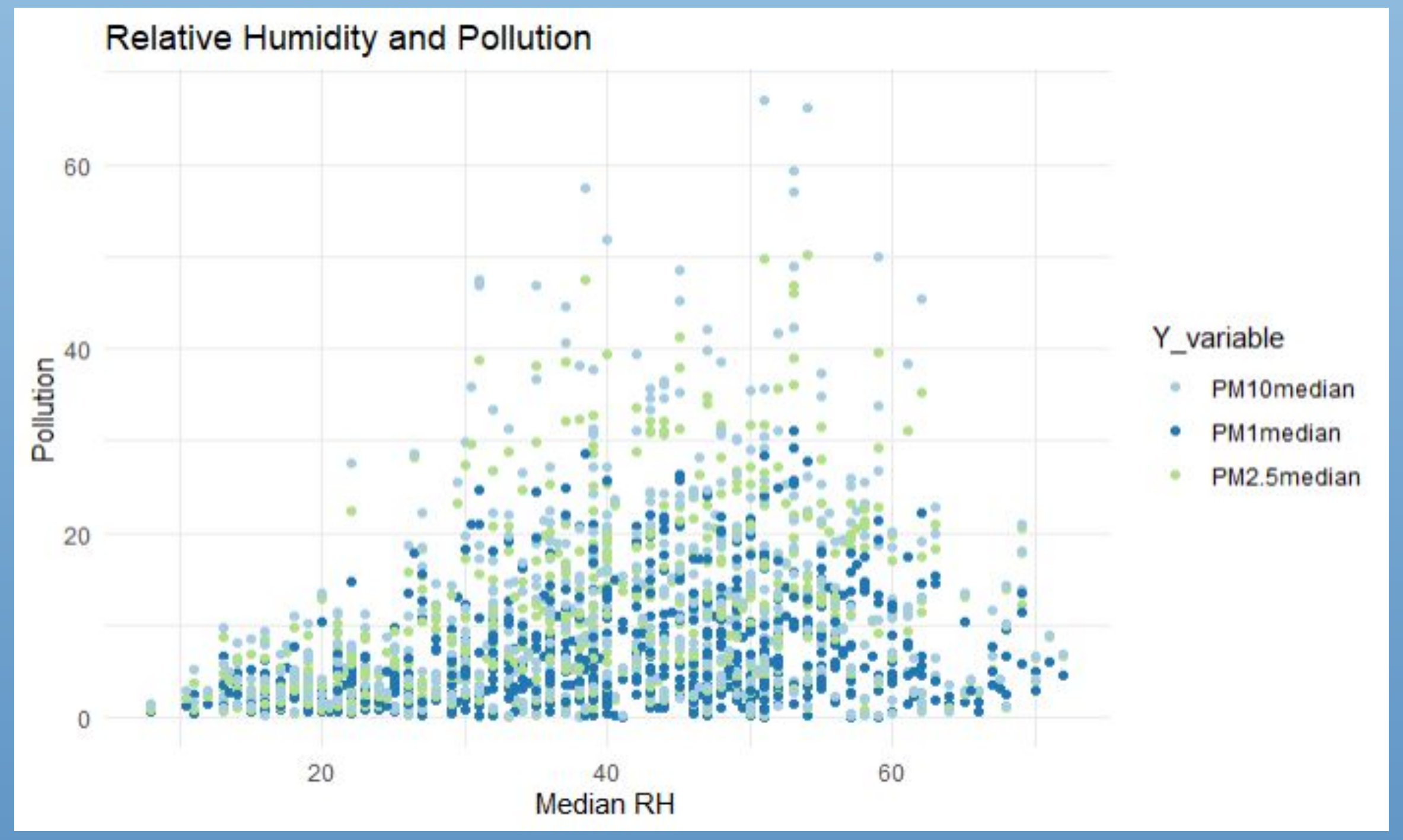
Analysis:

Based on our analysis and confidence intervals, we are 95 % confident that the true mean PM10 levels for Chicago (18.58), Denver (6.680), KOHS (8.424), and South Side (14.6993) are captured in their intervals and below the recommended annual average PM2.5 level of 45 μg/m³ as the guideline level for air quality because 45 μg/m³ is not a plausible value in any of the intervals. This indicates that the pollution metric of 10 is relatively safe in those regions. Chicago has the highest true mean PM10 levels out of all of the four regions which shows that there are more PM10 pollutants in the air. Denver has the lowest true mean PM10 Levels out of all four regions indicating there are less PM10 pollutants in the air. We are 95% confident that the true mean PM2.5 levels are above the accepted average levels for Chicago (16.360) while they are below accepted levels for Denver (5.480) , KOHS (8.424343), and South Side (14.140) . This suggests that Chicago has moderate pollution levels, whereas Denver, Dormont, and South Side maintain better air quality in terms of PM2.5. Therefore, it is unlikely that Denver locations are dangerously polluted, as the accepted dangerous metric is not contained within any of our intervals. Based on our 95 percent confidence interval, PM 1 in Denver has the narrowest confidence interval (3.474, 4.71) which shows that the estimated mean (4.0967) is relatively stable and if we were to perform many trials using the 95 % confidence level, we would get the similar results. The Chicago PM 10 confidence interval is the widest (16.664 to 20.506) which indicates that the uncertainty is greater, that maybe the mean could change if we did many tests with the same confidence level of 0.95. For all locations their PM 10 confidence intervals were the highest while their PM 1 confidence intervals were the lowest compared to the other pollution metrics. The PM levels for different regions overlap. For example, PM 1's confidence interval for Denver and KOHS overlap which suggests that despite being different locations, they experience comparable levels of PM 1 pollution, indicating potential similarities in pollution sources or environmental conditions affecting these areas.

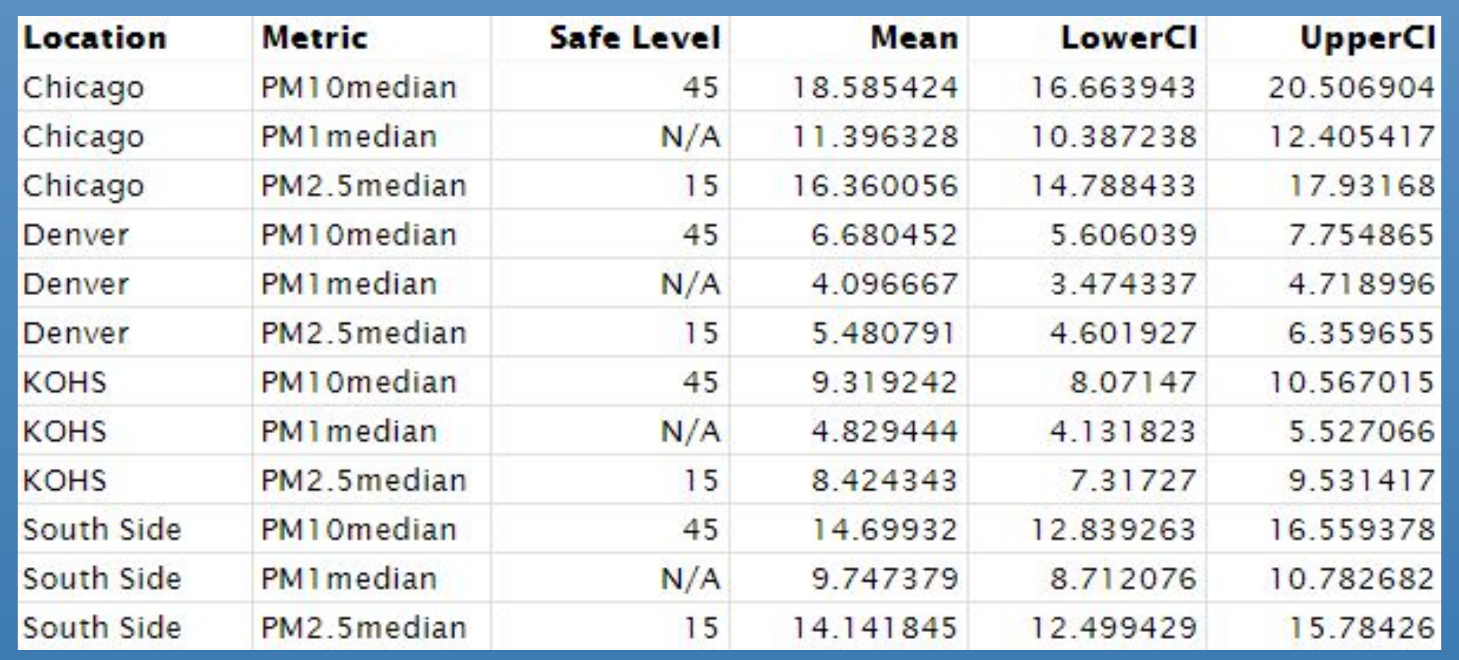


RHmedian	Temp_mediar	PM1median	PM2.5median	PM10median	Location	Elevation	Season
Min. : 8.0	Min. :-13.80	Min. : 0.050	Min. : 0.10	Min. : 0.10	Length:655	Min. : 184.0	Length:655
1st Qu.:30.0	1st Qu.: 11.10	1st Qu.: 2.350	1st Qu.: 3.50	1st Qu.: 4.10	Class :characte	1st Qu.: 184.0	Class :characte
Median :42.0	Median : 18.80	Median : 5.250	Median : 7.90	Median : 8.60	Mode :charact	Median : 339.0	Mode :charact
Mean :40.5	Mean : 19.74	Mean : 7.179	Mean :10.67	Mean :11.96		Mean : 625.4	
3rd Qu.:51.0	3rd Qu.: 28.80	3rd Qu.:10.45(	3rd Qu.:15.12	3rd Qu.:16.25		3rd Qu.:1622.0	
Max. :72.0	Max. : 41.10	Max. :31.200	Max. :50.20	Max. :66.90		Max. :1622.0	
NA's :3	NA's :3						



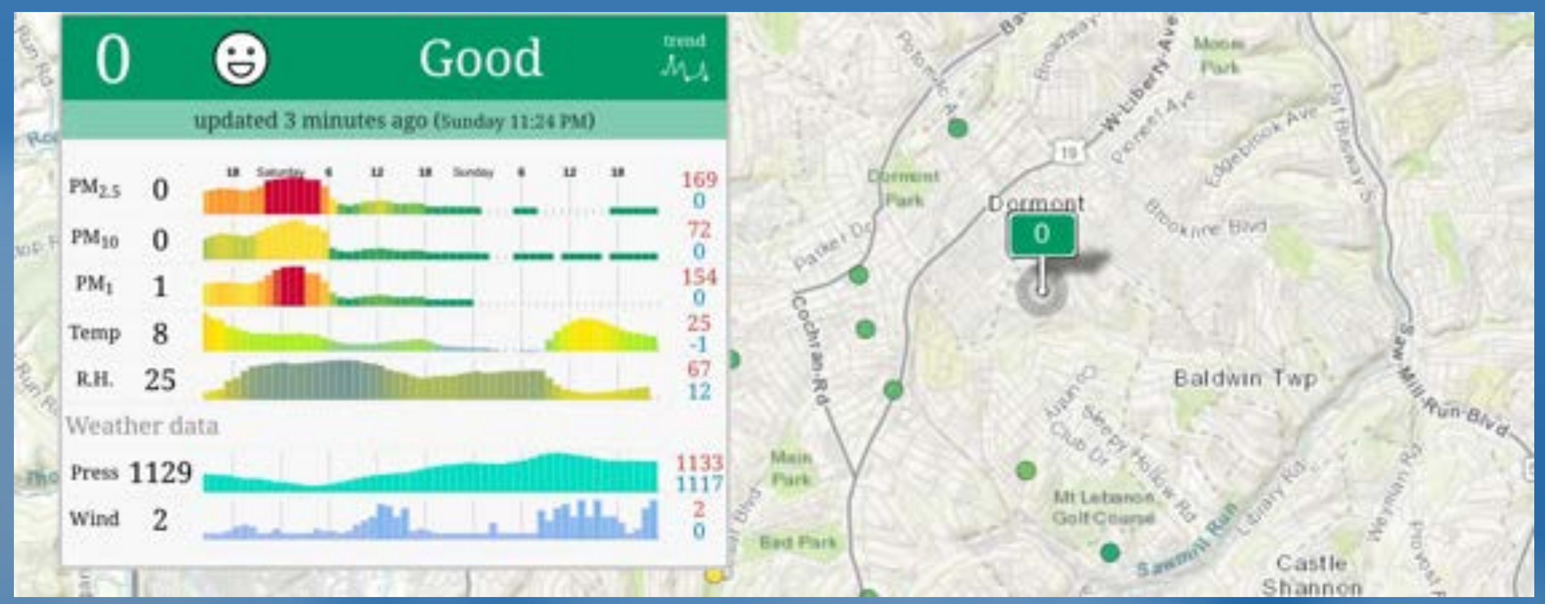


Confidence Intervals



Location	Metric	Safe Level	Mean	LowerCI	UpperCI
Chicago	PM10median	45	18.585424	16.663943	20.506904
Chicago	PM1median	N/A	11.396328	10.387238	12.405417
Chicago	PM2.5median	15	16.360056	14.788433	17.93168
Denver	PM10median	45	6.680452	5.606039	7.754865
Denver	PM1median	N/A	4.096667	3.474337	4.718996
Denver	PM2.5median	15	5.480791	4.601927	6.359655
KOHS	PM10median	45	9.319242	8.07147	10.567015
KOHS	PM1median	N/A	4.829444	4.131823	5.527066
KOHS	PM2.5median	15	8.424343	7.31727	9.531417
South Side	PM10median	45	14.69932	12.839263	16.559378
South Side	PM1median	N/A	9.747379	8.712076	10.782682
South Side	PM2.5median	15	14.141845	12.499429	15.78426

Purple Air AQI. Green dots on the map represent other Purple Air stations.



Conclusions:

Using significance tests, summary statistics, and graphs from various cities in the U.S., we did not discover convincing evidence to support our hypothesis that elevation is the most important factor in predicting pollution levels. We found weak negative correlations between elevation and pollutant concentrations, we found that elevation has the highest negative correlation out of all the three variables (temperature, elevation, and relative humidity) across all three PM pollutants; however, it's not an accurate predictor of pollution because its correlation is very low. While temperature and relative humidity also show some influence on pollution levels, their effects are even smaller than the effect elevation has on pollution. Additionally, the summary statistics reveal that pollutants like PM10, PM1, and PM2.5 tend to have higher concentrations, as shown by their right-skewed distributions. We found that PM levels are collinear, which means that to predict one PM metric, the other PM metrics are required. Understanding air pollution involves considering various factors beyond elevation, temperature, and humidity. Factories, vehicle emissions, and natural disasters, all affect air quality. Addressing these factors is important for developing strategies to reduce air pollution and protect public health and the environment.

Effect of High School Sports on Academics

Problem

What are the effects of participating in extracurricular sports on adolescents' academic performance?

Hypothesis

Sports and general physical activity have a positive impact on students education. They teach the students time management and discipline as well as also keeping students physically healthy which also helps their mental health.

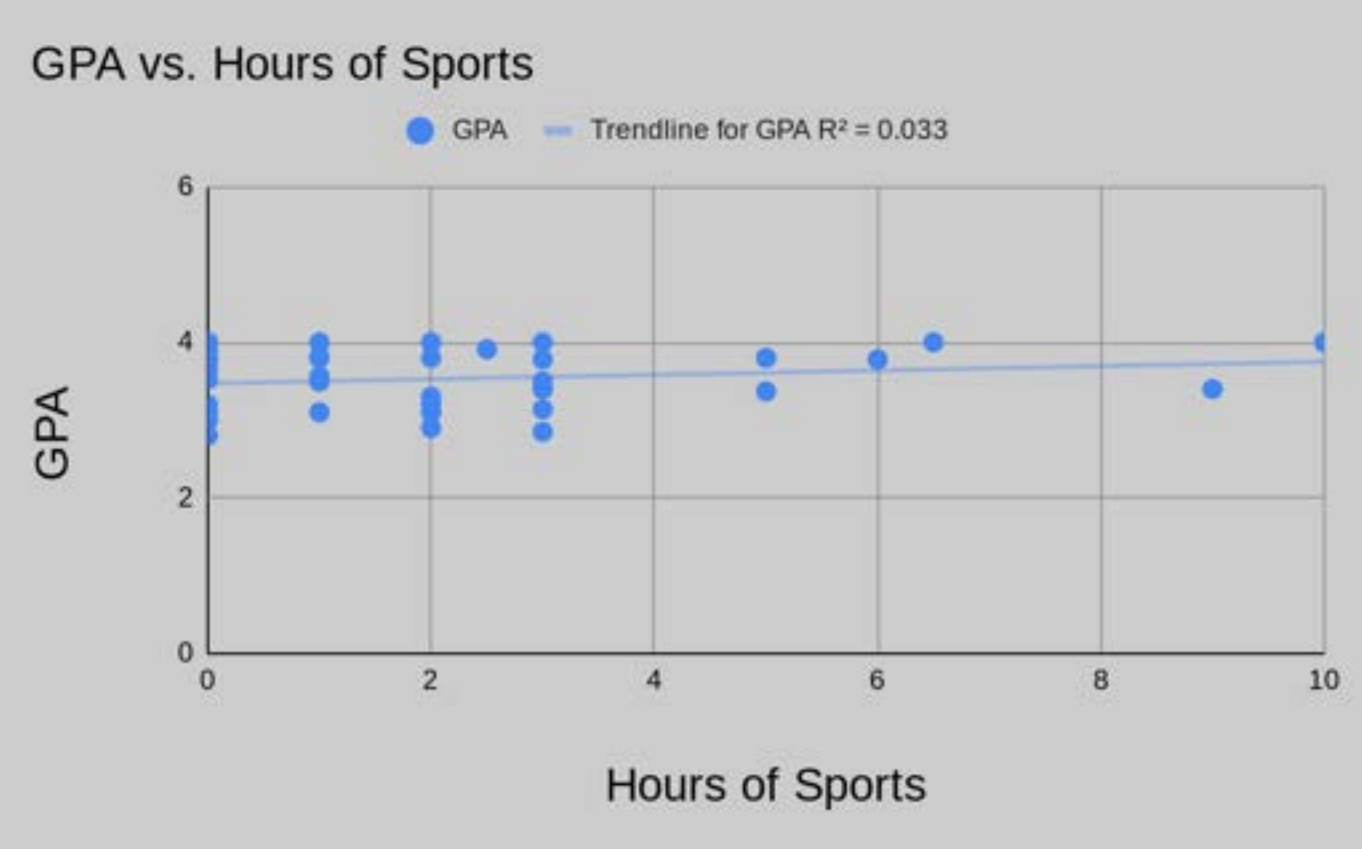
Importance

With millions of student athletes in America, it is important to know how it affects their grades and education. With this information, we could better help student athletes manage their heavier workload.

Plan

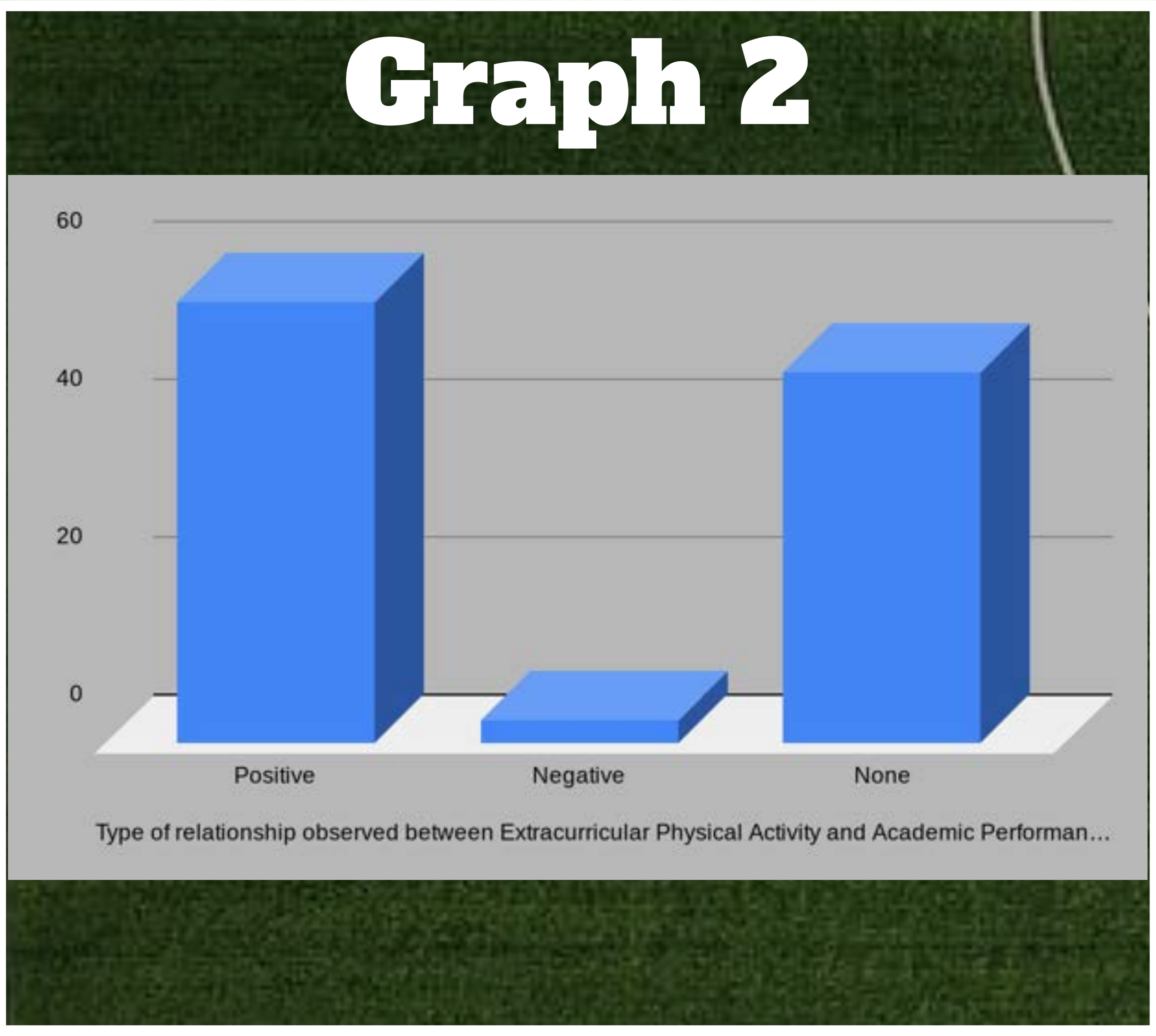
We will look at raw data sets that shows the relationship between physical activity and education. We will look at visuals that show the relationship. We can then develop and compare the data sets from the resources we've found to create a mean data set. Finding then a relative causation relationship to either support or deny our hypothesis.

Graph 1



Graph 1 shows a slightly positive correlation between the GPA of students and the number of hours spent playing sports. R = 0.182

Graph 2

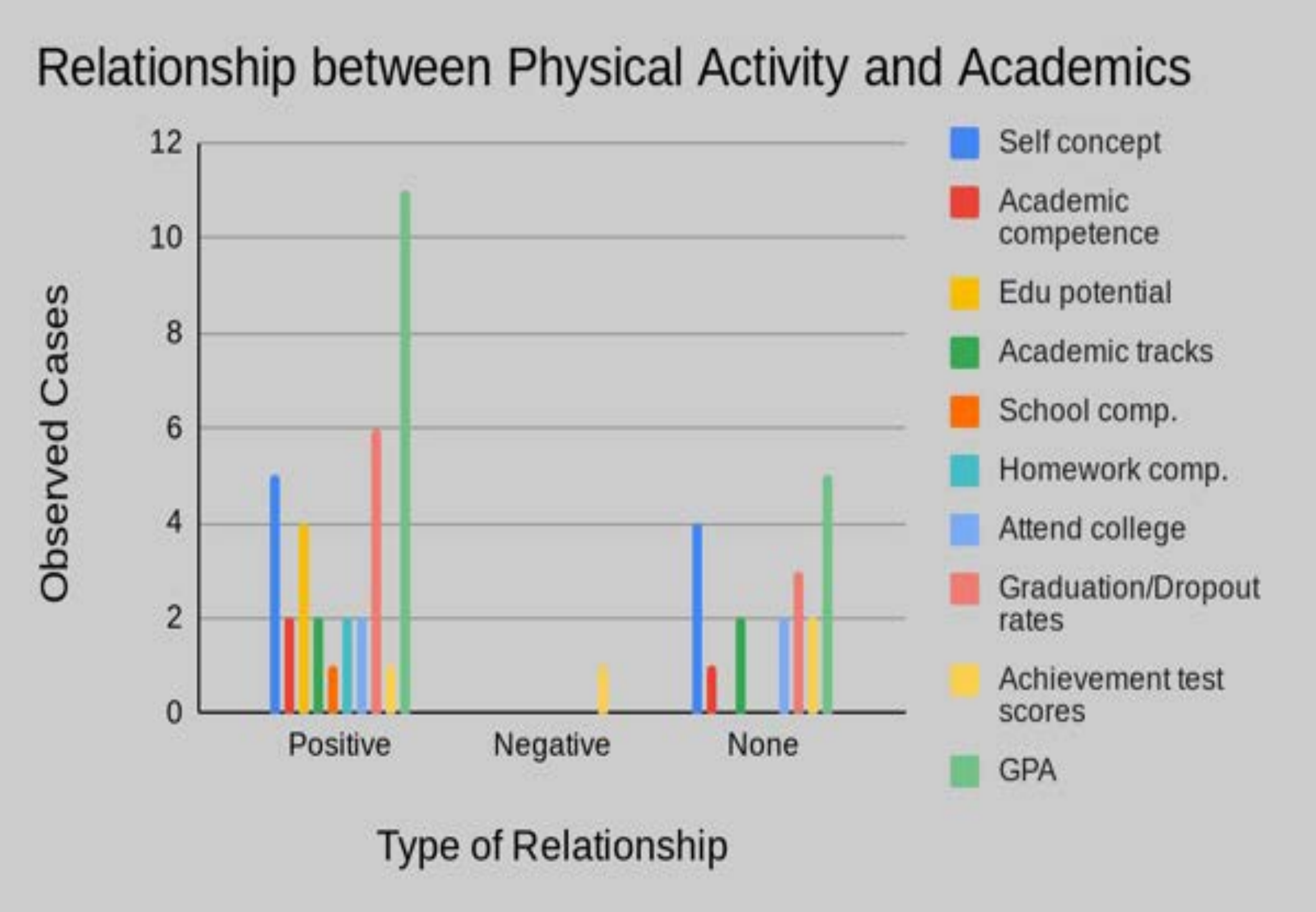


Challenges/Confounding Variables

- Kids in multiple sports would be even more busy which could affect their grades
- Kids have varying levels of commitment to their sport which could mean they put less effort into school
- Some teachers could be less tough on student athletes if they like sports or if they think they already have a lot on their plate or if they want a kid to remain academically eligible

Graph 2 reveals the effects (positive relationship, negative relationship, or no relationship) that extracurricular physical activities have on academic achievement. The intervention studies resulted in a total of 56 positive associations, 3 negative associations, and 47 no-associations.

Graph 3



Graph 3 shows the relationships of many different types of academics with physical activity.

Conclusion

There is not a negative correlation between participation in extracurricular sports and adolescents' academic performance, but rather a negligible or a positive correlation.

Ashley Schrecengost, Sean Verbosky, Trey Demink, Claire Price, Kolden McCall, Jack Mansfield, Nathan Cresanti

Maple Grove 2024 Team #2

Short Term Content and Attention Span

Maple Grove Jr. Sr. High School Team 1
Fletcher D., Jack B., Caleb T., Brennan M., Ellery Y., Sunnhi S., Lily B.

Problem

How has the introduction of various short term content platforms (less than a minute) affected the average viewer's attention spans and tolerance for longer-term content?

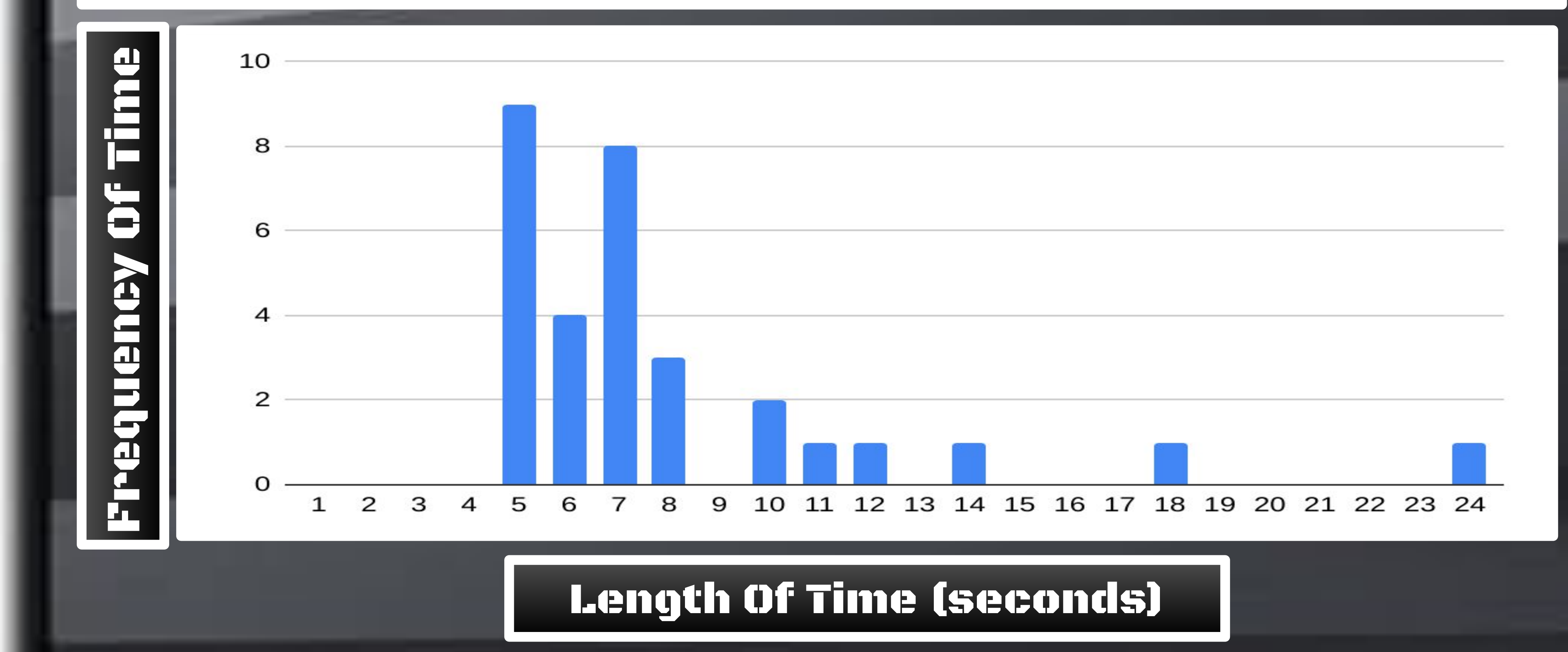
Hypothesis

As kids become more used to short videos they will have less tolerance for watching longer videos.

Why is it Important?

These days, most of us live our lives attached to our computers and smartphones, which are obvious sources of distraction. With this increase of short term content, our attention spans and tolerance for longer videos are suffering, limiting our ability to learn, communicate, and interact with society.

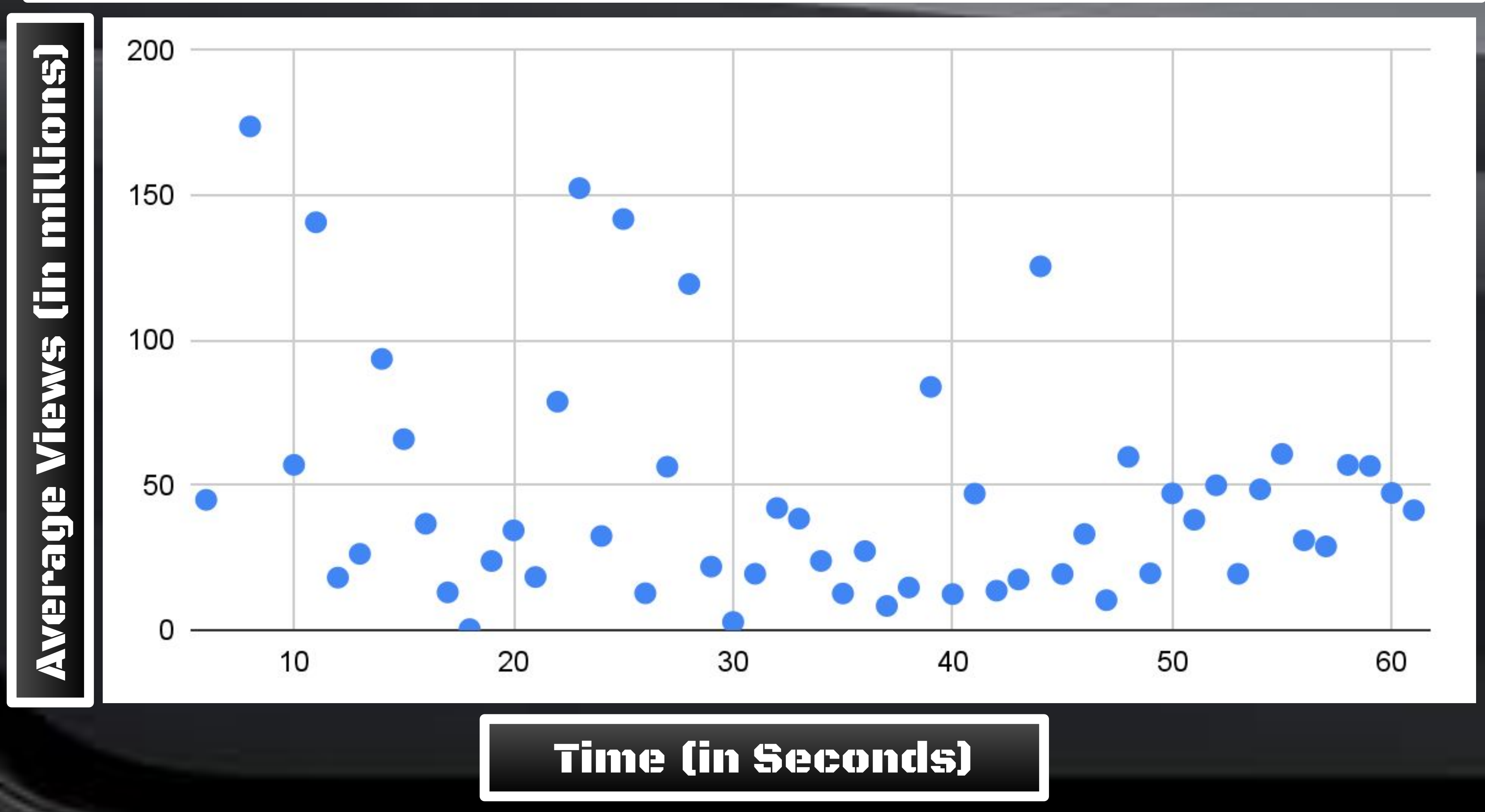
Top 30 videos watched on TikTok & YouTubeShorts 2022



Analysis Plan

In order to analyze our data we will be using several strategies to extract information and form our presentation. As a group we plan on looking through multiple resources about technology and its impact on our attention spans whilst comparing and contrasting their information. As we accumulate the most prominent details we will add them to our presentation. In order to add credibility to our sources we will use data from popular platforms, such as tiktok, youtube, and instagram from the Kaggle database to analyze their reels and see how they're affecting the average person's tolerance for longer videos.

Duration of Short V.S. Average Views



Data Required

Of the top 567 most viewed videos on Youtube shorts since 2017, 37% of the videos were only 1 minute and 7% were 59 seconds or less.

People have a less of a tolerance for longer content with availability of instant gratification on these reels.

The top 37 videos and trends on both Youtube and Tikotok in the 2022 summer were recorded in another data set.

12/37 videos were 5-7 second long making these the most liked.

As previously stated, people are developing a weaker tolerance for longer length videos. 11/37 videos were 7- 9 seconds long.

Challenges

Whilst attempting to gather information surrounding or topic on attention span, we struggled to find exemplary data that expound on our hypothesis.

Another challenge was that there was only data for Youtube Shorts and not for the other platform.

Conclusion

Our initial problem was to extract information about how the average viewers attention span was affected from short term content and how their tolerance for longer term content was affected. Our hypothesis predicted that kids would become more used to short videos (0-60 seconds) and have less tolerance for longer content (60+ seconds). However, when gathering information we did in fact determine that kids were susceptible to watch shorter content over time.

ervices.

Population in San Diego Vs "Cost of Living"

Miguel Torres, Bruno Perez, Fernando Galindo, Jenny Campos, and Alberto Garcia

san diego county office of EDUCATION FUTURE WITHOUT BOUNDARIES℠

monarch school
education. opportunity. transformation.

Research Question

Is the economy driving natives out of San Diego?

Background

It's always sunny in San Diego. San Diego is an amazing city to live in it has a lot of tourist attractions, beautiful beaches, fun parks, and best of all we have Sea World! Right now San Diego's economy has been thriving and although people from the outside think that that would be a good thing, San Diego natives feel differently. The impact that the economy has had in its residents can be classified as good and bad. For example, San Diego's economy improving has lead to bigger city improvements and active road work on neighborhoods who have previously been neglected by the city, but for a majority of its residents it has gotten too expensive to live here. As San Diego natives, we, as a team, have all felt the repercussions of the rise of cost of living and wanted investigate further.

Hypothesis

We believe that with the growth of the cost of living, population rates are going down (causing a negative net domestic migration rate in San Diego).

Definition

The negative net domestic rate is when the number of people moving out of the state in a year exceeds the number moving in.

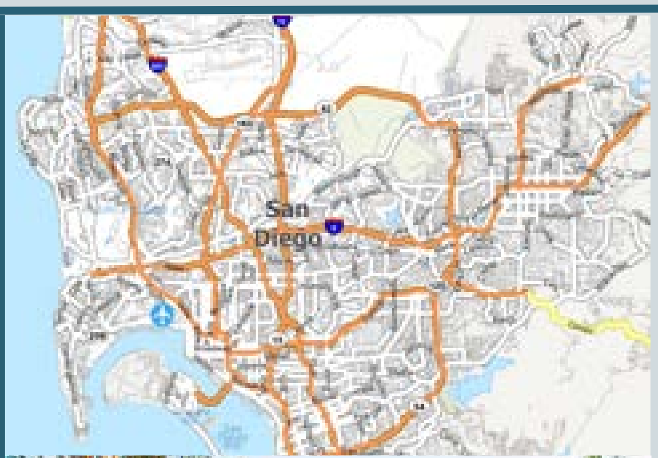
Methodology

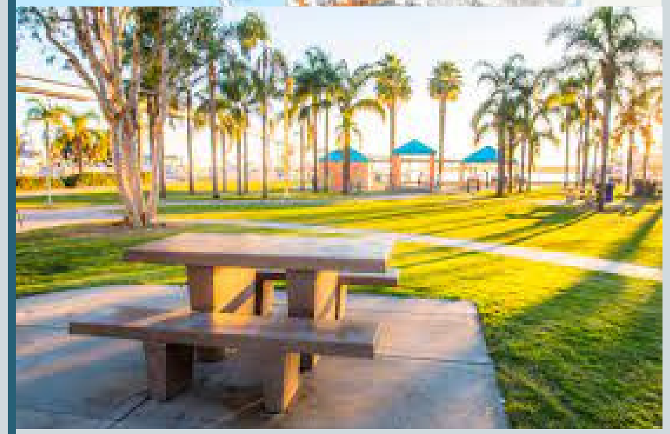
1.Data collection: We collected information from data commons.org and, by searching through multiple websites and comparing the numbers to search for similarities.

2. Data Filtering: We imported and carefully inspected the data, used the information that was most relevant to our topic.

3. Analysis: Using google sheets and information from The County of San Diego Government Site we utilized:

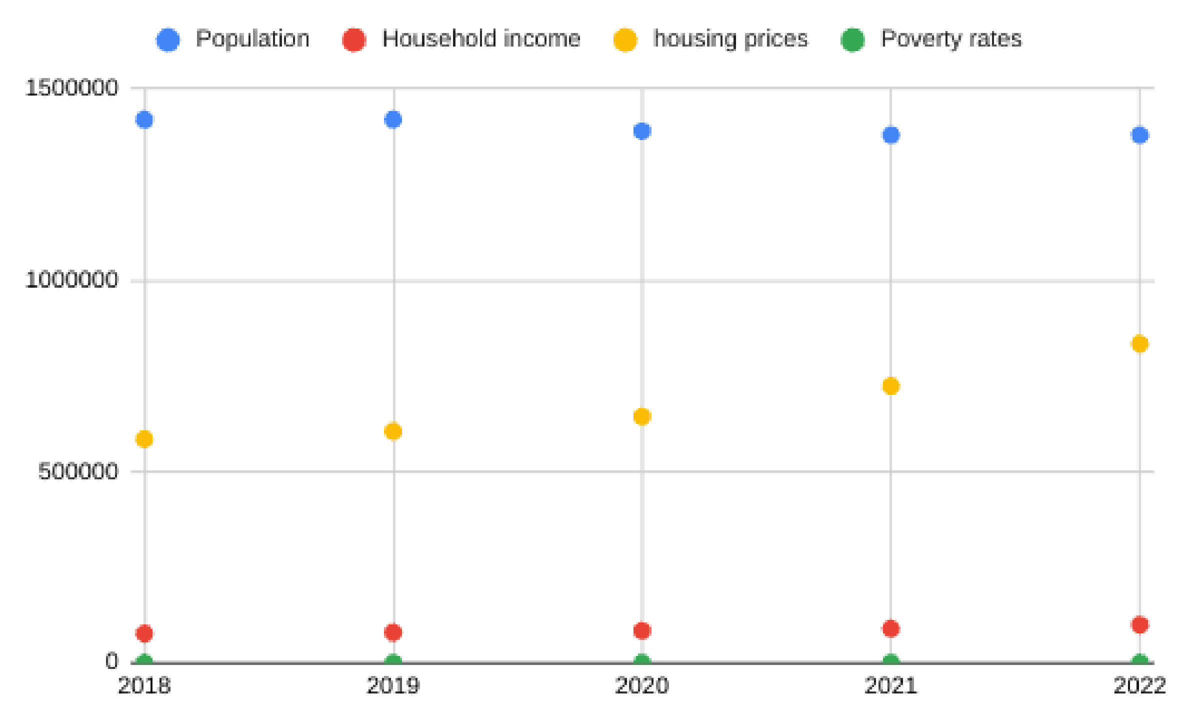
- Correlation analysis
- Regression analysis

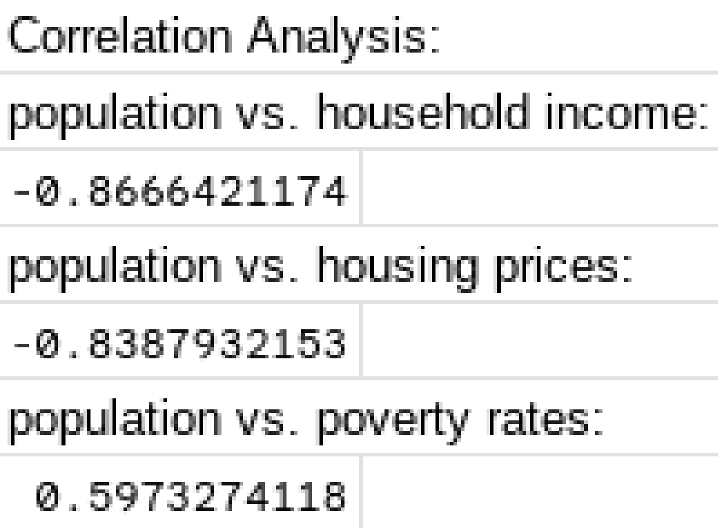




Datasets

Our data sets contain information from various sources all within the area span of San Diego.





Correlation Analysis:

population vs. household income:

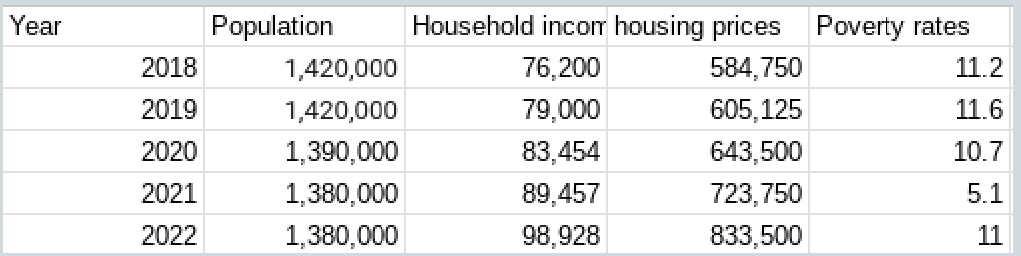
-0.8666421174

population vs. housing prices:

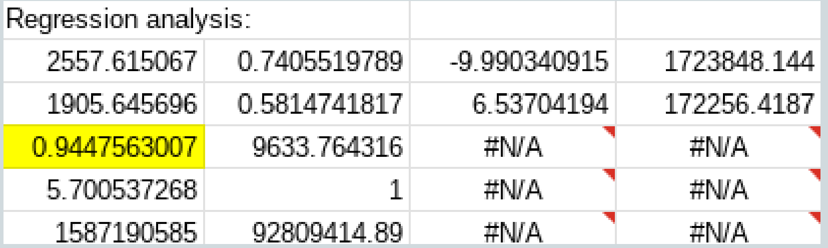
-0.8387932153

population vs. poverty rates:

0.5973274118



Year	Population	Household incom	housing prices	Poverty rates
2018	1,420,000	76,200	584,750	11.2
2019	1,420,000	79,000	605,125	11.6
2020	1,390,000	83,454	643,500	10.7
2021	1,380,000	89,457	723,750	5.1
2022	1,380,000	98,928	833,500	11



Regression analysis:			
2557.615067	0.7405519789	-9.990340915	1723848.144
1905.645696	0.5814741817	6.53704194	172256.4187
0.9447563007	9633.764316	#N/A	#N/A
5.700537268	1	#N/A	#N/A
1587190585	92809414.89	#N/A	#N/A

Challenges

Our main challenge was to find reliable data sources and connect all of the information we found at first it was very hard to find the information for the cost of living because a lot of the websites only had percentiles and not viable numbers that we can extract and use for the maps

Conclusion

In conclusion, based on our findings, research shows that in the last couple years poverty rates and housing prices have increased but the population has decreased to 40,000. Why does it happen? Well, it's simple. The more people move out there's less demands for services and goods, which causes the prices to go up. In addition, a smaller population means fewer workers, which can result in a shortage of skilled labor and drive up wages. This can contribute to the overall increase of cost of living. So, it's important to consider the impact of population changes when looking at cost of living. Furthermore, another factor to consider is that when a population goes down there may be a decrease in economic activity. This can lead to a decline in businesses and services available in an area, resulting in limited options and potentially higher prices for goods and services.

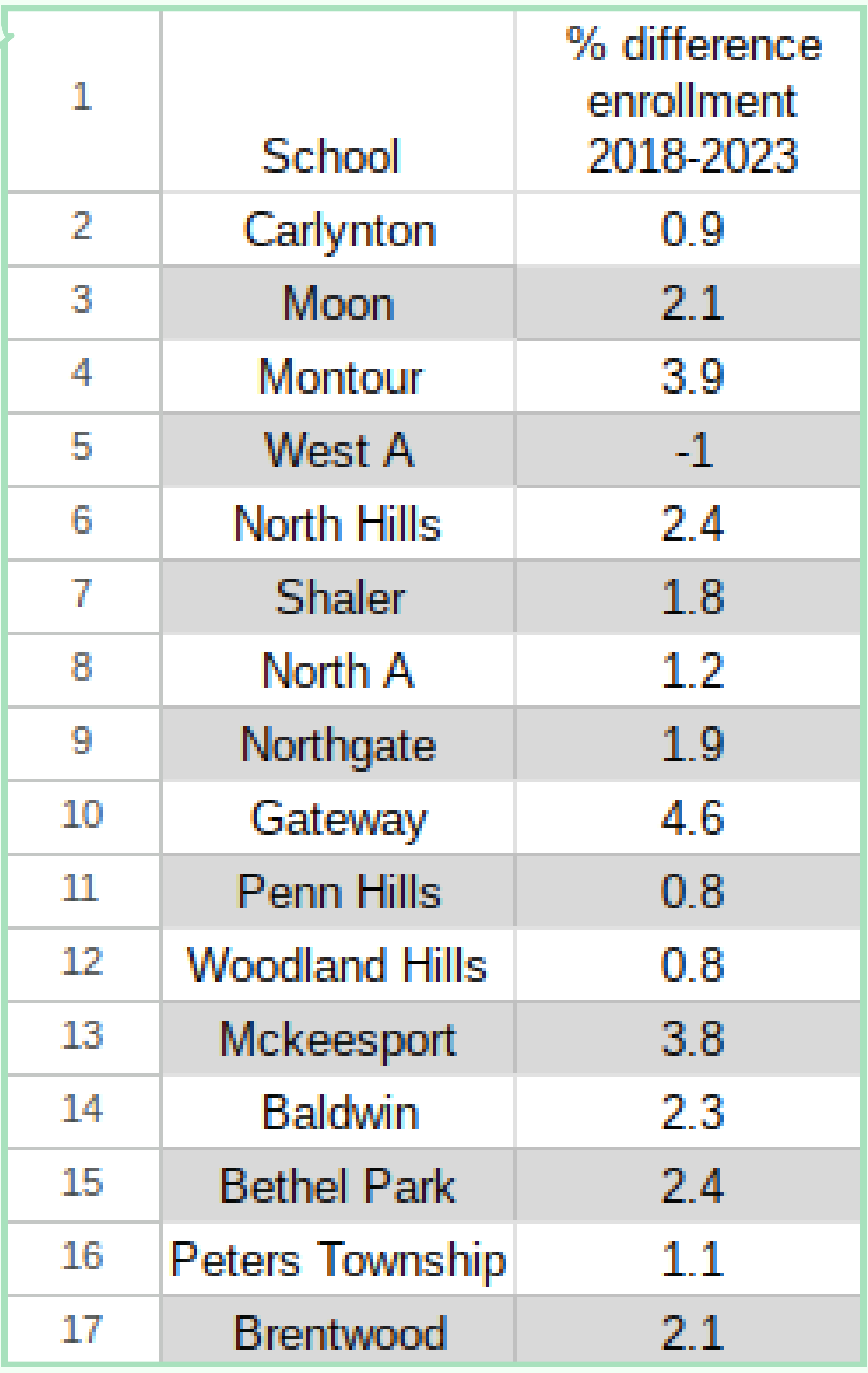
Special Education and its Expenditures: did COVID cause a difference?

Carlynton HS - Sean Hart, Ella Snyder, Bailey Vetter

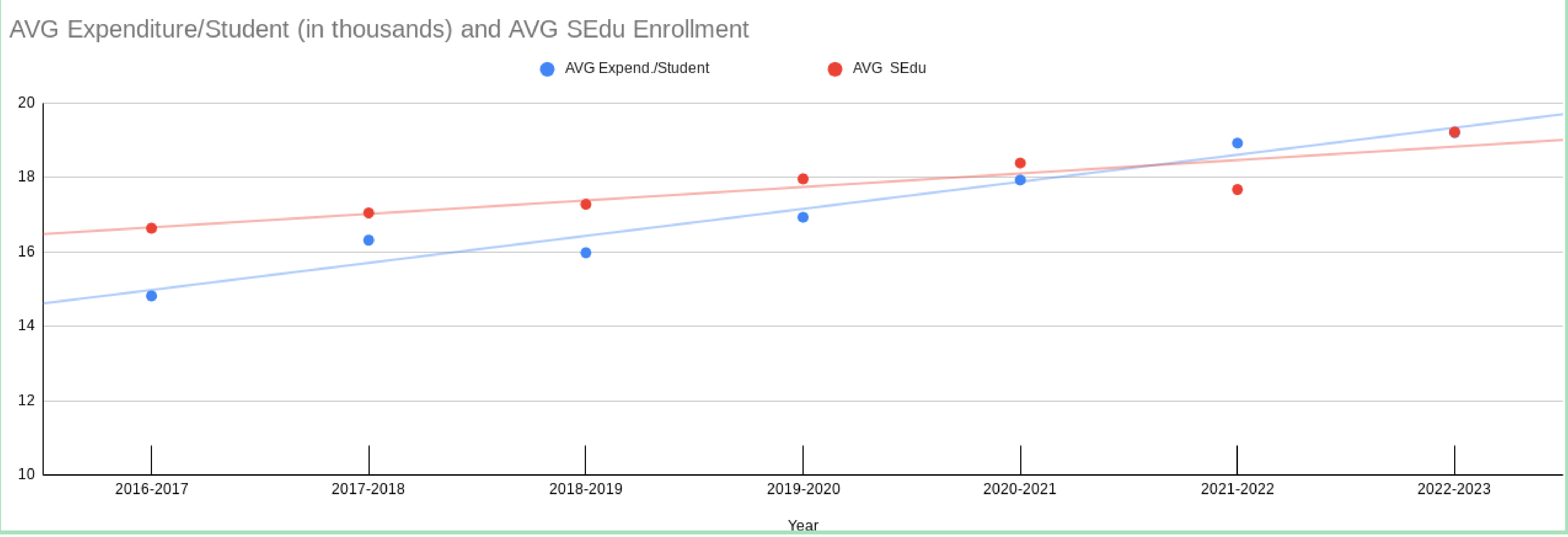
Problem: How has registration in special education programs in Pennsylvania been affected by changes in public school budgeting due to the pandemic?

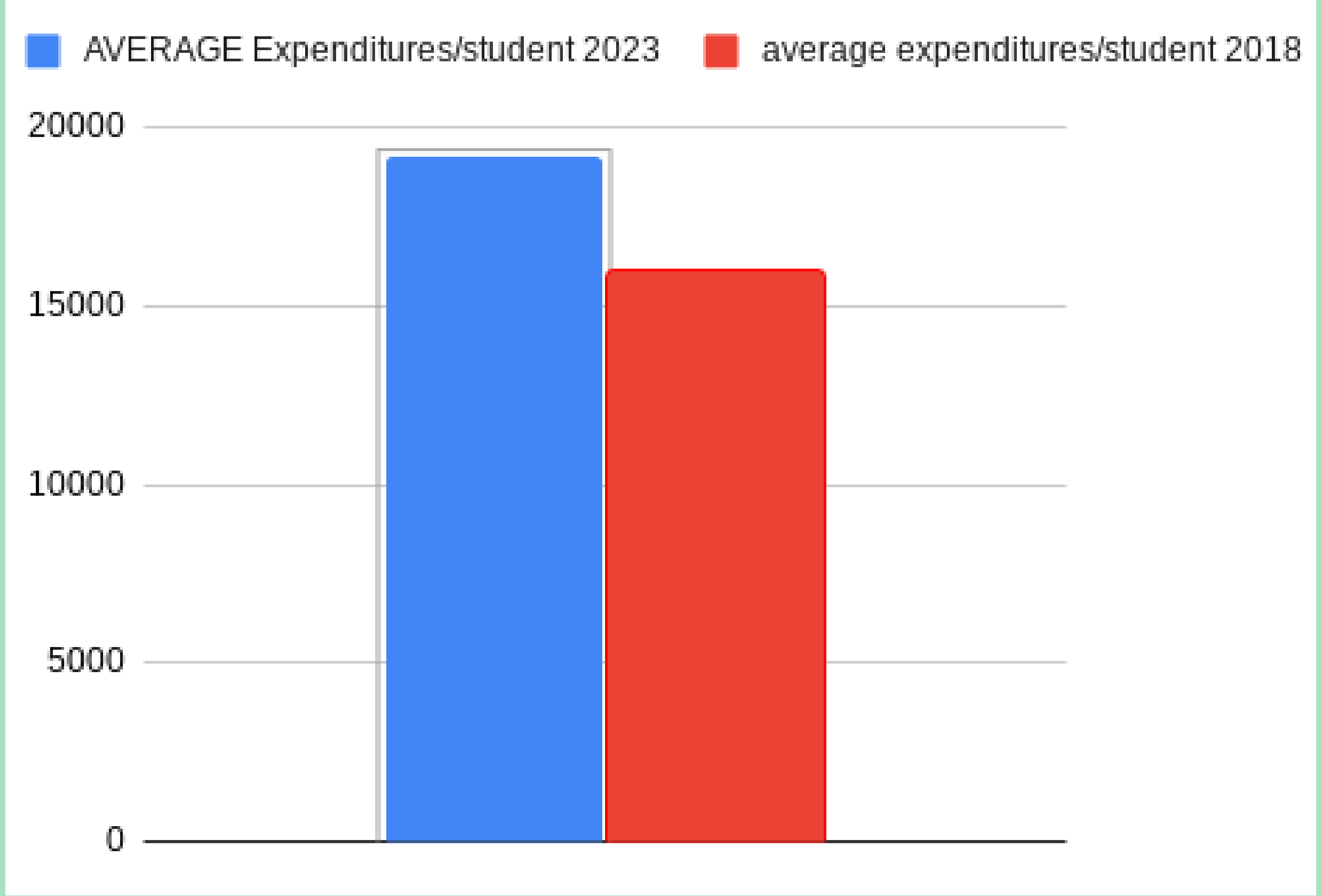
Why is it important?: While the subject may seem inconsequential at first, understanding the underlying causes behind visible shifts in trends among youth--especially in the education sector--holds crucial importance. The impact is not just limited to students with special needs; it extends to all students, prompting educators to adjust, or even redefine, their methods in providing effective education.

Hypothesis: The pandemic has depleted the special education program budgeting despite already-present upward trends in enrollment.



1	School	% difference enrollment 2018-2023
2	Carlynton	0.9
3	Moon	2.1
4	Montour	3.9
5	West A	-1
6	North Hills	2.4
7	Shaler	1.8
8	North A	1.2
9	Northgate	1.9
10	Gateway	4.6
11	Penn Hills	0.8
12	Woodland Hills	0.8
13	Mckeesport	3.8
14	Baldwin	2.3
15	Bethel Park	2.4
16	Peters Township	1.1
17	Brentwood	2.1





Analysis Plan:

Utilizing a sample of Allegheny County public schools , we developed a time series chart to illustrate increases in SEdu enrollment percentage and average expenditures/student. To confirm whether these changes were statistically significant, we performed hypothesis tests exploring if average expenditures/student increased and if average SEdu enrollment percentage increased. Both tests were performed with a p-value approach. The enrollment trends were analyzed from 2016-2023 and the funding compared the 2018-19 year and the 2022-23 year

Datasets:

We primarily used the PVAAS website to find the special ed and total enrollment numbers for all the schools throughout the years. The state of Pennsylvania compiles all school General Fund Budget data within the Department of Education website. From those datasets, each school district's expenditures on special education can be found either within a spreadsheet or as a PDF of each district's budget.

Challenges:

- **One challenge was finding exactly how much the school district budgeted for special education.**
- **One figure we found was local special ed funding, and there was a different figure for actual expenditures.**
- **Some schools had itemized budgets that we had to analyze.**

Summary and Conclusion: Our two tests showed that while the percent of students enrolled in SEdu per school is trending upward, the average funding per student has remained consistent. Based on our sample, the school districts in Allegheny County have adequately kept up with SEdu funding in the post-pandemic era regardless of enrollment increases.

THE EFFECTS OF HIGH SCHOOL ACADEMIC FACTORS ON STUDENTS' POST-GRADUATION PATHS

Central Dauphin HS - Nguyen Ngo, Jasmin Echeverria, McKenna Mumma, Lauren Waldner, Anya Damdin

Question & Hypothesis

How do scholarships, IEPs, gifted programs, 504 plans, class ranks, cumulative weighted GPAs, credits earned/attempted, and total advanced classes affect the post-graduation placements of students at Central Dauphin High School (CDHS)?

<u>Hypothesis</u>: School performance key metrics predict initially stated post-graduation path.

H_o: $B_x = 0$

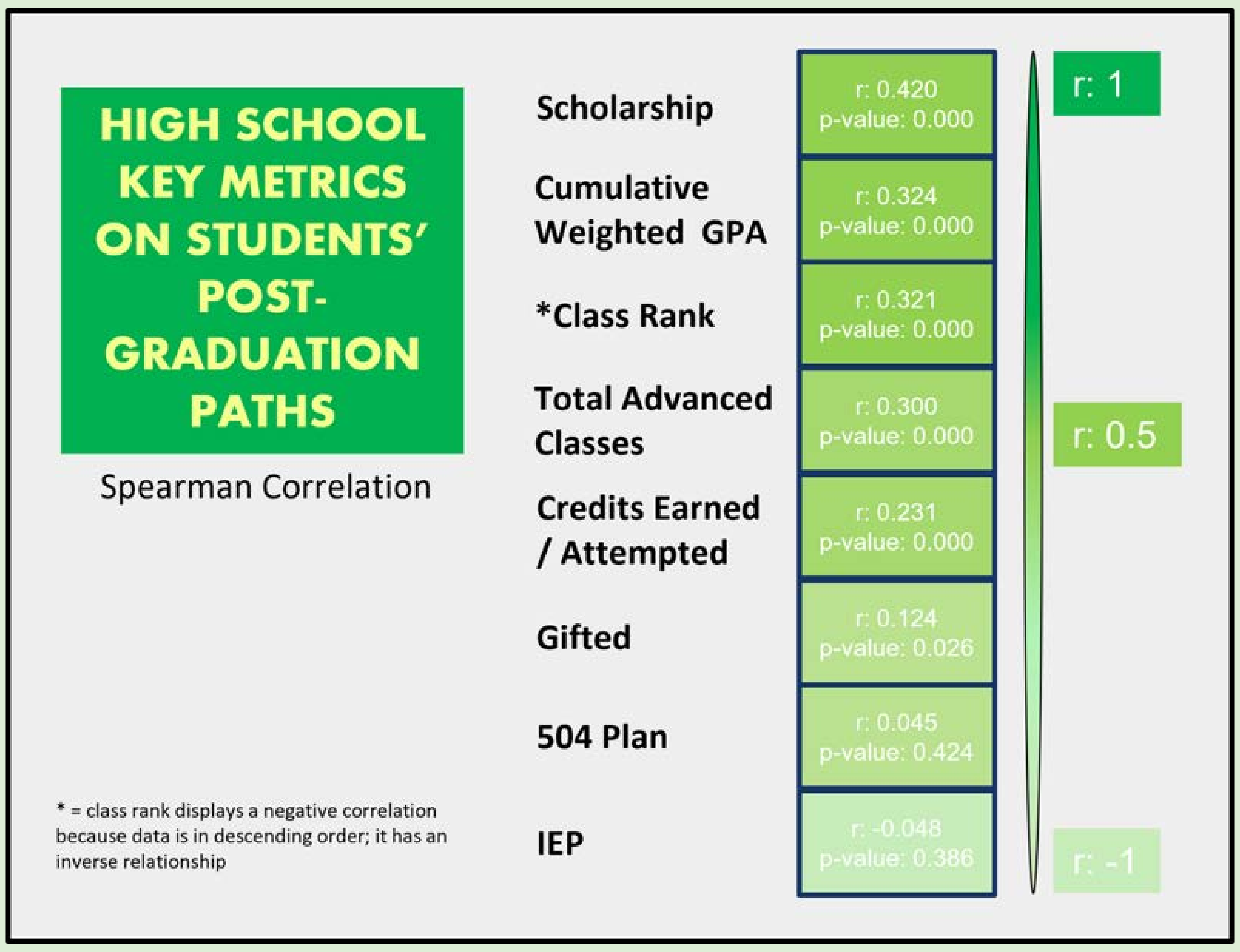
H_a: $B_x \neq 0$

Where x represents each predictor.

Data Sets

Our school provided us with data from the 2023 school year's graduating senior class. The data came from a survey that students could optionally fill out. Our teacher blinded the data to protect the privacy of the students.

Results



HIGH SCHOOL KEY METRICS ON STUDENTS' POST-GRADUATION PATHS

Spearman Correlation

Scholarship	r: 0.420 p-value: 0.000
Cumulative Weighted GPA	r: 0.324 p-value: 0.000
*Class Rank	r: 0.321 p-value: 0.000
Total Advanced Classes	r: 0.300 p-value: 0.000
Credits Earned / Attempted	r: 0.231 p-value: 0.000
Gifted	r: 0.124 p-value: 0.026
504 Plan	r: 0.045 p-value: 0.424
IEP	r: -0.048 p-value: 0.386

* = class rank displays a negative correlation because data is in descending order; it has an inverse relationship

<u>KEY</u> (based on our data)

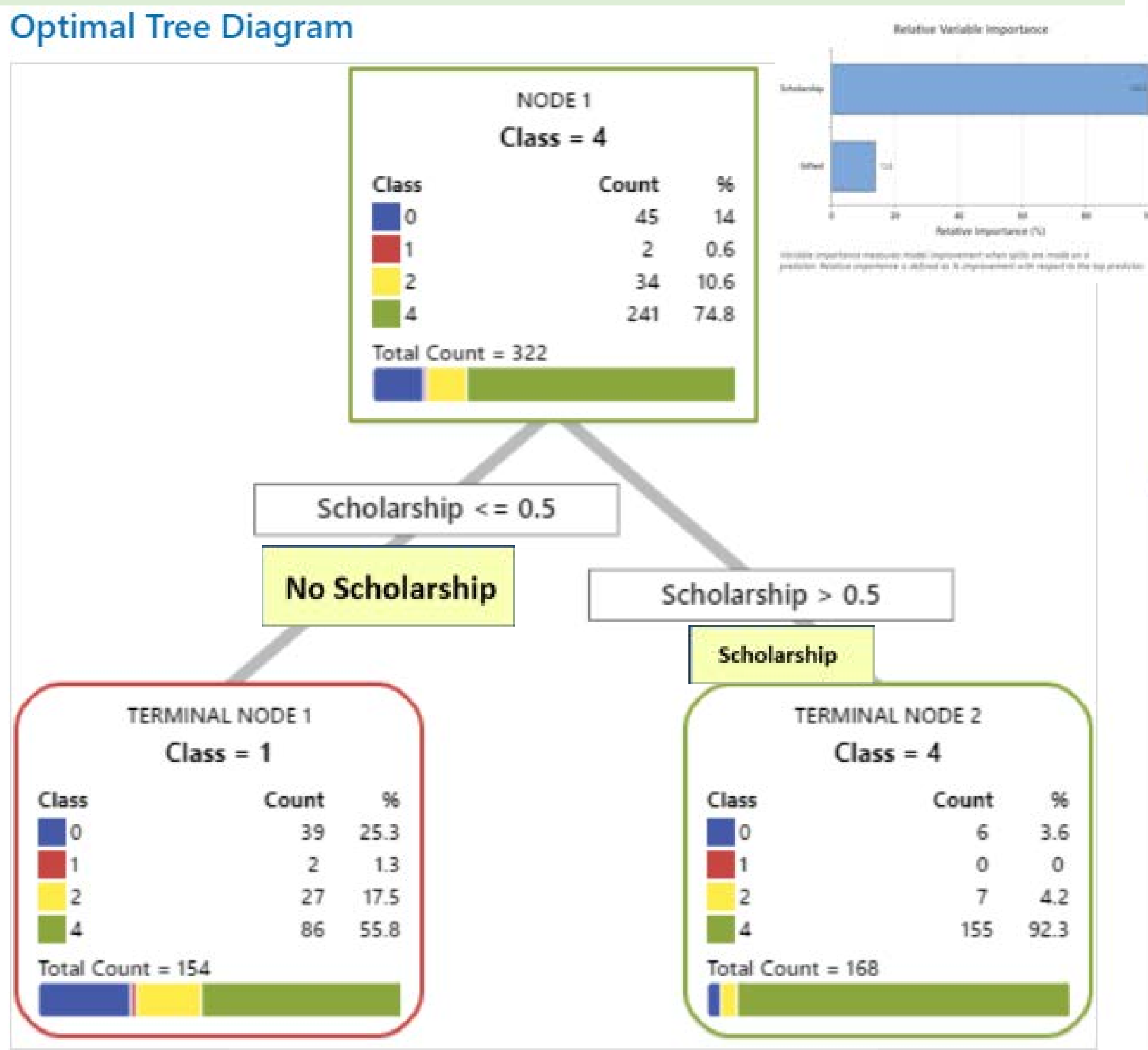
0 – respondents who skipped question, unsure, answered unrelated content

1 – Apprenticeship

2 – Community College

4+ – 4 Years Uni, Military Related Activities

notes: for students who answered at least one answer, the first choice listed was chosen for classification



Optimal Tree Diagram

NODE 1 — Class = 4

Class	Count	%
0	45	14
1	2	0.6
2	34	10.6
4	241	74.8

Total Count = 322

Scholarship <= 0.5 — No Scholarship

TERMINAL NODE 1 — Class = 1

Class	Count	%
0	39	25.3
1	2	1.3
2	27	17.5
4	86	55.8

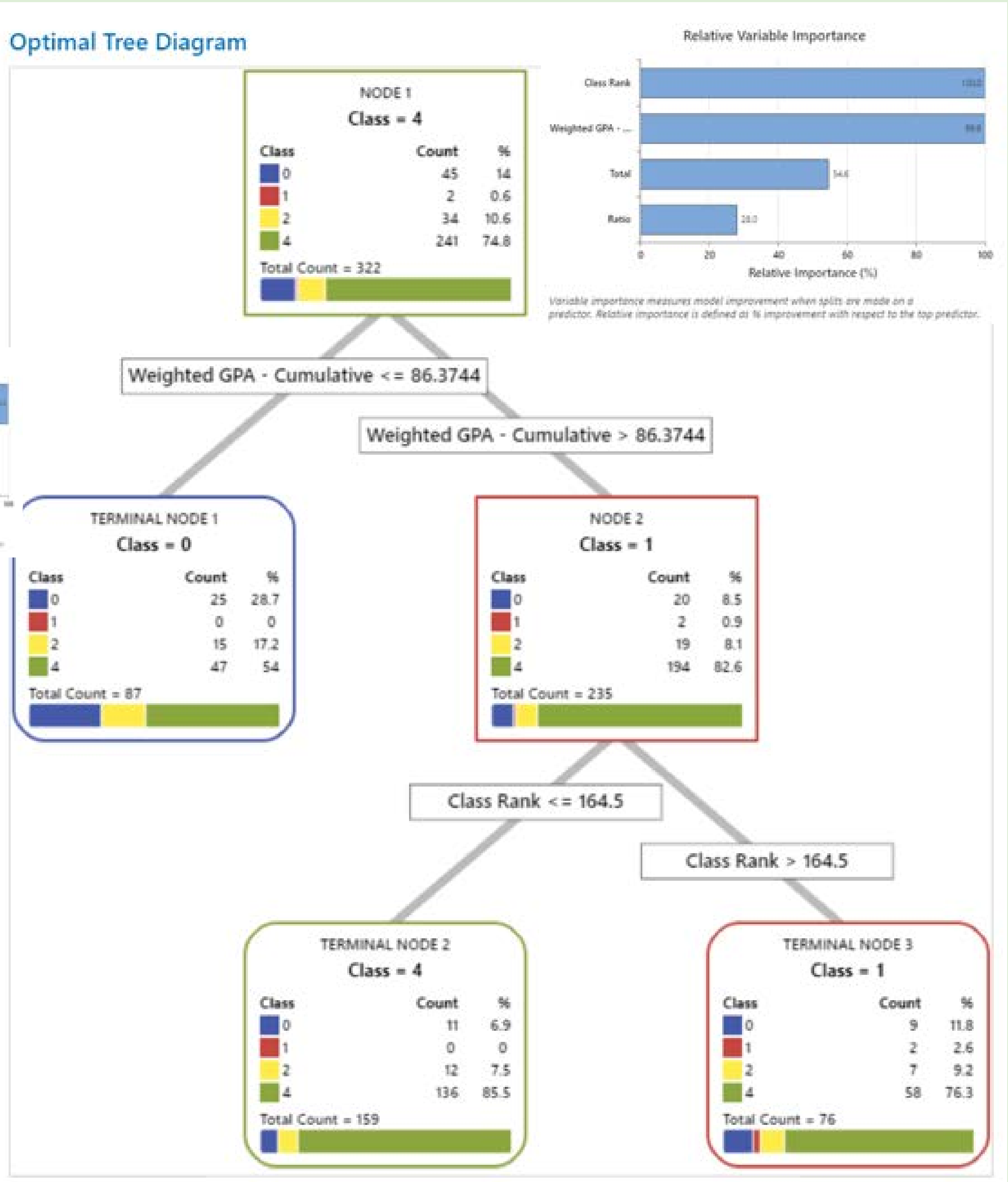
Total Count = 154

Scholarship > 0.5 — Scholarship

TERMINAL NODE 2 — Class = 4

Class	Count	%
0	6	3.6
1	0	0
2	7	4.2
4	155	92.3

Total Count = 168



Optimal Tree Diagram

NODE 1 — Class = 4

Class	Count	%
0	45	14
1	2	0.6
2	34	10.6
4	241	74.8

Total Count = 322

Weighted GPA - Cumulative <= 86.3744

TERMINAL NODE 1 — Class = 0

Class	Count	%
0	25	28.7
1	0	0
2	15	17.2
4	47	54

Total Count = 87

Weighted GPA - Cumulative > 86.3744

NODE 2 — Class = 1

Class	Count	%
0	20	8.5
1	2	0.9
2	19	8.1
4	194	82.6

Total Count = 235

Class Rank <= 164.5

TERMINAL NODE 2 — Class = 4

Class	Count	%
0	11	6.9
1	0	0
2	12	7.5
4	136	85.5

Total Count = 159

Class Rank > 164.5

TERMINAL NODE 3 — Class = 1

Class	Count	%
0	9	11.8
1	2	2.6
2	7	9.2
4	58	76.3

Total Count = 76

Variable importance measures model improvement when splits are made on a predictor. Relative importance is defined as % improvement with respect to the top predictor.

Methods

1. <u>Data Sorting</u>:
   a. We classified the types of post-graduation paths using the key below
   b. Found the ratio of credits earned / credits attempted, found the total of advanced classes taken by reorganizing courses into categories of STEM, ELA, Social Studies, Language, Others
   c. For analysis, we used Minitab to transform binary responses (YES/NO) into (1/0) respectively
   d. Used Google Sheets to create visual representations of the data
2. <u>Data Analysis in Minitab</u>:
   a. Used the Anderson-Normality Test to test normality (95% level significance) for each predictor ($H_o = 0$; $H_a \neq 0$)
   b. Used Spearman Correlation because:
      i. <u>Assumptions</u>:
         ☑ 1. Ordinal, Ratio Variables
         ☑ 2. Paired Observations
         ☑ 3. Unsure of a Monotonic Relationship
         ☑ 4. No Assumption of Normality
   c. Used Cart Classification (Decision Tree) to demonstrate the "if-then" relationships between academic factors and the post-graduation types by creating nodes (groups) with similar characteristics that may have an effect on our response
      i. The Relative Variable Importance ranks the level of importance each predictor has which corresponds to our findings in the matrix (e.g. Cumulative Weighted GPA is 99.9% as important as Class Rank)

Challenges

1. <u>Bias</u>: There was a reporting bias where ONLY 69.4% students that were a part of the senior class of 2023 filled out the survey which causes skewness.
2. <u>Analysis</u>: Since our data did not pass the the normality test and it displayed outliers, statistical testing options were limited.

Conclusion

There is sufficient evidence to conclude that Scholarship, Cumulative-Weighted GPA, Class Rank, Total of Advanced Classes taken, Credits Earned, and Gifted Program can predict students' post-graduation path with an alpha level of 0.05 (p-values: 0.000, 0.000, 0.000, 0.000, 0.000, 0.026, respectively). On the other hand, there is little-to-no relationship with IEP and 504 Plans on students' post-graduation paths. With a boost in GPA and Class Rank, it can be seen from the decision tree that the percentages of students pursuing a 4 years university or military related activities are slightly higher (GPA: 82.6% > 54%; Class Rank: 85.5% > 76.3%) . In addition, students could possibly be sure of their post-graduation paths more if changes in our school district occur (as suggested to the right). With Scholarship, if more are offered to students, the likelihood of them pursuing a higher type of education is higher (99.3% > 55.8%). Adding on to that, the percentage of unsure students may decrease if more scholarship opportunities are given (25.3% down to 3.6%).

Suggestions

As our data indicates, Cumulative-Weighted GPA and Scholarship have the strongest correlation to post-graduation paths students at Central Dauphin take. To increase the number of students who receive scholarships, the College and Career Center at CDHS can better promote local scholarships and provide more inclusive opportunities for students to apply. It is important to note that scholarship offers usually require transcripts. So, as for increasing students' GPA and Class Rank, this can be improved by re-evaluating grading policies at our district as well as promoting office hours and tutoring programs for students.

$ MAJOR MOOLAH: NAVIGATING DEGREES AND DOLLAR SIGNS $

Keystone Oaks Team #1

Tina Tran & Naudia Booker

CONTEXTUALIZATION

The question every little kid is asked is what do they want to be when they grow up. Most kids answer with I want to be a rock star! I want to be a firefighter! I want to be a vet! As the kids turn into young adults they begin to question which occupation they want to pursue and they begin questioning how much money will this job make. Money plays one of the most important rules in our day to day life and we aim to predict which job will allow us to make the most money we can.

HYPOTHESIS

Qualitative research question: What are the most important factors to predict if you are a high earner?
Hypothesis: The number of years of education is the most influential in predicting if someone's a high earner.
Quantitative research question: What are the most important factors to predict someone's median salary out of college?
Hypothesis: Majoring in a STEM related field is the best indicator of an above average median salary.

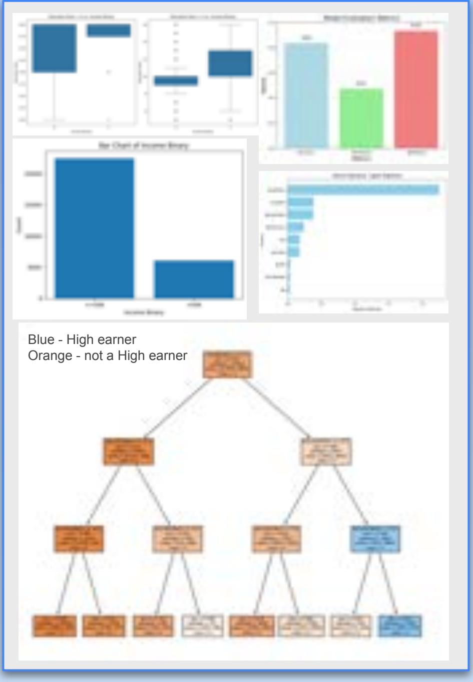
DATA SETS

We used data sets from the US Census Bureau(1990s) in order to conduct research while using a credible source. We also used kaggle.com(2010s) which is a reliable source.

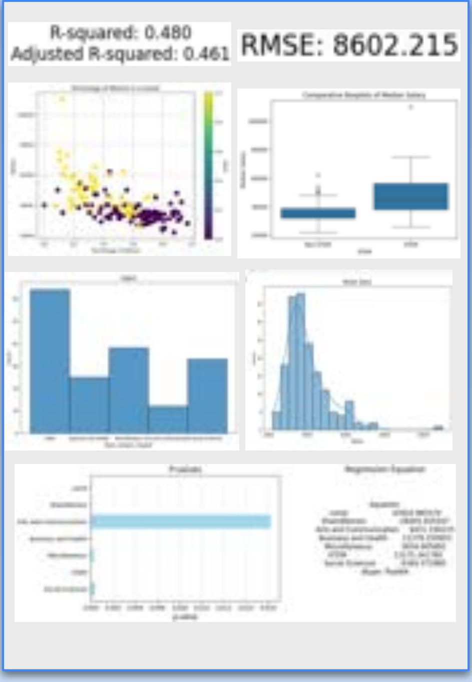
CHALLENGES

- Our initial dataset was limited to more labor intensive careers and did not help answer our question(s). Our initial intention was to focus on more college oriented occupations.
- Preparing the data for modeling was time consuming when deciding how we wanted to feature engineer certain variables like Country of Origin and Major Types.
- Variable Selection when fine tuning both our Linear and Logistic Regression models .
- Making our Graphs more efficient (choosing the right 3 variable combinations) .
- Attempting additional model solutions such as Decision Trees to improve on our Logistic Regression Accuracy.

QUALITATIVE



QUANTITATIVE



ANALYSIS

- According to our data set involving majors and their median salary the major with the lowest percent of women make more money than majors with a higher percent of women pursuing that field.
- Our p-values shown on the left indicate a strong correlation between majors and their median salary since the p value is less than .05. For example being a man has no relation to what your median salary will be because the p-value is greater than .05 so there is no significant evidence to prove it has an impact.
- Therefore college is worth the money if you major in a field with less women in it while if you want to be a high income earner your marriage status and years of education are the biggest indicators.

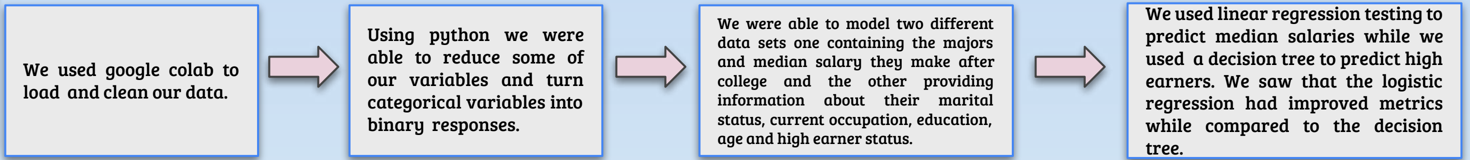
RESULTS SUMMARY

Fields with a low percentage of women exhibit a statistically significant association with higher post-college earnings, as indicated by p-values below the standard significance level of 0.05. Additionally, the dataset highlights a robust and significant relationship between years of education and marital status in predicting higher salaries.

RECOMMENDATIONS

High School students and parents can use this knowledge to make a more cognizant decision towards their future. Pursuing a higher education within a STEM related field shows the greatest relationship between being a high earner.

OUR PROCESS



We used google colab to load and clean our data.

Using python we were able to reduce some of our variables and turn categorical variables into binary responses.

We were able to model two different data sets one containing the majors and median salary they make after college and the other providing information about their marital status, current occupation, education, age and high earner status.

We used linear regression testing to predict median salaries while we used a decision tree to predict high earners. We saw that the logistic regression had improved metrics while compared to the decision tree.

IN-TUITION

KEYSTONE OAKS TEAM #2

EXAMINING THE FACTORS OF COLLEGE COST IN PA

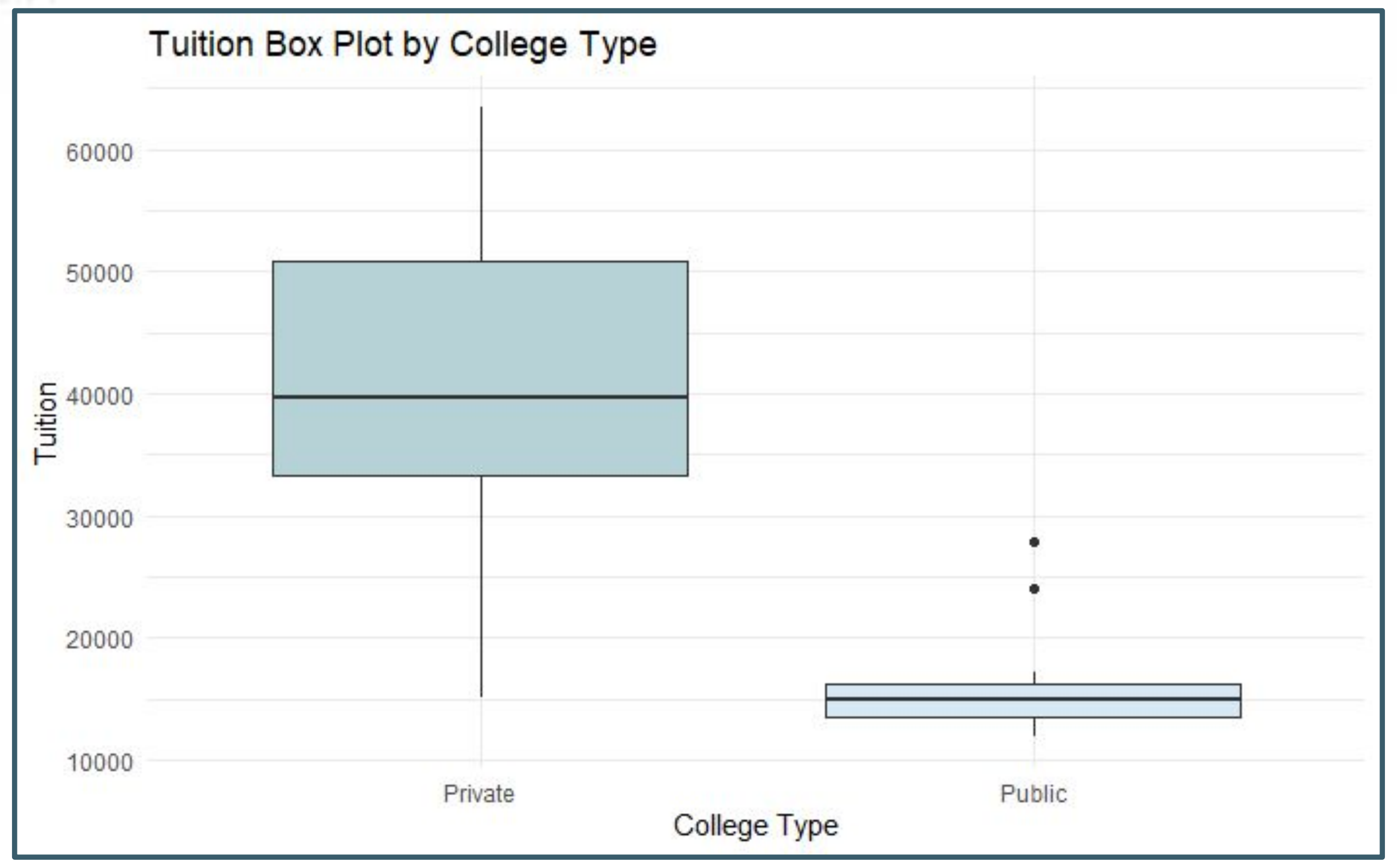
1 - QUESTION

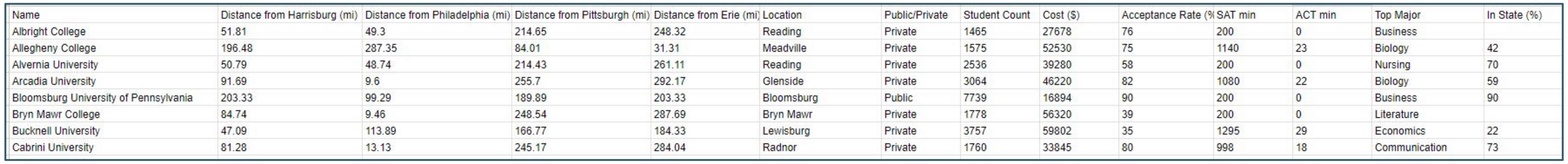
WHAT FACTORS OF COLLEGES BEST PREDICT TUITION COST?

We believe location and the school-type contribute most to tuition cost.

2 - OUR DATASET

Our dataset consists of 90 colleges and their location, school-type, student population, cost, acceptance rate, SAT and ACT min., top major, and percentage of in-state students. We found this information via /mycollegeselection.com

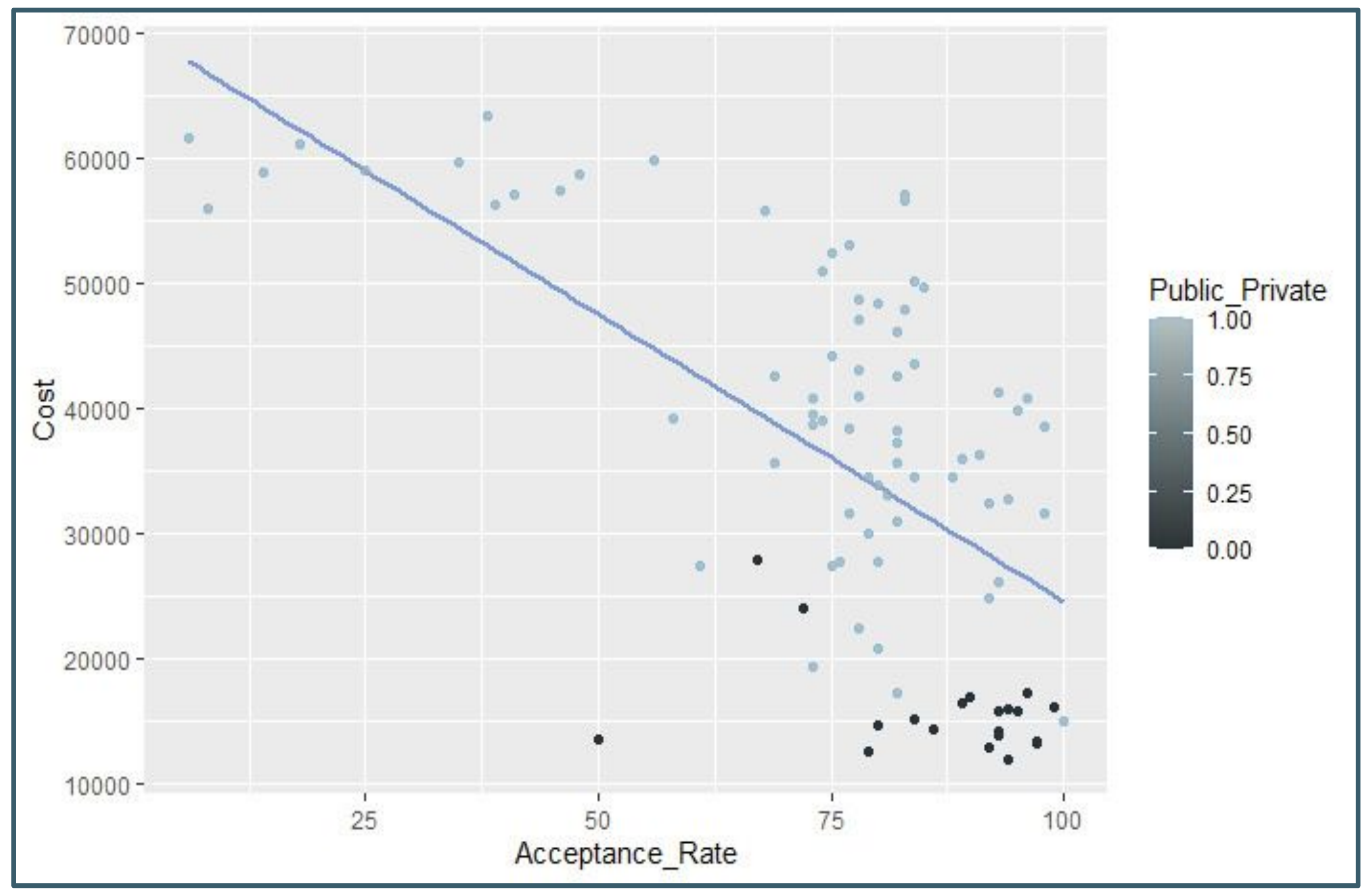




Name	Distance from Harrisburg (mi)	Distance from Philadelphia (mi)	Distance from Pittsburgh (mi)	Distance from Erie (mi)	Location	Public/Private	Student Count	Cost ($)	Acceptance Rate (%	SAT min	ACT min	Top Major	In State (%)
Albright College	51.81	49.3	214.65	248.32	Reading	Private	1465	27678	76	200	0	Business	
Allegheny College	196.48	287.35	84.01	31.31	Meadville	Private	1575	52530	75	1140	23	Biology	42
Alvernia University	50.79	48.74	214.43	261.11	Reading	Private	2536	39280	58	200	0	Nursing	70
Arcadia University	91.69	9.6	255.7	292.17	Glenside	Private	3064	46220	82	1080	22	Biology	59
Bloomsburg University of Pennsylvania	203.33	99.29	189.89	203.33	Bloomsburg	Public	7739	16894	90	200	0	Business	90
Bryn Mawr College	84.74	9.46	248.54	287.69	Bryn Mawr	Private	1778	56320	39	200	0	Literature	
Bucknell University	47.09	113.89	166.77	184.33	Lewisburg	Private	3757	59802	35	1295	29	Economics	22
Cabrini University	81.28	13.13	245.17	284.04	Radnor	Private	1760	33845	80	998	18	Communication	73

3 - ANALYSIS

We found that school type, acceptance rate, and student enrollment are the most important, each having p-values of less than 0.05, meaning they have a significant impact on our multivariate linear model. We also found that STEM majors and the distance from Harrisburg have a p-value of about 0.06, which are not ideal, but we kept them in our final model. Location and SAT scores both had high p-values, so we removed them from our final model. Our full model has an r^2 of 69% with a $8,980 residual standard error (RSE). After minimizing to the five best, the r^2 increased to 70%, and the RSE decreased to $8,313. We believe we have a jumping point to further expand upon this topic.



BRAINSTORM

We met together and shared ideas for possible projects, and ended up with college (since most of our members are thinking about college).

FIND DATA

We spent severely days manually imputing data from *My College Selection*. Everyone contributed to make this process more efficient.

HYPOTHESIZE

While inputting data, we thought about the results of our project, and everyone was confident that our hypotheses would be true.

CODING

We used R to create box plots, bar graphs, scatter plots, and linear regression models.

FINALIZE

Finally, we took our findings and interpreted them. Our models turned out well, making this process easier than anticipated.

RESULTS/SUMMARY

4 CHALLENGES

We faced challenges including insufficient data, lack of information for certain satellite campuses, and quantifying location data. We also feature-engineered and manually imputed the distances from the 4 main cities in PA. This was extremely time consuming and frustrating, since those features ended up not being influential for the dataset.

5 RECOMMENDATIONS

We recommend that those looking to go to college use this data to better inform their college decisions, especially as it pertains to cost, acceptance rates, and majors.

6 NEXT TIME

It would be interesting to look into other schools, rather than just PA colleges. We could also find more data on each college too, like student age, gender, ethnicity, etc. Predicting acceptance rate would be another aspect we'd like to explore as well.

Lia Scott, Alisha Thapa, Sierra D'Eramo, Madison Williams-Powell, Lael Nowlin, Katherine Cesario, Anthony Cerminara

Does healthcare accessibility impact the prevalence of Long COVID in a state?

How does this apply to San Diego's Post-COVID conditions?

Team North San Diego

Rajan Tavathia, Dylan Cairns, Benjamin Lee, Sumana Nandipati, Anika Tiperneni, Neha Srinivasan

Problem Statement and Hypothesis

Problem Statement: Do indicators of healthcare accessibility correlate with the prevalence of Long COVID in a state?

Alternative Hypothesis

Long COVID prevalence will be higher in regions with lower access to healthcare, as measured by five leading metrics.

Null Hypothesis

The indicators of healthcare accessibility will not correlate with the rate of Long COVID in a state.

Background

- **<u>Long COVID is a complex, potentially debilitating series of symptoms which continue or develop after COVID-19 infection</u>**.
- As **<u>limited data</u>** exists for Long COVID in San Diego, we decided to analyze whether there was a correlation between healthcare accessibility in a region and Long COVID prevalence. This way, we could recommend certain regions of San Diego county to be targeted for **<u>epidemiological surveillance and data collection efforts</u>**.
- Regardless of the findings of the project, we could make suggestions about the nature of Long COVID in the United States. If we found a correlation between healthcare accessibility and Long COVID, one could ultimately conclude that **<u>Long COVID may be preventable with reasonable clinical intervention efforts</u>**.

Datasets

- **County Health Rankings (for Healthcare Accessibility Date):** https://www.countyhealthrankings.org/health-data
- **U.S. Post COVID conditions/Long COVID data from HealthData.gov:** https://healthdata.gov/dataset/Post-COVID-Conditions/dx42-bzzu/about_data
- We compiled relevant data in an Excel sheet: https://docs.google.com/spreadsheets/d/1MG1oB1RQbyeGrPLFjvbWY4JVDN_Tsyoxy7pBlZL2ua8/edit?usp=sharing

Methodology

- We compiled data of **<u>5 leading indicators</u>** of the level of access to clinical care in a state: the percentage of uninsured adults, the number of adults per primary care provider (i.e. PCP density), the number of preventable hospital stays per 100,000 people, the percentage of women receiving mammography screening, and the percentage of adults receiving a flu vaccine.
- A higher value in some of these factors indicate lower healthcare accessibility, while some are directly related to healthcare accessibility.
- This was necessary as we wanted to analyze a potential cause of Long COVID, and not accounting for the rate of COVID cases in each state previously would create a confounding variable.
- **<u>We employed MATLAB and Python tools to analyze our data and generate visualizations, and employed further statistical tests to confirm our findings. This included linear and multiple regression, ANOVA, and Pearson correlation.</u>**

Findings (Linear Regression)

- We used linear regression to analyze how much of the variance in Long COVID rates could be attributed to each of the indicators of healthcare accessibility. An example analysis for percentage of insured adults is below:

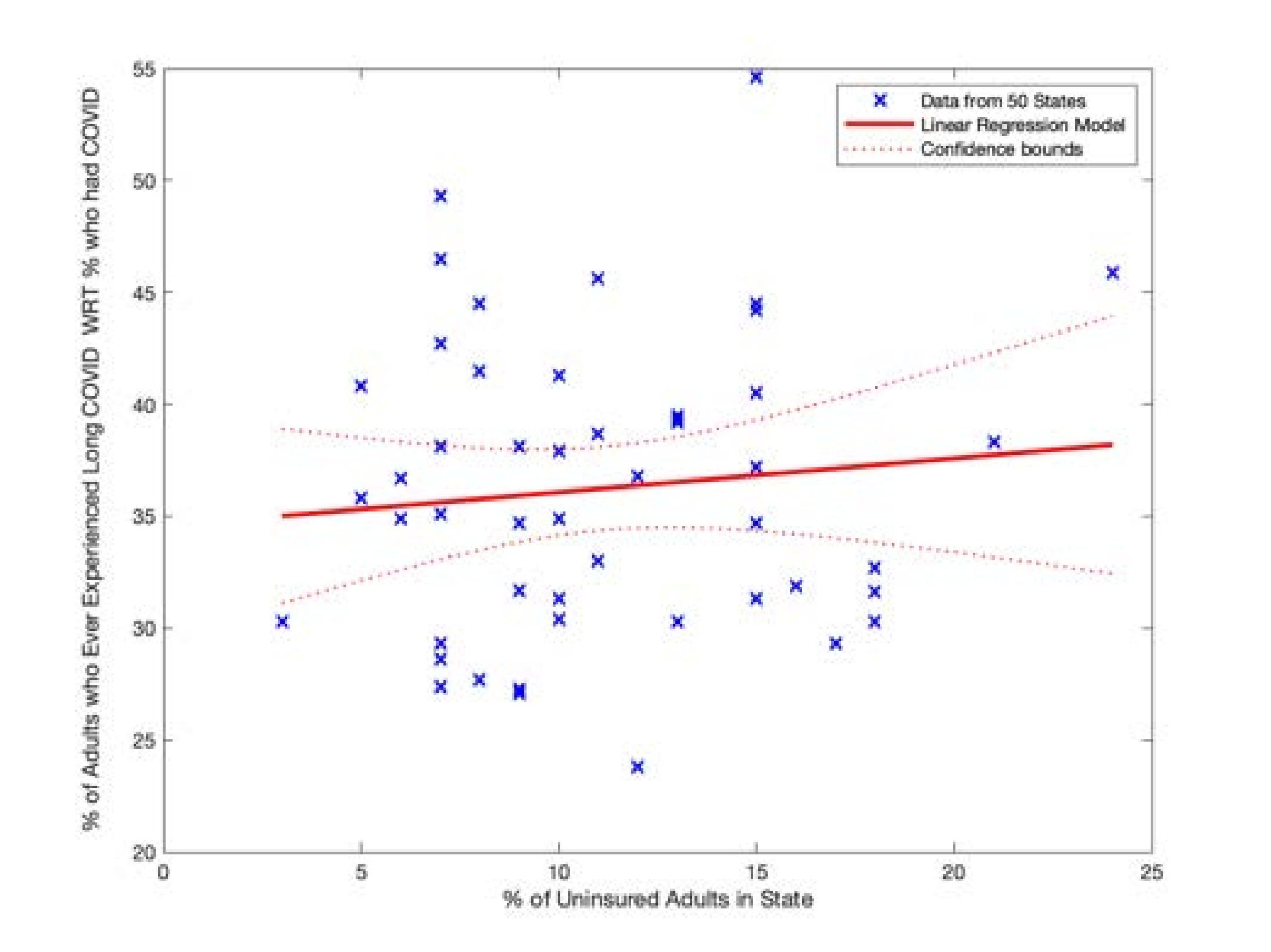
- We found that the Percentage of Uninsured Adults in Each State only explained 1.07% of the variance, or change, in Long COVID rates. An ANOVA test was used to determine whether this was statistically significant different from zero. Using an F value of 0.52, we calculated a p-value of 0.475. Since this was greater than the significance level of 0.05, we failed to reject our null hypothesis and found that the statistical relationship between the Percentage of Uninsured Adults in a state and Long COVID prevalence was not significant as the coefficient of regression was not statistically different from 0.

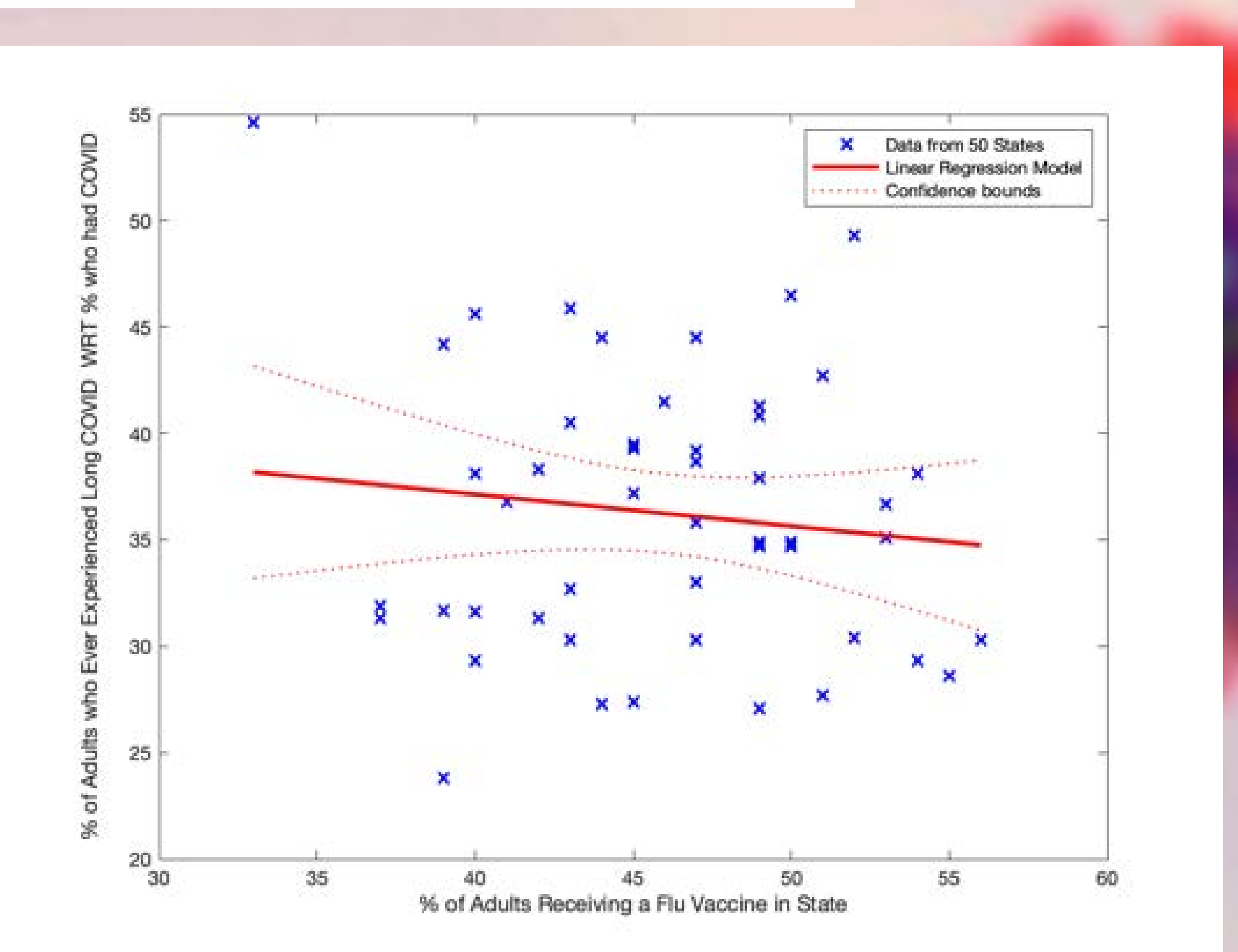
- A residual plot for this data demonstrated homoscedasticity, and thus a linear regression model was appropriate.

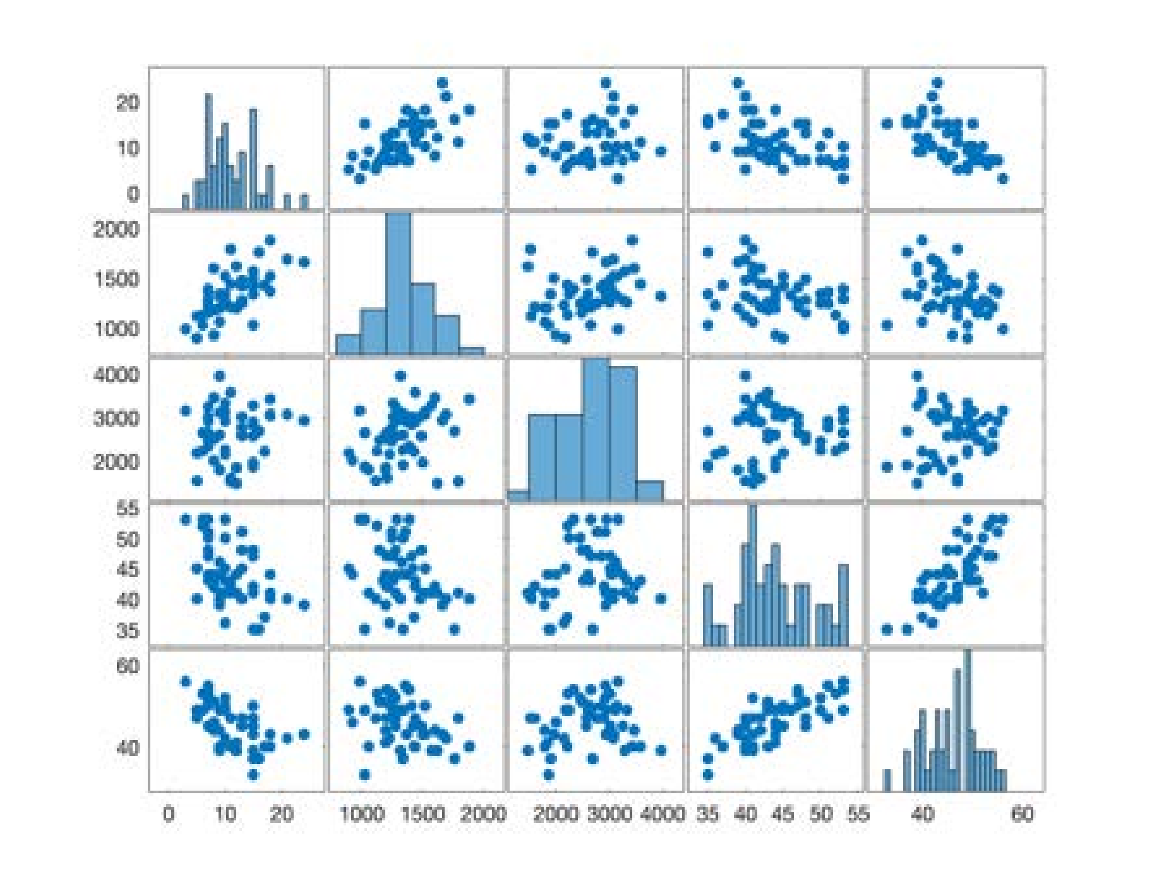
- A Pearson correlation test was ran on the data and we found that the correlation coefficient r=0.1 was not statistically different from 0, and thus the low, positive correlation was not significant enough to establish a trend between the percentage of uninsured adults in a state and Long COVID prevalence.

- **<u>ALL 5 indicators of healthcare accessibility demonstrated no statistically significant correlation or regression coefficient, and thus we failed to reject the null hypothesis in all test cases. However, non-significant correlation were seen in all of the indicators which rejected our null hypothesis.</u>**

Visualizations







- Top left: Regression model of % of Uninsured Adults in State versus Long COVID Prevalence
- Top right: Regression model of % of Adults with Flu Vaccination in State versus Long COVID Prevalence
- Bottom left: Python correlation chart between each of the variables of healthcare accessibility. Visually, we saw that most of the variables were somewhat correlated and thus could be used to indicate healthcare accessibility.

•A multiple linear regression test was ran to establish whether the indicators of healthcare accessibility collectively impact ed the rate of Long COVID in a state.

•We found that the indicators of healthcare accessibility in each state only explained 8.65% of the variance, or change, in Long COVID rates. We employed an ANOVA test to determine whether this was statistically different from zero, indicating that the IV of indicators of healthcare accessibility had no impact on the DV of Long COVID prevalence. Using an F value of 0.83, we calculated a p-value of 0.533. Since this was greater than the significance level of 0.05, we failed to reject our null hypothesis and found that there was no statistical relationship between indicators of healthcare accessibility and Long COVID prevalence as the coefficient of regression was not statistically different from 0.

Conclusion and Recommendations

- We failed to reject our null hypothesis, and found that healthcare accessibility does not affect the prevalence of Long COVID in a region. Thus, Long COVID is likely only related to the number of COVID-19 cases in an area, and since San Diego does not have a very high incidence of COVID-19 currently, Long COVID research efforts should not be specifically directed to our community pending further research.

The Severity of Accidents Involving Large Passenger Vehicles in the US

Studying automotive accidents involving larger passenger vehicles as opposed to smaller vehicles in relation to the reported severity of injury to passengers, with secondary analysis on the aforementioned relation at vvarying speeds.

Norwin High School: Aaron Berger, Dmitri Berger, Simone Pal, Rex Wu

Background

In a nation as geographically expansive as the United States, passenger vehicles are inherently an integral component of modern society. In recent years, however, the particularities of American automotive culture have generated controversy, with regard to larger passenger vehicles. That larger vehicles, such as light trucks and vans/minivans, now comprise a majority of passenger vehicles on the road and a supermajority of passenger vehicle sales has drawn questions as to the safety of larger vehicles on the road, for motorists and pedestrians alike. In this study we will analyze the relationship between the proportion of high-severity vehicular accidents (2015) involving exactly two vehicles and the involvement or absence of large passenger vehicles at varying accident speeds. We will also analyze whether or not there is an appreciable difference in severity of injury in accidents with and without large passenger vehicles. For the purpose of this analysis, large passenger vehicles are classified as either "passenger trucks," or "vans/minivans." No larger vehicles, such as straight-unit trucks or buses, were included in the analysis. Injury was measured on the KABCO scale, in this study limited to from 0 to 4, 4 being fatal injuries.

Research Question

Is there a significant relationship between the presence or absence of large passenger vehicles in accidents to the proportion of such accidents with high severity of injury, and how does this relationship manifest at varying speeds?

Hypotheses

We predicted that accidents involving large passenger vehicles would have a greater propensity for high severity and fatality to passengers than accidents with such vehicles absent, at equal speeds.

Data Sources

National Accident Database 2015:
https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/GES/GES15/
Motor Vehicle Registrations:
https://www.fhwa.dot.gov/policyinformation/statistics/2015/mv1.cfm

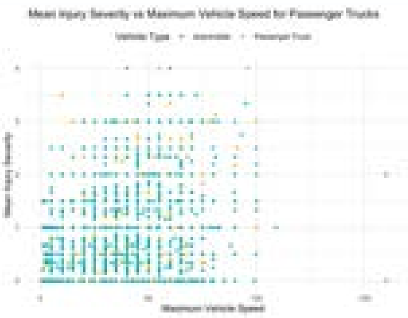
Methodology

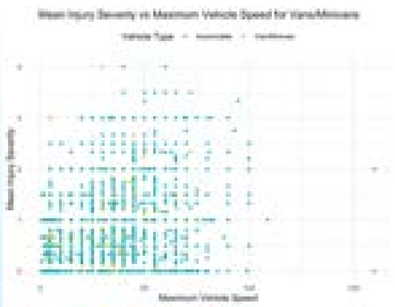
We imported our datasets into MacOS Numbers and used a Python script to remove irrelevant and incomplete data. The script also concatenated our data for each accident, whereafter we imported it into R Studio. The bounds of these analyses covered accidents involving only passenger vehicles, and either only two vehicles or one vehicle and pedestrians. Accidents with unknown information were excluded. We created several models in R Studio, described below.

- Multiple Linear Regression Models
  - Dependent variable: mean injury severity of persons in crash
  - Independent variables: maximum vehicular speed and large or small vehicle, for vehicle types (VT) of large vehicles
  - p-values for large vehicle: Passenger Trucks
    - $p_{speed} < 2(10)^{-16}$, $p_{type} = 3.83(10)^{-4}$
  - p-values for large vehicle: Vans/Minivans
    - $p_{speed} < 2(10)^{-16}$, $p_{type} = .0648$
- Linear Regression Models
  - Dependent variable: difference in proportion of mean serious injuries caused
  - Independent variable: rounded vehicle speed
  - p-values for large vehicle: Passenger Trucks
    - $p = 0.0115$
  - p-values for large vehicle: Vans/Minivans
    - $p = 0.00165$

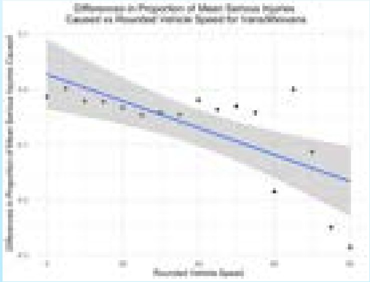
Challenges

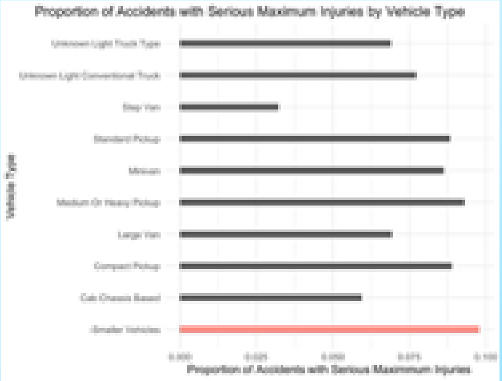
- Our data included vehicle types and classifications that could be inconsistent or outdated. Additionally, state-by-state variations stymied the utility of some sources.
  - Our 2015 motor vehicle registration dataset did not include New Hampshire
  - The NHTSA FARS dataset had classifications inconsistent with the GES dataset
- Our NHTSA GES dataset was separated across multiple datasets (i.e. vehicles involved, people involved, etc. and their accompanying statistics were in separate sheets), making it very unwieldy to work with.

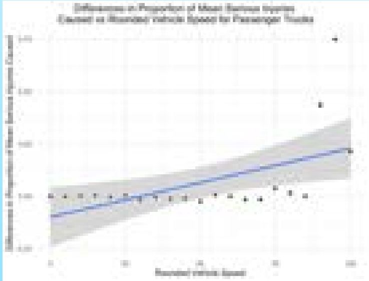




Graphs







Analysis and Conclusion

Taking into account the calculated p-values, we concluded that, at $\alpha = 0.05$, the relationship between the difference in proportion of mean serious injuries caused (between large and small passenger vehicles) and speed is statistically significant, for both passenger trucks and vans/minivans. Additionally, we found that the relationship between the presence of mean serious injuries caused, speed, and the presence of passenger trucks, was statistically significant. However, under the same case but for the presence of vans/minivans is NOT statistically significant.

Our results on injury severity in accidents suggest that passenger trucks are more dangerous to motorists involved in accidents with said vehicles, and that this pattern is consistent across varied speeds. This is in line with our predictions. As for the other category of larger passenger vehicles, we found that while there is a demonstrable difference between vans/minivans and smaller passenger vehicles in injury severity caused in accidents, this relationship does not hold at varied speed. A further topic of research could be exploring the reasons for this difference between vans/minivans and passenger trucks.

Overall, our results show that on the road, larger passenger vehicles pose a not-insignificant threat to the health and safety of other motorists. More research is needed to determine remedies to the dangers posed by larger vehicles, and the reasons behind their proliferation.

NUTLEY HIGH SCHOOL DATA JAM TEAM

Factors Affecting Congenital Heart Disorders in Newborns

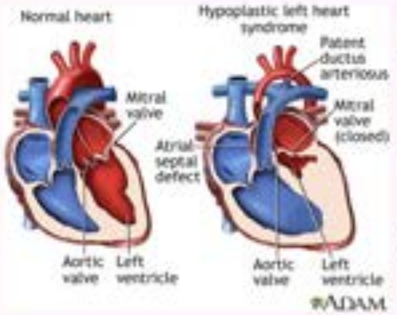
Abigail Puleo, Emily Kean, Camila Loikova, Alexis Fontanilla, Sindi Gjonbocari, Aashi Bhandari

1 PROBLEM

The rates of babies born with Congenital Heart Diseases, such as Hypoplastic Left Heart Syndrome, has grown in recent years and may be attributed with the lifestyle of a mother before and during their pregnancy.

RESEARCH QUESTION

What factors of a mother's lifestyle pre-pregnancy influence the prevalence of Hypoplastic Left Heart Syndrome in newborns?



4 DATA SOURCES

- NJSHAD Website
  - HLHS Data 2010-2019
  - Binge Drinking 2017-2020
- CDC Website
  - Environmental Public Health Tracking Network
- NJ.gov
  - Diabetes Action Plan Report 2013-2015

2 Background

The prevalence of Congenital Heart Disorders in newborns can be influenced by a variety of factors in a mother's pre-pregnancy lifestyle. Understanding these factors is important for prevention, early detection, and possible intervention. Identifying these risk factors, mothers and healthcare providers can implement changes into their lifestyle to reduce the risk of CHDs. It also gives way for the development of educational information for prospective parents. Therefore, maternal lifestyle and its impact on CHDs is crucial for research and enhancing treatment and prevention plans.

3 NULL HYPOTHESIS:

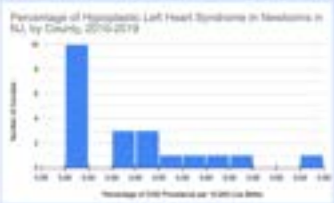
A mother's lifestyle pre-pregnancy will not increase the presence of Hypoplastic Left Heart syndrome in their newborns (no correlation: r = 0).

ALTERNATIVE HYPOTHESIS:

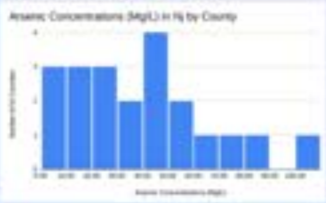
A mother's lifestyle pre-pregnancy will have an impact on the prevalence of Hypoplastic Left Heart Syndrome in their newborns (there is a correlation: r≠0)

We predict that there will be specific factors that directly correlate with the prevalence of Hypoplastic Left Heart Syndrome. For example, we anticipate the factors of a mother's genetic predisposition, measured by diabetes, and a mother's actions during pregnancy, measured by alcohol abuse, can influence the human body and therefore increase how common Hypoplastic Left Heart Syndrome is in children. We would also like to measure how exposure to arsenic may affect the presence of HLHS.

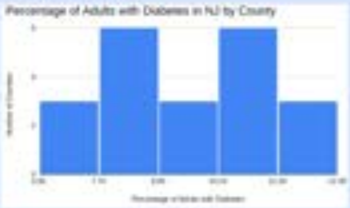
5 DATA DISTRIBUTIONS



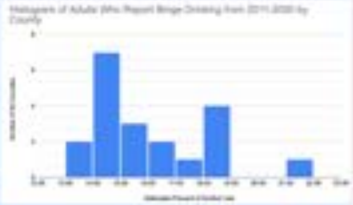
Hypoplastic Left Heart Syndrome



Arsenic Concentrations



Diabetes



Binge Drinking

6 Methodology

In order to determine the distributions of our data sets, we used the histogram feature on Google Sheets and created 3 graphs. Since each histogram showed a non-normal distribution, we decided our test must be a non-parametric test with discrete variables.

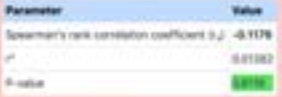
In order to run the Spearman's rank correlation test, we organized our data by county onto Google Sheets and then inputted the corresponding data for each variable we were analyzing. Next, we used the Google Sheets RANK AVG function to rank each set and then find the difference. This allowed us to square that value and add them all up to find the total sum. We used that value for each calculation and inputted them into the formula to find the Spearman Correlation Coefficient, which led us to our three different p-values. These would then be used to make a determination about the significance of our data.

We analyzed the percentage of various population in NJ counties with diabetes, alcoholism, and percentage of babies born with HLHS by county and not the individual level. This prompted us to then include arsenic concentrations along with our other data sets because it better corresponded with our data by county.

7 ANALYSIS

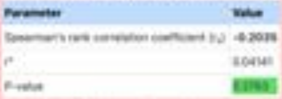
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Hypoplastic Left Heart Syndrome and Diabetes



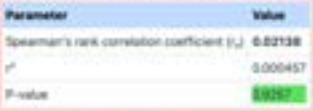
Parameter	Value
Spearman's rank correlation coefficient (r_s)	-0.1176
r^2	0.01382
P-value	[illegible]

Hypoplastic Left Heart Syndrome and Binge Drinking



Parameter	Value
Spearman's rank correlation coefficient (r_s)	-0.2035
r^2	0.04141
P-value	[illegible]

Hypoplastic Left Heart Syndrome and Arsenic



Parameter	Value
Spearman's rank correlation coefficient (r_s)	0.02138
r^2	0.000457
P-value	0.9267

8 Recomendations

Given that all three of our tests led to insignificant results, we would like to further investigate other factors of a mother's lifestyle to understand what impact they may have on HLHS. This includes tobacco and drug abuse, as well as other environmental factors such as exposure to radon and pesticides. With these results, we recommend that additional action be taken by critiquing the type of data found. Specifically, we would like to shift our focus from a county level to an individual parent level since that may correspond better with how frequently a birth defect is present in newborns. Contrastingly, we would like to investigate additional environmental toxins, but keep those specific to counties since it matches better when studied through a geographic lens. Although these results were not significant, we would like to further our research and explore more facets of a mother's lifestyle to truly address the growing prevalence of HLHS.

9 CONCLUSIONS

Using $\alpha = 0.05$ and the calculated p-values, the relationship between adults with diabetes and newborns with HLHS, adults who binge drink and newborns with HLHS, and adults exposed to arsenic and newborns with HLHS, is not statistically significant. We found that adults with diabetes and newborns with HLHS and adults who binge drink and newborns with HLHS both have a weak, negative Spearman's correlation; meanwhile, adults exposed to arsenic and newborns with HLHS has a weak, positive correlation. Overall, we do not have convincing evidence that diabetes, alcohol intake, and arsenic have a linear relationship with hypoplastic left heart syndrome (HLHS) and all of their correlations were not significant.

Based on the results from our HLHS analysis with different pre-pregnancy factors, it can be assumed that diabetes, binge drinking, and arsenic, are not measures of pre-pregnancy factors that could contribute to newborns having HLHS. However, we can assume that other factors are still contributing due to other pre-pregnancy lifestyles considering that despite there being low variance, more tests would have to be run to find if there truly is no correlation between HLHS and these pre-pregnancy factors.

Therefore, the data we gathered, suggests that pregnant women who give birth to a newborn with HLHS are likely not going to have to worry about any of these factors contributing to their baby's risk of getting HLHS. Though HLHS is known to be rare, no child should be put at risk. If we, as a society, care about our future generations, we must remain cautious to protect and prevent more cases from occurring. Despite not finding any factors that would show convincing evidence of a correlation between HLHS and a pre-pregnancy factor, there could still be other factors contributing to the prevalence of HLHS cases in NJ that must be further researched. We can deduce that looking for a correlation by county is probably not the way to look for a relationship between pre-pregnancy factors and HLHS because the characteristics we are looking at are individual level factors it would be beneficial in the future to find data from individual subjects to see if within an individual what the characteristics are of the mother and newborn with HLHS (or other forms of CHD).

10 CHALLENGES

One of the challenges we encountered was finding an appropriate statistical test to fit our data. Our team decided to use Google Sheets, as it was the most beginner-friendly platform, but this also limited the variety of tests we could run with our discrete and non-parametric variables. Initially, we considered the chi-square test due to its compatibility with non-parametric data. However, upon closer examination, we realized that this test required our discrete data to be turned into continuous variables. Eventually, we found that Spearman's Correlation would work best with the type of data we used.

Another challenge we ran into revolved around sourcing and aligning the data. We required datasets organized by county and spanning consistent timeframes to conduct our analysis effectively. We found that there were a limited amount of sources offering data at the county level; they often covered different years or used different formats. After looking into a multitude of sources, we found datasets that met our criteria in terms of geography and time periods.

Moreover, the test we ran took a geographic lens to analyze our topic, given that our data sets were organized by county. This was a challenge for us because parts of our research question were better fitted for data that represents individual mothers to see whether a correlation can be made to HLHS. We attempted to remedy this with our inclusion of data regarding arsenic concentrations, since an environemntal factor corresponds well with geographically categorized data. In the future, this is a facet of our project we would like to correct by changing the scope of our data sets so that it better answers the research question at hand.

The Link Between Disunity and Development

North Allegheny Senior High School: Kelly Tai, Joseph Widjaja, Audrey Zheng, Mihir Sharma, Haresh Muralidharan, Rushil Surti

Research Question: How are political polarization and economic welfare correlated in the United States?

Background

Political polarization in the United States has become more prevalent in recent years with the advent of social media and rising political tensions. A divide in the public is mirrored in Congress, which in turn impacts policymaking. We hope to answer how political polarization affects the economic welfare of the country, a question that is all the more important to consider given the upcoming election cycle.

Hypothesis

We hypothesized that political polarization would have a negative impact (negative correlation) on the economic welfare of the United States, as it reduces the quality and effectiveness of governance.

Challenges

1. Measures of sentiment analysis can vary widely depending on the model being used to generate a sentiment value and the specific data being processed by the model. These could have potentially greatly affected whether we saw a correlation. In fact, the sentiment analysis data was measured based on analyzing specifically Twitter tags and messages alone, meaning that it is possible that the true relative political polarization of the country is not captured in its entirety by the dataset as other social media platforms like Reddit or Threads may give a more comprehensive demographic.
2. We were not able to find any datasets for general political party-based polarization past 2014.

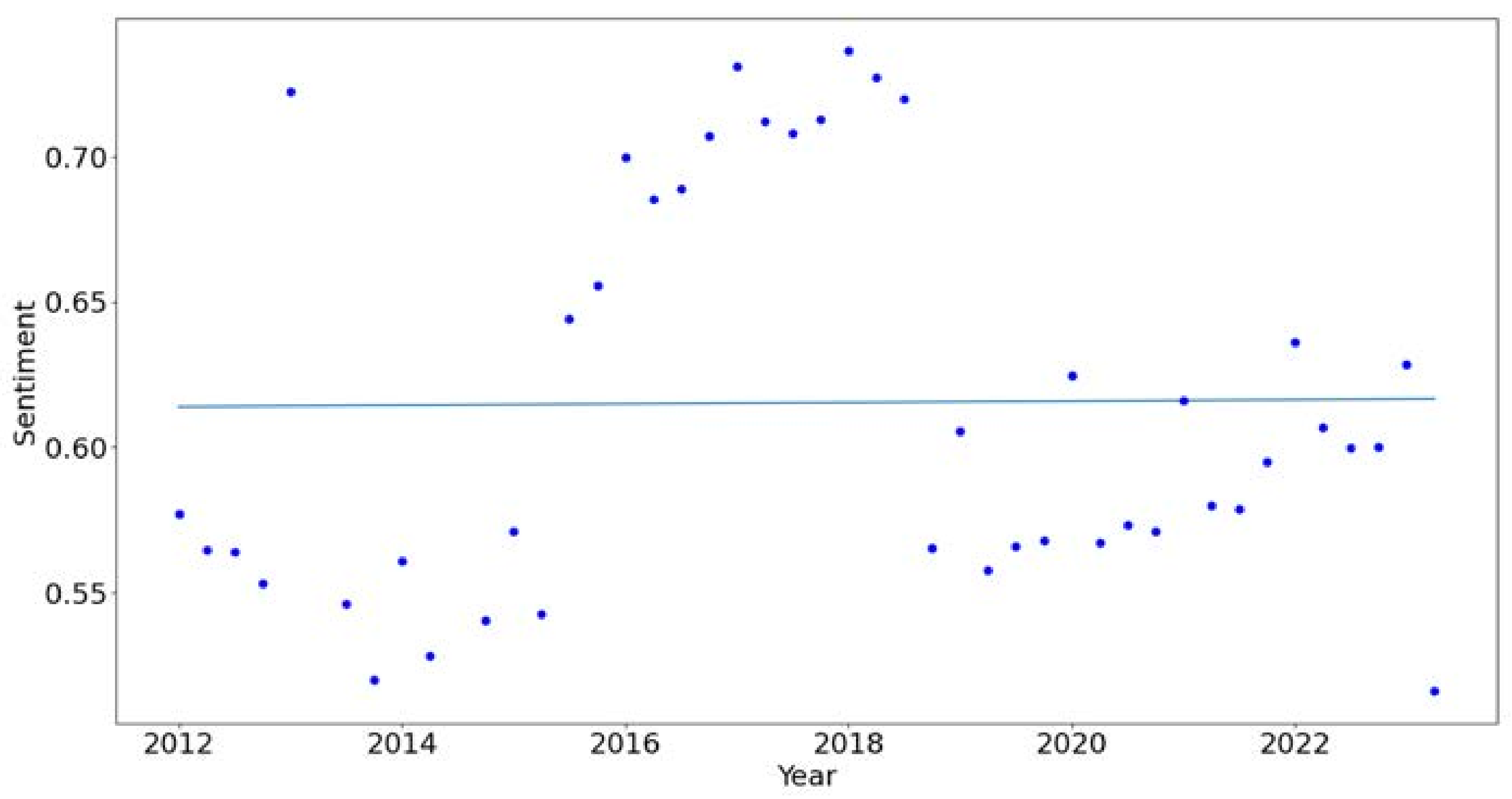
Method

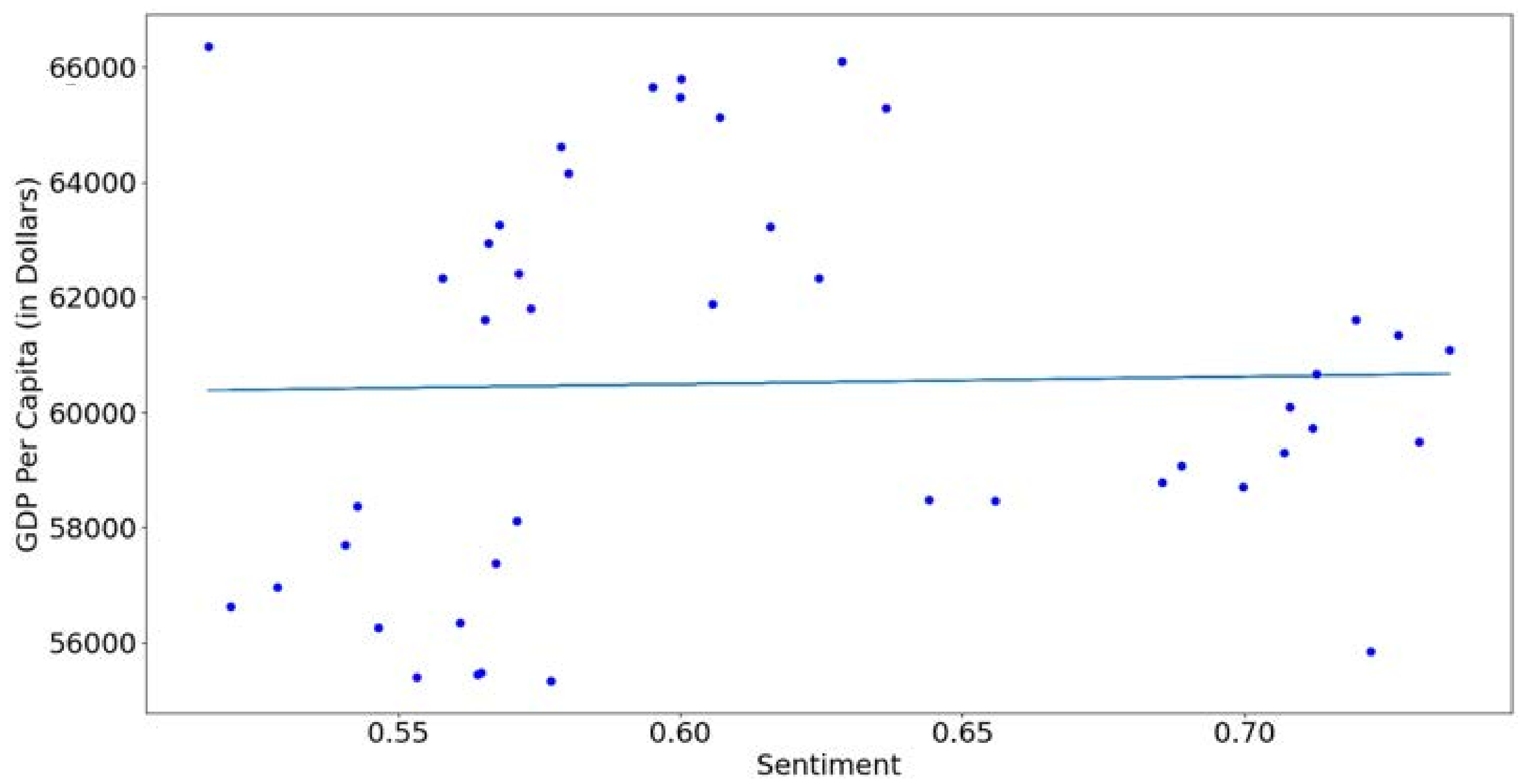
We found datasets for real GDP per capita and sentiment analysis of Twitter data. From each dataset, we took four data points from each year, compromised of measurements from January, April, July, and October, dating from 2012-2023. We plotted the datapoints for real GDP per capita and polarization against each other, using matplotlib to generate the following figures and Python to calculate the correlation coefficient.

Data Sources

FRED GDP Per Capita: https://fred.stlouisfed.org/series/A939RX0Q048SBEA
Harvard Sentiment Analysis Dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3IL00Q
(Sentiment Analysis: Harvard used a neural network to analyze Twitter messages and assign a relative measure of how polarized they are)

Results





Conclusion

We obtained a correlation coefficient of r^2 = 0.0006781 (*n* = 44). Performing a rho-test, we obtained a test statistic of 0.1688 and a *p*-value of 0.5666 (because we hypothesized a negative correlation). Thus, we have significant evidence against our initial hypothesis. Our results suggest that there is no significant correlation between political polarization in the U.S. and economic welfare. This may be due to external sources such as the variability between machine learning analyses of sentiment, but it could also be due to the high variability in sentiment over the years itself, rising with election years and gradually falling over time. Perhaps an analysis that captures more of the local time behavior or utilizes different measures of economic welfare would yield more insightful results.

CANCER VS. CAPITAL: EXAMINING THE EFFECTS OF INCOME ON 3 MAJOR CANCERS

NORTH ALLEGHENY HIGH SCHOOL: ASHLEY JANG, MEGUMI KAWABE, ADITI MOTHA, SAMUEL XIAO

BACKGROUND

PROBLEM

Cancer has become the second leading cause of death in the United States. Due to inequities in healthcare, cancer cases are disproportionately affecting the socioeconomically disadvantaged.

IMPORTANCE

As a definitive cure for cancer has not yet been found, analyzing the relation between the socioeconomic factor of income to find what type of assistance people need to lower cancer risk will be incredibly vital in the battle against cancer.

GOALS

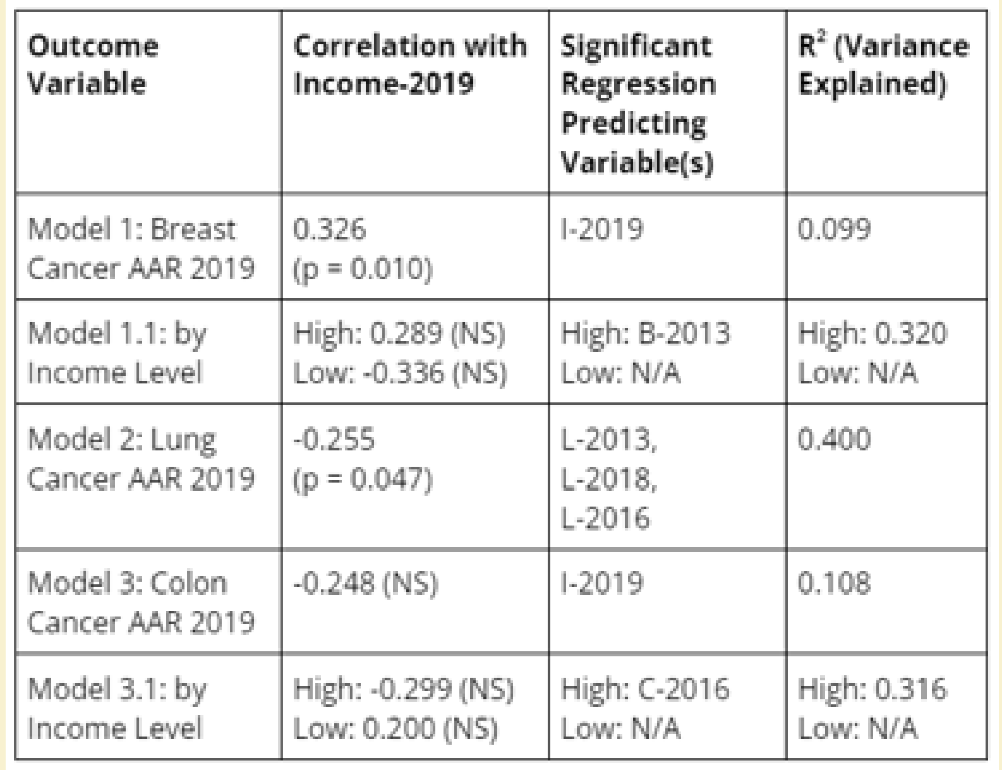
RESEARCH QUESTION

In Pennsylvania, what effect does income have on cancer (specifically breast, lung, and colon) incidence rates?

HYPOTHESIS

Counties with lower per capita incomes will see higher rates of all three cancers due to less access to healthcare resources.

RESULTS TABLE



Outcome Variable	Correlation with Income-2019	Significant Regression Predicting Variable(s)	R^2 (Variance Explained)
Model 1: Breast Cancer AAR 2019	0.326 (p = 0.010)	I-2019	0.099
Model 1.1: by Income Level	High: 0.289 (NS) Low: -0.336 (NS)	High: B-2013 Low: N/A	High: 0.320 Low: N/A
Model 2: Lung Cancer AAR 2019	-0.255 (p = 0.047)	L-2013, L-2018, L-2016	0.400
Model 3: Colon Cancer AAR 2019	-0.248 (NS)	I-2019	0.108
Model 3.1: by Income Level	High: -0.299 (NS) Low: 0.200 (NS)	High: C-2016 Low: N/A	High: 0.316 Low: N/A

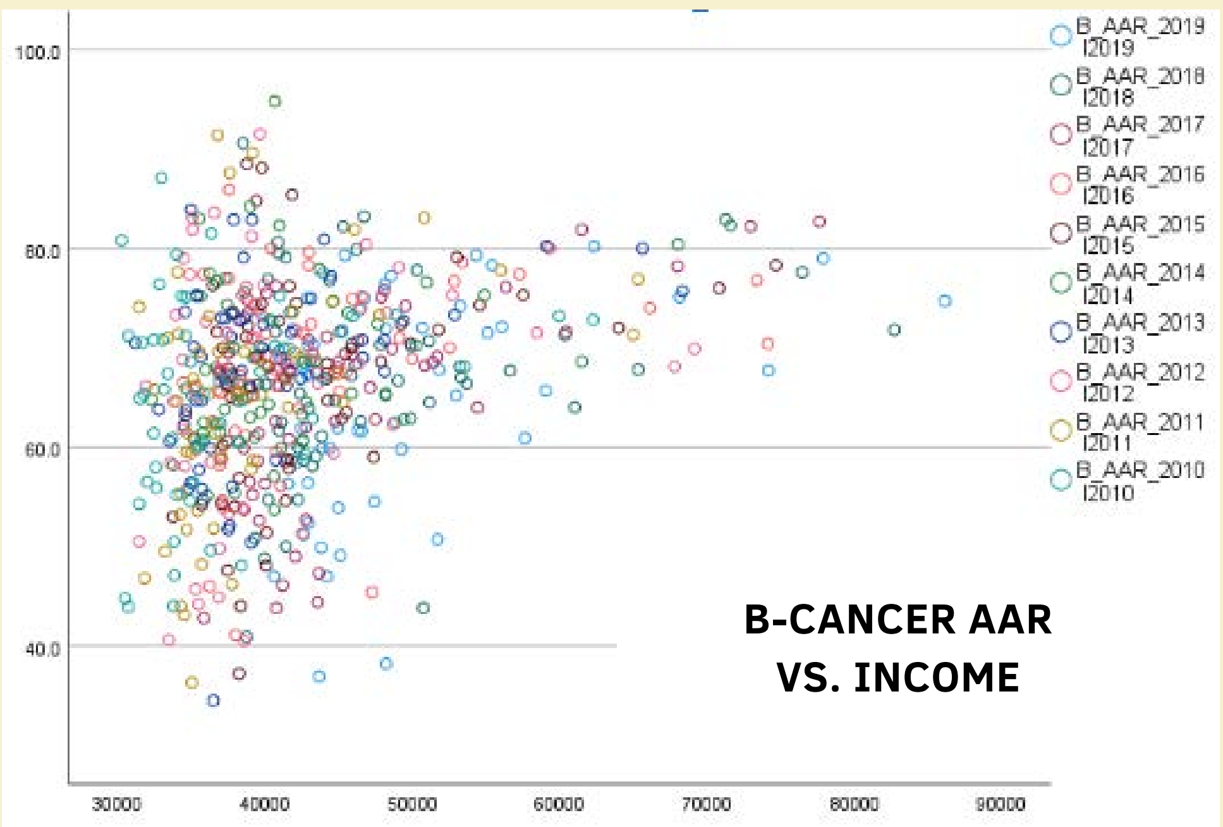
KEY: B = BREAST, L = LUNG, C = COLON, I = INCOME,
AAR = AGE-ADJUSTED RATE (PER 100,000 POPULATION),
NS = NOT SIGNIFICANT, N/A = NOT APPLICABLE

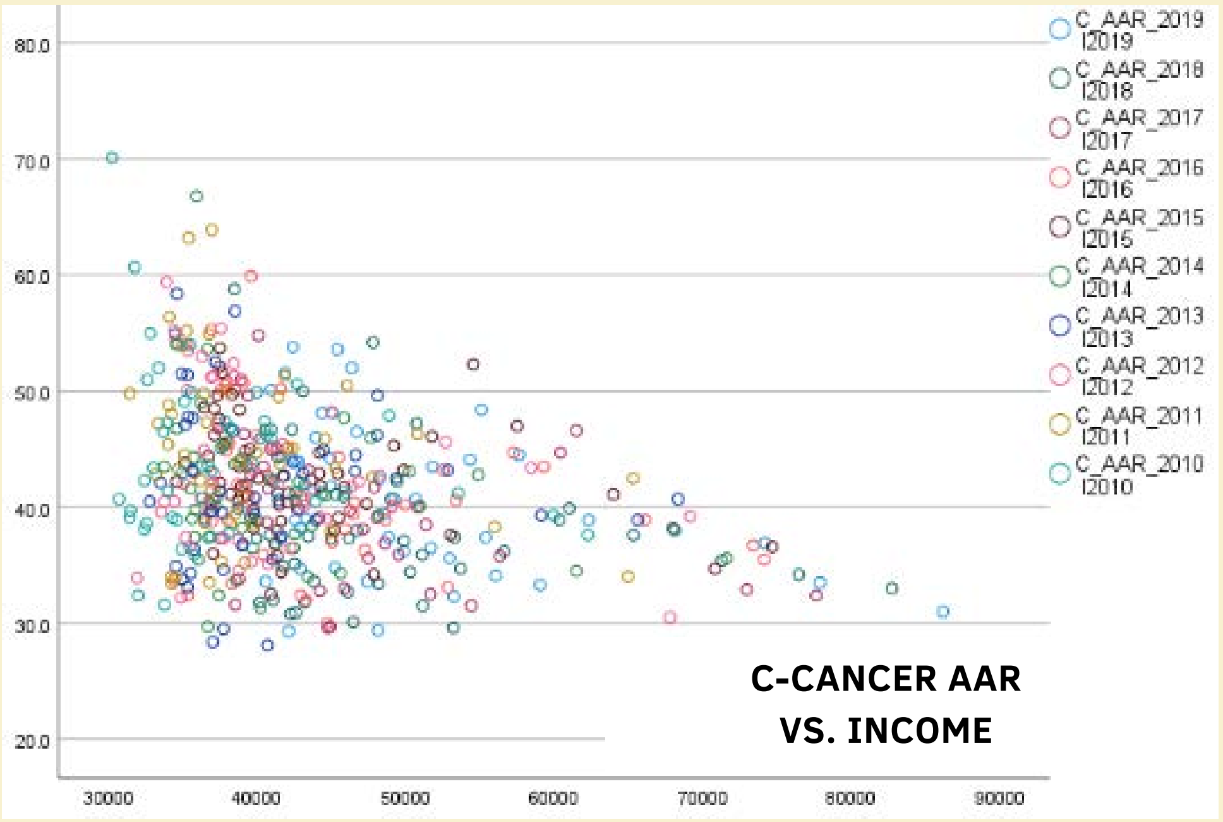
METHODS

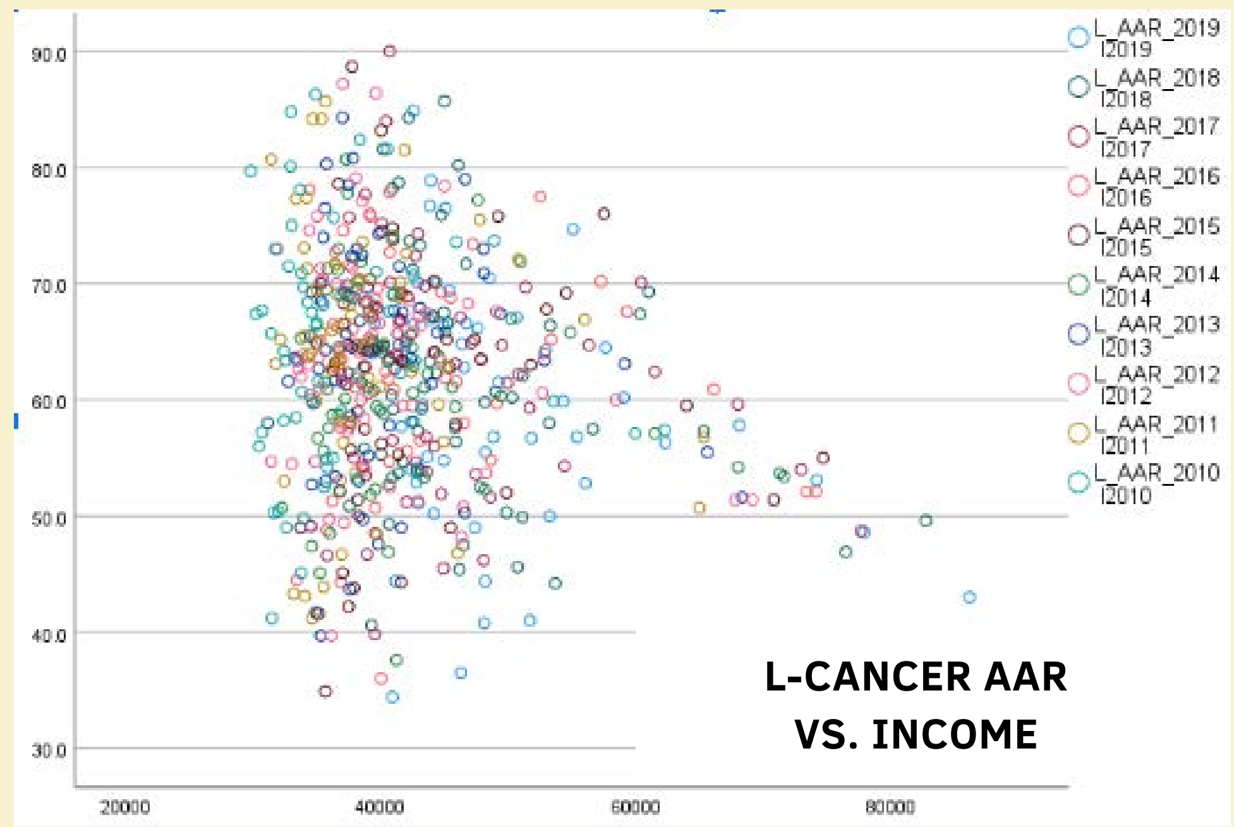
1. Longitudinal data in the ten-year span of 2010-2019 was gathered from the PA Department of Health's Enterprise Data Dissemination Informatics Exchange (EDDIE) program for B/L/C type cancers (the 3 most common types). Average county per capita income data came from the St. Louis Fed Bank database. The data was imported into Excel, sorted by county, and cleared of irrelevant information. AAR values have been adjusted to account for county population.
2. Four sets of Bivariate Correlations were performed within income and the B/L/C cancer rates. Income had strong (>0.94) correlations with itself (p<0.001), implying its stability over the 10-year period, while the three cancer rates had almost no correlation with themselves.
3. A Stepwise Linear Regression was performed using 2010-18 rates and 2019 incomes as independent variables and 2019 cancer rates as dependent variables. The 2019 income set was chosen because it was the most recent available and because of the stability previously mentioned.
4. Separate Regressions were performed for B/C cancers based on low/high income groups (using $45K as an approximately median cutoff).

All analyses was performed using IBM's Statistical Package for the Social Sciences (SPSS Version 22).

GRAPHS







CHALLENGES

It was difficult to find datasets that had exactly the type of cancers we were looking for and broke down their reporting by county.

In addition, many data sources were actually embedded within PDF files or other larger reports, making the data hard to extract and often unworkable. We had to search carefully for a timeframe in which all datasets had sufficient data, as the Department of Health stopped releasing data at the start of the COVID-19 pandemic.

In addition, all available data was county-level, not individual-level, so this made it much more difficult to explore relationships between cancer rates and income at more detailed levels.

RESULTS

Regression results for cancers B/C show a statistical paradox; income is a significant predictor of B/C cancers using the entire sample, but this effect disappears when low/high income groups are tested. For lung cancer, income is not significant at any scale. For lung and colon cancer, previous years' incidence rates serve as a better predictor variable than income. Lung and colon cancer show moderate negative correlations with income, but for colon cancer the correlation is insignificant. However, contrary to the hypothesis, breast cancer shows a stronger *positive* correlation with income, implying that higher income is associated with higher rates of breast cancer at the county level. For lung and colon cancer, previous years' incidence rates serve as a better predictor variable than income.

Results are summarized in the above table.

CONCLUSIONS

While Lung cancer had a significantly negative correlation to income, Breast cancer had a significantly positive correlation. The results for breast cancer may be partially explained by the fact that higher-income individuals tend to have more access to medicines/contraceptives that can increase breast cancer risk. In addition, low-income individuals likely have less access to annual cancer screenings in the first place, therefore lowering cancer rates for this population in all three types of cancer.

Recommendation: Because higher income does not decrease all risks of cancer, raising overall awareness about the risk factors of cancer that people overlook is one step that must be taken. In addition, further investigating what factors that result from higher/lower income affect the cancer rates should be done to test the speculations above.

Mentor
Daniel Hufnagle

BREATHE EASY: EXPLORING ASTHMA TRIGGERS

North Allegheny Team C
Andrew Li, Angelina Li

How do different types of social, economic, and environmental factors influence hospitalization and diagnosis rates of children's asthma in Pennsylvania?

BRAINSTORMING

- Identified major issue
- Drew from our personal lives
- Extensive literature review
  - ID'ed more factors of interest
- Narrowed scope to **PA only**
- PA's asthma statistics are one of the most concerning

DATA GATHERING

- **Main Data:** Asthma School Data Report (PA Department of Health)
- **Additional Data:** US Census, opendata PA

METHODS AND ANALYSIS

Preparation

- Narrow factors for multi-linear through **linear regression** (Fig 2,Fig 3)
  - Factors with $p < 0.05$ viable for multi-linear regression
- Co-linearity test ➤ further narrows down factors

Multi regression model

- 0.1*N independent variables to test (around 5-8) most ideal
- **Step-wise regression:** eliminates factors with no significance per turn/step until an ideal number
- Conduct test + check multi-colinearity

Logistic Regression

- Using results from multi-linear regression, choose select factors included in model
- Binary Variable: High/Low Risk Area based on asthma rate per county
  - Threshold set at 11.3%
  - Determined by 1 sample z-test using PA data ($\mu = 10.3$, $\sigma = .0051$)
- Split county data into train (50 counties) and test (17 counties) data
- Conduct tests through R
- Check for multi-colinearity
- Use RMSE, AIC, and AUC values to check the accuracy

HYPOTHESIS

Null: No factors of interest impact diagnosis and hospitalization rates of children's asthma in PA.

Alternative: There are factors of interest that impact diagnosis and hospitalization rates of children's asthma in PA.

LINEAR REGRESSION

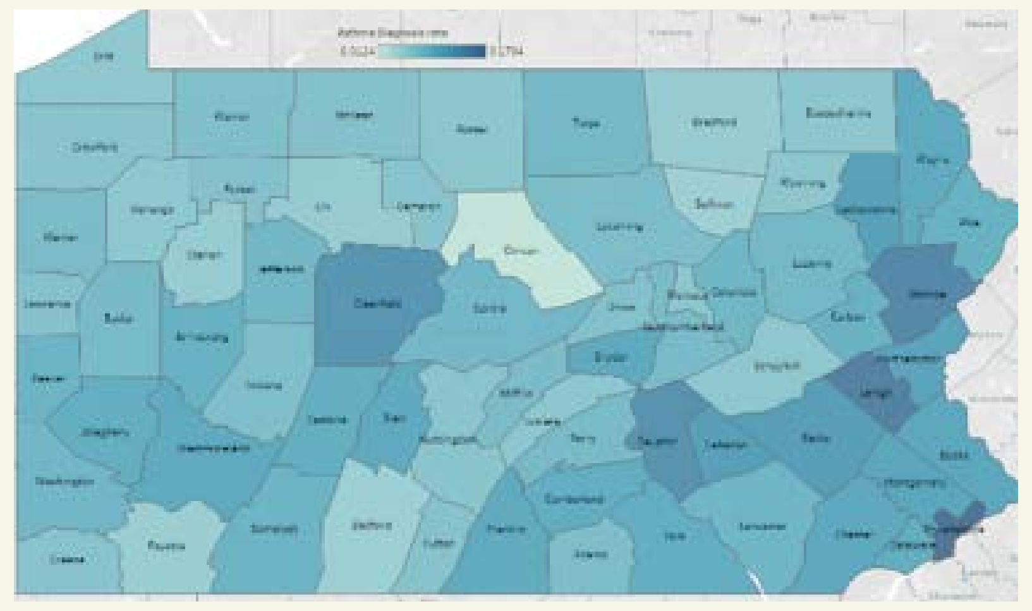


Figure 1: map of diagnoses rates in 100k children in PA for asthma (2019-20), with Philadelphia being the highest

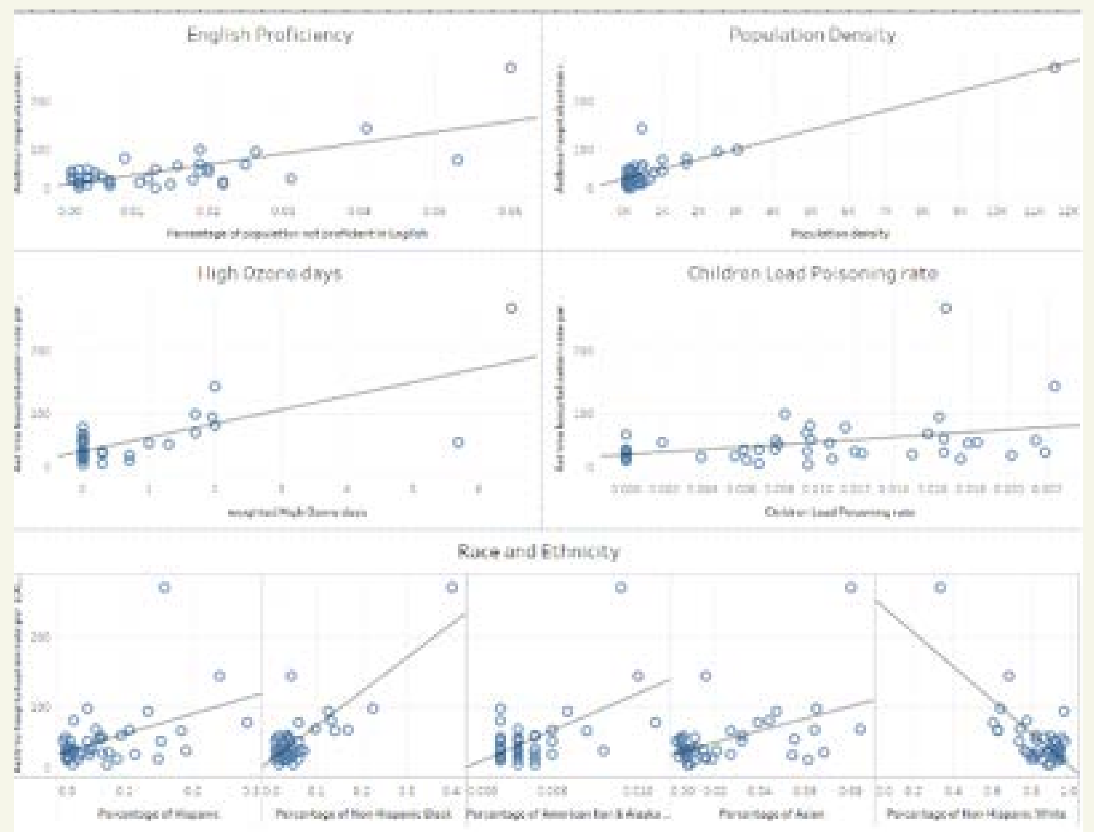


Figure 2: linear regression scatterplots of PA child Asthma hospitalization rate by factor

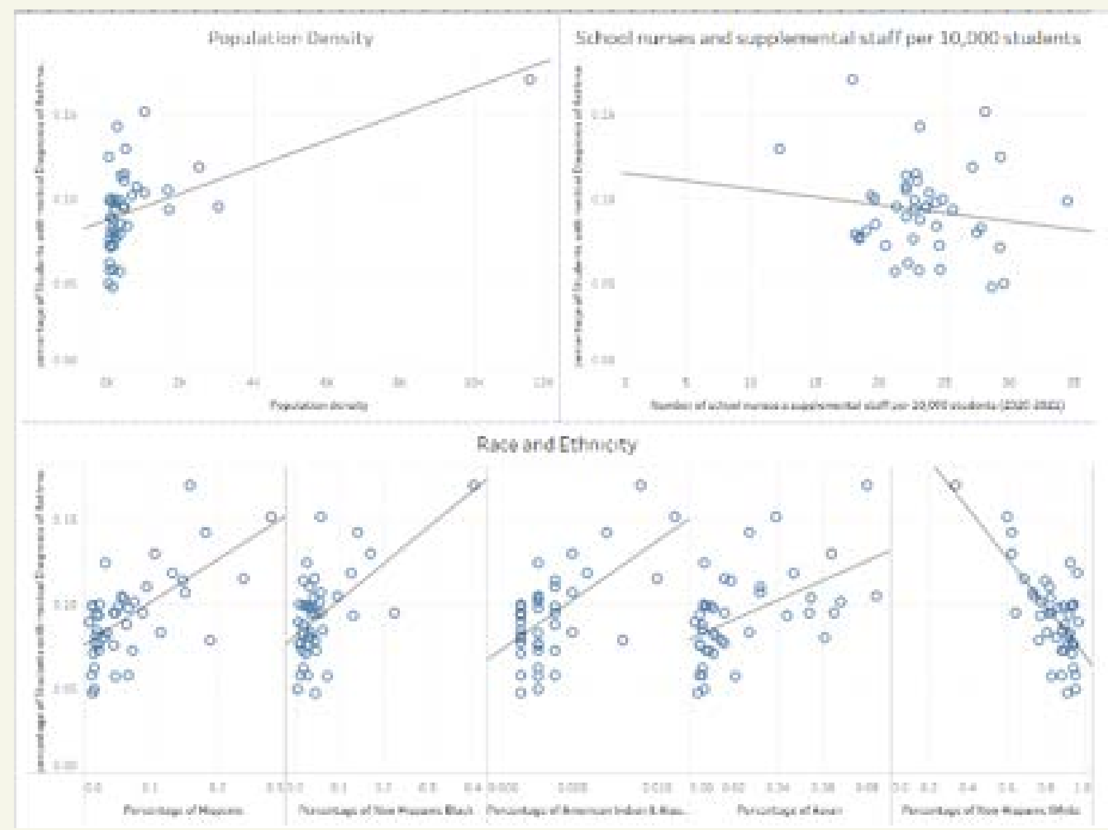


Figure 3: linear regression scatterplots of PA child Asthma diagnosis rate by factor

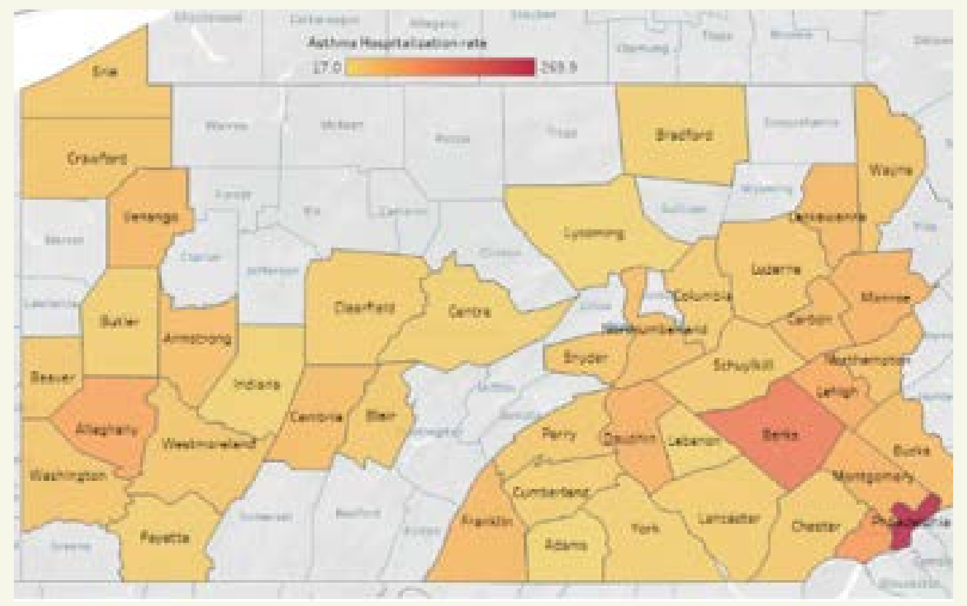


Figure 4: map of PA hospitalization rates for children's Asthma (from 100,000 children), with Philadelphia being the highest

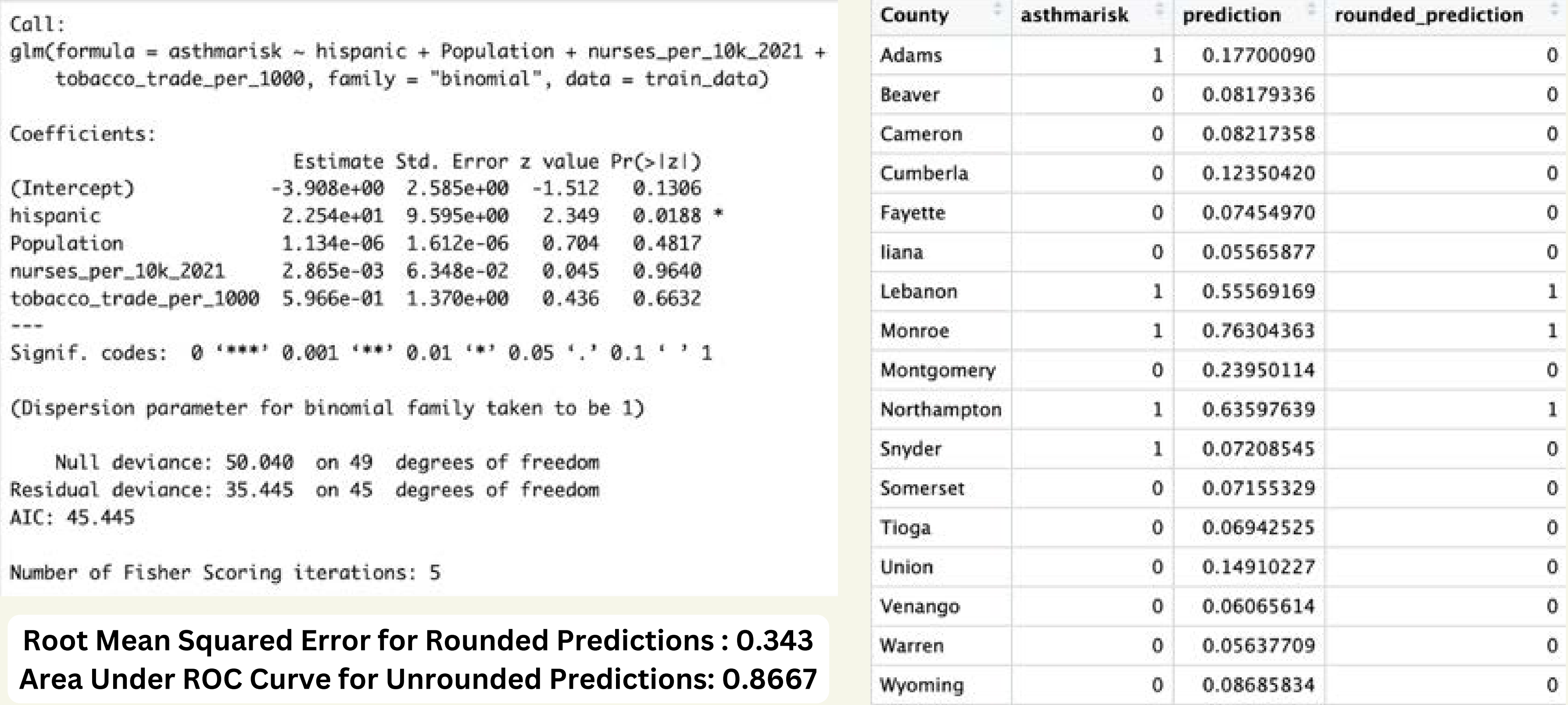
MULTI-LINEAR REGRESSION

Table 1: Hospitalization rates	Coefficient	Error	tStat	P-value	Test Decision
Intercept	(18.400)	14.184	(1.297)	0.203	
Population Density	0.015	0.002	6.321	0.000	Reject
Hispanic	130.993	40.770	3.213	0.003	Reject
High Ozone Days	6.897	2.928	2.355	0.024	Reject
Children Lead Poisoning Rates	951.853	416.218	2.287	0.028	Reject
Tobacco Retail Density	24.535	11.526	2.129	0.040	Reject
Ratio of Population to PCP	0.007	0.004	1.635	0.111	Fail to Reject

Table 2: Diagnosis Rates	Coefficient	Error	tStat	P-value	Test Decision
Intercept	8.365E-02	0.0097	8.5993	0.0000	
Hispanic	2.400E-01	0.0414	5.7912	0.0000	Reject
Population Density	5.871E-06	0.0000	3.5920	0.0006	Reject
nurses per 10k kids	-6.826E-04	0.0003	2.0080	0.0489	Reject

Note: Table 2 does not display all factors included in the test because of spacing issues, but none were significant
Tobacco retail density: number of tobacco sale locations / 1000 lives
Hospitalization: admission in a hospital for a minimum period of 24 hours for asthma concerns
Hispanic - Population proportion of hispanics in county

LOGISTIC REGRESSION



```
Call:
glm(formula = asthmarisk ~ hispanic + Population + nurses_per_10k_2021 +
    tobacco_trade_per_1000, family = "binomial", data = train_data)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.908e+00  2.585e+00  -1.512   0.1306
hispanic               2.254e+01  9.595e+00   2.349   0.0188 *
Population             1.134e-06  1.612e-06   0.704   0.4817
nurses_per_10k_2021    2.865e-03  6.348e-02   0.045   0.9640
tobacco_trade_per_1000 5.966e-01  1.370e+00   0.436   0.6632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50.040  on 49  degrees of freedom
Residual deviance: 35.445  on 45  degrees of freedom
AIC: 45.445

Number of Fisher Scoring iterations: 5
```

Root Mean Squared Error for Rounded Predictions : 0.343
Area Under ROC Curve for Unrounded Predictions: 0.8667

County	asthmarisk	prediction	rounded_prediction
Adams	1	0.17700090	0
Beaver	0	0.08179336	0
Cameron	0	0.08217358	0
Cumberla	0	0.12350420	0
Fayette	0	0.07454970	0
liana	0	0.05565877	0
Lebanon	1	0.55569169	1
Monroe	1	0.76304363	1
Montgomery	0	0.23950114	0
Northampton	1	0.63597639	1
Snyder	1	0.07208545	0
Somerset	0	0.07155329	0
Tioga	0	0.06942525	0
Union	0	0.14910227	0
Venango	0	0.06065614	0
Warren	0	0.05637709	0
Wyoming	0	0.08685834	0

Figure 5: Logistic Regression summary of train data results (top left), Calculated prediction values for test data (right), Accuracy Checks (bottom left) of 67 county's asthma risk (diagnosis rate) in Pennsylvania
Note: asthmarisk value of 1 represents High Risk, 0 represents Low Risk

OBSERVATIONS AND CONCLUSION

Our study provides valuable insight into the most significant triggers of asthma using multi-linear regression, and puts to use such results in a logistic regression model. We are able to conclude, based on p-values in Table 1 and Table 2 that **Population Density and Hispanic Population Percentage are by far the 2 most important predictors of children's asthma diagnosis and hospitalization within PA.** The common thread between these two factors is urban living, which leaves populations more susceptible to allergens, pollution, and the unhealthy behavior of others around us. Past the surface level, it is crucial to be aware of the underlying genetic risk associated with race, as its interplay with other determinants causes the observed significance within Hispanic populations. The other statistically significant factors exist in every aspect, from socioeconomic to behavioral to environmental factors, further underlining just how widespread and unpredictable a disease like asthma truly is. The logistic regression model utilizes the results of multi-linear regression to select the ideal factors to include when training the model. By testing on 17 counties (25%), our model returns accurate results (RMSE = .343, AUC = 0.8667) and the opportunity to improve and expand upon further. **In conclusion, self-management regarding all possible triggers of asthma is essential, and a good place to begin is with your PCP, asking and learning about your own genetic risk, urban environment, and additional preventative practices.**

CHALLENGES

- **Limited + Unstructured data**
  - Ex: Second-hand smoke:
    - **No data** at the county level
    - Combine # tobacco retailers and population to estimate tobacco retail density
    - Provides better data on county-level smoking/tobacco use
- **Eliminating factors:**
  - **Ideal power** for multi-linear requires limited variables
  - Needed an efficient, well-backed method to eliminate variables
- **Logistic Regression:**
  - Understanding procedure/theory
  - Original data provides rates. We needed to **change that to a binary variable**
    - Done through research into typical asthma rates
    - Using statistics, determine % at which $p < 0.05$ or significantly different from normal healthy rate (11.3%)

DIVING DEEPER

- **Why is the percentage of the Hispanic population so significant?**
  - Impacted by all aspects
  - **Genetics** - underlying risk, increased predisposition
  - **Cultural** - Language barrier regarding asthma education (PCP/healthcare (interaction)
  - **Socioeconomic** - Tend to live in warmer climates, urban areas
    - Greater exposure to pollution: dust, allergens, mold, lead poisoning, contamination
- **Taking Logistic Regression further:**
  - Our study predicts based on one year of data. Future work in studying additional years
  - Increase the amount of data by studying more states or expanding to national level
    - Builds stronger reinforced model

SMOKE BUSTERS: THE RISE OF SMOKERS

The Preuss School UC San Diego

Gael Herrera, Vy Ho, Diego GB, Lia Le, Cody Pham, Dillon Huynh, Angel Lugo

Hypothesis

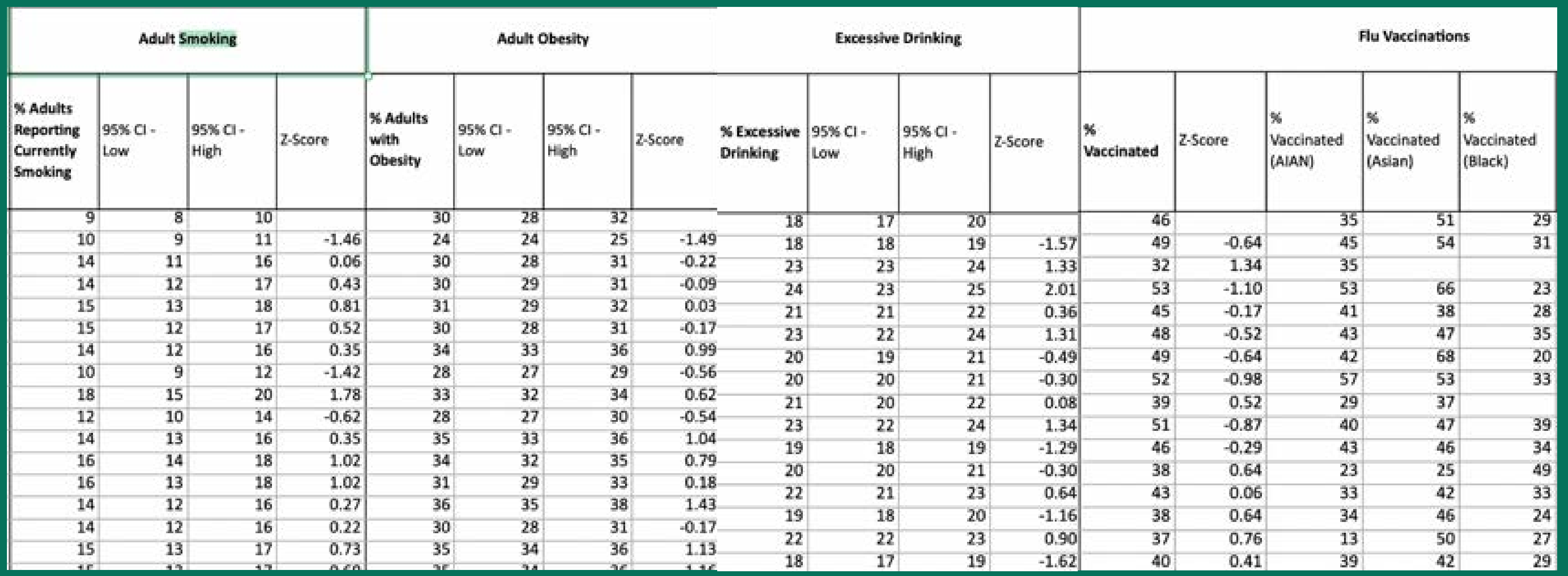
We predict that there will be a correlation with the percentage of adults who smoke in California and adults with obesity, adults excessively drinking, and flu vaccination rates.

Challenges

- Finding uninterpreted raw data.
- Identifying reliable sources of data with no gaps in time and information.
- Having to adapt our project focus depending on the actual data available.
- Being flexible in returning to early stages of the data science process.

Datasets

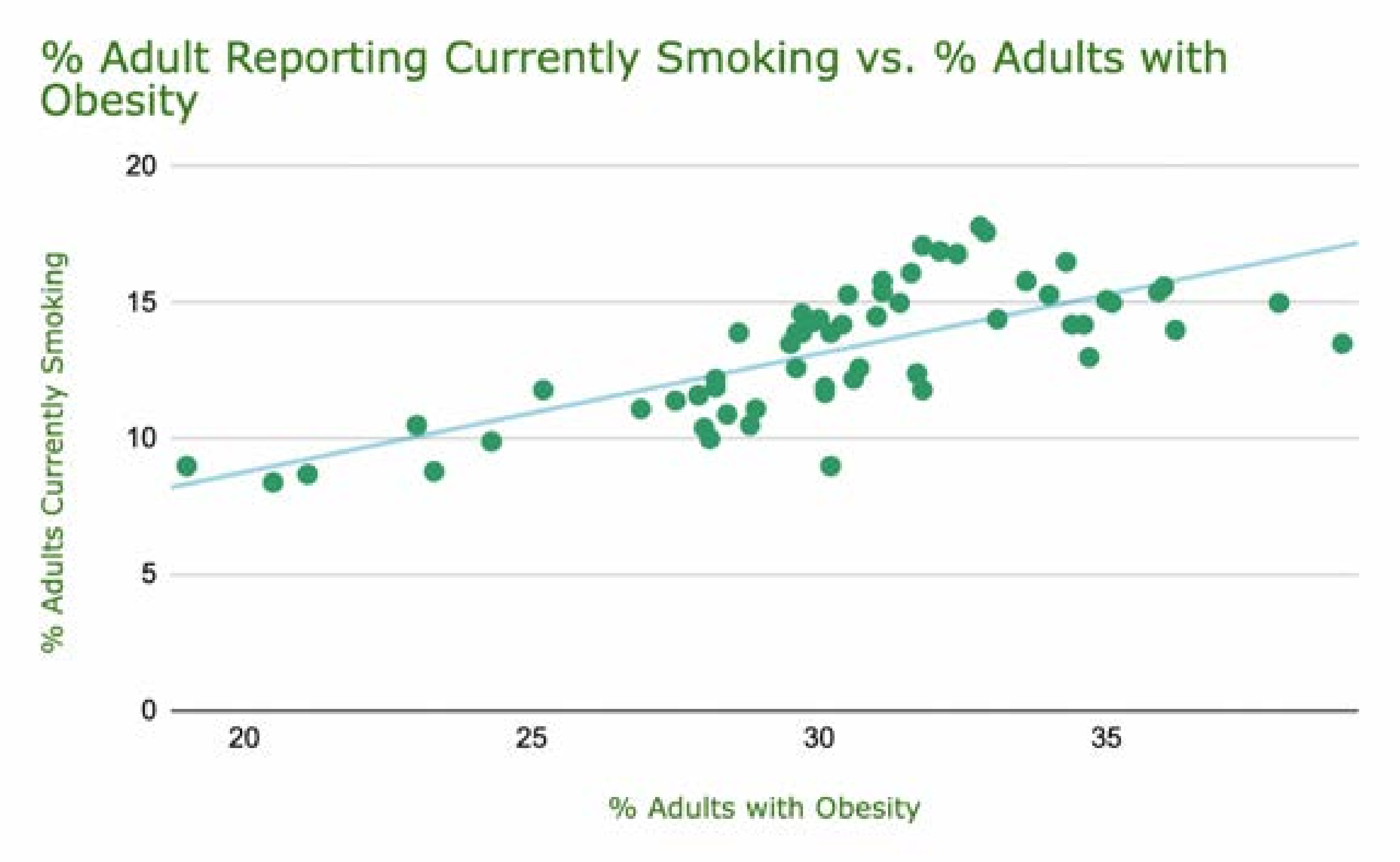
County Health Rankings - California 2019-2023 Data Sets https://www.countyhealthrankings.org/health-data/california/data-and-resources

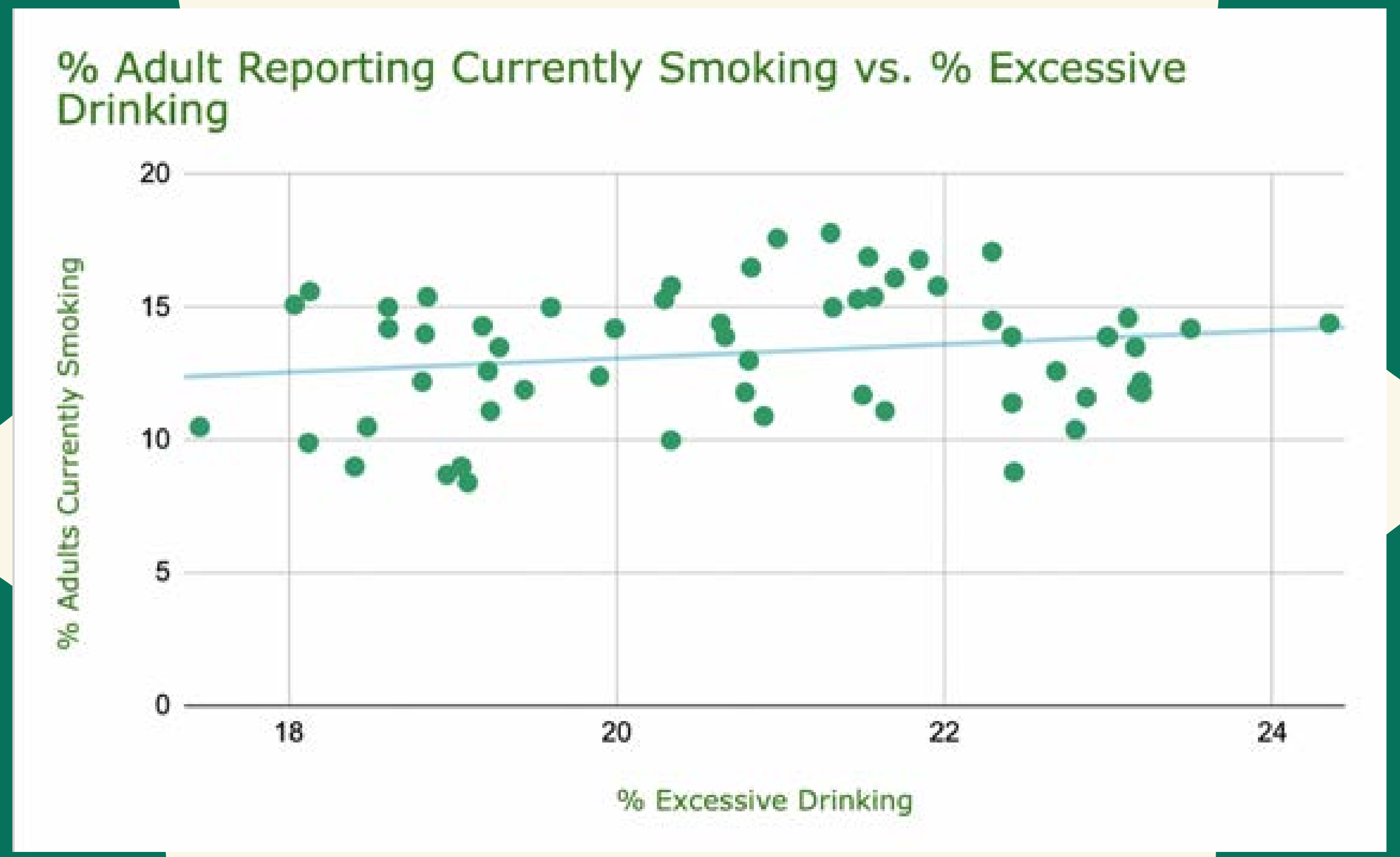


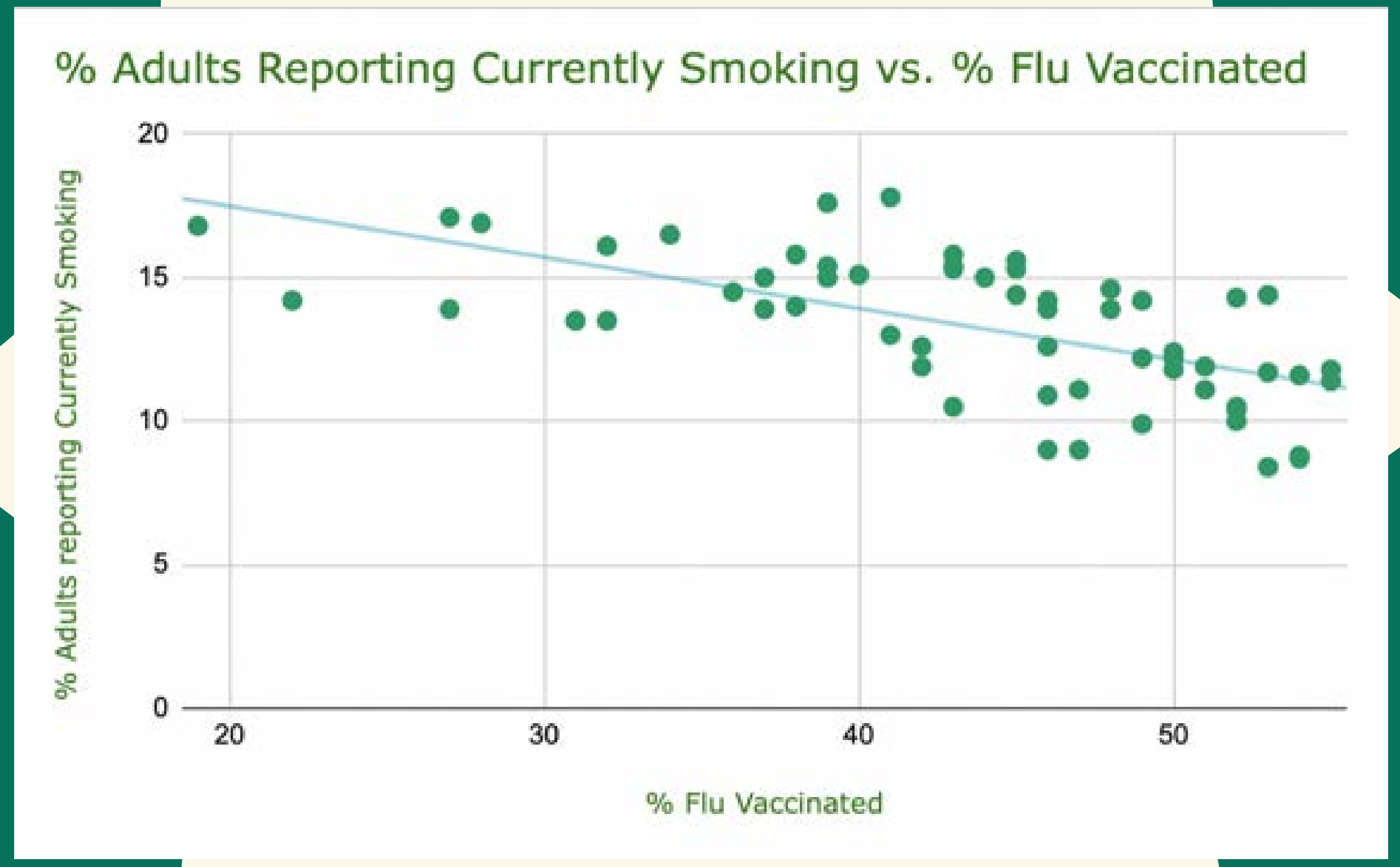
Adult Smoking				Adult Obesity				Excessive Drinking				Flu Vaccinations				
% Adults Reporting Currently Smoking	95% CI - Low	95% CI - High	Z-Score	% Adults with Obesity	95% CI - Low	95% CI - High	Z-Score	% Excessive Drinking	95% CI - Low	95% CI - High	Z-Score	% Vaccinated	Z-Score	% Vaccinated (AIAN)	% Vaccinated (Asian)	% Vaccinated (Black)
9	8	10		30	28	32		18	17	20		46		35	51	29
10	9	11	-1.46	24	24	25	-1.49	18	18	19	-1.57	49	-0.64	45	54	31
14	11	16	0.06	30	28	31	-0.22	23	23	24	1.33	32	1.34	35		
14	12	17	0.43	30	29	31	-0.09	24	23	25	2.01	53	-1.10	53	66	23
15	13	18	0.81	31	29	32	0.03	21	21	22	0.36	45	-0.17	41	38	28
15	12	17	0.52	30	28	31	-0.17	23	22	24	1.31	48	-0.52	43	47	35
14	12	16	0.35	34	33	36	0.99	20	19	21	-0.49	49	-0.64	42	68	20
10	9	12	-1.42	28	27	29	-0.56	20	20	21	-0.30	52	-0.98	57	53	33
18	15	20	1.78	33	32	34	0.62	21	20	22	0.08	39	0.52	29	37	
12	10	14	-0.62	28	27	30	-0.54	23	22	24	1.34	51	-0.87	40	47	39
14	13	16	0.35	35	33	36	1.04	19	18	19	-1.29	46	-0.29	43	46	34
16	14	18	1.02	34	32	35	0.79	20	20	21	-0.30	38	0.64	23	25	49
16	13	18	1.02	31	29	33	0.18	22	21	23	0.64	43	0.06	33	42	33
14	12	16	0.27	36	35	38	1.43	19	18	20	-1.16	38	0.64	34	46	24
14	12	16	0.22	30	28	31	-0.17	22	22	23	0.90	37	0.76	13	50	27
15	13	17	0.73	35	34	36	1.13	18	17	19	-1.62	40	0.41	39	42	29

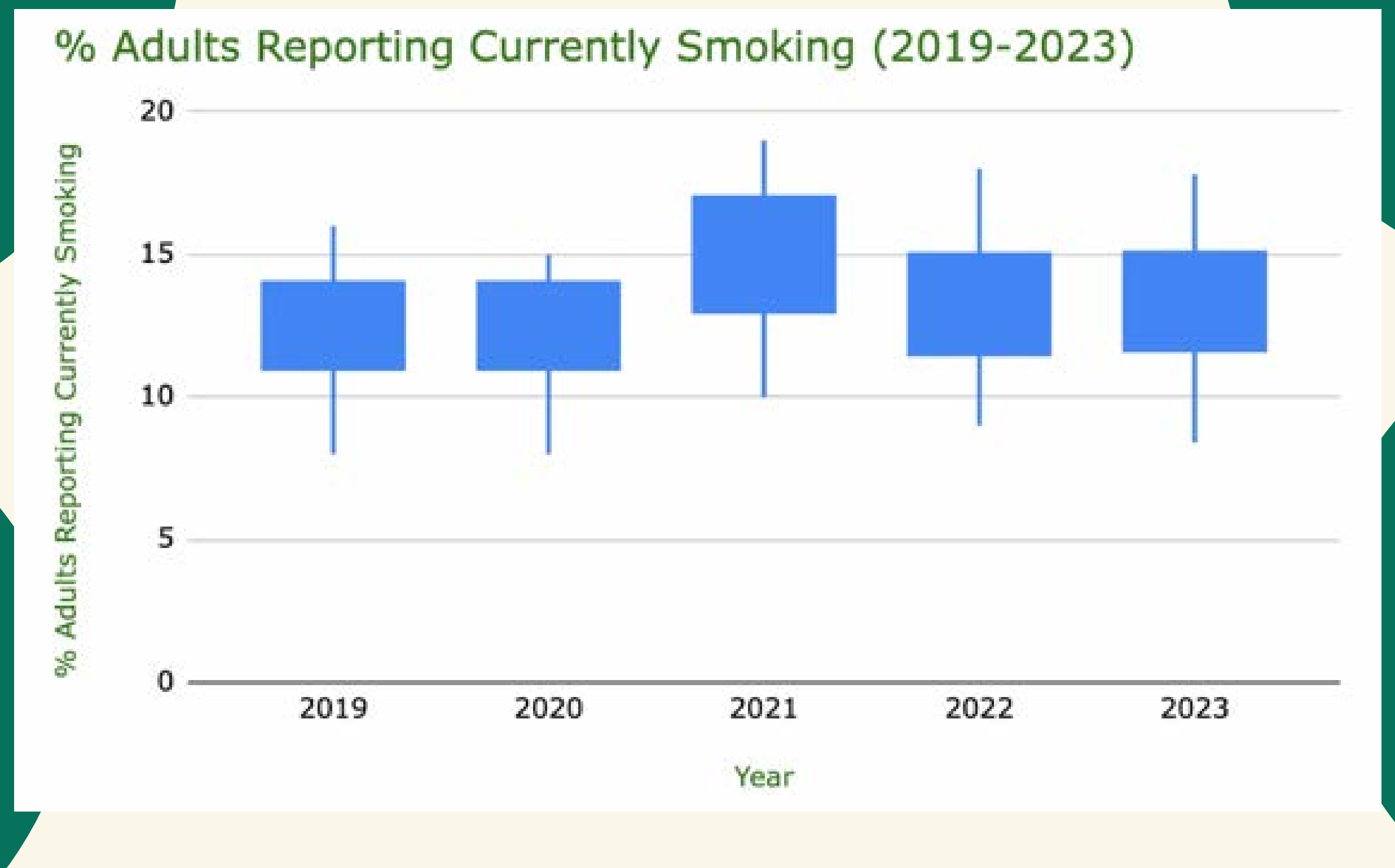
Research Question

What is the correlation between adult smoking and other societal factors in California?









Analysis

- We used multiple linear regression to find the r squared value of 0.67, which means that about 2/3 of the variability in smoking rate could be explained by our multiple linear regression.
- The graph comparing percentage of adults smoking rate to obesity rate demonstrates a moderate positive correlation of r = 0.73.
- The graph comparing adults smoking rate to excessive drinking rate demonstrates a low positive correlation of r = 0.19. It has a smaller r value and higher p-value so it's a weaker predictor to smoking rates.
- The graph comparing adults smoking rate to flu vaccination rate demonstrates a moderate negative correlation of -0.63. There may be a confounding variable affecting the calculated correlation, such as overall trust in medical science which could explain adults not smoking and getting vaccinated.
- The rate of smokers remained increasingly steady from 2019-2023 across the counties in California, with a peak in 2021. We suspect that the pandemic may have contributed to this spike.

Results

- The higher the percentage of smokers, the more likely they are to be obese.
- The higher the percentage of smokers, the higher the likedhood of excessive drinking.
- The higher the percentage of smokers, the less likely they are to be vaccinated. This means that the higher the percentage of smokers, the less likely they are to be vaccinated.

INCOME LEVEL & ACCESSIBILITY TO HEALTHY FOOD

THE PREUSS SCHOOL (TEAM SOUP-ERIOR)
Anthony Mendias, Nicole Nguyen, Nova Solan

HOW DOES THE INCOME LEVEL OF DIFFERENT COMMUNITIES IN SAN DIEGO COUNTY IMPACT THEIR ACCESSIBILITY TO HEALTHY FOODS?

DEFINITIONS & BACKGROUND

HPI Score: Used by the **Public Health Alliance of Southern California** to indicate the healthiness of communities (the higher the better).

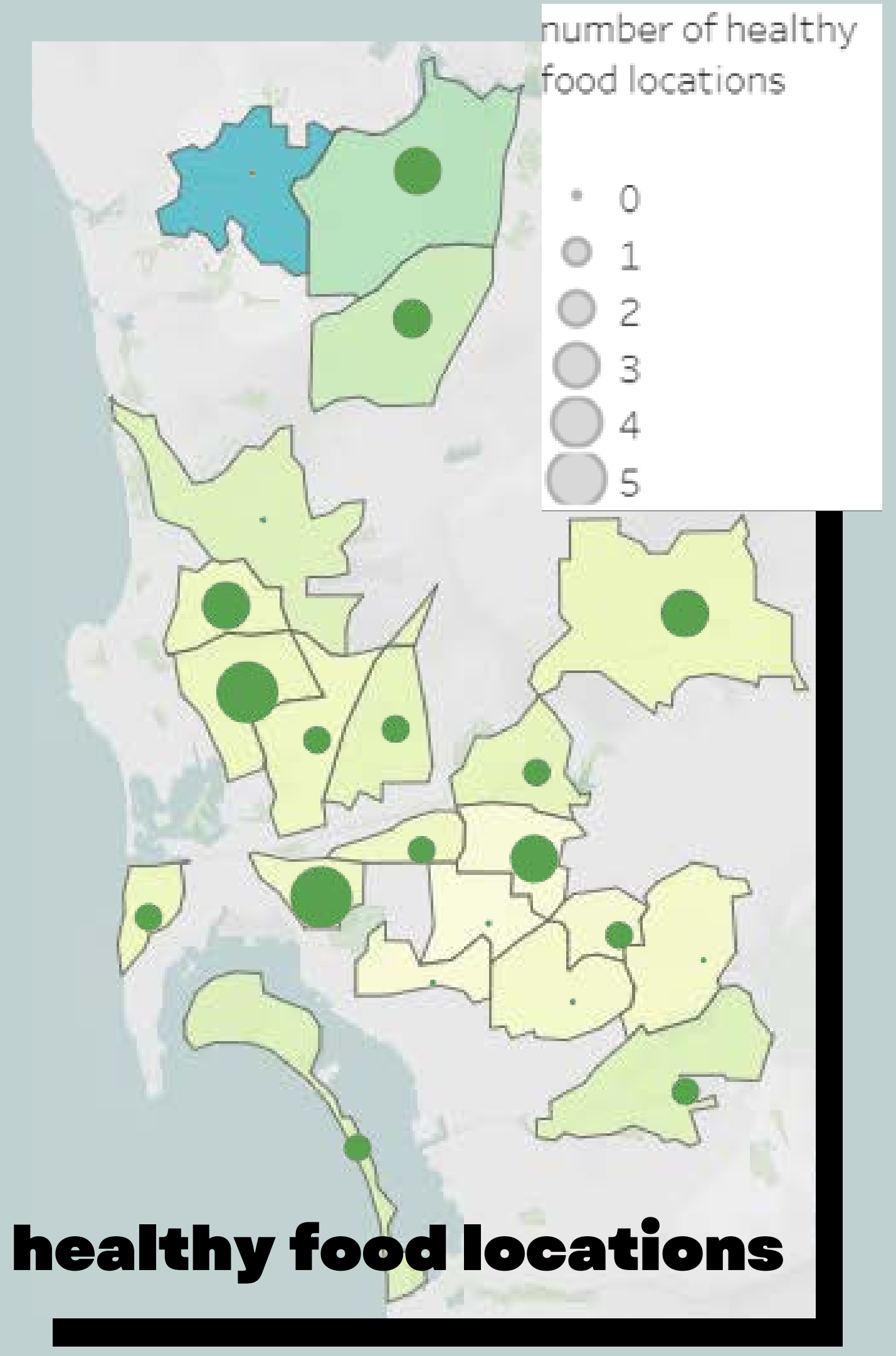
COMMUNITIES IN SD: We looked at 21 zip codes with information about median household income available at **data.census.gov.**
(92105, 92115. 92102, 91945. 92114, 91977, 92116, 92103, 92111, 92122, 92107, 92117, 92071, 92123, 92120, 91902, 92121, 92118, 92129, 92127, 92067)

ACCESS TO HEALTHY V. UNHEALTHY FOOD: We considered healthy food locations as places with affordable fresh produce like grocery stores and unhealthy food locations as places like fast food restaurants. We looked at city maps (**Google Maps**) to create reference points of the ratio of these two in different zip codes using 5 brands of each. The data in our charts show the total of each.
HEALTHY: Sprouts, Vons, Whole Foods, Trader Joe's, Ralph's
UNHEALTHY: McDonald's, Starbucks, Jack in the Box, In N Out, Little Caesar's

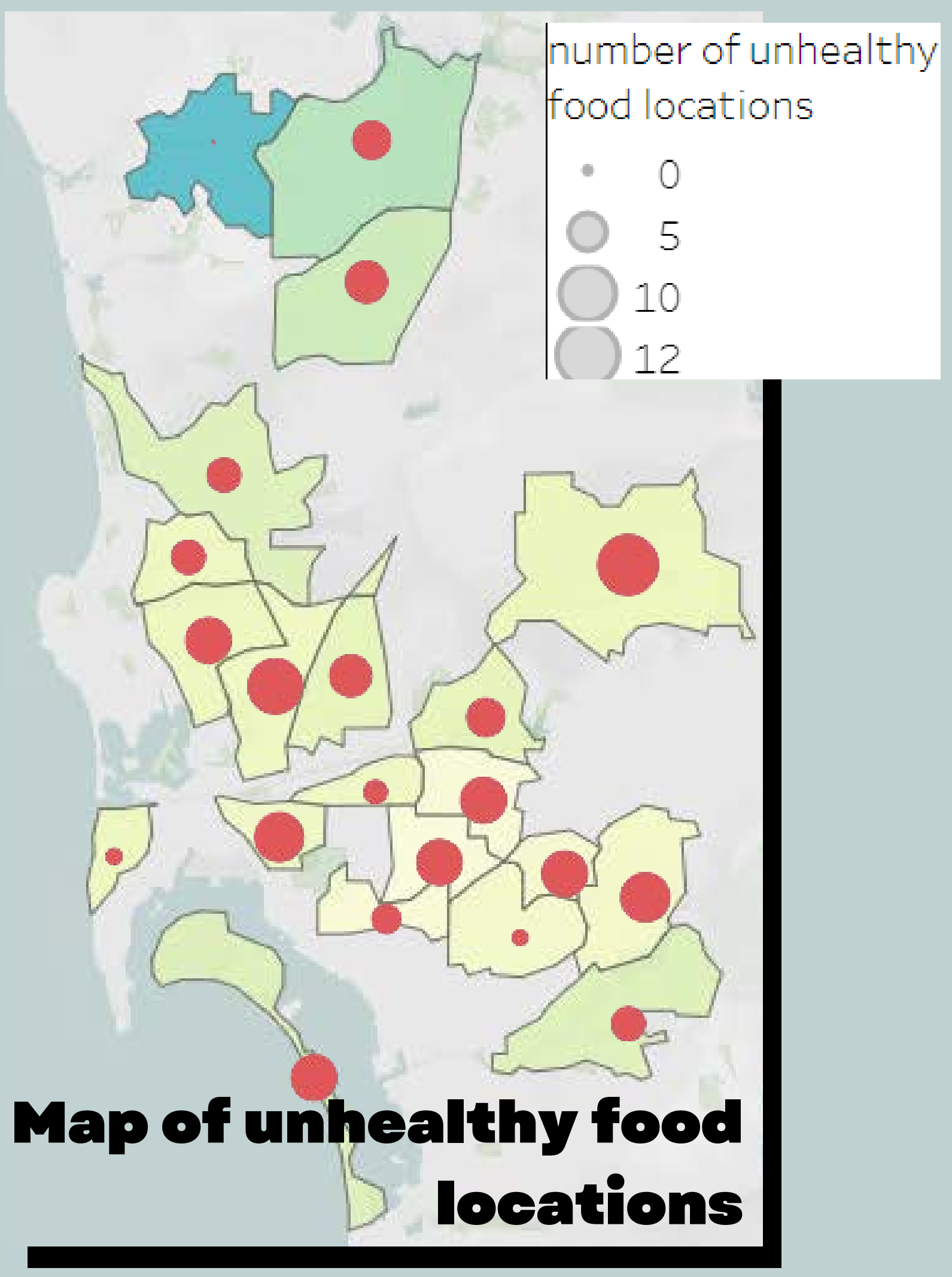
HYPOTHESIS

THERE WILL BE A POSITIVE CORRELATION BETWEEN INCOME AND HEALTHINESS. LOWER INCOME COMMUNITIES WILL HAVE MORE ACCESS TO UNHEALTHY FOOD LOCATIONS AT A HIGHER RATE THAN HIGHER INCOME COMMUNITIES.

DATA AT A GLANCE

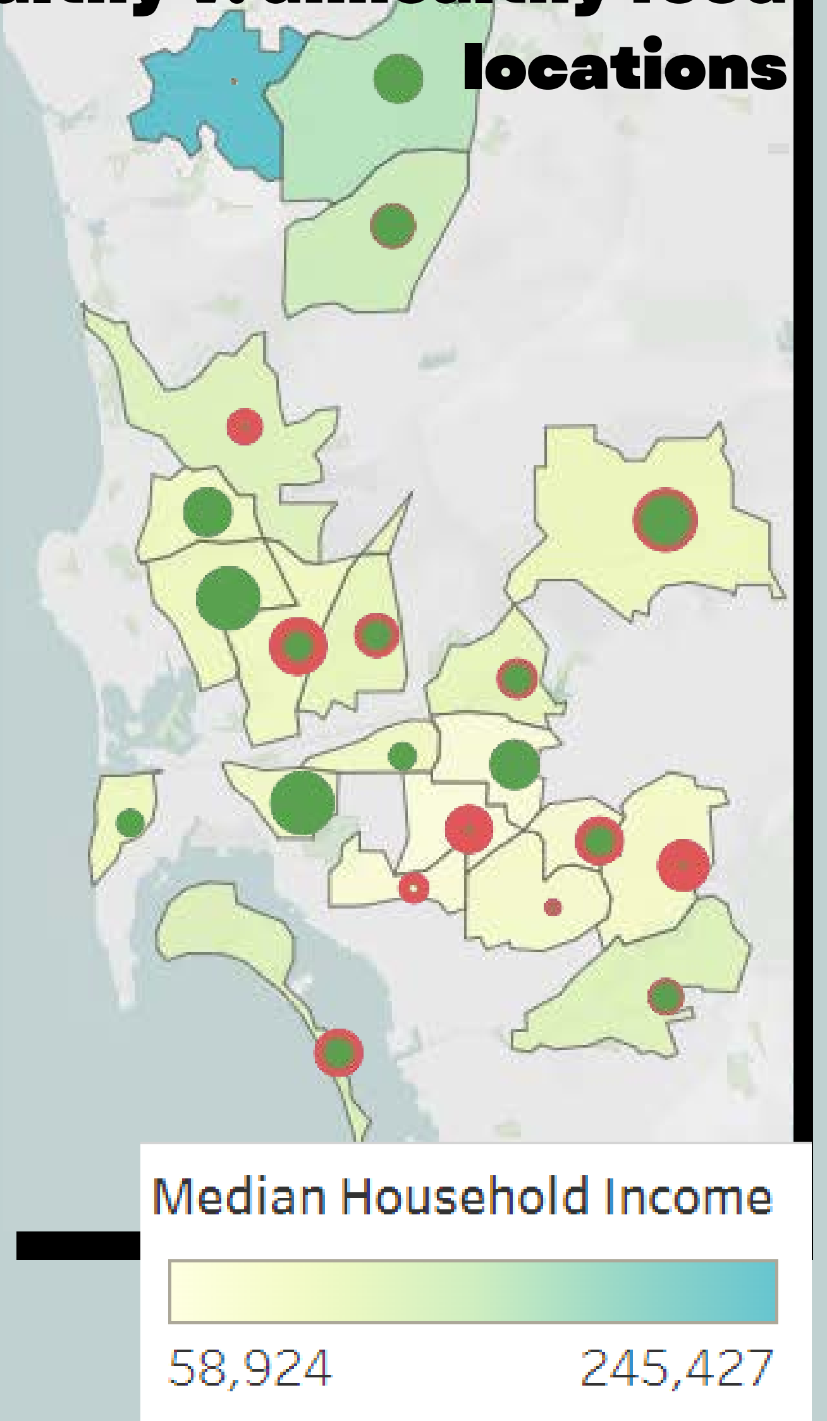


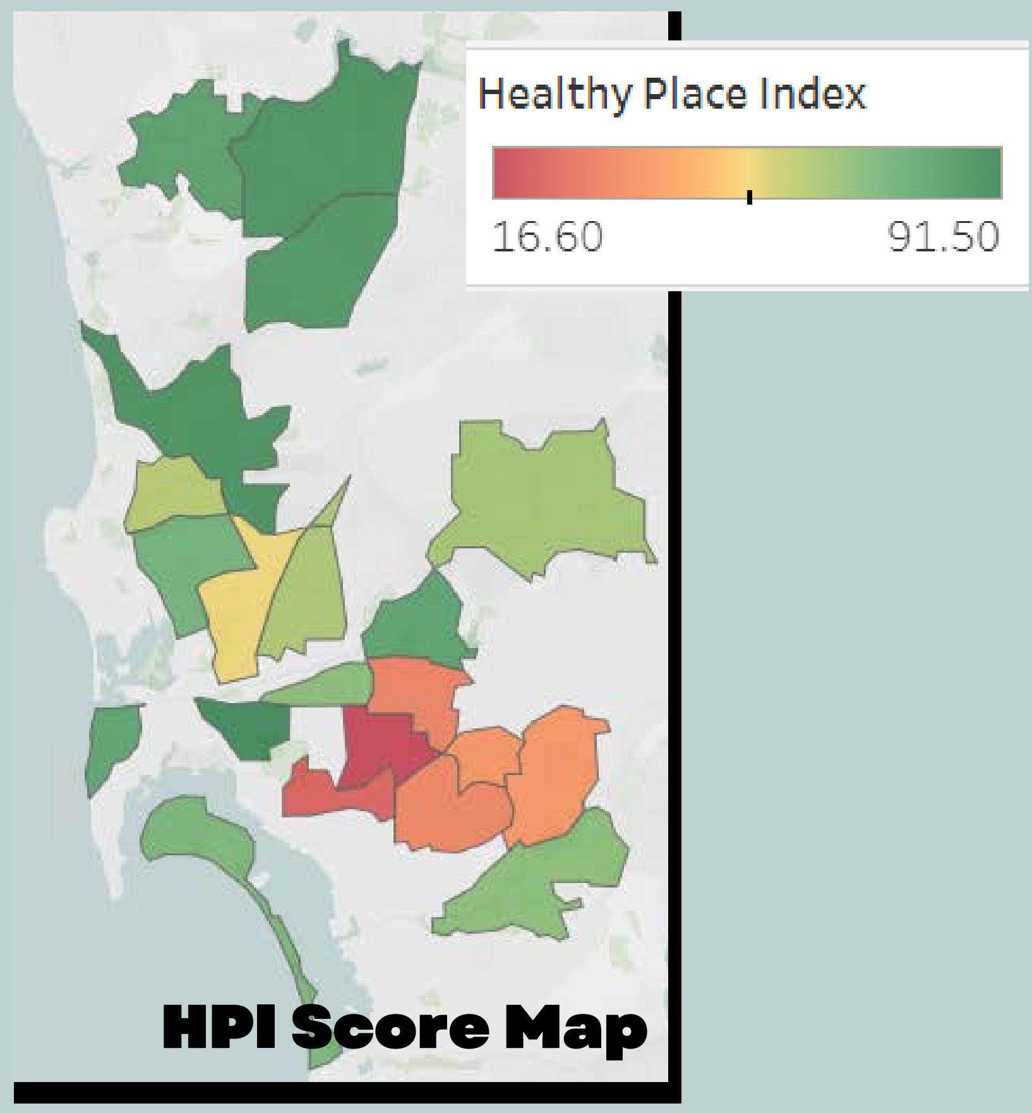
Map of healthy food locations



Map of unhealthy food locations

Map of healthy v. unhealthy food locations



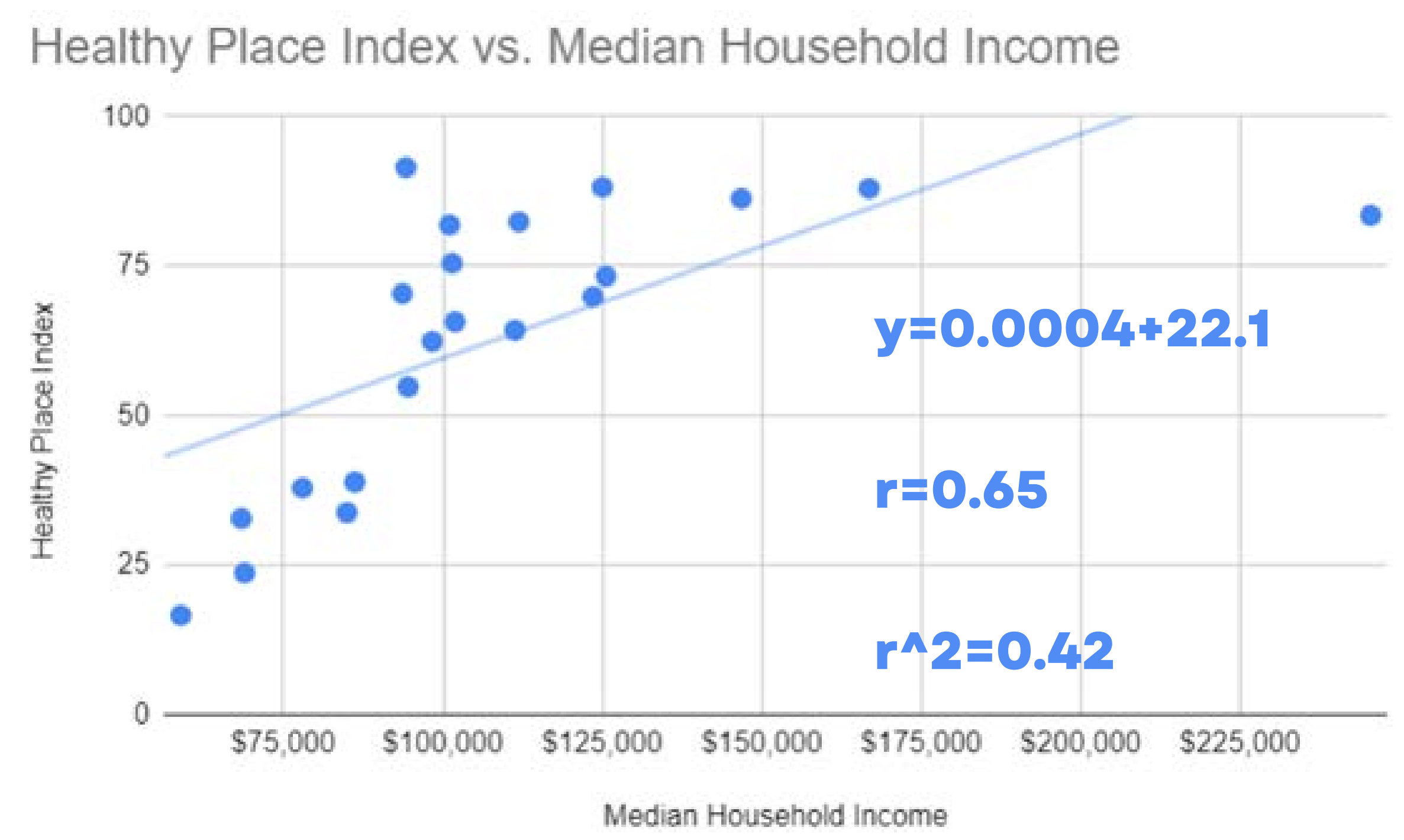


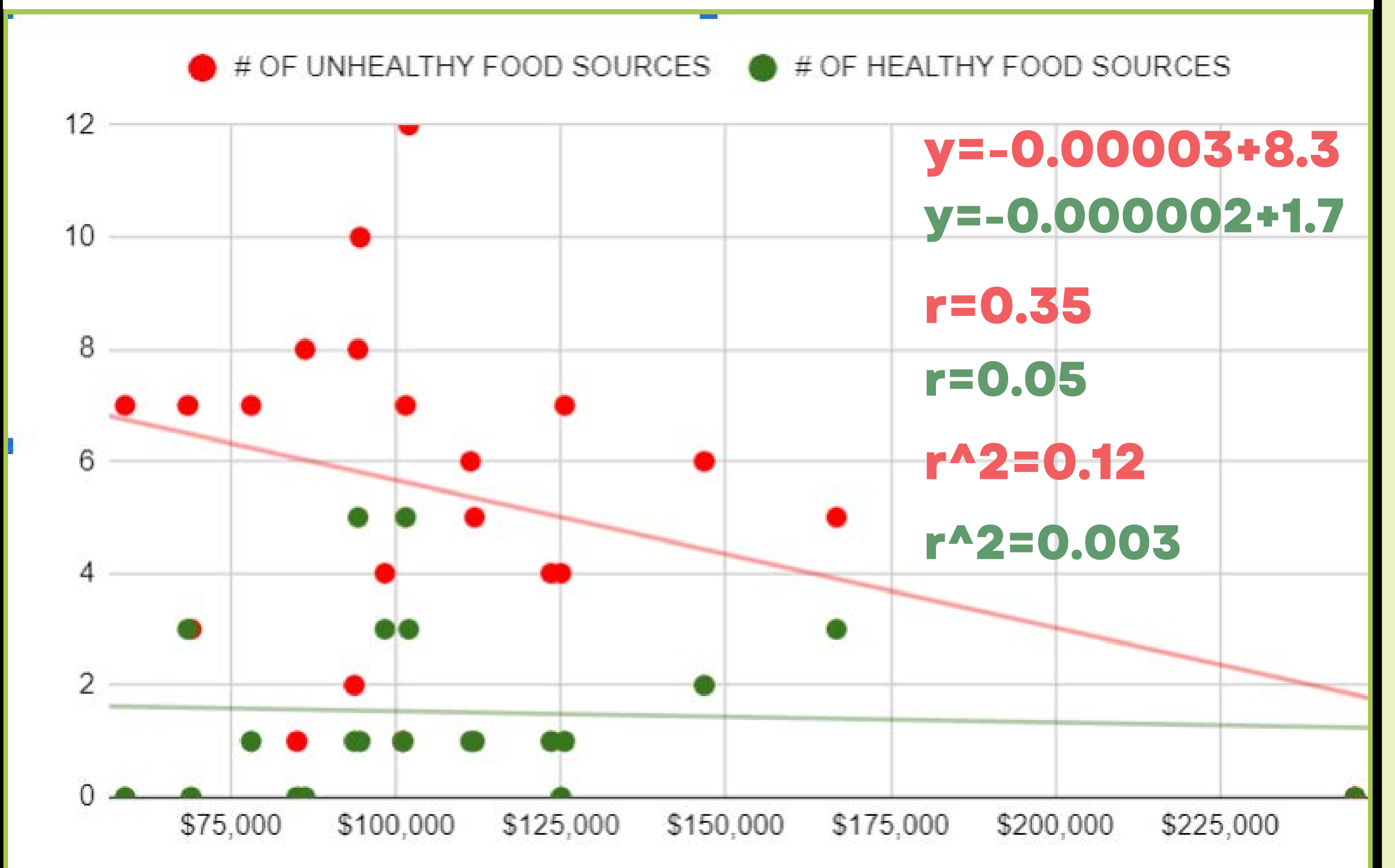
HPI Score Map

- The HPI scores have a positive correlation with income when comparing maps and using linear regression graph. The r^2 of 0.65 means that this is a relatively good correlation.

- According to the graphs, both healthy and unhealthy food locations have a negative correlation with income. BUT, looking at the map that compares the two, the places that show red (meaning as a ratio it outweighs the healthy food at higher rates) are typically in lower income areas.

Healthy Place Index vs. Median Household Income





CONCLUSION

Income in a community in San Diego has a relatively good correlation with health scores (the higher the income, the better the health). However, the correlation with the accessibility to both healthy and unhealthy food locations through our experiment is not strong enough to determine whether the income level of these communities gives them more or less access to healthy food locations. That being said, there was somewhat of a negative correlation to the number of unhealthy foods v. income (r^2 = 0.35), and we speculate that if we had looked at all fast food locations, the r^2 might have been higher. We also observed the emergence of a possible pattern where in most locations, the number of unhealthy food locations were greater than healthy food locations (this difference was seemingly greater in lower income communities). If we were to gather more data to see a stronger correlation and take all locations into consideration, this problem could mean more efforts need to be implemented to improve accessibility to healthy food in all locations.

CHALLENGES

- Using maps to quantify restaurant/store locations was time consuming and could have caused inaccuracies.
- The census did not provide income information for all zip codes in SD county.
- Missing zip codes means we did not consider all the locations of these sites in SD.
- The 10 named locations are not all the healthy/unhealthy food locations in SD, so it is a less accurate measure.

Clean or a Crime Scene: What are your Pittsburgh Parks like after dark?

Roisin Tsang, Kennedy Waffensmith, Kaitlyn Miller, Veronica Karpov, Kendall Reilly

The Problem:

Crime rates have been high and rising in recent years, so our project seeks to understand the complicated factors that may be affecting this issue.

Null Hypothesis:

There is no relationship between time after park cleanings and crime rates.

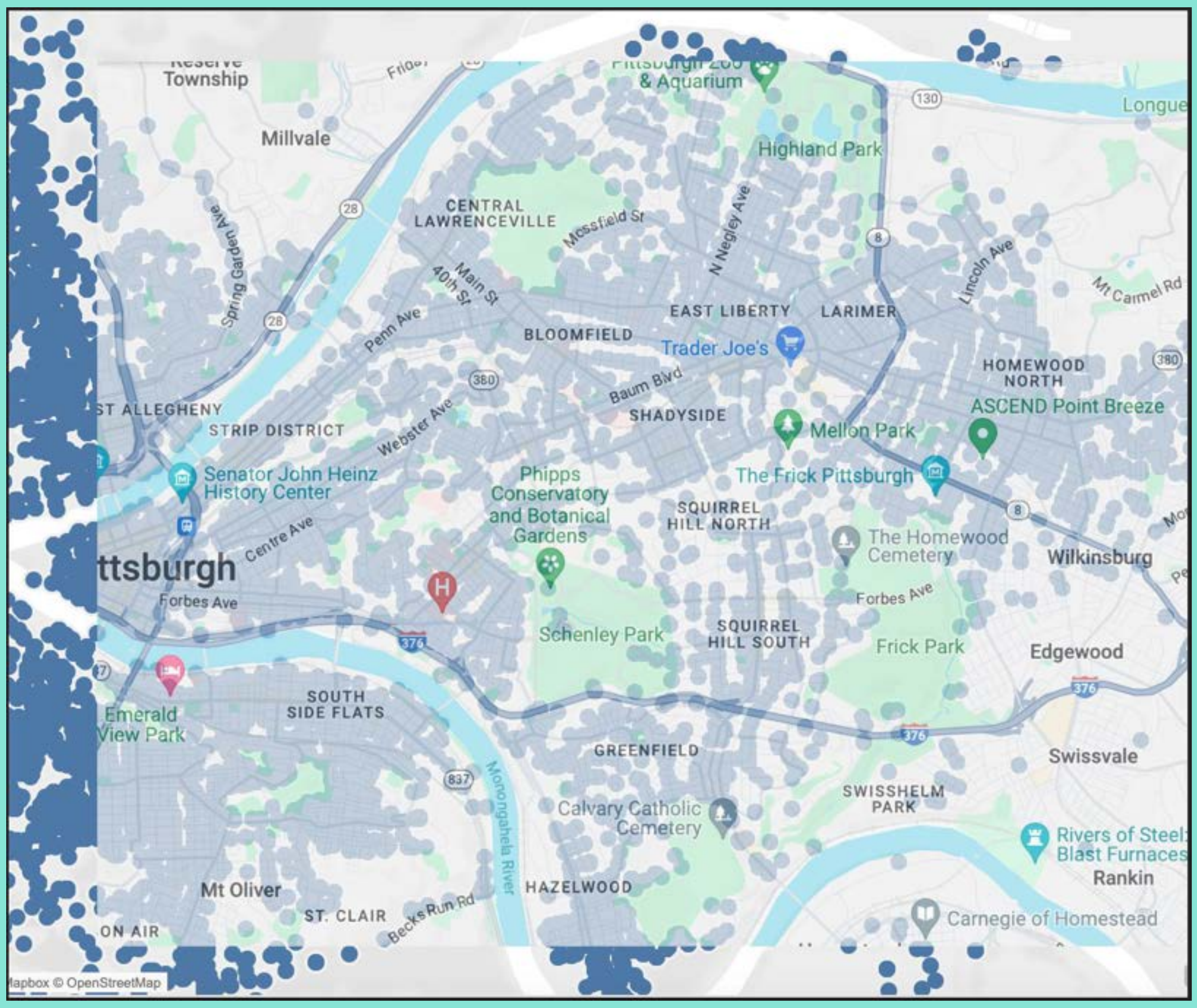


Figure 1: crime locations imposed over a map of Pittsburgh parks

Challenges:

We are a small team, so there was difficulty with large workloads for each member. There was also no available database for the geographic locations of parks, so we had to create our own.

Alternate Hypothesis:

There is a negative correlation between time since park clean - ups and surrounding crime rates

Preliminary Results:

As can be seen in Figure 1, crimes generally do not happen within parks, but the same cannot be said for areas surrounding parks.

Conclusions:

Based upon this preliminary research, it seems that the more parks there are, no matter how small, will cause less crimes to be committed within a neighborhood

Recommendations:

More research into this topic must be done, but preliminary research shows that green spaces are correlated with less crime, therefore we would reccommend greater green spaces

Sources:

City Parks and Addresses: https://www.pittks.org/community/city-parks/ , Clean Up Calendar and locations: https://pittsburghpa.gov/dpw/volunteer-apps/ , Parks and Crime Grade(Rating based on crimes committed): https://www.drkattorneys.com/blog/neighborhood-crime/ , Some Park Locations and Quick Information: https://www.anyplaceamerica.com/directory/pa/allegheny-county-42003/parks/?page=2 , Arrest Data: https://data.wprdc.org/dataset/arrest-data

Political Afilliation and Health

Tivadar Torrez, Juan Vega Sanchez , Xzavier Aguilar, Emely Tejeda Tejada

Passaic Academy for Science and Engineering

Mentor:Ajeet K Subramanian

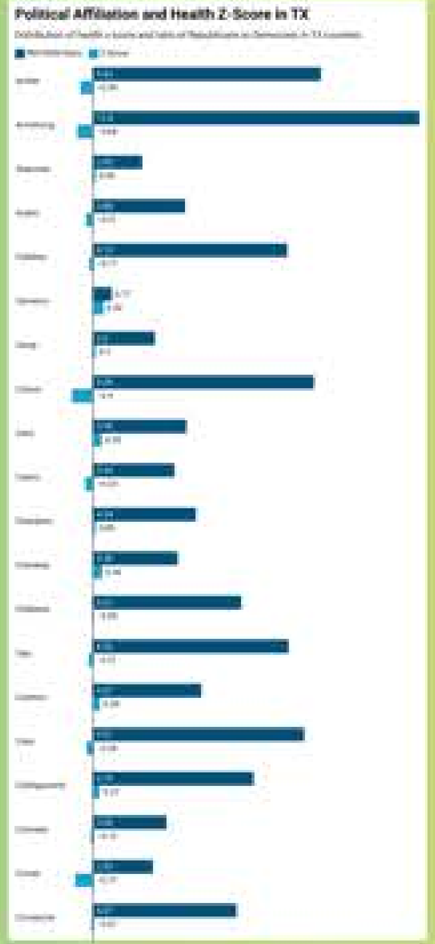
Project Development:

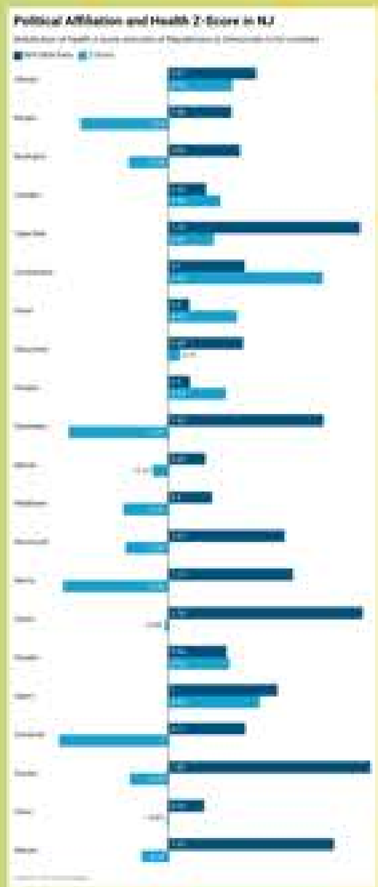
- When we first began the project, our hypothesis was that people in certain political parties or affiliations, will have a slightly higher probability of facing health issues due to their conservative beliefs or resistance to established scientific fact.
- At first we began looking into different categories or subjects we believed could have some sort of relationship. Sooner or later politics came to mind.
- We found this topic interesting because people's political stance and beliefs change and that sometimes could affect their health.
- To find datasets we used links cited on the DataJam website for New Jersey dataset, and mostly just google searches
- Finding raw datasets wasn't easy, but we were able to find good sources to analyze
- Due to the large amount of counties in Texas, we ran into some issues and decided to get rid of most of its counties and focus on specific counties

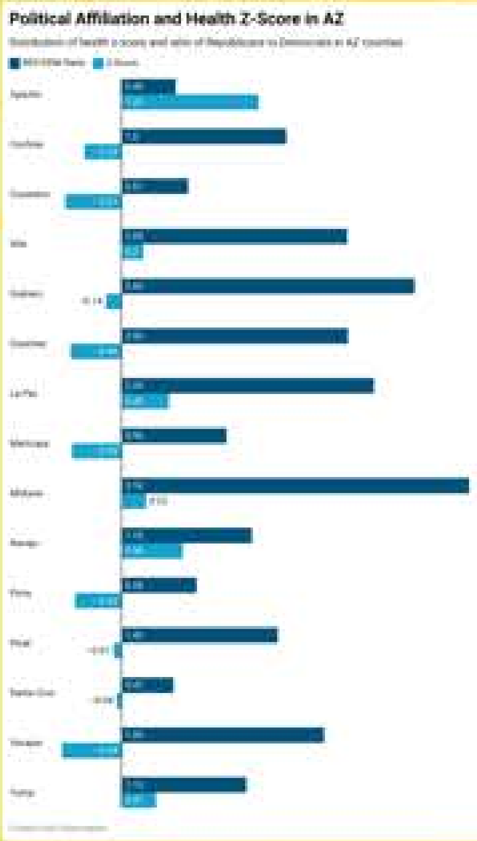
Data:

- Datasets specifically for NJ Counties or NJ in general wasn't easy to acquire. When finding these datasets we needed raw numbers and certain types of files to actually download them.
- We used NJ health datasets and political party affiliations to compare the counties in NJ (A Democratic led state) and their health rankings
- We also used Politico to gather some of the most recent voting in elections from the states of Arizona, New Jersey, and Texas
- We used the state websites of Arizona and Texas to gather their health data

Data visualizations and Findings:







interpretations and Recommendations:

Based on these graphs we can conclude how in states with different political party preference and different ratio of republicans to democrats in their respective tend to have no correlation between their affiliation and health ranking. This trend can be seen in all of the three states, which shocked us as the data revealed political affiliation doesn't affect the health of the population living in the state.

References:

PositCloud/Rstudio, Official site of the state of New Jersey, Texas, and Arizona (Voter info by county), Politico

High School Students: Balancing School and Work

Ryan Balerio, Leonardo Martinez
Passaic Academy for Science and Engineering

Mentors: Ajeet K Subramania & Kris-Jolen Lara
University of Pittsburgh *Caldwell University*

Project Development

When we started our hypothesis was going to be weather students that have a job would do better or worst than those that don't have a job. The project evolved from us wanting to find out who did better at school to weather students get better at managing their time by balancing a job and their education. We decided to do this for our project because we generally wanted to know if having a job affects a students educational life.

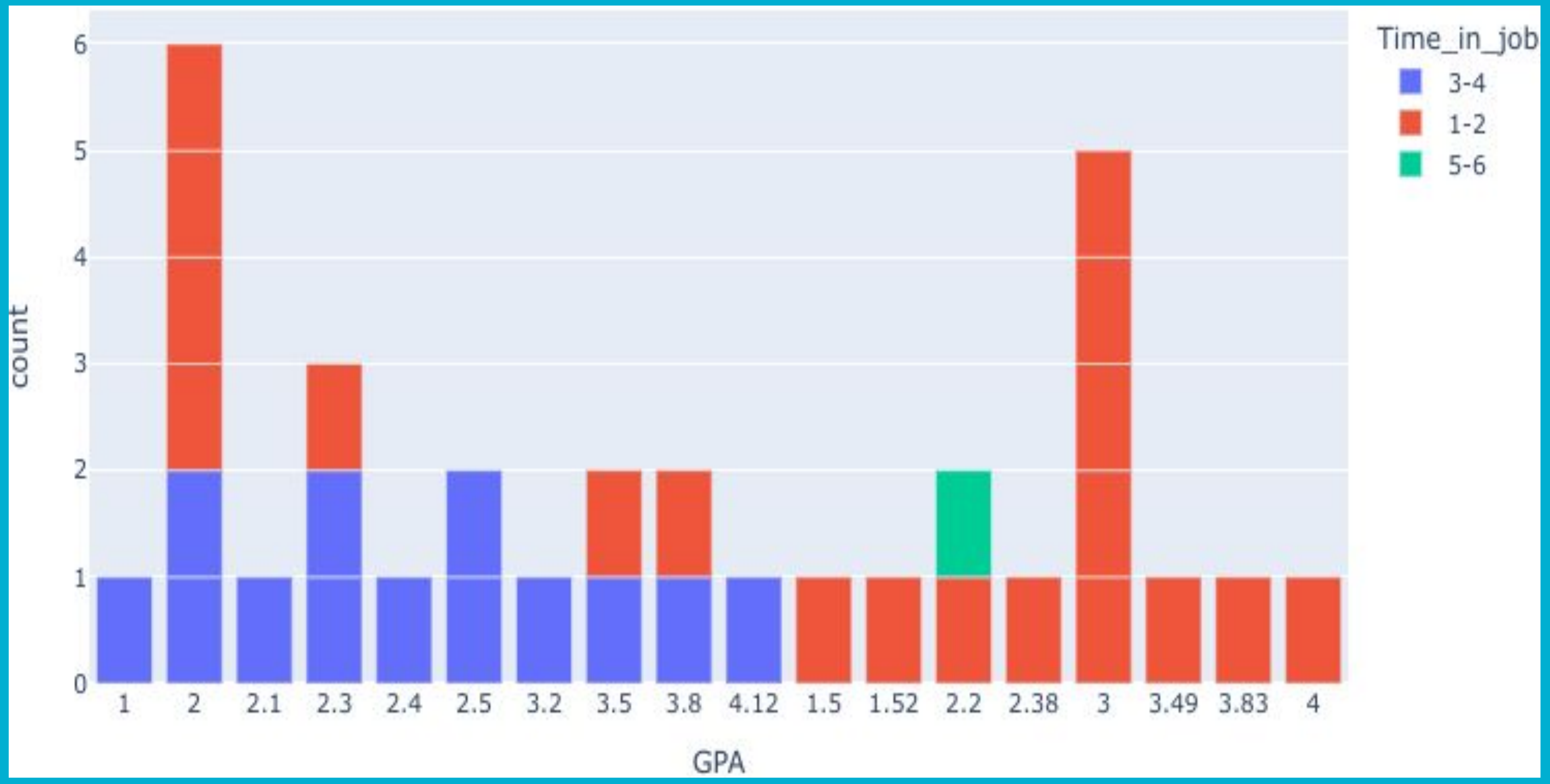
Data

We made a survey to put in our school announcements with the approval of the principal for highschool students to do, it was overall an easy thing to do we just had to make the students aware that this survey was up.Some of the data had to be cleaned up for one of the questions since there were many different answers so we had to round up the numbers.

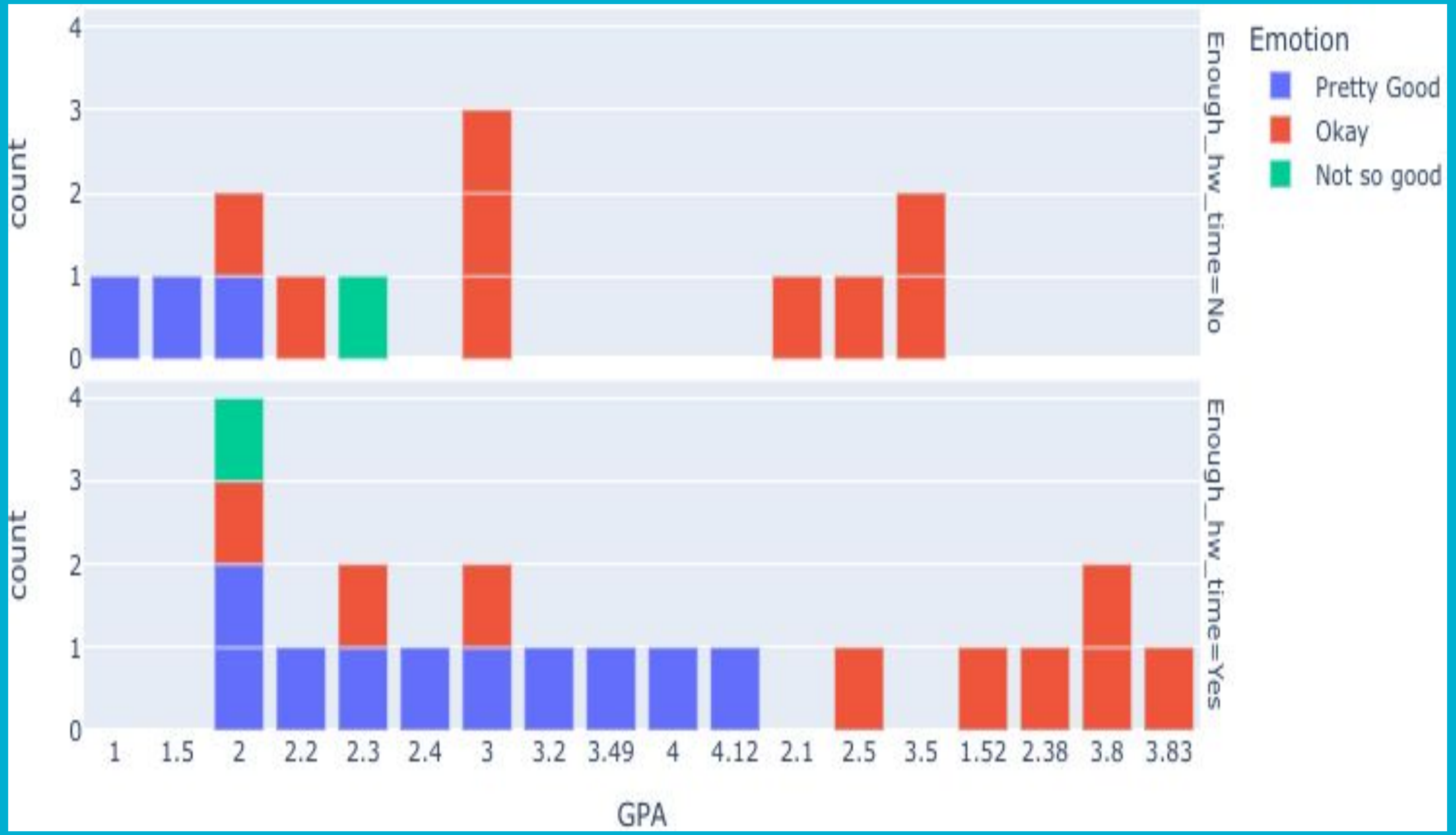
Interpretations and Recommendations

Based on our results we are unable to give certain recommendations due to the lack of data we collected.

Data Visualizations and Findings



This bar uses the GPA 1-4 from the students that have a job and color codes them by how many times they work a week. This is able to display the GPA and its correlation to how much they work a week.



This bar also uses the GPA from the students that have a job and color codes them by how they are doing emotionally, and also separates it into two bar graphs where one is for those that think they have enough time to do their homework and those that think they don't.

References

- Starting Data
- Ending Data
- Survey With Questions Used

Thanks to DataJam, Pittsburgh DataWorks, our teacher Mr. Chomko, and our awesome Mentors for this awesome opportunity!

School Improvement with a Focus Area on the Foreign Language Department

RAMAN Team

Robert Eberly

Aidan O’neill

Michael Chotiner

Adan Mai

Neehaan Patel

Question: Why is there a severe drop in the number of enrolled students as they progress through the different levels of foreign languages?

Hypothesis: Fewer students continue their language education because of concerns about teacher quality, class relevance, educational policies, social trends, and cultural shifts.

Central Dauphin High School

With the help of:
Bob Moreland
Andrew Lindros

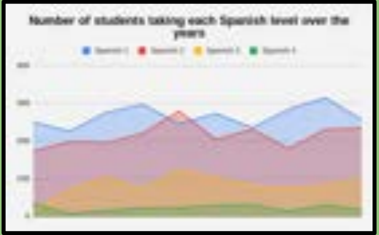
Challenges:

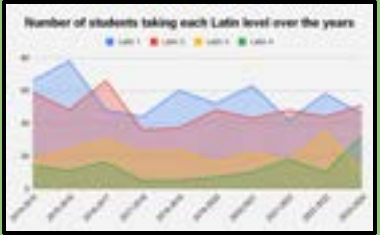
-Obtaining Data
-Communicating with the teachers
-Time
-Smoothing out internal politics with language department

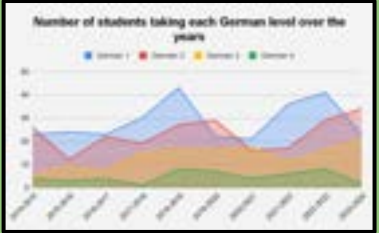
Side Investigation

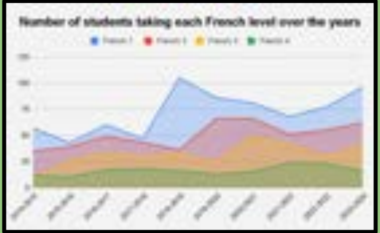
We looked at test score data of our district and school. We gathered data from a source(School Digger) that had data from the DOE(Department of Education). Using this data, we analyzed the increase in students compared to the drop off in test scores.

Data Sets:









Analysis:

If we had reliable data, we would have used an ANOVA test to compare results from the survey. For example, when comparing how students rated their teachers, the null hypothesis would be that all teachers are rated the same. An ANOVA test will test this null hypothesis against the alternate hypothesis, that one language is different.

Results

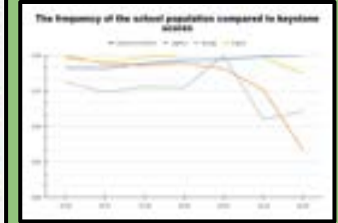
Using an ANOVA test with the unreliable data we had, we rejected the Ho, meaning that there was a difference in the mean when rating the teachers. However, due to insufficient data, we are unable to make any conclusions that describe the decrease in students.

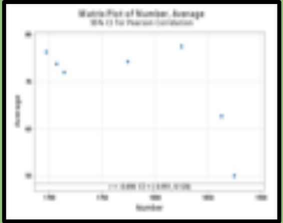
Conclusions and Recommendations

We are unable to draw conclusions because of the state of the data. If we were able to obtain a census or use a random sample, we would have been able to analyze the survey and determine possible trends

Next Steps

We would plan to revise the survey to better satisfy the teacher’s interests. Then use this improved version of the survey to make reliable conclusions about why students are not taking higher levels of their language courses. After obtaining this data, we would use a similar analysis plan to fully understand the data collected. Once we figure out what a possible problem is, we will investigate it further.





General School Improvement

The graphs to the left show that with an increase in the school population, test scores drop. This could be due to teachers not being able to give attention to all of their students. The same could possibly be said with the language department.

Promoting Social Inclusion Through Intro, a Novel Computationally Driven Matchmaking Platform

Can a computational platform that groups individuals via compatibility survey to facilitate structured interactions promote social inclusion within a culturally diverse, traditionally underserved, and socioeconomically disadvantaged student population?

THE PREUSS SCHOOL UC SAN DIEGO:
Yared Fente
Maryamawet Debele
Brian Sanchez
Nathan Cherinet
Aron Ekubaselase
Ruth Kibrom
Yenatfanta Hailemariam
Nathineal Getachew
Samson Lemma
Hao Nguyen
Andy Phan

01. Introduction

Friendship is essential for mental and physical health, providing emotional and psychological support and reducing stress. However, the advent of social media has complicated friendship for members of Gen Z, for whom social media often provided echo chambers that exacerbate societal divisions and provide unrealistic standards for comparison. Fortunately, however, fostering in-person interaction and community-based social network development can empower individuals to form meaningful connections and gain an improved and realistic sense of social identity. Our approach to promoting healthy socialization involves utilizing Intro, a computational platform that quantitatively measures social connection, identifies friend groups, and utilizes common interests and preferences to foster new in-person interactions. Through these efforts, we aspire to cultivate a culture of empathy, understanding, and connection within our community.

02. Methodology

To assess existing friendships within a group of 30 students comprising a single advisory class, each student in the advisory was asked to rate the potential for friendship and/or productive collaboration (five-point scale: 1 (low potential/current conflict or disagreement) to 5 (high potential/active friendship or productive working relationship)) with each of the other students in the advisory. Those data produced a 30 x 30 square matrix on which several analyses were performed.

03. Data Summary and Descriptive Statistics

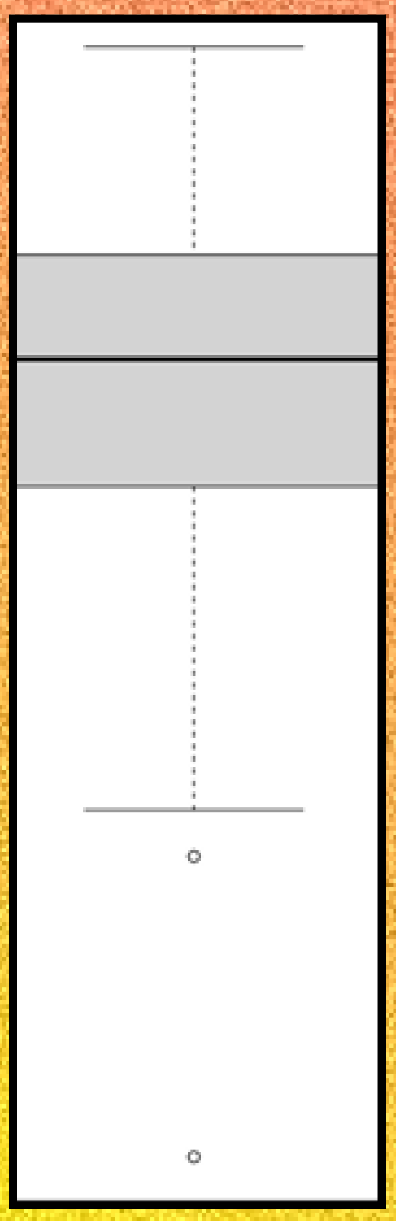
The mean friendship rating provided by peers was calculated for each advisory student (note: names were replaced with numbers, 1 through 30, to preserve students' anonymity). A five-number summary was produced to describe the set of mean values; additionally, those values were displayed by a box plot with outliers identified by the 1.5 IQR criterion represented as open circles. Further, the mean values were used to produce a histogram that adopts an otherwise approximately normal distribution with two (or three, if slightly less strict criteria than 1.5 IQR are used) low outliers.

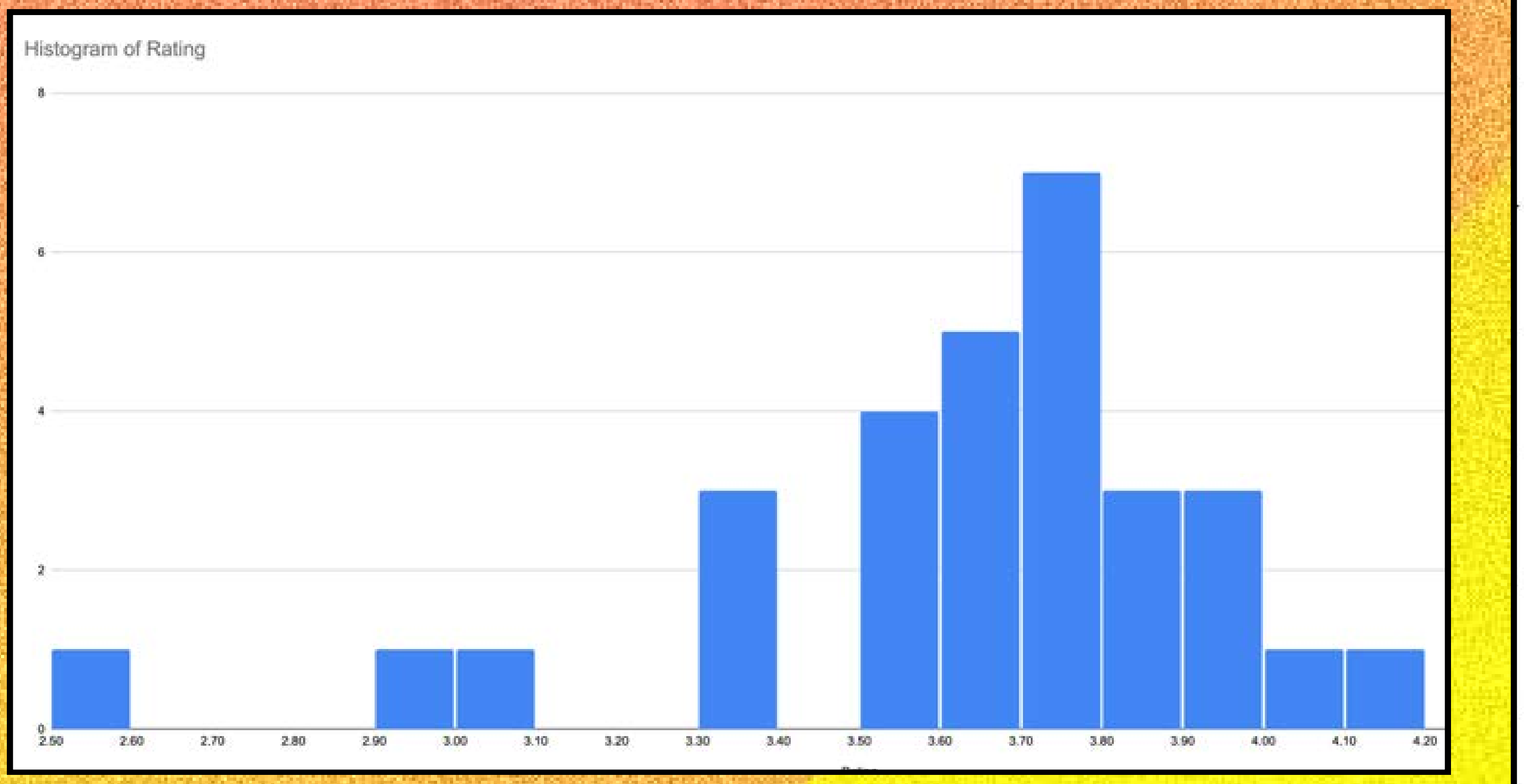
Multivariate Analyses:
Principal component analysis (PCA) was performed on the full data set (30 x 30 square matrix of friendship potential ratings). K-means clustering was then performed on the PCA results. Together, these analyses provide, roughly, a radial distribution of individuals (PCA) that can be subdivided to reveal three social groups (k-means clusters).

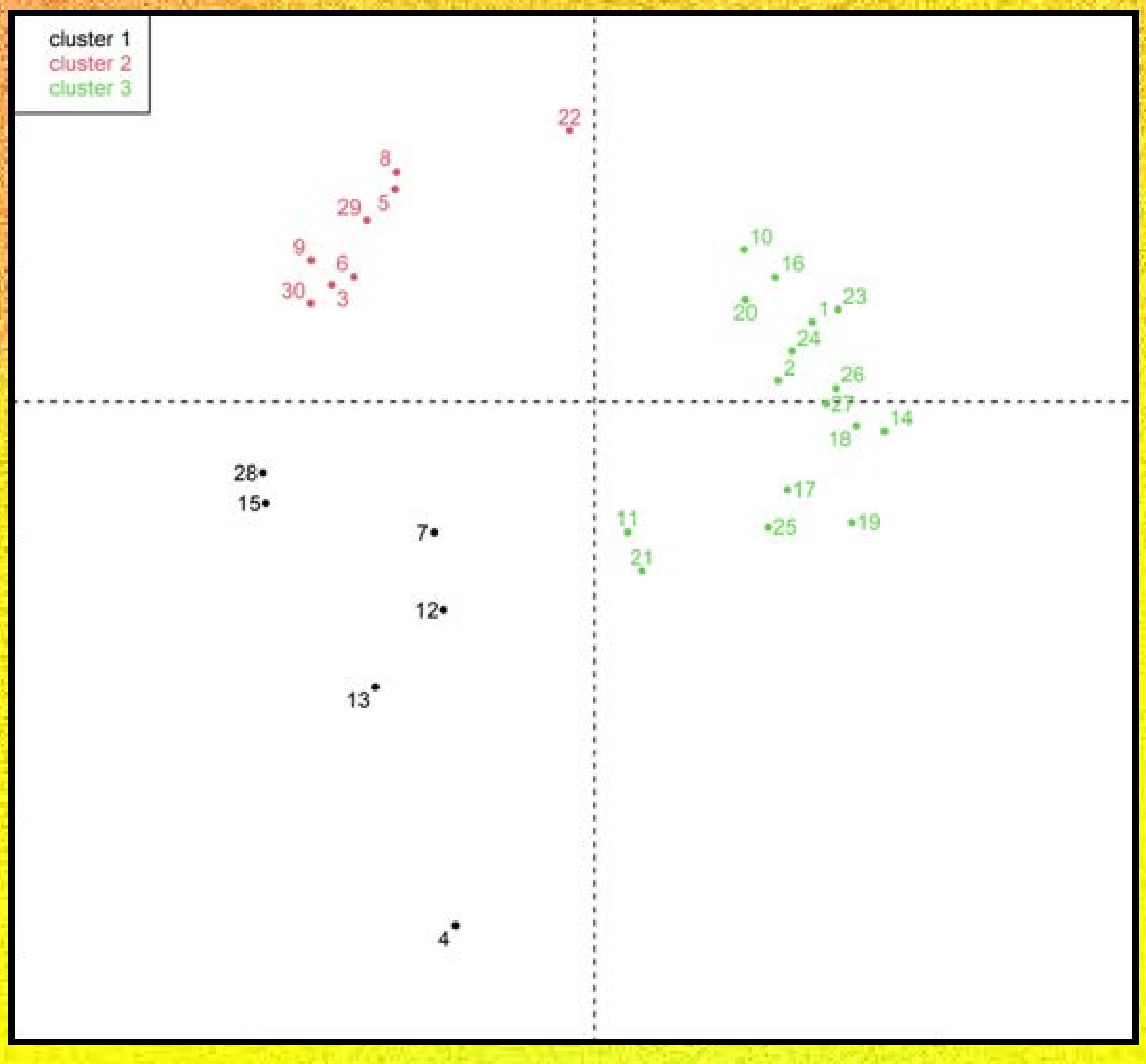
Five-Number Summary	
Minimum	2.53
Q1	3.5
Median	3.68
Q3	3.8
Maximum	4.13

Friendship Potential

Student	Mean Rating
1	3.80
2	3.60
3	3.80
4	2.53
5	3.50
6	3.67
7	3.53
8	4.03
9	3.90
10	3.83
11	3.03
12	3.33
13	2.97
14	3.60
15	3.37
16	3.90
17	3.93
18	3.80
19	3.77
20	3.97
21	3.50
22	4.13
23	3.93
24	3.60
25	3.67
26	3.70
27	3.70
28	3.53
29	3.80
30	3.37







04. Results/Findings

Descriptive statistics of mean friendship potential (FP) ratings reflect a group of students who share, on average, reasonably positive interactions and relationships with one another. Notably, the first quartile (3.50), median (3.68), and third quartile (3.80) are all above the neutral friendship rating of 3.0. However, while the histogram and box plot appear to represent a group of students who are uniformly moderately positively associated with one another, the PCA and k-means clustering tell a different story. From PCA and k-means, we can see that the students occupy three discernable clusters whose (relatively but with some variation) tight dispersions demonstrate that the students in those clusters associate positively with one another (note: this conclusion is supported by elevated intracluster mean FP ratings, not shown, and by friendships known within the advisory). Further, these analyses reveal that students are generally unfamiliar with those outside their friend groups (neutral intercluster mean FP ratings, not shown, reflected by separation along the first two principal components as shown in the combined graph from PCA and k-means clustering). This result was surprising, given that the majority of the students in this advisory group have been together for five years/grade levels. Additionally, the known identities of the students (again, hidden to preserve anonymity) reveal that the default friend groups present in the advisory have been established by means consistent with association based on gender and race (first principal component capturing 33% of the variance in the data) and introversion/extroversion or, if preferable for its quantitative basis, mean FP rating (second principal component capturing 20% of the variance in the data). However, as we will discuss in the next section, the revealed unfamiliarity provides an opportunity to foster connection.

05. Reccomendations

The data collected demonstrate that the students in the advisory group studied are further socially siloed within that group. Within the context of the advisory group, students associate with a small subgroup of friends but otherwise have little knowledge of or familiarity with those outside their friend groups despite having existed in the same cohort for five years. These results represent preferential association with members of the same gender, race, and/or inclination to socialize (introverts with introverts and extroverts with extroverts) and are consistent with clique/echo chamber formation promoted by social media. Thus, to promote new connections, we recommend (and are in the process of testing) fostering interaction among members of different social groups.

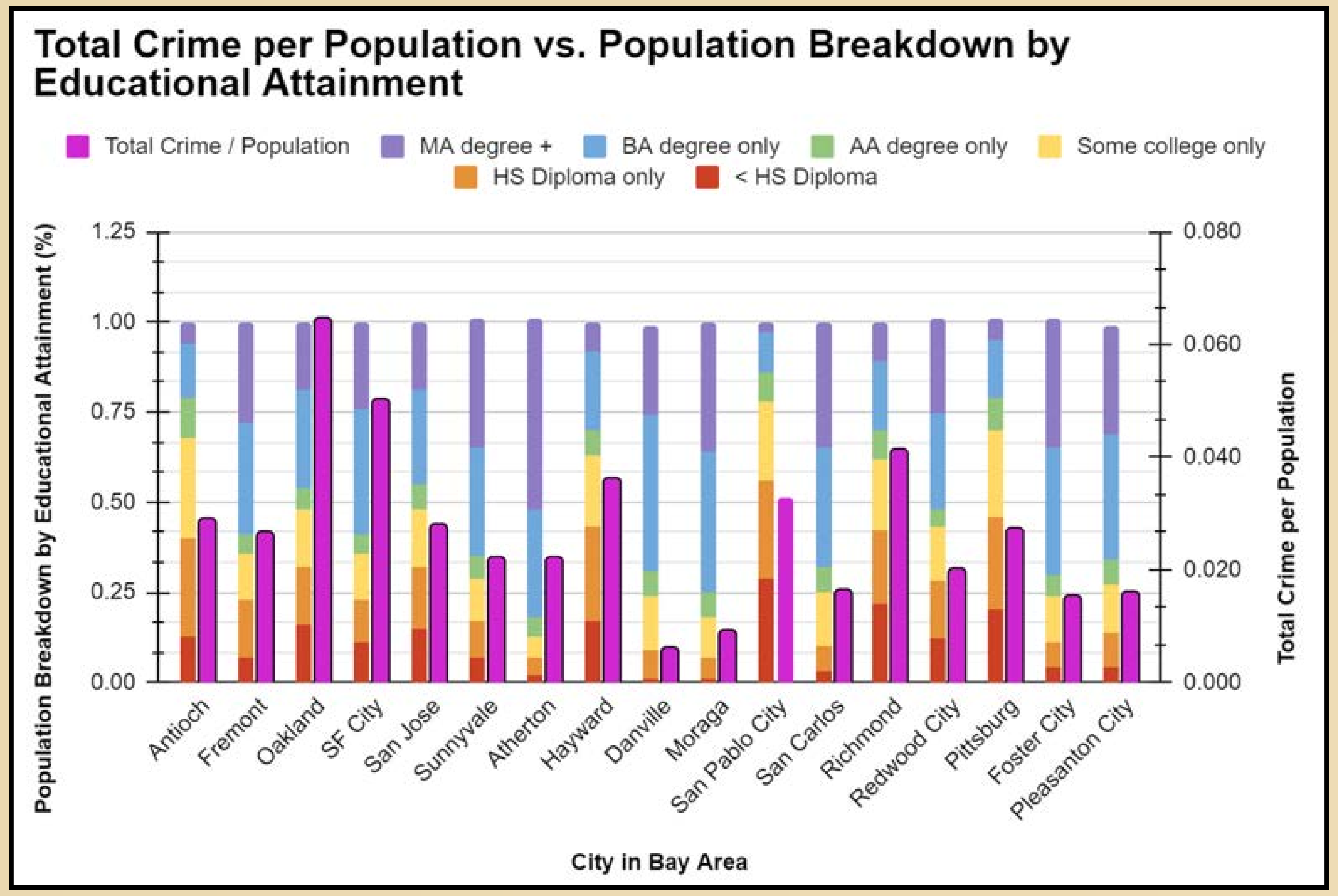
Indeed, each advisory student is participating in a social growth project titled Share the Fun. In this project, students were initially introduced to twelve different members of a different social group (two k-means clusters were used for group creation preceding this activity) through an active-listening icebreaker activity. Then, an inventory of personal preferences and interests was used to group pairs of individuals from different social groups (two students from one social group paired with two from another) sharing a set of identified common interests (multiple correspondence analysis, not shown). Thus far, early observational data (lively interactions, positive informal student feedback, and student-expressed intrinsic motivation to continue with the project) indicate that students are enjoying the process of making new connections; we hope to provide follow-up results during our presentation.

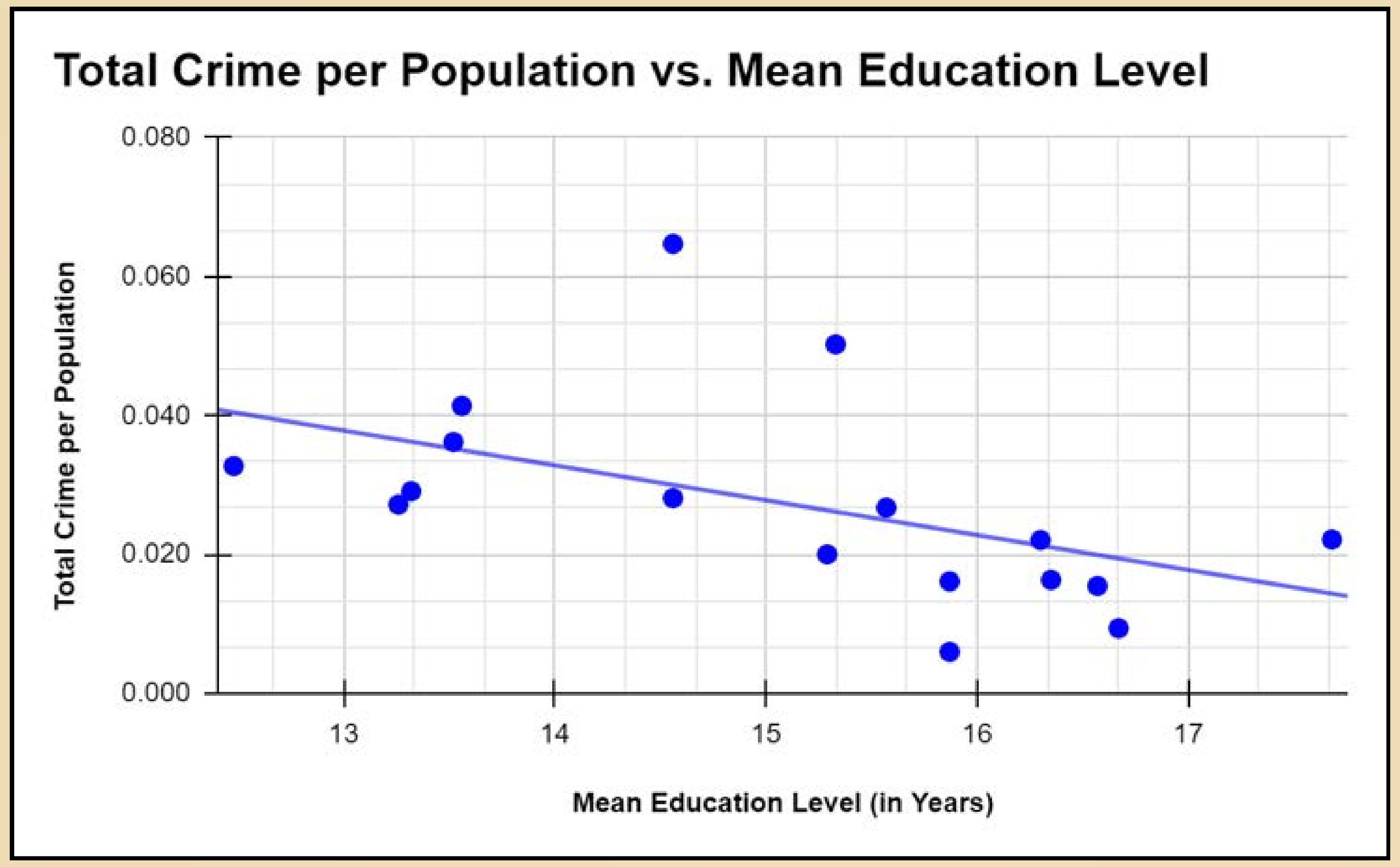
Education: the Solution to Bay Area Crime?

A study on the relationship between educational attainment, median income, and crime in Bay Area cities.

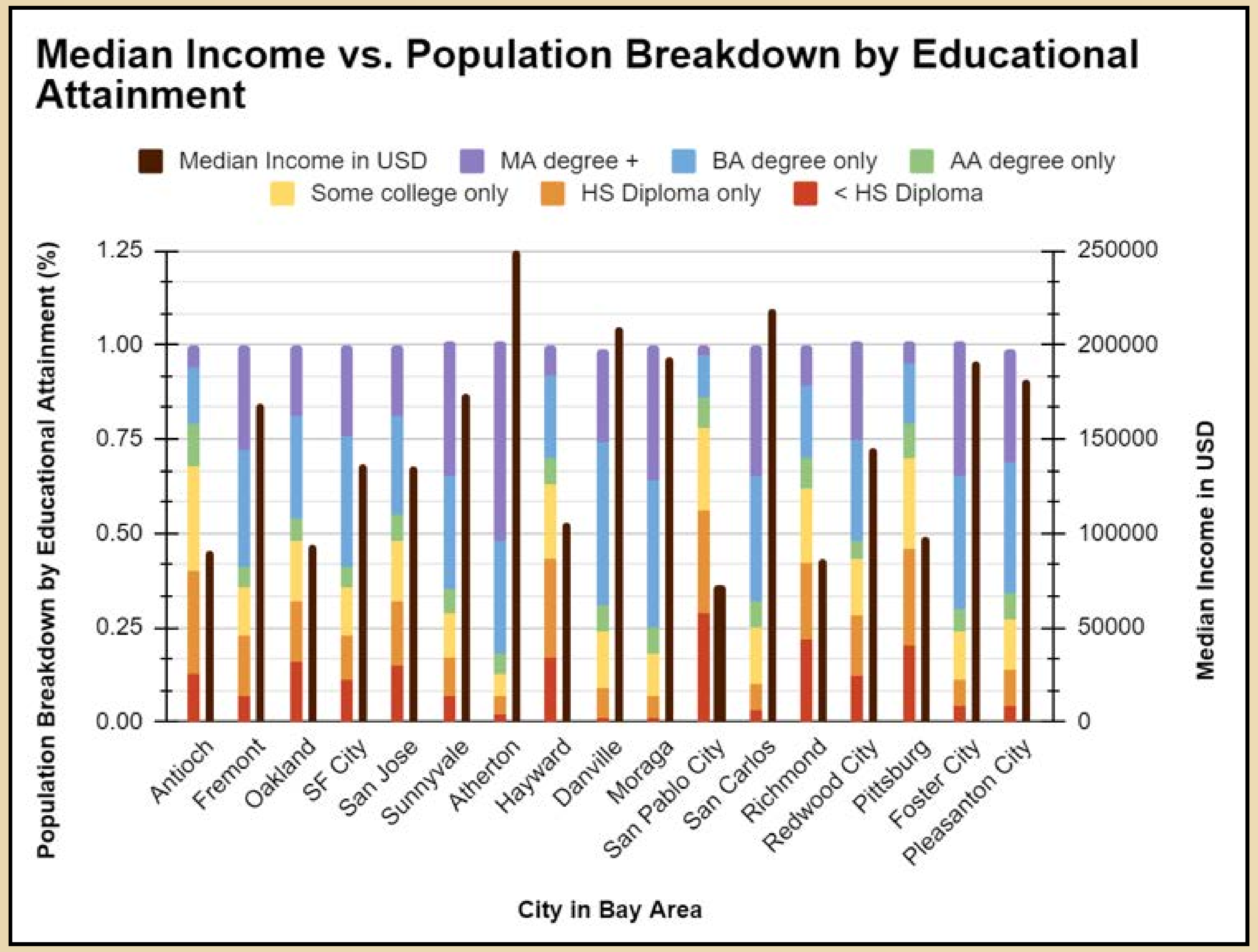
San Ramon Valley High School: Vaishnavi Akella

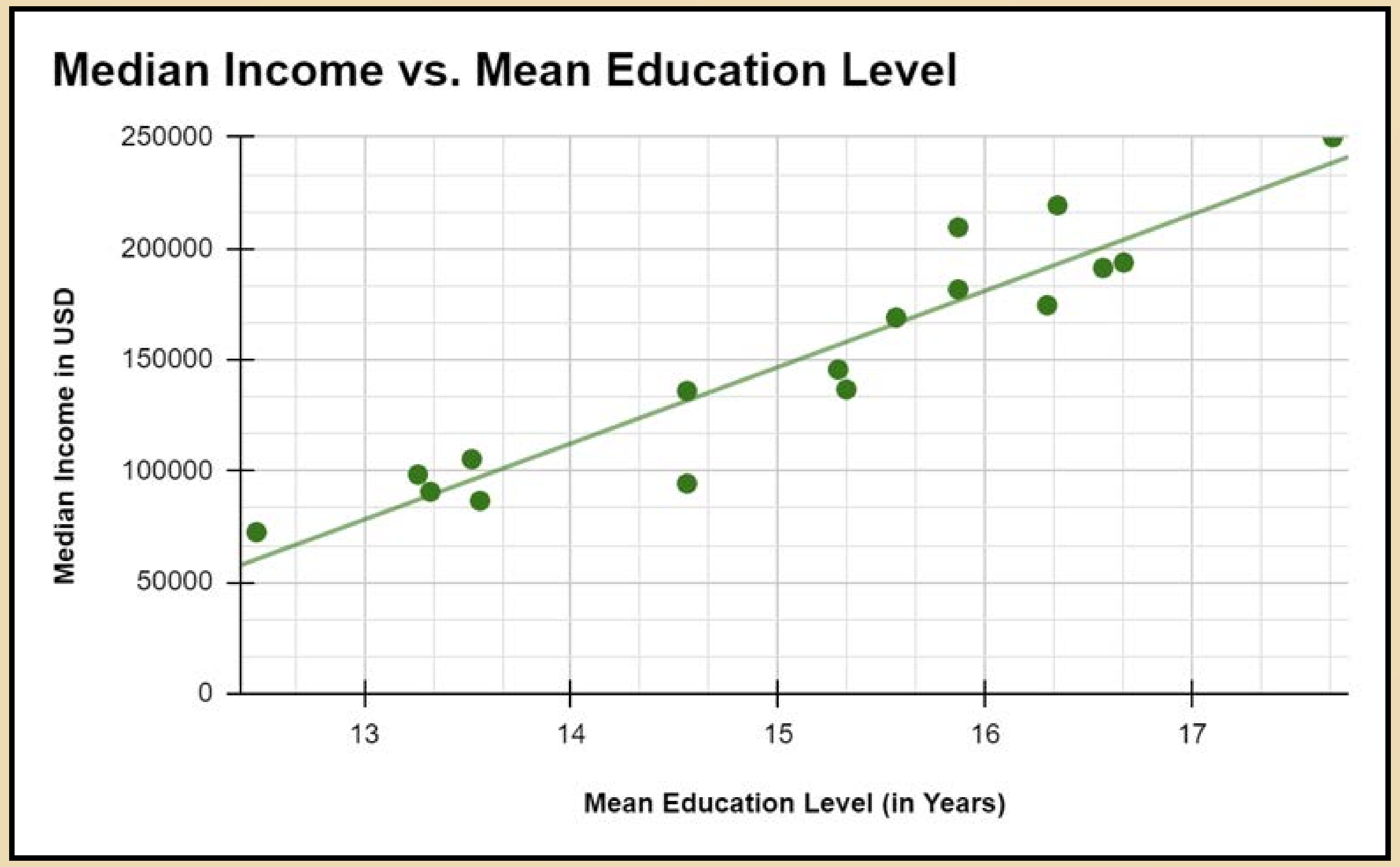
Graphs and Results





In the Linear Regression Analysis, educational attainment and total crime per population have a statistically significant Pearson correlation (*r* = -0.501, *p* = 0.041). The negative association means that a relatively lower education is associated with a relatively higher total crime per population, and vice versa, which is what I hypothesized. In the Correlation Matrix, many other variables are also highly correlated with crime, which explains why the correlation here is not as strong as the one below.





In the Linear Regression Analysis, educational attainment and median income have a strong statistically significant Pearson correlation (*r* = 0.945, *p* < 0.001). The positive association means that a relatively lower education is associated with a relatively lower median income, and vice versa, which is also what I hypothesized.

Conclusion

This study finds that educational attainment and crime have a negative correlation, and the former and median income have a positive correlation, which means that my hypothesis was correct. Education is a solution to reducing crime; if cities support their people as much as possible in advancing their education, they can get a petter-paid job, earn a stable income, and lead a financially secure life, with less incentive to commit crime to obtain necessities.

Research Question

Is there a significant relationship between adult educational attainment and the crime rate in cities of the Bay Area?

Hypothesis

Because higher educational attainment allows for jobs with higher salaries, I hypothesized that cities with a relatively higher educational attainment among adults would have a higher median income per household and consequently less incentive to commit crimes and thus a lower crime rate.

Datasets

- Bay Area Equity Atlas: Educational Attainment
- Crimes & Clearances - Open Justice
- California (USA): State, Major Cities, Towns & Places - Population Statistics, Maps, Charts, Weather and Web Information
- California - U.S. Census Bureau QuickFacts
- Police Department Size Calculator

Challenges

- Data from different sources was organized in different ways
    - I had to standardize the variables
- Learning how to do the Linear Regression Line Graph

Methodology

Data used included:

- Educational attainment in 14 Bay Area cities in the following categories: less than high school diploma (coded as 10 years), high school diploma only (12 years), some college only (coded as 13 years), Associate Degree only (14 years), Bachelor's Degree only (16 years), and Master's Degree or higher (coded as 20 years).
- Crime categories: violent crime, property crime, and arson.
- Median income, Percent poverty

Using Python, I did the two analyses:

- Linear Regression Analyses between educational attainment (independent variables) and income and crime (dependent variables)
- Correlation Matrix (all variables)

FORCES DRIVING MENTAL ILLNESS

Upper St. Clair High School
Sofia Alfredson-Themudo, Augusta Bottonari, Rohan Inampudi, Rohan Mehta, Harshini Sivakumar, John Unice, William Whitman

QUESTION:

Is there a correlation between access to **health care** and rates of **mental illness** in the United States?

HYPOTHESIS:

If the rates of health insurance access are high in a certain state, then the rates of mental illness in that state should be low (in respect to other areas), and if insurance rates are low, mental illness should be high (in respect to the opposite).

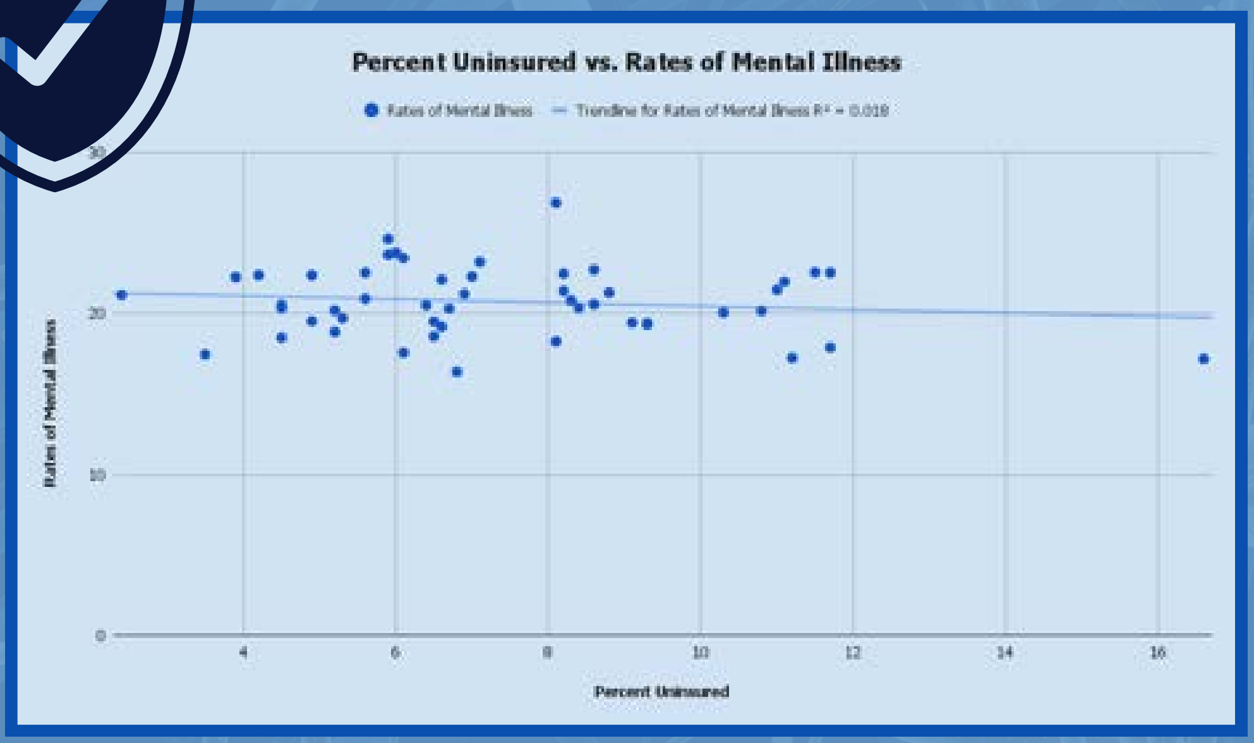
OUR PLAN:

We planned to use chronologically parallel data regarding access to healthcare and rates of mental illness from 2022 (by state) to create an accurate regression between these data values. We would use the regression to statistically deduce the correlation between the independent variable (access to healthcare) and the dependent variable (rates of mental health illness). Other statistical methods could be employed depending on our initial regression analysis.

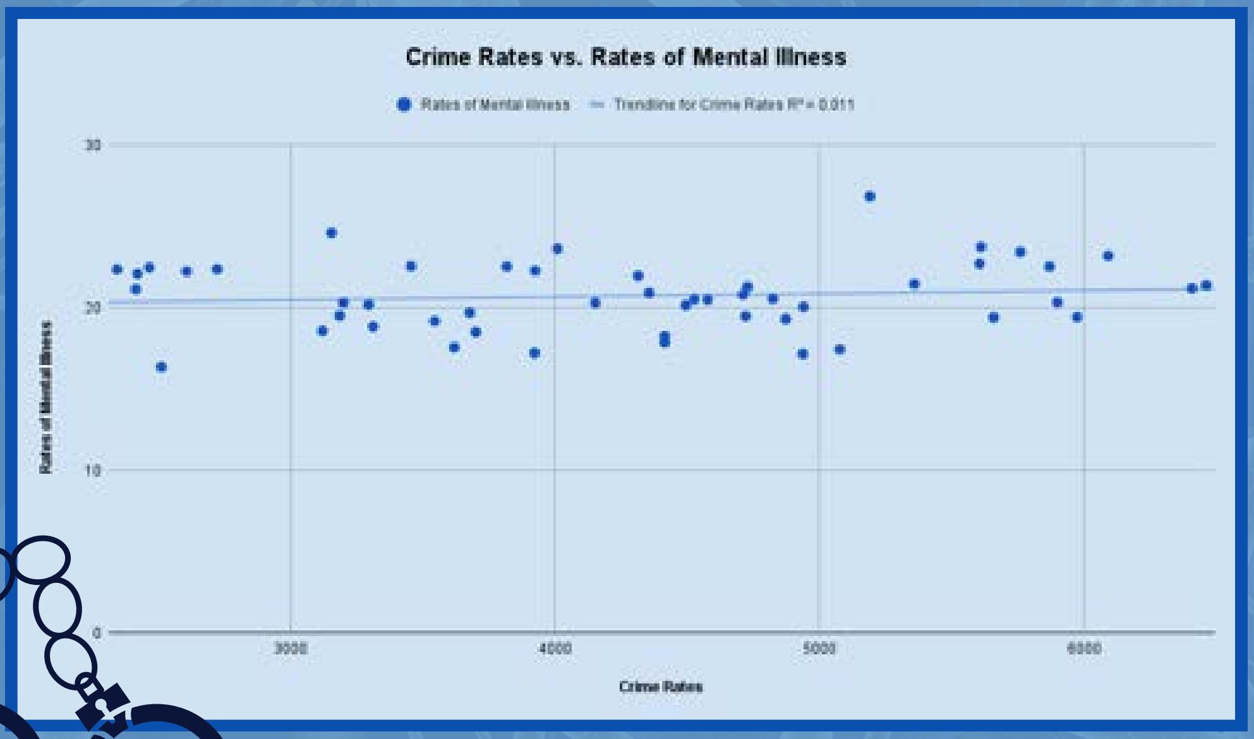
DATASETS USED:

Depression Rates (by state) from the **Center for Disease Control**
Mental Illness Rates (by state 2022) from **Mental Health America**
Mental Illness Rates (by state 2008) from **SAMHSA**
Population without Health Insurance (by state 2022) from the **US Census**
Population without Health Insurance (by state 2008) from the **US Census**
Average Salary (by state) From **Sofi.com**
Crime Rates (by state) from **Datapandas.org**

THE FORCES:



FORCE #1: INSURANCE



FORCE #3: CRIME

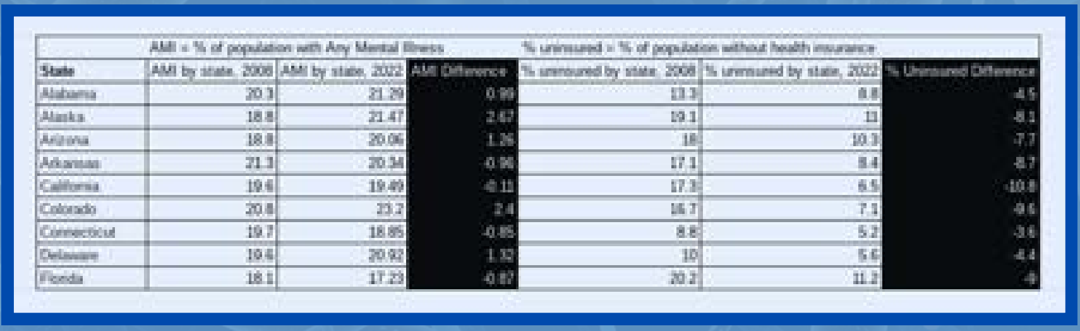
FORCE #1 - Health Insurance: We predicted that there would be a statistical correlation between percent of population without health insurance vs. rates of mental illness (by state). We found that there isn't a reliable correlation between the datasets because the **R²** value is **0.018**.

FORCE #2 - Salary: It is reasonable to predict less mental illness in states where average salary is higher. To calculate the correlation, we made a graph with average salary vs. rates of mental illness by state. However, there did not seem to be a strong correlation between mental health and salary, with an **R²** value of **0.024**.

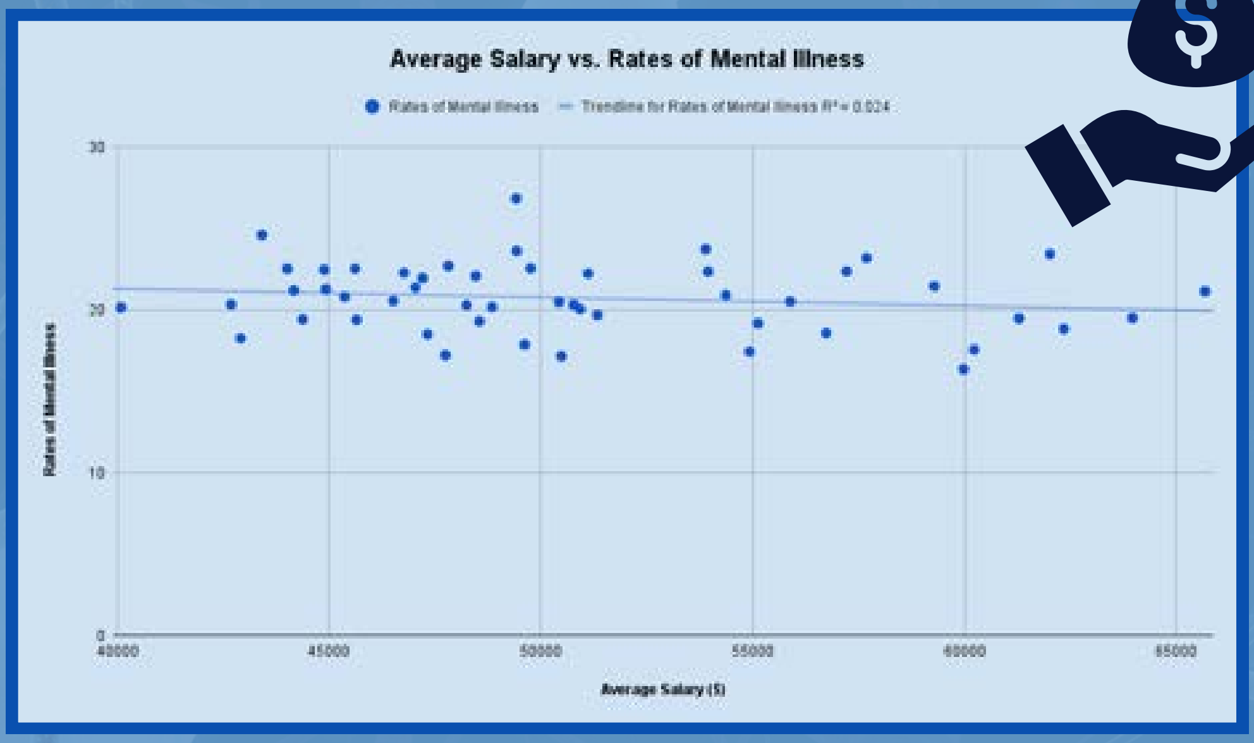
FORCE #3 - Crime: Not only may crime drive mental illness, but mental illness could also affect crime rates. We tested this by putting our mental health data against 2022 minor felony rates by state. However, we got an **R²** of about **0.011**, well below the threshold for statistical correlations.

#4 - Growth/Decay: Instead of looking for direct factors affecting current mental illness rates, we looked at the change (growth/decay) of our previous variables and tested it against the change (growth/decay) of mental illness rates. All of our previous datasets from 2022 were used, but we found new datasets for a few variables from 2008. We used an Excel data table to calculate the difference between the 2008 data and 2022 data. We put the changes of the variables (the graph shown is growth of health insurance rates by state) and the growth of mental illness rates by state against each other, but to no avail, as our highest **R²** value from this test was **0.043**. There is minimal correlation between the growth of these variables from 2008 to 2022.

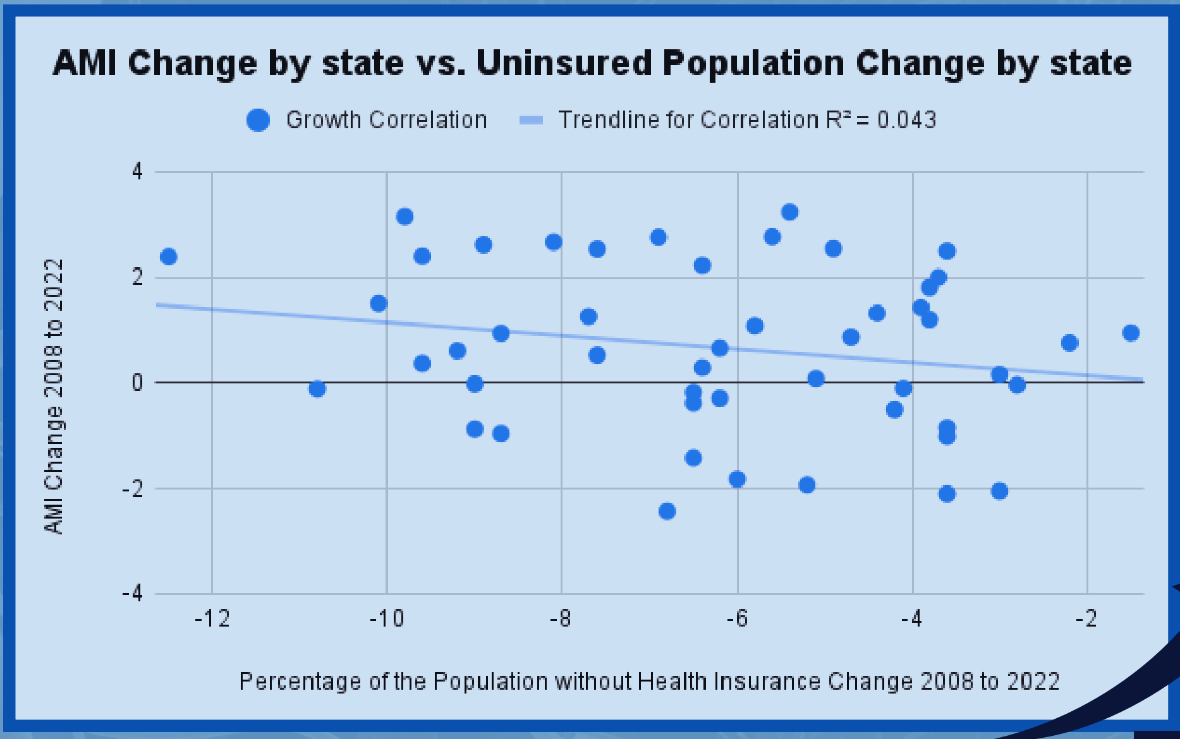
Here is a snapshot of part of our growth data table:



	AMI = % of population with Any Mental Illness			% uninsured = % of population without health insurance		
State	AMI by state, 2008	AMI by state, 2022	AMI Difference	% uninsured by state, 2008	% uninsured by state, 2022	% Uninsured Difference
Alabama	20.3	21.29	0.99	13.3	8.8	-4.5
Alaska	18.8	21.47	2.67	19.1	11	-8.1
Arizona	18.8	20.06	1.26	18	10.3	-7.7
Arkansas	21.3	20.34	-0.96	17.1	8.4	-8.7
California	19.6	19.49	-0.11	17.3	6.5	-10.8
Colorado	20.8	23.2	2.4	16.7	7.1	-9.6
Connecticut	19.7	18.85	-0.85	8.8	5.2	-3.6
Delaware	19.6	20.92	1.32	10	5.6	-4.4
Florida	18.1	17.23	-0.87	20.2	11.2	-9



FORCE #2: SALARY



#4: GROWTH/DECAY

RESULTS + CONCLUSION:

Results:

- Each R² value was under 0.4, the minimum threshold for statistical correlations.
- There is no **single quantitative variable** that solely or dominantly drives mental illness.
- The cause is linked to multiple **qualitative origins**.
- The **main causes** vary from **county to state to region (**no unifying factor).

Conclusion:

- A linear regression model was used to analyze variable correlations.
- Access to healthcare, crime rates, and others were **independent variables**; mental health illness was the universal **dependent variable**.
- We expected to find a **direct correlation** between the two variables; however, **minimal correlation** was observed.
- Proceeded to test growth and decay from 2008 to 2022 with same variables; **R² value was comparable** to original tests
- Concluded that no **singular quantitative input** directly caused mental illness; instead, **many factors** are responsible for this trend.

CHALLENGES FACED:

- Our progress was delayed because we spent too long developing a question. We were too concerned about potential datasets not working in our favor.
- We experienced difficulty finding accurate, reliable, and raw data pertaining to our question. However, we kept looking and eventually found data usable for our project.
- There was no significant correlation between relevant data sets, so we decided to utilize many different variables to explain our conclusion.
- After we found that we couldn't get a strong correlation between our variables, we settled on shaping our conclusion from the data rather than from our hypothesis.

OUR RECOMMENDATION:

- Our minimal correlation suggested that there is **no one factor driving** America's mental illness crisis
- The US government should budget its resources among **multiple potential variables** responsible for poor mental health
  - ex. making insurance more affordable or implementing additional crime-fighting initiatives
  - Don't put all your eggs in the same basket!

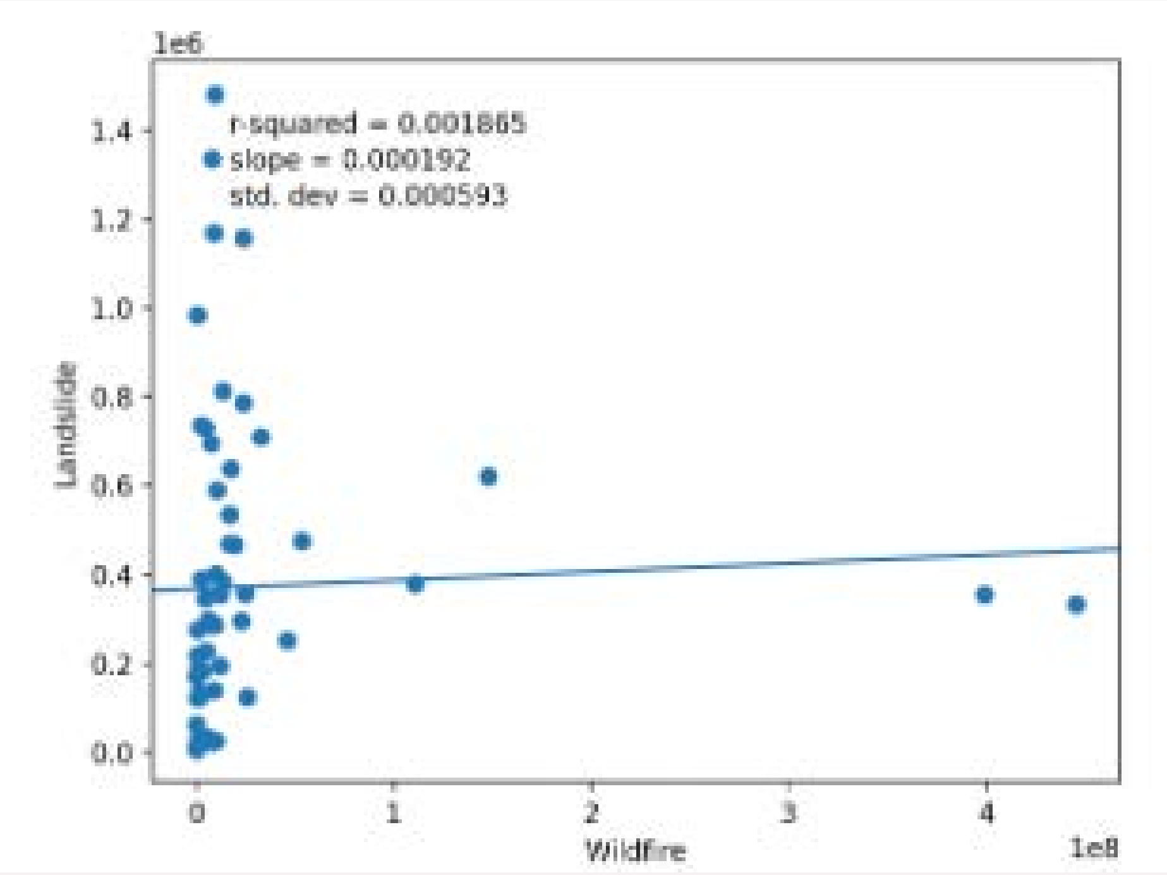
Wildfires and Landslides

Hypothesis: There is a correlation between wildfire risk and landslide risk.

San Lorenzo Valley High School
Kai Wildberger, Chaeyi Lee, Mikayla Casey,
Hudson McKinney, Jaice Williamson
Stacy Clark, Kartik Haribhai Patel

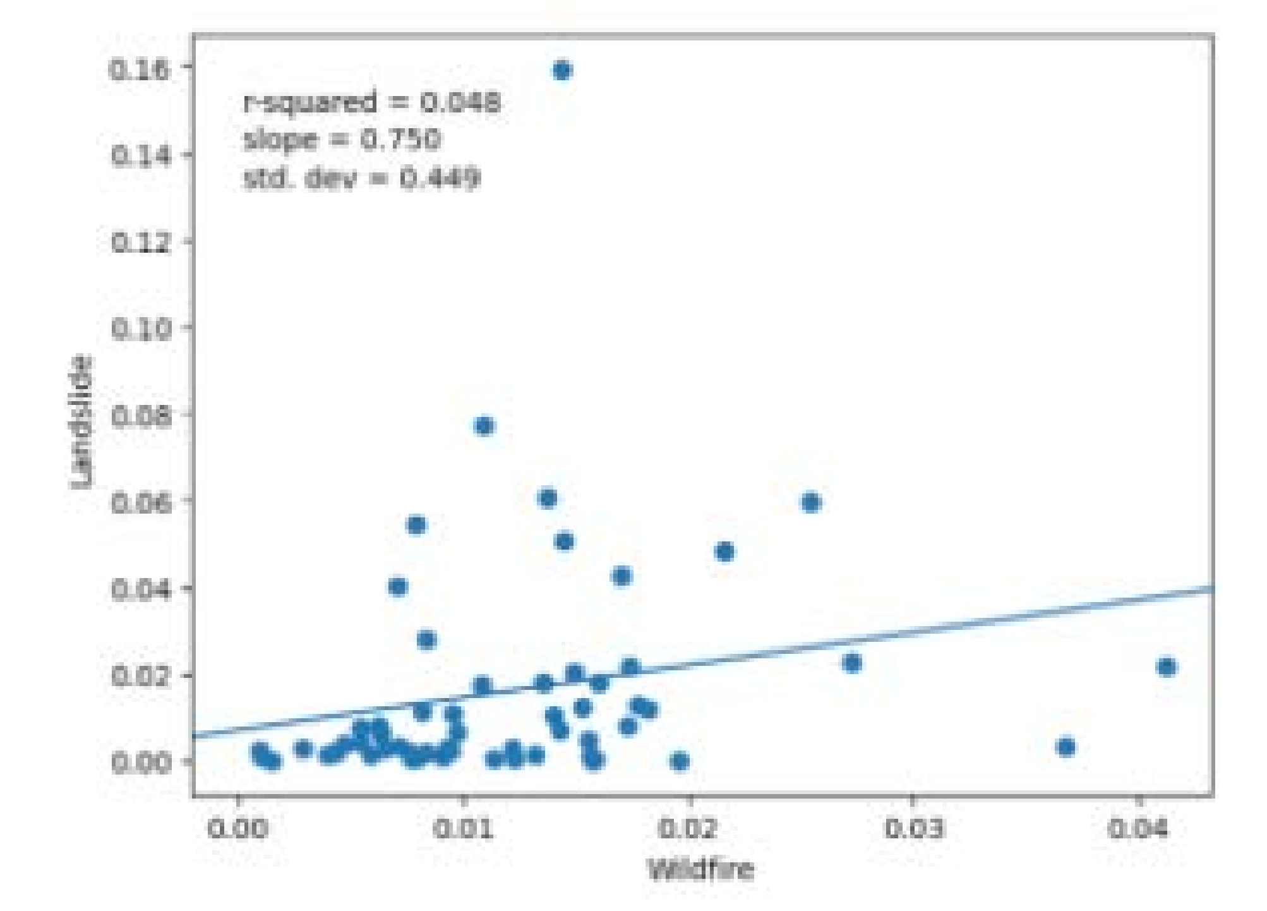
Analysis Question: Is there a relationship between wildfires and landslides in California?

We chose this question because our local Santa Cruz Mountains community has recently been heavily affected by both wildfires and landslides. In the summer months, extreme heat causes plants to dry out which leads to an increasing wildfire risk. In 2020, the CZU Lightning Complex burned close to 40,000 acres in the northern Santa Cruz and southern San Mateo counties. In winter, torrential rain saturates and destabilizes weak soil, which then causes substrate collapse. We wanted to know if the fires, which destroyed plant and tree roots that slopes rely on for stabilization, increased the risks of landslides.



Risk Value of Wildfires and Landslides

Explanatory variable: The original value of the wildfire risk in all 58 counties in California
Response variable: The original value of the landslide risk in all 58 counties in California
Cleaning data: From the datasets made by FEMA's National Risk Index (NRI), we chose a dataset about California. We used the variables of WFIR_RISKV and LNDS_RISKV, which contain the original value of the wildfire risk and the landslide risk in all 58 counties in California.
Modeling: We created a scatterplot relating the original values of wildfire risk and landslide risk. To identify the relationship between the original value of wildfire risk and the original value of landslide risk, we used the Least-Squares Regression Line (LSRL) as a model for our data.



Proportion of County Area
(using Exposure — Impacted Area from NRI divided by total sq mi)

E**xplanatory variable**: The area exposed by wildfire divided by the whole area of each county
Response variable: The area exposed by landslides divided by the whole area of each county
Cleaning data: From the datasets made by FEMA's National Risk Index (NRI), we chose a dataset about California. We used the variables LNDS_EXP_AREA and WFIR_EXP_AREA, which contain the areas that are exposed by wildfires and landslides in every county in California. To find out how much of each county would be impacted by wildfire or landslides relative to its total area, we divided the exposed impacted area for wildfire and landslides by the total area of each county to get the ratio.
Modeling: We created a scatterplot relating the proportion of exposed impacted areas by wildfire and the proportion of exposed impacted areas by landslides. To identify the relationship between the proportion of exposed impacted areas by wildfire and the proportion of exposed impacted areas, we used the Least-Squares Regression Line (LSRL) as a model for our data.

Challenges

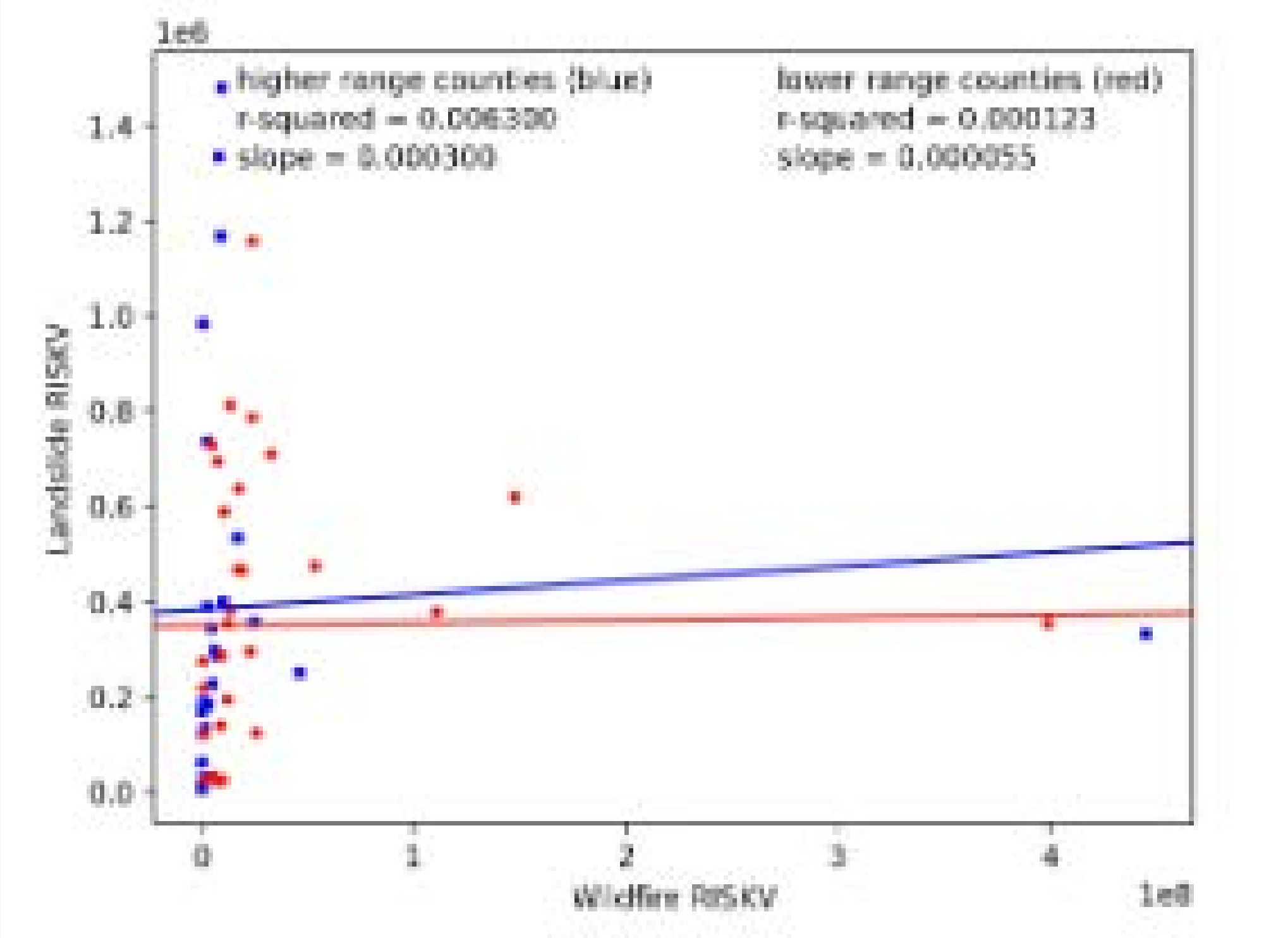
We found it difficult to use massive geodatabase formats, so we were limited to using CSV and GeoTIFF formats. This limited the data that were available to us, and meant that we were restricted to certain types of visualizations; for example, records of individual landslide and wildfire events are typically limited to small demonstration forests in Northern California, since the area covered is more manageable to survey. We could not find the dataset about the range of elevation for all counties in California, so we decided to use the datasets about the highest point and the lowest point of each county and subtracted the lowest elevation from the highest elevation using a Google spreadsheet function.

Data Sets Used

We used data from FEMA's National Risk Index (NRI), which provides information about natural hazards across the US. We downloaded CSV files that contained various variables, such as Risk Value (calculated by multiplying social vulnerability by expected annual loss and dividing it by community resilience) and Exposure-Impacted Area (as a percentage of land area for each county). We processed them using Python, selecting Risk Value and Exposure Total Area to include in our own reduced CSV.

To get elevation ranges for counties, we used data from the TessaDEM Digital Elevation Model project. We then split the county data into two groups based on elevation ranges: higher elevation range counties (> 7,000 meters) and lower elevation range counties (≤ 7,000 meters). The elevation range is calculated through the difference in highest point and lowest point. The 7,000 meter cutoff is determined by the mean of the elevations of all 58 counties in California, which is 7,000 meters.

TessaDEM
National Risk Index



Counties split by elevation

Explanatory variable: The original value of the wildfire risk in 27 counties in California that have a higher range of elevations (> 7,000 meters) and 31 counties in California that have a lower range of elevations (≤ 7,000 meters)
R**esponse variable:** The original value of the landslide risk in 27 counties in California that have a higher range of elevations (> 7,000 meters) and 31 counties in California that have a lower range of elevations (≤ 7,000 meters)
Cleaning data: From the datasets made by FEMA's National Risk Index (NRI), we chose a dataset about California. We used the variables of WFIR_RISKV and LNDS_RISKV, which contain the original value of the wildfire risk and the landslide risk in all 58 counties in California. To determine whether elevation range influences the relationship between wildfire risk and landslide risk, we used data from the TessaDEM Digital Elevation Model project and split the counties into two groups based on elevation ranges: higher elevation range counties (> 7,000 meters) and lower elevation range counties (≤ 7,000 meters). The elevation range is calculated through the difference between the highest point and the lowest point. The 7,000-meter cutoff is determined by the mean elevations of all 58 counties in California, which is 7,000 meters.
Modeling: We created two scatterplots relating the original values of wildfire risk and landslide risk in the counties with a higher range of elevations (blue) and the counties with a lower range of elevations. (red) To identify the relationship between the original value of wildfire risk and the original value of landslide risk depending on elevation range, we constructed the Least-Squares Regression Line (LSRL) for each scatterplot as a model for our data.

Recommendations

We believe that the best way to improve our process is to get access to more datasets, something which has been proven difficult. We could expand our scale from California counties to counties along the west coast so that a random sample could be taken and significance tests could be performed. This would fix many of the issues we encountered involving bad data or graphs, especially in regards to accurately representing our data. Another key thing we would like to try is obtaining our own data, which could greatly improve our analysis. Collecting data on individual wildfire and landslide events is difficult, considering both the remoteness of these events and the need for historical precedent. A team of our size would not have been able to do this, but we believe that it would be best practice to conduct our own surveys moving forward.

Summary

Based on the raw wildfire risk and landslide risk from the National Risk Index, there is no correlation between wildfire risk and landslide risk. It appears that the higher wildfire risk in a county does not affect the landslide risk, and we do not have enough evidence to prove otherwise. When looking at the proportion of county areas affected, there is a correlation, most likely because larger counties are expected to have larger impacted areas for any event due to their size. When split into groups of "flat" and "hilly" counties, the data shows that only counties with higher elevation show a noticeable correlation, but we cannot draw conclusions from it.