SOMETHING IN THE AIR

HOW IS AIR QUALITY INDEX RELATED TO SOCIETAL FACTORS?

DEFINITIONS

AQI: An index scale used to measure the extremity of air quality on a given date. The index spans from 0-500, where a lower value indicates cleaner air quality. AQI includes chemicals and particulate matter (referring to any harmful particles in the air of certain mm diameters).

HYPOTHESES

H_0: Air quality will have no effect on the number of asthma-related hospitalizations in Allegheny County.

H_A: The air quality will increase the number of asthma-related hospitalizations in Allegheny County.

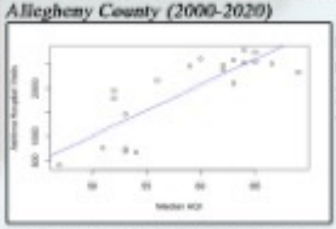
DATASETS

Air Quality Index: EPA & WPRDC

Societal Factors: Pennsylvania Department of Health (Asthma Hospital Visits, Fall Deaths, Bronchus Cancer, Cardiovascular Disease)

GRAPHS

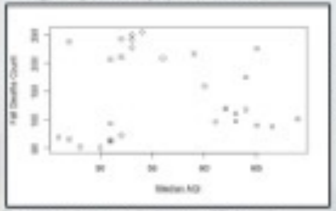
Yearly Median AQI v. Asthma Hospital Visits in Allegheny County (2000-2020)



r=0.817
r^2=0.668
y = 106.22x - 4303.12

Because the correlation coefficient is 0.668, Yearly Median AQI's correlation with asthma hospital visits is moderately strong.

Yearly Median AQI v. Fall Deaths in Allegheny County (2000-2020)



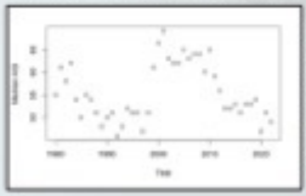
r=-0.0001
r^2=.0000002
y = -0.002x +142.31

Because the correlation coefficient is 0.0000002, the Median AQI has no correlation with fall deaths.

The weak correlation between Median AQI and Fall Deaths adds more significance to the moderate correlation between Median AQI and Asthma Hospital Visits.

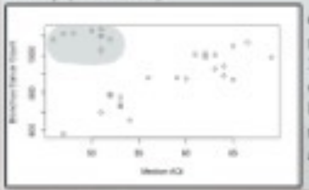
AQI TRENDS

Median AQI v. Year in Allegheny Country (1980-2020)



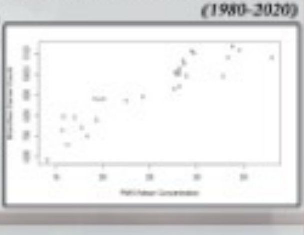
Median AQI was at a record low in 2020. PM 2.5 refers to any particulate matter sized at 2.5mm. Scientists have been recording PM2.5 since 1999. This likely accounts for the abrupt jump shown in the AQI per year graph. The trend in AQI per year between 1980 and 1999 indicated that AQI decreases with each following year. This trend is present again in the 2000 to 2022 AQI data.

Median AQI v. Bronchus Cancer Count in Allegheny County (1980-2020)



The data trend in the AQI v. Bronchus Cancer graph inedicates that PM2.5 not being recorded has resulted in a low AQI measurement. Despite this, cancer counts are still high, likely because of this change in data recording methods. We recognize this as a limitation that may effect our conclusions.

PM10 Mean v. Bronchus Cancer in Allegheny County (1980-2020)



CONCLUSION

Based on the data, we determined that median AQI moderately correlates with asthma, so as median AQI increases, so too does the number of asthma hospital visits.

To determine the relationship between the median AQI and count of hospital visits due to asthma, we used the program R to measure regression. Median AQI was found to be a significant factor in asthma hospital visits.

Although our AQI in Pittsburgh is decreasing over time, there is still a prevalance of PM2.5 in our air. We recommend further investigation on the impact AQI has on public health.

CHALLENGES

We had difficulty finding data sets relevant to our location, causing inaccurate statistical results due to a lack of adequate data points. Additionally, the particulate matter data set was difficult to filter, as it included many different chemicals. Many counts of asthma and other related diseases were extremely low, causing conclusions to be inaccurate.

AVONWORTH DATA JAM (LAUREL PURCELL, CATRINA RAICH, AMELIA HARDIMAN, JACKSON SHIELDS, COLIN CRAWFORD, BRAYDEN SIMMONS, ZOE TREXEL)

Standardized Testing: A Reflection on Intelligence or Economic Environment?

A study on standardized test scores in relation to economic factors in Allegheny County's public school districts, with an additional analysis on passing proportion discrepancies due to the COVID pandemic in relation to economic factors.

North Allegheny Senior High School: Max Fang, Angel Qu, Gautam Ramkumar, Aneri Shethji, Riddhima Singh, Risha Solanki, & Collin Wang

BACKGROUND

Standardized tests aim to act as a benchmark for academic achievement, yet they have been controversial since their formulation, with disputed issues ranging from test design to cost. However, the most heated controversy in recent years has surrounded the validity of testing in general. In this study, we researched whether standardized test scores are a measure of inherent intelligence or of economic environment and its educational resources (or lack thereof). Herein we report a study analyzing the relationship between MV/PI AR values (Market Value / Personal Income Aid Ratio) in Allegheny County and SAT test scores (2017 - 2019) and proportion of PSSA test-takers who passed the exams (2017 - 2022**) in addition to PSSA passing proportion discrepancies between tests administered before and after the COVID-19 pandemic in Allegheny County to determine whether certain public school districts within Allegheny County may inherently be at a disadvantage concerning standardized test scores as a result of economic disparities.

RESEARCH QUESTION

Is there a significant relationship between economic variables and SAT test scores or PSSA passing proportions, and has the COVID-19 pandemic unequally impacted PSSA passing proportions across public school districts in Allegheny County?

HYPOTHESIS

We predicted that school districts located in areas with better conditions (i.e. lower MV/PI AR) would be significantly associated with higher SAT scores and PSSA passing proportions as well as smaller PSSA passing proportion discrepancies. Furthermore, we anticipated that variations in average SAT scores across school districts would be more strongly correlated with economic factors than PSSA passing proportions.

CHALLENGES

- The Pennsylvania Department of Education (PDE) stopped releasing reports of SAT/ACT scores after the COVID-19 pandemic started.
  - Some high schools we reached out to refused to release their data, so we were unable to analyze the impact of COVID-19 on SAT scores.
- Additionally, we were unable to analyze the data for Duquesne and Wilkinsburg Borough school districts because their data were incomplete.

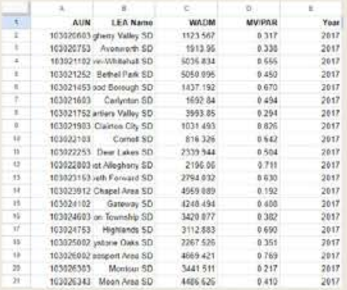
METHODOLOGY

1. We imported the data into Google Sheets and filtered them to only include the data relevant to our study. For the purposes of this study, each SAT score or PSSA passing proportion represents an average SAT score or proportion of PSSA test-takers who passed from a public school district in Allegheny County. A sample of the cleaned data that we used is shown below.
2. Using the data, we created seven models in RStudio:
   a. Using Linear Regression
      i. Average SAT scores vs MV/PI AR ($p < 2.2 * 10^{-16}$) (2017 - 2019)
      ii. Differences in PSSA passing proportions before and after COVID* vs MV/PI AR (one model per subject -- ELA, Mathematics, Science) ($p = 0.001, 0.824, 0.079$, respectively) (2017 - 2022**)
   b. Using Multiple Linear Regression
      i. PSSA passing proportions as the dependent variable and whether the test was administered after the COVID-19 pandemic and MV/PI AR as the independent variables (one model for each subject -- ELA, Mathematics, Science) ($p_{COVID} = 6.8 * 10^{-15}, < 2.2 * 10^{-16}, = 0.021$, respectively; $p_{MV/PI\ AR} < 2.2 * 10^{-16}$ for all subjects) (2017 - 2022**)
3. We then performed ANOVA tests on each model to determine the statistical significance of each variable (p-values above)

SAMPLE DATASET

DATA SOURCES:

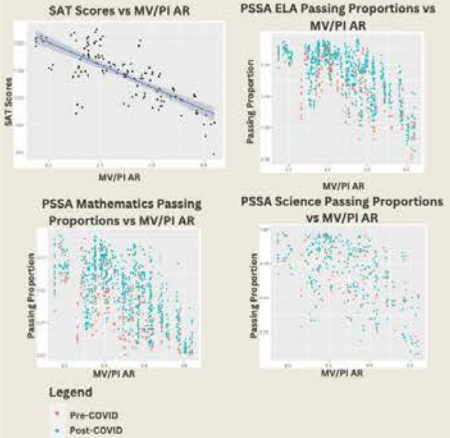
- **PDE Website**
  - **PSSA School Level Data (2017, 2018, 2019, 2021, 2022)**
  - **Public School SAT Scores (2017, 2018, 2019)**
  - **Aid Ratios (2017-18, 2018-19, 2019-20, 2021-22, 2022-23)**



	A	B	C	D	E
1	AUN	LEA Name	WADM	MV/PI AR	Year
2	103020603	gheny Valley SD	1123.567	0.317	2017
3	103020753	Avonworth SD	1913.95	0.338	2017
4	103021102	vin-Whitehall SD	5036.834	0.555	2017
5	103021252	Bethel Park SD	5050.095	0.450	2017
6	103021453	ood Borough SD	1437.192	0.670	2017
7	103021603	Carlynton SD	1692.84	0.494	2017
8	103021752	artiers Valley SD	3993.85	0.294	2017
9	103021903	Clairton City SD	1031.493	0.826	2017
10	103022103	Cornell SD	816.326	0.642	2017
11	103022253	Deer Lakes SD	2339.944	0.504	2017
12	103022803	ast Allegheny SD	2196.06	0.711	2017
13	103023153	eth Forward SD	2794.032	0.630	2017
14	103023912	Chapel Area SD	4959.089	0.192	2017
15	103024102	Gateway SD	4248.494	0.400	2017
16	103024603	on Township SD	3420.877	0.382	2017
17	103024753	Highlands SD	3112.883	0.690	2017
18	103025002	ystone Oaks SD	2267.526	0.351	2017
19	103026002	eeport Area SD	4669.421	0.769	2017
20	103026303	Montour SD	3441.511	0.217	2017
21	103026343	Moon Area SD	4486.626	0.410	2017

**Differences were calculated by subtracting PSSA passing percentages in 2021 from those in 2019*
***2020 scores excluded*

GRAPHS

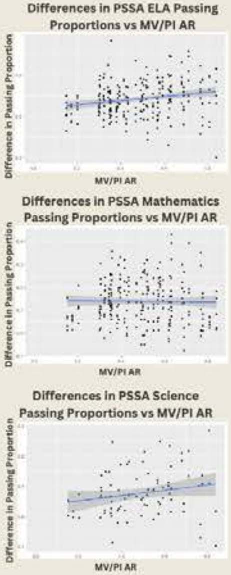


SAT Scores vs MV/PI AR

PSSA ELA Passing Proportions vs MV/PI AR

PSSA Mathematics Passing Proportions vs MV/PI AR

PSSA Science Passing Proportions vs MV/PI AR



Differences in PSSA ELA Passing Proportions vs MV/PI AR

Differences in PSSA Mathematics Passing Proportions vs MV/PI AR

Differences in PSSA Science Passing Proportions vs MV/PI AR

ANALYSIS & CONCLUSION

With α = 0.05 and using the calculated p-values, we concluded that the relationships between SAT scores and Allegheny County MV/PI AR values from 2017 to 2019, as well as between PSSA passing proportions and MV/PI AR values from 2017 to 2022**, were statistically significant. We also found that the relationship between the differences in PSSA passing proportions (before vs after the COVID-19 pandemic) was significant for the ELA exam; however, we found that *the correlations were not significant for the Mathematics and Science exams*.

The results from our SAT analysis suggest that SAT exams may not be a measure of intelligence, but rather of the resources available to test takers, such as tutors or prep classes. However, the results from the PSSA analyses indicate that not only are fewer resources available to students, the inherent *quality of education* is lower in areas with lower MV/PI AR values, i.e. poorer areas, assuming that PSSA test-takers do not use resources to prepare for the exam. Though we found that the COVID-19 pandemic had a significantly larger impact on PSSA ELA passing proportions associated with high MV/PI AR values, we were surprised to see that it did not have a significant impact on Mathematics or Science exam proportions. Further research would be needed to explore these unusual findings.

Overall, our results suggest that the one-size-fits-all approach to these so-called "standardized" tests does not truly fit all. Though our exams may be standardized, our quality of education is not. The facts are blatantly obvious -- education is not equal across diverse demographics. If we truly want to measure intelligence or education, we must prioritize equity over equality and choose one of two options: provide sufficient support for poorer school districts to improve the quality of their education or adapt our standardized exams to fit the needs of each district.

diverSEty: The correlation of diversity and publications

By: Shuchir Jain, Edward Yang, Andrew Feng, David Wang, and Sameer Gosalia from North Allegheny Intermediate High School

Research Question: As the diversity of the workforce in **Science and Engineering** related fields changes over time, how has the rate of academic publication correlatively changed in the U.S?

Our Hypothesis: There is a positive correlation between diversity and publication output, for the greater difference in backgrounds contributes to a wider range of perspectives and potential from the IQ curve and thus more research studies.

Step 1: Find the Data

- All data from **NCSES** (National Center for Science and Engineering Statistics)
- **The National Survey of College Graduates**–for information regarding the population of demographics in the S&E workforce
- **The Publications Output: U.S. Trends and International Comparisons**–about the peer-reviewed scientific publications created by year.

Step 2: Clean the Data

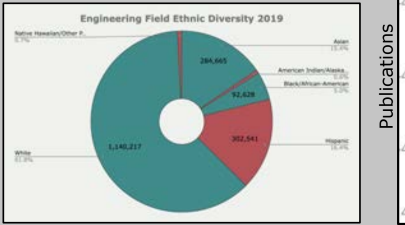
- We compiled a **multi-sheet publication data** into a **single sheet** across the years 2010, 2013, 2015, 2017, and 2019–common to both sets.
- Then, we created **Individual** sheets for **each S&E field,** each pulling workforce data to calculate **diversity scores** (expanded in analysis).

Step 3: Analyze the Data

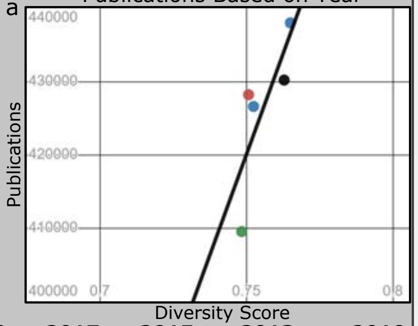
- To quantitatively measure diversity, we used the **custom diversity score formula.** This formula finds the standard deviation of every demographic from each other and subtracts the result from 1, outputting a 0-1 number where 0 is no diversity and 1 is complete diversity.
- Using this formula, we calculated the **diversity scores** for each field and combined them together using **weighted average** with each fields' publication percentages as weights.
- The diversity scores, plotted on the x-axis, and the publications, plotted on the y-axis, in Graph 1 show a linear regression. This graph seems to affirm the hypothesis, but due to the regression F statistic having a p-value of 0.11, the **correlation is statistically insignificant**.

$$D = 1 - \sqrt{\frac{\sum_{k=1}^{n} \frac{\sum_{i\in[1,k)\cup(k,n]} (P_i - P_k)^2}{n-1}}{n}}$$

D : the diversity score
n : number of demographics
P_i : Percent of a demographic in a population



Diversity Score and Publications Based on Year



S&E Diversity Scores	2019	2017	2015	2013	2010
Engineering	62.03%	61.45%	60.10%	59.74%	58.59%
Compsci & Math	69.81%	70.06%	69.38%	68.21%	66.97%
Bio, Agriculture, Life	83.68%	83.11%	81.88%	82.59%	82.47%
Physical	72.82%	70.93%	68.76%	70.00%	68.12%
Social	73.11%	76.54%	76.27%	74.60%	76.33%
Weighted Average	**76.51%**	**76.29%**	**75.08%**	**75.25%**	**74.85%**

Concluding Statements

Though the correlation was insignificant...

This research can still provide actionable value. Because diversity does not correlate with publication output, the demographics of an individual likely does not correspond to their academic output and therefore should not affect one's chances at things like admission and hiring. Further cases from the STEM field could also be studied, like patents, to further confirm or refine our results.

Effects of Cancel Culture on the Attention a Celebrity Gets on Social Media

MacKenna Healy, Anika Balog, Hailey Kurylo, Griffin Grushow

South Hunterdon Regional High School

Question: Does cancel culture affect the social media following of celebrities in regards to before and after being canceled?

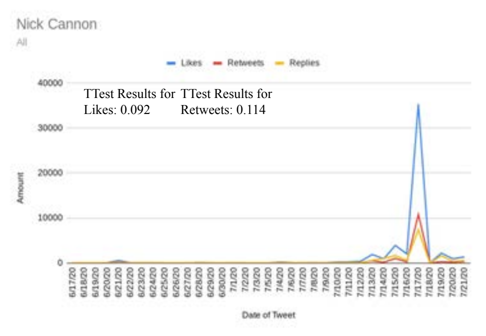
Introduction:

Cancel culture is a method of holding someone accountable for their opinions. This subject is important because people have been determining others' views on social media for a long time. In our research, we looked at the likes, retweets, and replies of posts made by Nick Cannon and Piers Morgan on Twitter. We collected this information from before and after their week of posts that got them canceled. Our main goal was to observe the level of fluctuation among their internet attention before and after cancellation.

Challenges

Along the way, we faced several challenges including:

- Data not being readily available;
  There was not a downloadable dataset so we had to find the data ourselves
- We had to decide on a subset of celebrities;
  We settled on two from the same field but canceled for different reasons.
- Wanted to look at Net Worth We couldn't find any reliable data



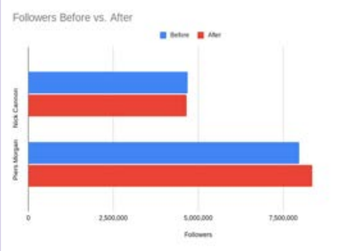
TTest Results for Likes: 0.092

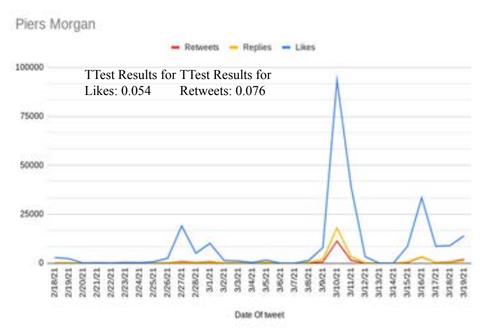
TTest Results for Retweets: 0.114

Nick Cannon Twitter Data

(6/17/20 - 7/21/20)

Followers Before & After





TTest Results for Likes: 0.054

TTest Results for Retweets: 0.076

Piers Morgan Twitter Data

(2/18/21 - 3/19/21)

Data Sources/Ways of Analysis:

- Twitter was the main source that was used to find data about Piers Morgan and Nick Cannon. We used the advanced search in Twitter to help us find older dates which made our data collection manually intensive.
- The charts show the number of tweets, likes, and followers before and after the cancellation.
- We ran T-tests to check the statistical significance of the differences before and after.

Conclusion

The cancellation of Nick Cannon and Piers Morgan has increased their retweets, likes, and replies on Twitter for a short period of time. Their retweets, likes, and replies on Twitter varied. While we can see the effect, it was not always statistically significant and not necessarily long-term impactful. It is much harder to assess the impact of cancel culture than we thought it would be.

CARLYNTON JR./SR. HIGH SCHOOL

HOW TECHNOLOGY AFFECTS TEENS IN CAR CRASHES

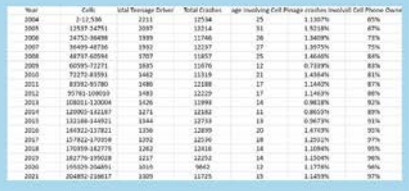
ALYSE CROWN, MARY DOUGHERTY, SIMON SCHRIVER, & ELAINE ZHANG

HYPOTHESIS: HIGHER RATES OF TEENAGE CELL PHONE USAGE HAS BEEN A SOURCE OF INCREASE IN CAR ACCIDENTS.

PROBLEM

DOES CELL PHONE USAGE HAVE A CORRELATION TO AN INCREASE IN CAR ACCIDENTS? TECHNOLOGY HAS DRASTICALLY EVOLVED OVER THE PAST SEVERAL YEARS, BUT CAR ACCIDENTS HAVE ALSO INCREASED SEVERELY. ALTHOUGH THESE TWO SUBJECTS DON'T SEEM TO HAVE A CORRELATION WITH EACH OTHER, A DEEP DIVE REALLY SHOWCASES HOW MUCH THESE NEW DEVICES ARE AFFECTING THE NEW GENERATION OF DRIVERS. TEENS NOWADAYS MAKE "TEXTING & DRIVING" THEIR DAILY ROUTINE, AS WELL AS CHECKING SOCIAL MEDIA HOURLY, WHICH OVERALL FORMS A LIFE-THREATENING DISTRACTION WHEN DRIVING.

DATA:



Year	Cells	otal Teenage Driver	Total Crashes	age Involving Cell Ph	age crashes involvi	Cell Phone Owne
2004	2-12,536	2211	12534	25	1.1307%	65%
2005	12537-24751	2037	13214	31	1.5218%	67%
2006	24752-36498	1989	11746	26	1.3409%	73%
2007	36499-48736	1932	12237	27	1.3975%	75%
2008	48737-60594	1707	11857	25	1.4646%	84%
2009	60595-72271	1635	11676	12	0.7339%	83%
2010	72272-83591	1462	11319	21	1.4364%	81%
2011	83592-95780	1486	12188	17	1.1440%	87%
2012	95781-108010	1483	12229	17	1.1463%	88%
2013	108011-120004	1426	11993	14	0.9818%	92%
2014	120005-132187	1271	12182	11	0.8655%	89%
2015	132188-144921	1344	12733	13	0.9673%	91%
2016	144922-157821	1356	12899	20	1.4749%	95%
2017	157822-170358	1292	12536	18	1.2931%	97%
2018	170359-182778	1262	12416	14	1.1094%	95%
2019	182779-195028	1217	12252	14	1.1504%	96%
2020	195029-204851	1016	9842	12	1.1776%	96%
2021	204852-216617	1309	11725	15	1.1459%	97%

PROCESS AND MISTAKES:

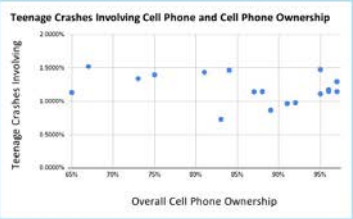
OUR GROUP STARTED BY LOOKING FOR DATA SETS, WHICH PROVED TO BE A PROBLEM ITSELF, AS THERE WERE FEW THAT SEEMED TO RELATE TO OUR QUESTION. AFTER A MEETING WITH OUR MENTOR, WE BEGAN TO QUESTION THE VALIDITY OF OUR ORIGINAL STANCE, AND HOW TO SUPPORT IT. WE TWEAKED OUR HYPOTHESIS AND FOUND MORE APPLICABLE DATA SETS. FROM THERE, WE WERE ABLE TO CALCULATE THE CORRELATION BETWEEN CRASHES AND CELL PHONE OWNERSHIP. THE RESULTS SEEMED TO BE ANOTHER PROBLEM ON THEIR OWN, AS THEY WERE THE COMPLETE OPPOSITE OF WHAT WE WERE EXPECTING.

WHY IS IT IMPORTANT?

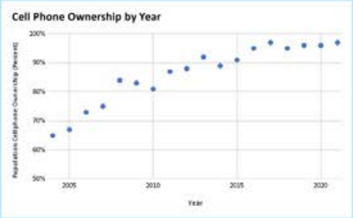
AS TEENAGERS WHO DRIVE, WE REALIZE HOW MUCH TECHNOLOGY HAS BECOME A MAJOR DISTRACTION TO OUR DAILY LIVES. A TEACHER AT CARLYNTON JR./SR. HIGH SCHOOL, MR. COLONNA, WHO DOES DRIVER'S ED AS A SIDE JOB, HAS OBSERVED THE CHANGES OF BEHAVIOR IN TEENAGERS FOR THE PAST DECADE. WHEN INTERVIEWING HIM ABOUT HIS EXPERIENCES IN TEACHING TEENS TO DRIVE, HE STATES, "THE NUMBER ONE CAUSE OF CAR ACCIDENTS AMONGST TEENS IS DISTRACTED DRIVING; THE PRIMARY REASON FOR DISTRACTED DRIVING IS BECAUSE OF CELL PHONES."

ANALYSIS

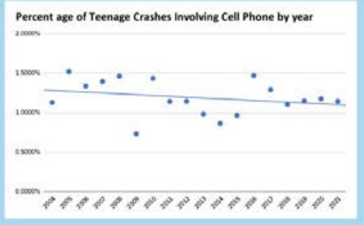
OUR GROUP REJECTED OUR INITIAL HYPOTHESIS, BECAUSE OUR TESTS SHOWED THAT THERE WAS NO RELATIONSHIP BETWEEN TEENAGE CRASHES IN ALLEGHENY COUNTY, AND CELL PHONE USAGE. WE FIRST USED A CORRELATION TEST, WHICH CAME OUT TO .290 MEANING THAT THERE WAS NO SIGNIFICANT CORRELATION BETWEEN CRASHES AND CELL PHONE USE. ADDITIONALLY, WE USED A TRADITIONAL HYPOTHESIS TEST, WHICH ALSO SHOWED THAT THERE WAS NO RELATIONSHIP. ALTHOUGH OUR HYPOTHESIS WAS REJECTED FOR TEENAGERS IN ALLEGHENY COUNTY, THE HYPOTHESIS COULD STILL BE TRUE FOR TEENAGERS IN OTHER COUNTIES, OR EVEN ADULTS IN ALLEGHENY COUNTY. ONE REASON THE HYPOTHESIS WAS REJECTED, COULD BE THAT PA HAS A STRICTER DRIVING LAWS FOR TEENAGE DRIVERS. ANOTHER REASON COULD BE NEWER SAFETY LAWS THAT THAT HAVE BEEN ADOPTED, OR ADDITIONAL SAFETY MEASURE THAT CAR MANUFACTURERS HAVE NOW ADDED.



Teenage Crashes Involving Cell Phone and Cell Phone Ownership



Cell Phone Ownership by Year



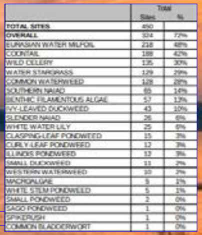
Percent age of Teenage Crashes Involving Cell Phone by year

Eurasian Milfoil's Abundance in the Central & South Basins of Chautauqua Lake in Proportion to Other Invasive Species After Being Sprayed with Herbicides

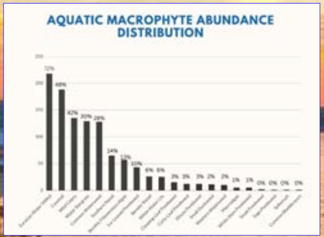
PROBLEM: What invasive species in Chautauqua Lake has the most abundance after being sprayed with herbicides and does the region of the lake affect the amount of plants remaining?

IMPORTANCE: Plant growth like seaweed and algae are overwhelming in Chautauqua Lake. Here in Bemus Point, NY, the lake is a common pastime for those who live in the area, and its condition is important. The aquatic plant growth in this area tends to get out of control (this includes dangerous blue-green algae) and herbicides could allow us to manage the growth to keep the lake healthy. We are trying to see what species is most dense in our lakes after being sprayed.

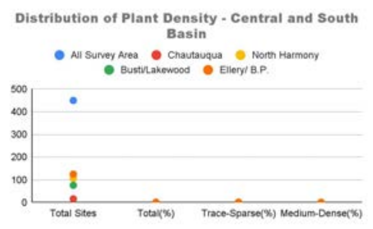
HYPOTHESIS: Herbicides are effective at controlling the density of invasive Eurasian Milfoil found in lakes, based on sampled areas.



	Total	
	Sites	%
TOTAL SITES	450	
OVERALL	324	72%
EURASIAN WATER MILFOIL	218	48%
COONTAIL	188	42%
WILD CELERY	135	30%
WATER STARGRASS	129	29%
COMMON WATERWEED	128	28%
SOUTHERN NAIAD	65	14%
BENTHIC FILAMENTOUS ALGAE	57	13%
IVY-LEAVED DUCKWEED	43	10%
SLENDER NAIAD	26	6%
WHITE WATER LILY	25	6%
CLASPING-LEAF PONDWEED	15	3%
CURLY-LEAF PONDWEED	12	3%
ILLINOIS PONDWEED	12	3%
SMALL DUCKWEED	11	2%
WESTERN WATERWEED	10	2%
MACROALGAE	5	1%
WHITE STEM PONDWEED	5	1%
SMALL PONDWEED	2	0%
SAGO PONDWEED	1	0%
SPIKERUSH	1	0%
COMMON BLADDERWORT	1	0%



	Total		Trace		Sparse		Medium		Dense	
	Sites	%	Sites	%	Sites	%	Sites	%	Sites	%
TOTAL SITES	450									
OVERALL	324	72%	50	15%	62	19%	129	40%	83	26%
EURASIAN WATER MILFOIL	218	48%	130	60%	67	31%	18	8%	3	1%



PLAN: We plan to analyze the statistical correlations between location of sampling zone and invasive species density in proportion to all other species recorded. We will do this by using data from the CLA (Chautauqua Lake Association) and other lake data in the Northeast. The method of analysis will be determined by The Point Interception Method which tracks the extent of aquatic growth within certain littoral zones (nearshore areas). The data we will use will consist primarily of Chautauqua Lake Basins located in the North and South zones.

Challenges

- Pesticides aren't proportionally used in each section of the lake meaning that plant growth rates differ depending on where the data was collected.
- Chautauqua Lake is a small lake located in a low populated suburban area and therefore not much research has been done on the effect of herbicides on plant growth.
- Hard time finding before and after data lead to changes in our original problem.

CONCLUSION: Our data shows that the invasive species with the highest abundance in proportion to other invasive species is Eurasian Milfoil. It also shows that after being sprayed with herbicides, Eurasian Milfoil is the most resistant to the treatment when analyzing the data from the Central and Southern Basins of Chautauqua Lake.

MAPLE GROVE 2023

Elizabeth Quadt, Alexandra Gren, Caleb Barton, Keegan Rishel, Rachael King, Emma Schrecengost, Eli Moore, Madeleine Wadsworth, Daniel Quattrone

Improvements to Pittsburgh's city parking lot distribution

Hampton High School

Aaron Peng, Darren Wang, Sebastian Villalba, Vitaliy Pikalo

Introduction

Many areas in the city suffer from increased congestion during high traffic hours, especially parking lots, where at times the capacity is pushed to the limit. Meanwhile, other lots in more sparsely used areas stay mostly empty for days on end. This inefficient use of space resources negatively impacts the functionality of our city. We believe that by strategically distributing parking spaces in high traffic areas, and reusing or improving existing parking spaces that are underutilized, we will be able to benefit the traffic flow of our city. We decided that the best way to analyze and find under- and over-used parking lots is to find a correlation between traffic rates and the usage of lots around the area. Additionally, by finding available and unused lots near areas of activity, we can identify places for improvement or construction. After examining the data at hand, we came to our research question:

To what extent is there a correlation between traffic rates and the usage of parking spaces near those locations, and what lots can we improve based on the analysis obtained?

The purpose of this analysis is to locate overused lots, underused lots, and empty spaces where parking lots could be developed. We will then propose of a list of lots to be renovated or constructed in strategic locations that can help relieve the flow of traffic or congestion due to lack of parking spaces.

Methodology

The first dataset we used is the Parking Data Dashboard, pulled from the Western Pennsylvania Regional Data Center. We used said dataset to analyze parking lot usage across parking lots across Pittsburgh. Specifically, we looked at the number of transactions, available parking spaces, payments, and utilization (a metric of purchases made during a time interval). These statistics helped provide an estimate of parking lot availability.

The next dataset we utilized is the City of Pittsburgh Traffic Count, pulled from the Western Pennsylvania Regional Data Center. We used the data from this dataset to assess automobile traffic volume and speed across Pittsburgh.

The last data set we used is Zone and Lot Attributes pulled from the Western Pennsylvania Regional Data Center. We used this data to determine maximum occupancy and parking location (on street/off street) for parking lots in zones across Pittsburgh.

We also analyzed the relationship between traffic rates and the parking spaces with linear regression to find a correlation between the two factors.

Limitations

One limitation of our project is that some parking lots didn't have any traffic counters nearby or lacked data on spaces/rates. Because of this, we had to exclude some data from our analysis. Furthermore, there were some conflicting accounts on the amount of spaces for some lots. We addressed this issue by taking data from the same source.

Additionally, the traffic count data we used was collected in different years (2019-2021), so some changes in traffic trends may not have been captured by the data properly.

Furthermore, the dataset for parking lots did not list latitude and longitude, so we had to manually find these values. Because we relied on Google Maps in this process, we could not find latitude and longitude with pinpoint accuracy.

Lastly, we did not account for other factors that may impact demand for parking, including proximity to commercial areas, parking rates, and on-street vs off-street parking. Thus, our analysis of the relationship between traffic and parking likely does not paint a full picture of how parking lots in Pittsburgh should be revised. We suggest expanding the scope of future analysis to get more conclusive results.

Results

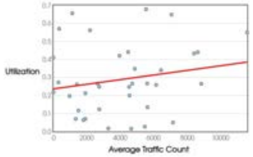
We managed to pull 63 pieces of data for parking lots and garages, and over 400 pieces for traffic points. Then, we calculated the traffic flow by averaging the 5 nearest traffic data points within half a mile. We began by graphing traffic flow to lot spaces, in hopes that this will yield results that show some lots having heavy traffic, yet with little spaces to accommodate for said traffic. We found that there is no significant correlation between the two factors, however, as shown in Figure 3. The r-squared value was 0.0085.

We also analyzed the relationship between average traffic count and lot utilization through linear regression. With a r-squared value of 0.0364, we concluded that there is no statistically significant relationship between those factors (Figure 1). In further support of this conclusion, the residual plot for this regression showed no clear pattern, which all the points being randomly dispersed.

For our utilization metric, we analyzed 36 different lots across the city of Pittsburgh, which also yielded some interesting results. Foremost, many lots seem to be heavily underutilized, with utilization numbers reaching as low as two percent. The mean and median for our utilization data were both between 25-30 percent. Our highest recorded utilization value was 67.9 percent. Notably, the two lots with the greatest number of spaces, Downtown 1 (299 spaces) and Downtown 2 (373 spaces) had among the lowest utilization (Table 1).

Figure 1.

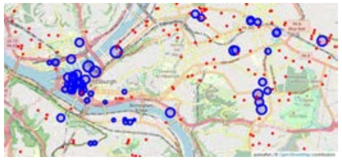
Linear Regression of Average Traffic Count and Utilization



Note. This regression depicts the relationship between average traffic count and utilization. The correlation for this regression was 0.1909 and the r-squared value was 0.0364.

Figure 2.

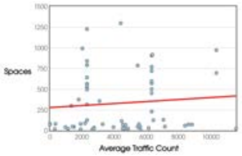
Distribution of traffic lots (blue) and traffic count data (red)



Note. The size of the circles were determined by the average of the 5 closest traffic counts within half a mile, taken to a (4/7)th root-relationship

Figure 3.

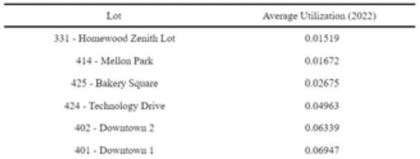
Linear Regression of Average Traffic Count and Number of Spaces



Note. This regression depicts the relationship between average traffic count and the number of spaces. The correlation for this regression was 0.0922 and the r-squared value was 0.0085.

Table 1.

Lots and their Respective Utilization Values



Lot	Average Utilization (2022)
331 - Homewood Zenith Lot	0.01519
414 - Mellon Park	0.01672
425 - Bakery Square	0.02675
424 - Technology Drive	0.04963
402 - Downtown 2	0.06339
401 - Downtown 1	0.06947

Note. This table depicts the six lots with the lowest average utilization overall for 2022

Conclusions & Recommendations

When looking at the connection between nearby traffic count and lot utilization, there is no significant relationship. Factors that could affect this include: no real correlation between traffic and need for parking spaces, sparsity of traffic count data for some lots, and low sample size. However, this does not detract from the fact that some lots have low utilization numbers. In tables 1, Homewood Zenith Lot is shown to be the lowest in terms of utilization, closely followed by Mellon Park and Bakery Square lots. These, according to the utilization values, are the least used parking lot that we have data for. On the other hand, Tamello Beatty lot(not shown), holds the greatest utilization number.

Therefore, we recommend that the city consider development projects around the parking lots with the least utilization to bring more activity and elevate parking demand to conserve space. For overused lots, such as Tamello Beatty, we recommend that either expansions are made, or non parking lot spaces around the area be converted to extra parking space to reduce congestion and improve traffic flow. The combination of these policies would maximize resource efficiency while still meeting the demands of drivers in need of parking.

In terms of future research on this topic, we believe these areas would be of significant interest to the City of Pittsburgh:

- Progress of how our research affected the future- years down the road
- Nationwide instead of just Pittsburgh, larger cities that may compare to Pittsburgh in terms of lot utilization
- Analyzing the relationship between utilization and other aforementioned factors, such as proximity to entertainment hotspots
- Determining a healthy target level for lot utilization

However, we acknowledge that our data is not comprehensive, and our sample size is small. Therefore, our results may be inaccurate compared to real life circumstances. Most notably, we have little traffic data in the downtown area, which may be more congested than available numbers communicate.

References

- WPRDC. (2018). Zone and lot attributes - datasets - WPRDC. Western Pennsylvania Regional Data Center. Retrieved from https://data.wprdc.org/dataset/zone-and-lot-attributes
- WPRDC. (2018). Aggregated parking transactions - datasets - WPRDC. Western Pennsylvania Regional Data Center. Retrieved from https://data.wprdc.org/dataset/parking-transactions
- WPRDC. (2018). Parking Data Dashboard. Parking data reports. Retrieved from https://tools.wprdc.org/parking/
- WPRDC. (2019). City of Pittsburgh Traffic Count - datasets - WPRDC. Western Pennsylvania Regional Data Center. Retrieved from https://data.wprdc.org/dataset/traffic-count-data-city-of-pittsburgh
- Brandt, Eric. "New Study Shows How Much Time and Money We Waste on Parking." The Drive, 12 July 2017, https://www.thedrive.com/article/12389/new-study-shows-how-much-time-and-money-we-waste-on-parking.
- Silderhuis, Herbert. "About Parking." Parking Network, 20 Oct. 2022, https://www.parking.net/about-parking.

Improving Central Dauphin's (G)IEP Program

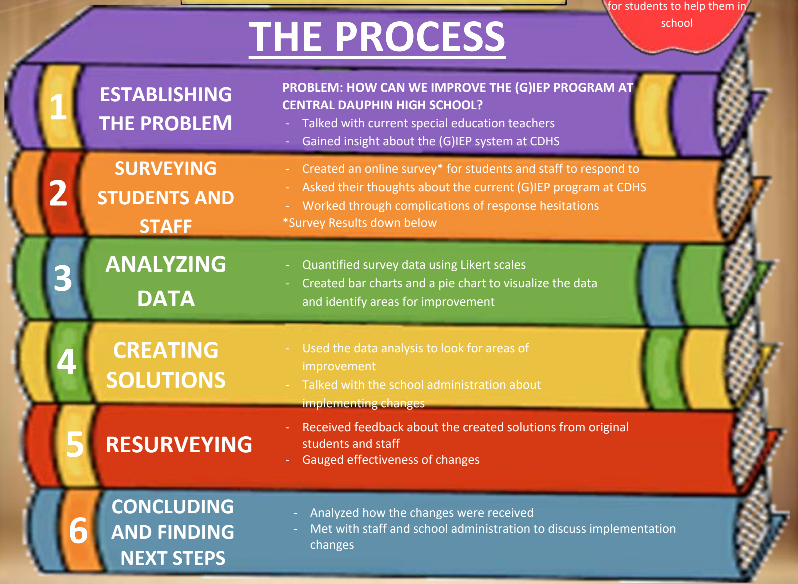
KEY

IEP - Individualized Education Plan

GIEP - Gifted IEP

504 Plans - Accommodations for students to help them in school

THE PROCESS



Phase 1: Surveying School

Phase 2: Implementing Changes

1 ESTABLISHING THE PROBLEM

PROBLEM: HOW CAN WE IMPROVE THE (G)IEP PROGRAM AT CENTRAL DAUPHIN HIGH SCHOOL?

- Talked with current special education teachers
- Gained insight about the (G)IEP system at CDHS

2 SURVEYING STUDENTS AND STAFF

- Created an online survey* for students and staff to respond to
- Asked their thoughts about the current (G)IEP program at CDHS
- Worked through complications of response hesitations

*Survey Results down below

3 ANALYZING DATA

- Quantified survey data using Likert scales
- Created bar charts and a pie chart to visualize the data and identify areas for improvement

4 CREATING SOLUTIONS

- Used the data analysis to look for areas of improvement
- Talked with the school administration about implementing changes

5 RESURVEYING

- Received feedback about the created solutions from original students and staff
- Gauged effectiveness of changes

6 CONCLUDING AND FINDING NEXT STEPS

- Analyzed how the changes were received
- Met with staff and school administration to discuss implementation changes

FIRST SURVEY

SNR**= 31/1

Questions Included:

1) How do you currently feel about the (G)IEP/504 program? (Likert Scale)
2) How could the (G)IEP/504 system be improved? (Open Ended)

Process:

We conducted a randomly sampled anonymous online survey, asking students and teachers for a key problem with our school's IEP program. We then made a Pareto (bar graph) of the responses. We chose the variable where we could make the greatest impact based on the Pareto effects. Then, we conducted an experiment based on the chosen problem.

Approval:

We received approval from the school administration to conduct the survey. We then included students' parents in the email with the survey to receive their approval.

Overall Problems:

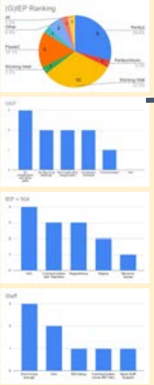
GIEP: No collaboration or scheduled meetings

IEP/504: Communication and program stigma

Staff: Communication and Specially Designed Instruction (SDI) confusion

SNR= <u>Sound Noise Ratio</u>: This ratio displays the number of responses to the number of no responses. We originally sent out 34 surveys, but 2 responses were claimed as "dead", so we removed them. We have since recorded **30 responses.

Graphs:



Solutions

GIEP:

- Create weekly meetings during connections° time slot.

IEP/504:

- Reduce Stigma
- Increase Communication

Staff:

- Create summary SDI reports

What We Did:

We created a 10 minute GIEP presentation and showcased our idea to all 5 principals.

GIEP students' follow-up survey is shown on the right.

Future Steps:

Future project explorers should discuss the communication issue with current IEP teachers and students.

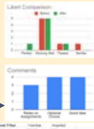
Long-term goals should include: School Board Presentations and K-12 program rerouting.

° Connections is a weekly ungraded class that all students must take

SECOND SURVEY

SNR*= 7/0

<u>GIEP</u>



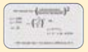
<u>Summary</u>

Upon resurveying the GIEP students, we found that the majority°° approved of our proposed solution to the flawed system. The change is awaiting approval from the school administration.

°°Note: One response was a potential anomaly.

<u>Minimal Sample Size</u>

To the **right**, there are calculations to find the min. sample size to detect a change of 0.5.



<u>Paired T-Test</u>

Using the Likert Scale, we compared the two surveys to construct how accurate of a change we found. The calculations can be seen **below** (excel sheet). We found with an **80% confidence that there was a change**.

CHALLENGES

- Social Engineering
  - The project's greatest challenge was to navigate the local political and administrative barriers involving confidential student information and even the studying of students itself. We had to establish mutual trust and respect.
- Data Collection
  - Throughout the first round of data collection, we faced initial nonresponse bias.

SKILLS USED

- Survey Design
  - We worked to create a clear and efficient survey with effective questions.
- Data Communication

Linear Algebra Team:
Central Dauphin High School

Oakland Catholic Data Jam 2023: Róisín Tsang, Maura Schorr
Mentor: Zhen Wu

Detention or Detainment: Does higher enrollment in schools impact crime rates?

School Enrollment in PA

Around the world there has been a rise in children skipping school. The pandemic greatly affected school enrollment rates, leading to an overall decrease in school attendance.

Credit: penndot.pa.gov

Crime in PA

Like many other states, PA has seen a rise in crime rates since the pandemic. We wanted to look into the possible factors affecting this ever present issue.

Null Hypothesis

There is no relationship between school enrollment and arrest rates in a neighborhood.

Alternate Hypothesis

Higher enrollment in schools has a negative linear correlation to arrest rates in a neighborhood.

Results:

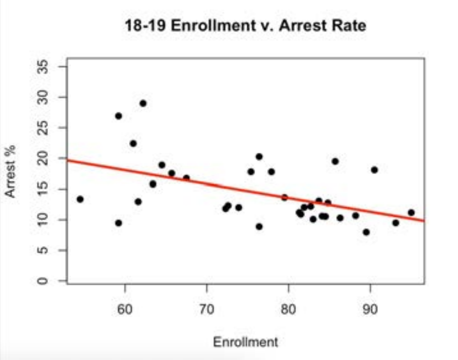


Figure 1: A graph of Arrest Rates over School Enrollment Rates for 18 to 19 year olds in PA in 2018. p-value = 0.003

Challenges we faced

We had to manually input the number of arrests and calculate the percent of the population arrested for our arrest dataset. We also changed mentors halfway through the year, so we had to catch up our new mentor with the state of our project.

Conclusion

The p-value of 0.003 far lower than the the alpha value of 0.05. This means that the relationship between our two variables is statistically significant and we can reject the null hypothesis that there is no relation.

Recommendations

Knowing what factors impact high crime rates could be greatly beneficial to the citizens of Allegheny county. Although further research of our hypotheses is required, any relationship between crime rates and school enrollment would possibly incentivise government officials to adjust the funding of our schools or create alternative programs to deter crime. A decrease in crime rates of our state would greatly benefit students and all citizens of our state which is why it is such a pertinent issue to investigate.

Sources

Arrest Rates by County in 2018: https://public.tableau.com/app/profile/stephen.st.vincent/viz/PSPUCR-TableBArrests/ArrestsbyAgeandSex
School enrollment in 2018:
https://data.census.gov/table?q=american+community+survey+school&t=School+Enrollment&g=0400000US42,42$0500000_1600000US4261000&y=2018
Age of County in 2018:
file:///private/var/folders/rj/grtzplsx1tdgr4hjf4m7z79h0000gn/T/com.microsoft.Excel/Microsoft%20Office%20Send%20Mail/AGE_BY_COUNTY_2018_CENSUS_BUREAU.pdf

Virtual vs Reality

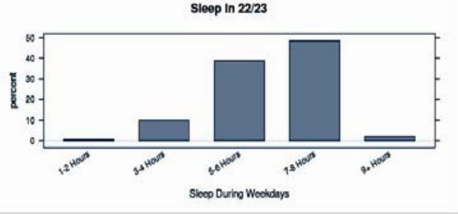
Kevin Hernadez, Yahir Martinez, George Sandoval, & Randy Juarez
Passaic Academy for Science and Engineering
Mentors: Connor Woods, Caldwell University & Taylor Mathis University of Pittsburgh

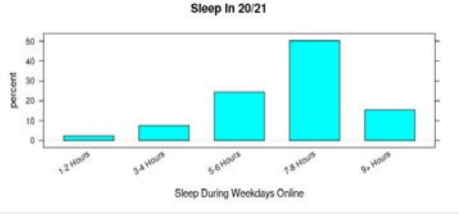
Project Development

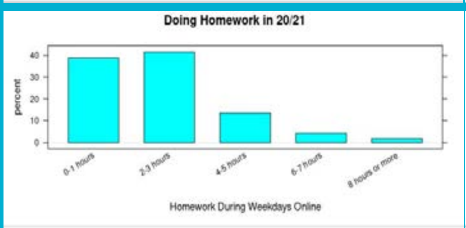
- Our hypothesis that we are trying to answer is: How did the pandemic affects students education during online learning?
- The project evolved by creting a survey for the whole school to fill out and analyze our results and make conclusions.
- We picked this project because we got curious about how virtual learning performances can or how it affected in person learning performances in this present time period

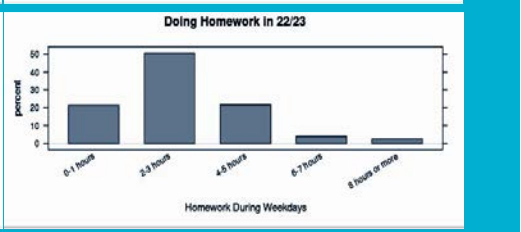
Data

- The data set we have used was a survey. In reality the only issue was getting people to actually respond to the survey
- The Data would have not needed to be cleaned as it was only needed to organize the data
- One of the main challenges while facing the data set how it was coded as it limited how ability to interpret the data and place them into graphs
- The focus would have not been limited during the process of interpreting the data.
- When it comes to things like time spent sleeping it was relatively the same compared to both online and in person.
- One of major limits of the data set was on the sample size and along with the fact that more data for things like mood would've been useful.
- Yes, the factors that constrained was how the values could only have been set to numerical.









Interpretations and Recommendations

- A recommendation could be teachers providing physical homework & online homework to support students into completing their responsibilities on time and faster.
- Ones results could make suggestions to be more responsible & manage their time more efficiently.

References

- Data Jam Survey designed and given at our school: https://docs.google.com/forms/d/e/1FAIpQLSdOIlmexeO6g7LD5TDDdWnBvOa0XJ27kpHlpLeI7_sB2_5AGw/viewform

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

Empty Streets, Cleaner Sheets

Zion Colon, Fernando Martinez, Brandi Ramirez, Joel Tlale
Passaic Academy for Science and Engineering

Mentors: Taylor Mathis, University of Pittsburgh
Anusha Pandey, Caldwell University

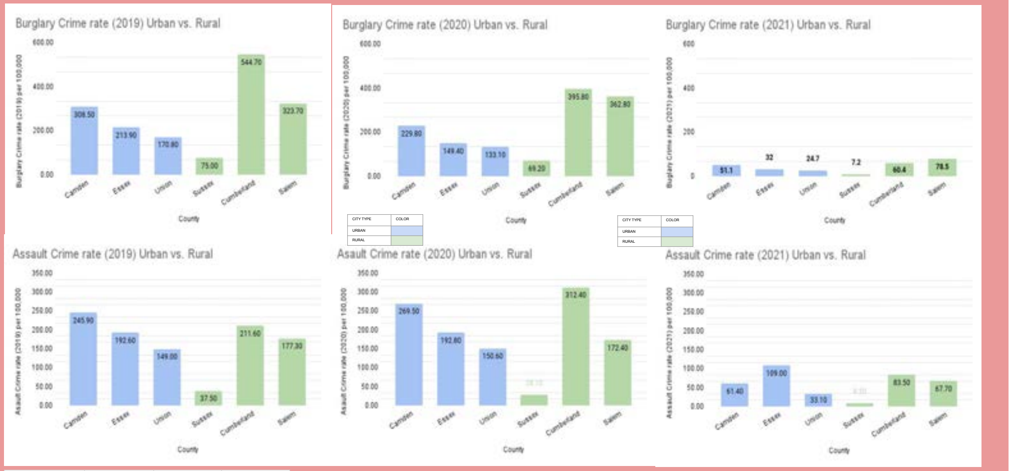
Project Development

- Our hypothesis was that crime rate in assault and burglary grew since the end of COVID-19 and compared to pre-COVID levels..
- The project started with looking for data going from Pre-COVID (2019) and more recent data (2022/2023), but after we could not find the data needed for 2022, we used data from 2019-2021.
- We decided on doing this because we noticed more crime in areas around us after covid. That is also the reason why we found it interesting.

Data

- We used the Uniform Crime Reports from the NJ State Police for 2021, 2020, and 2019.
- The data needed to be cleaned/reduced because there was extraneous data that we were not looking for inside of the Uniform Crime Reports. We did this by transferring the data to a new Google Sheet.
- We changed the years to reduce the range, and then we decided on the types of crime we would include.
- Protests and Political and Social Outrage due to events such as the Death fo George Floyd in 2020 may have affected the 2020 rise in assault for urban areas.
- A rise in depression and fiscal disparity due to Covid-19 killing jobs may have caused a rise in the crimes we focused on.

Data Visualizations and Findings



County	Population (2019)	Population (2020)	Population (2021)
Camden	503,145	523,485	503,145
Essex	800,305	863,728	800,305
Union	559,751	575,345	559,751
Sussex	138,714	144,221	138,714
Cumberland	149,815	154,152	149,815
Salem	61,473	64,837	61,473

Interpretations and Recommendations

- Following the trend depicted by the burglary crime rate. The assault crime rate was up between 2019 and 2020 and then fell dramatically post-lockdown.
- For all counties in NJ we found that the burglary crime rate was up between 2019 (before Covid-19) and 2020 (during Covid-19, but then the burglary crime rate fell dramatically after lockdown (2021).

Our Conclusions

The crime rate for burglary and assault crimes after covid has lowered compared to how it was before covid. Covid-19 had an effect on all populations everywhere that had contributed to a rise in crime during 2020

References

- 2021_UCR_Jan - march.xlsx https://docs.google.com/spreadsheets/d/1ivJuhtVGM4A9CllHO4yk-gPOfEEQtzUB/edit#gid=1524344962
- "Uniform Crime Reports." State of New Jersey Web Site, https://www.nj.gov/njsp/ucr/uniform-crime-reports.shtml.

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

Social and Health Determinants of Obesity in Pennsylvania

Bethel Park High School - Team #2

Ryan Patterson, Leo Devine, Alex Burt

Research Question

How do different social and health factors affect obesity rates within Pennsylvania?

Challenges

1. Data from multiple Sources: Some organizations store various data in different ways. This caused us to have to reorganize our data.

2. Data was not specific or organized enough.

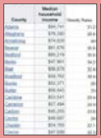
3. Visual data maps were not accurate or not easily read.

4. Data sets not being accurate or not having enough data.

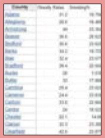
Data Sources

1. Western PA Regional Data Center (WPRDC)
   - Center Health data- Obesity Rates Statewide Data (2017)
   - Fast food restaurants density (2016)
2. Obesity rates of cities (MedJam & cdc.gov) (2021)
3. Population density of Pennsylvania (2018)
4. ArcGIS income map (2019)

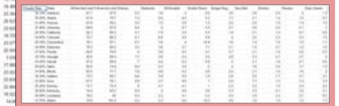
Data Set Examples



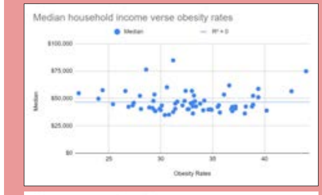
County	Median household income	Obesity Rates
Adams	[illegible]	[illegible]
Allegheny	[illegible]	[illegible]
Armstrong	[illegible]	[illegible]
Beaver	[illegible]	[illegible]
Bedford	[illegible]	[illegible]
Berks	[illegible]	[illegible]
Blair	[illegible]	[illegible]
Bradford	[illegible]	[illegible]
Bucks	[illegible]	[illegible]
Butler	[illegible]	[illegible]
Cambria	[illegible]	[illegible]
Cameron	[illegible]	[illegible]
Carbon	[illegible]	[illegible]
Centre	[illegible]	[illegible]
Chester	[illegible]	[illegible]
Clarion	[illegible]	[illegible]

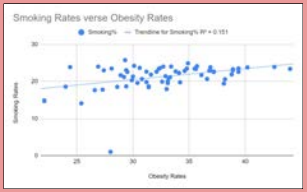


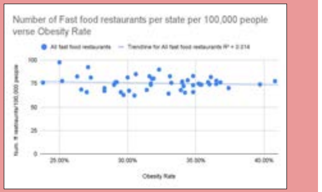
County	Obesity Rates	Smoking%
Adams	[illegible]	[illegible]
Allegheny	[illegible]	[illegible]
Armstrong	[illegible]	[illegible]
Beaver	[illegible]	[illegible]
Bedford	[illegible]	[illegible]
Berks	[illegible]	[illegible]
Blair	[illegible]	[illegible]
Bradford	[illegible]	[illegible]
Bucks	[illegible]	[illegible]
Butler	[illegible]	[illegible]
Cambria	[illegible]	[illegible]
Cameron	[illegible]	[illegible]
Carbon	[illegible]	[illegible]
Centre	[illegible]	[illegible]
Chester	[illegible]	[illegible]
Clarion	[illegible]	[illegible]
Clearfield	[illegible]	[illegible]

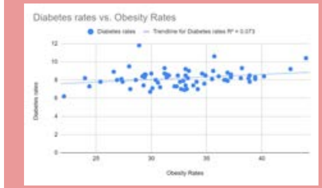


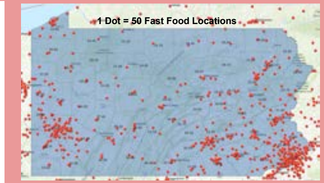
Visualizations

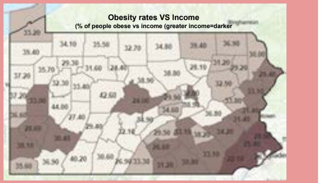












Summary

★ There is a minimal correlation between smoking rates, diabetes rates, and obesity rates. There is a negligible correlation between obesity rates and fast food density, population, and income. The data suggests that social (smoking) and health factors (diabetes) affect obesity rates more than those of an individual's environment.

Policy

★ In conclusion, there was no obvious correlation among obesity rates and fast food density, population, and income. There was a small but noticeable correlation between obesity rates and smoking rates as well as diabetes rates. Since there are only minimal correlations appearing, we suggest an overall education campaign that focuses both on counties with high obesity rates and low income areas. County health departments, hospitals and other non-profit organizations might partner to deliver this valuable health education to raise awareness of the health challenges associated with obesity.

Effects of Social Media on Teenage Sleep

Martina Tatalias, Leah Hartman, Leah Armstrong, Liz Alacce
Bethel Park Team #4

Question: Does an increase in social media usage have an impact on the number of hours teenagers sleep at Bethel Park High School?

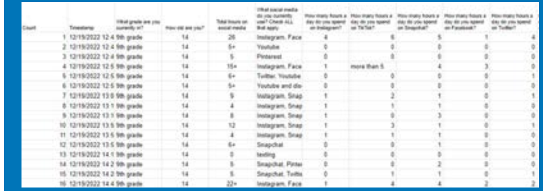
Challenges

Some challenges we faced during our project was that we had trouble getting a lot of people to take our survey. We also could have problems with people responding to the survey inaccurately giving us inaccurate data.

Data Sources

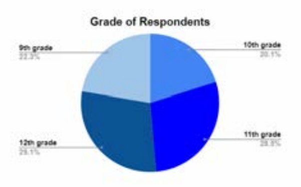
For our data source we created a survey and got it approved by our principal to be sent to all students at Bethel Park High School. In total we collected 183 responses. To get students to take our survey we sent out a school wide email to all the students. We also created posters with QR codes to the survey and posted them around the school and in classrooms From this we were able to gather all the data necessary to find any correlation.

Data Set Examples

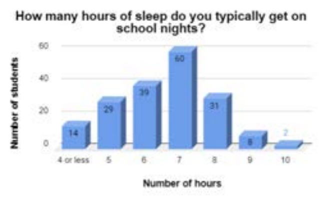


Count	Timestamp	What grade are you currently in?	How old are you?	Total hours on social media	What social media do you currently use? Check ALL that apply.	How many hours a day do you spend on Instagram?	How many hours a day do you spend on TikTok?	How many hours a day do you spend on Snapchat?	How many hours a day do you spend on Facebook?	How many hours a day do you spend on Twitter?
1	12/19/2022 12:4	9th grade	14	26	Instagram, Face	4	6	5	1	4
2	12/19/2022 12:4	9th grade	14	5+	Youtube	0	0	0	0	0
3	12/19/2022 12:4	9th grade	14	5	Pinterest	0	0	0	0	0
4	12/19/2022 12:5	9th grade	14	15+	Instagram, Face	1	more than 5	4	3	0
5	12/19/2022 12:5	9th grade	14	6+	Twitter, Youtube	0	0	0	0	1
6	12/19/2022 12:5	9th grade	14	5+	Youtube and dis	0	0	0	0	0
7	12/19/2022 13:0	9th grade	14	9	Instagram, Snap	1	2	1	0	1
8	12/19/2022 13:1	9th grade	14	4	Instagram, Snap	1	1	1	0	0
9	12/19/2022 13:1	9th grade	14	8	Instagram, Snap	1	0	3	0	0
10	12/19/2022 13:5	9th grade	14	12	Instagram, Snap	1	3	1	0	1
11	12/19/2022 13:5	9th grade	14	4	Instagram, Snap	1	1	1	0	0
12	12/19/2022 13:5	9th grade	14	6+	Snapchat	0	0	1	0	0
13	12/19/2022 14:1	9th grade	14	0	texting	0	0	0	0	0
14	12/19/2022 14:2	9th grade	14	5	Snapchat, Pintei	0	0	2	0	0
15	12/19/2022 14:2	9th grade	14	5	Snapchat, Twitte	0	1	1	0	1
16	12/19/2022 14:4	9th grade	14	22+	Instagram, Face	1	4	4	2	2

QR code to Google Form Survey

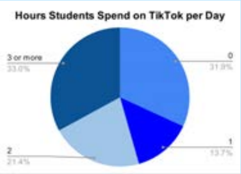


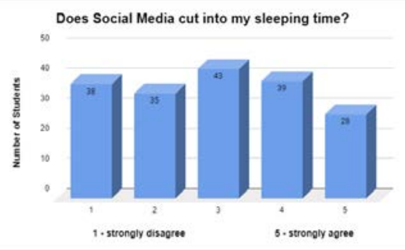
Analysis

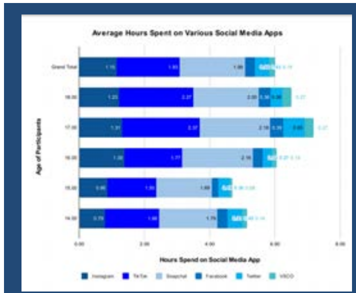


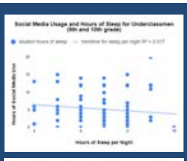
Teens ages 13-18 are supposed to get at least 8-10 hours of sleep a night, which means over 75% of the students are not getting enough sleep.

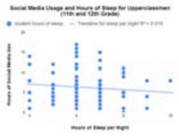
1/3 of students at Bethel Park High School spend 3 or more hours a day on just TikTok alone.

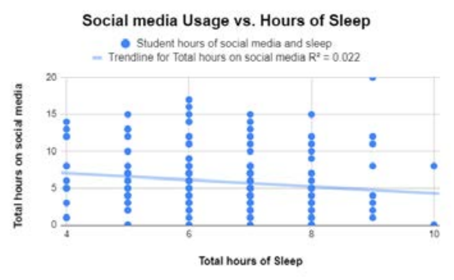












Analysis

Conclusion

Based on the high school students who chose to participate in the survey, we found a direct relationship between the amount of time teens spend on social media and the number of hours teens sleep per night. Although there are many outliers in our research the general trend suggests this relationship.

Potential Policy

Based on these findings, the companies that own these social media applications should encourage tracking time spent of them. The data can also be used to educate teens to spent less time on social media. With this data, schools can see the effect of social media usage and adjust school start times so that students perform better by getting more sleep. We will also let our school administrators know about our results so the school might be able to address the issue by educating students on how it is affecting their education.

Comparing Western Bluebird Populations Throughout Southern California

A DataJam Project with Pala Band of Mission Indians Learning Center and Torrey Pines High Bluebird Club

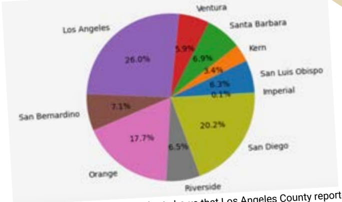
Amara Sanchez[1], Diana Durro[1], Doretta Musick[1], April Cantu[1], Sierra Kriss[1], Martina Calac[1]
Lily Bruch[2], Sneha Lele[2], Minchan Kim[3] ,Timothy Chu[3], Kimberly Mann Bruch[3]
[1]Pala Band of Mission Indians, [2]Torrey Pines High Bluebird Club,
[3]San Diego Supercomputer Center, UC San Diego

Who

Situated approximately 40 miles northeast of downtown San Diego and 30 miles inland from the Pacific Ocean, the Reservation of the Pala Band of Mission Indians is home to 1250 enrolled members– consisting of Cupeños and Luiseños. Meanwhile, Torrey Pines High School is situated in northwest San Diego county and serves more than 2000 students who live in the coastal communities of Del Mar, Solana Beach, and the surrounding areas.

What

A group of Pala youth collaborated with a group of Torrey Pines students to form a DataJam team and their goal was to study the variation in Western bluebird populations throughout southern California. They worked with the San Diego Audubon Society and used ebird.org (thanks to a generous donation from Cornell Laboratory of Ornithology) for their study. They were mentored by a UC San Diego data science student, funded by the National Science Foundation.



Compiled with Google Sheets, this pie chart shows that Los Angeles County reported more bluebird sightings to ebird.org than other southern California counties. (2023, Minchan Kim)

How

The Pala-Torrey Pines DataJam Team was thus formed; the students named their project "Comparing Western Bluebird Populations Throughout Southern California." To make the comparison of bluebird populations in the various counties of southern California, the students used data from ebird.org. Specifically, they used Google Sheets for analysis and found that San Diego County had more reports of bluebirds than nearby counties, as shown with the pie chart. Because the data was compiled by citizen scientists, we would like to note the bias issue as only uploads to ebird.org are shown with our work.

In addition to the data science aspect, the team also deployed bluebird boxes located near security cameras. The students placed one box near the Pala Learning Center and another one near the Torrey Pines High School. To date, they have not had any visits by bluebirds.

Future Work

Future work will involve the students continuing to monitor bluebird boxes. They also plan to prepare a proposal to present their data findings and future efforts at the July 2023 PEARC Conference in Portland, Oregon or the November 2023 Supercomputer Conference in Denver, Colorado.

Additionally, they are discussing future projects that can be accomplished over the summer as each student will have additional time for such work.

Project Lead Amara Sanchez, of the Pala Band of Mission Indians, determines where to place the bluebird box at the Pala Learning Center.

A recent effort to improve Western bluebird populations throughout Southern California appears to have been fruitful.

UC San Diego data science student Minchan Kim explains approaches for analyzing large sets of data. Not only did he work with the secondary students on simple analysis methods using Google sheets, but he also briefed them on the use of more complex tools such as Jupyter. (left)

UC San Diego computer engineering student Timothy Chu worked with the students on selecting birdboxes and camera systems for the hands-on aspect of the project. (right)

The project was funded by the National Science Foundation Northeast Big Data Innovation Hub (1916585) and the West Big Data Innovation Hub (1916481) with support also provided by the Pala Band of Mission Indians.

Is there a correlation between race, income level, or proximity to fast food restaurants with diabetes rates in Pennsylvania counties?

Hempfield Area High School Team
Adam Custer and Harrison Methven

Why are we studying diabetes?

- According to the World Health Organization (WHO), the number of people who have diabetes rose from 108 million in 1980 to 422 million in 2014
- Despite the patent for manufactured insulin originally being sold for $1 in 1923, prices today have skyrocketed to up to $350 per vial for this life saving medicine
- The purpose of this data collection was to find if there were systemic factors that correlate with the amount of diabetes cases

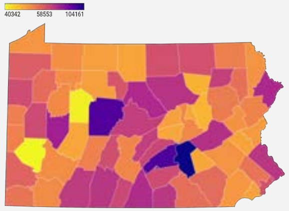
Hypotheses

- We suspected that there is a strong positive correlation between the number of fast food restaurants and percentage of population affected by diabetes in PA counties?
- We also suspected there is a strong negative correlation between average income level and percentage of population affected by diabetes in PA counties?
- We also assumed that African Americans are disproportionately affected compared to Non-Hispanic Caucasians

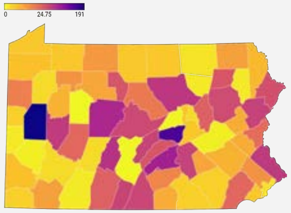
Challenges

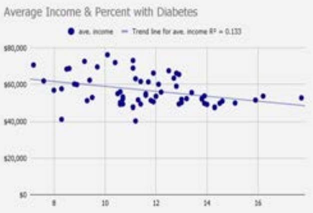
- Difficulty in obtaining data, especially with respect to numbers regarding race. Some sources used raw data and others were based on how individuals identified as a particular race.
- Type 1 and Type 2 Diabetes have different causes (genetics & diet respectively) which makes them hard to correlate with the variables studies here.

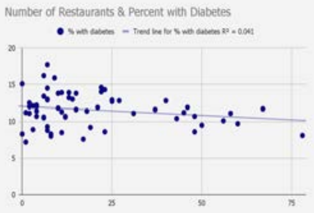
Average Income by County in Pennsylvania



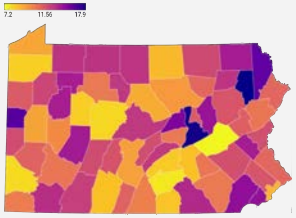
Number of Fast Food Restaurants per County in Pennsylvania



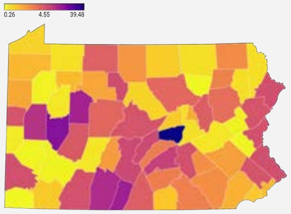


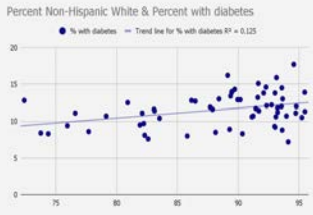


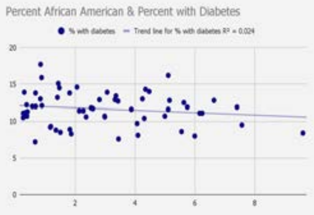
Percent of Population with Diabetes per County in Pennsylvania



Percent of African American per County in Pennsylvania







What did we find?

- As shown by the scatterplots on this poster, there is little to no correlation between our 3 tested variables and percentage of population with diabetes
- This was somewhat expected, because there are many possible hidden or lurking variables present when looking at data involving human beings
- Further analysis could include looking for pockets of increased diabetes rates in more populated areas within the counties as opposed to looking at rates simply across the counties

Conclusion

- There is little to no correlation between the number of fast food restaurants and percentage of population affected by diabetes
- There is also little to no correlation between income and diabetes
- The data suggests African-Americans and Caucasians have similar rates of diabetes
- Based on these findings, there is no easily definable societal factor for diabetes
- Further research into this topic should include smaller scale analysis as well as differentiating between the two main types of diabetes

Dataset Sources

- **Diabetes** (https://www.socialexplorer.com/a9676d974c/explore)
- **Race** (https://data.world/amberthomas/county-population-by-race-2020-census/workspace/file?filename=county_population_by_race.csv)
- **Fast-food Restaurants** (https://hub.arcgis.com/datasets/UrbanObservatory::fast-food-restaurants/explore?layer=0&location=37.237780%2C-80.119736%2C7.00&showTable=true)
- **Income** (https://www.socialexplorer.com/tables/ACS2020_5yr/R13337178)

Data Spreadsheet

(We will approve all requests to review the data)

Café GPA

Alison Leung, Salaha Suleyman, and Madeline McDine
Keystone Oaks High School

Question

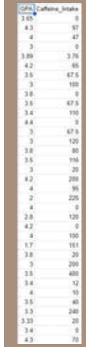
Does caffeine have an affect on your academic performance in school/GPA?

Null Hypothesis

There is no correlation between caffeine intake and GPA.

Alternate Hypothesis

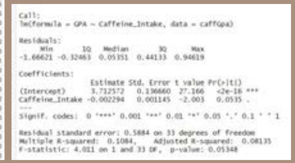
There is a correlation between caffeine intake and GPA.



GPA	Caffeine_Intake
3.65	0
4.3	97
4	47
3	0
3.89	3.76
4.2	65
3.5	67.5
3	100
3.8	0
3.5	67.5
3.4	110
4.4	0
3	67.5
3	120
3.8	80
3.5	116
3	20
4.2	200
4	95
2	225
4	0
2.8	120
4.2	0
4	100
1.7	151
3.8	20
3	200
3.5	400
3.4	12
4	10
3.5	40
3.3	240
3.33	20
3.4	0
4.3	70

Data

We created a survey and used it to crowdsource data from the students at our school.



```
Call:
lm(formula = GPA ~ Caffeine_Intake, data = caffGpa)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66621 -0.32463  0.05351  0.44133  0.94619

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.712572   0.136660  27.166   <2e-16 ***
Caffeine_Intake -0.002294   0.001145  -2.003   0.0535 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5884 on 33 degrees of freedom
Multiple R-squared:  0.1084,    Adjusted R-squared:  0.08135
F-statistic: 4.011 on 1 and 33 DF,  p-value: 0.05348
```

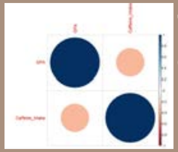
Background

Many high school students are stressed and tired so they use caffeine to help them stay awake and concentrate in school. High schoolers that take higher classes such as honors or AP have to stay up longer on school nights to study for tests or finish their homework since these classes demand more work and effort.

Challenges

We had trouble collecting accurate data because our questions were misunderstood by many. Most people don't pay attention to their caffeine intake so many of our testers estimated or guessed and it could've been incorrect. We also had trouble cleaning the data since there were unreasonable answers or answers that were not numerical, so we had to eliminate a few values from the overall dataset.

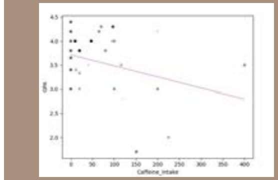
Correlation:

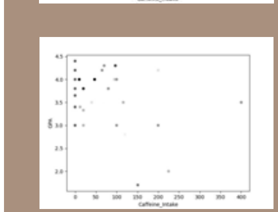


The correlation plot between gpa and caffeine intake shows that it has a weak negative linear correlation.

Correlation Coefficient: -0.32919

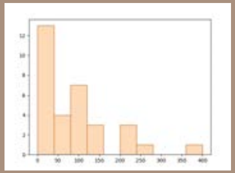
Caffeine Intake Vs GPA



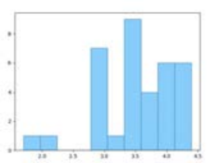


Line of best fit: y=-0.002294x + 3.713

Histograms

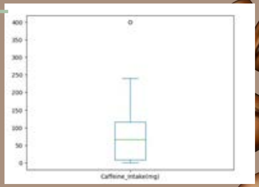


Caffeine intake has a strong skewed right distribution.

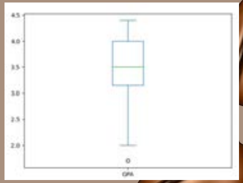


GPA has a bimodal distribution.

Box and Whisker Plots



Median caffeine intake: 67.5 mg
Mean caffeine intake: 81.836



Median GPA is 3.5
Mean GPA: 3.5

Conclusion & Recommendation:

Although we had a low negative correlation between GPA and caffeine intake, since our p value is slightly above 0.05, we fail to reject our null hypothesis. As a result, there is no significant statistical evidence that shows a correlation between a student's caffeine intake and GPA. A recommendation based on our results are for students to engage in healthy habits (for example, sleeping more) and to put in the work required for academic success. Caffeine intake most likely does not have a positive effect on their grades and may be detrimental to students in the long run.

STAT 1000 - Central Dauphin High School

Bob Moreland; Robert Eberly, Elijah Mackey, Marlayna Maurlanda-Rea, Vedant Patel, Sage Winters

Natural Disasters and Climate Change

Question: How is global temperature change related to the frequency of natural disasters and how will these responses escalate in the future?
Hypothesis: When global temperature anomalies increase, the frequency and intensity of natural disasters will also increase.

Summary of Analyses, Results, and Datasets Used:

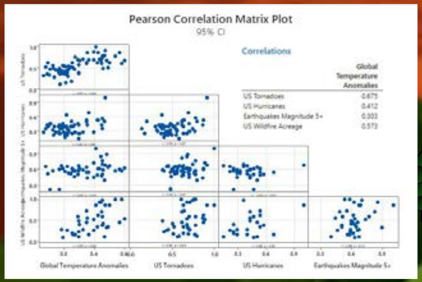
The process of answering our question began with gathering data on global temperature anomalies, tornadoes, hurricanes, earthquakes, and wildfires. The data for US tornadoes and hurricanes comes from the National Oceanic and Atmospheric Administration (NOAA), our earthquake data is provided by the United States Geological Survey, the wildfire data is from the National Interagency Fire Center, and the global temperature anomaly data is from NASA. Our first plan of action was to create a time series of all the predictors (natural disasters) and the response variable (temperature anomalies) to see if all data trended in an upward direction like we hypothesized. In order to do this we had to contract the tornado and hurricane data sets to be yearly instead of monthly and we also scaled all data to be in terms of frequency so it could be presented on one graph. The time series graph pictured here is the result and shows an upward trend for all datasets as time moves forward.

In the next part of our analysis we looked at the correlation between the natural disasters and global temperature anomalies to see if there was a strong, moderate, or weak correlation between each natural disaster and the global temperature anomalies. US tornadoes and wildfires show a moderate, positive correlation while US hurricanes and earthquakes show a positive but low correlation.

Challenges

- Finding reliable data sets that fit a specific time frame and range
- Getting the frequency of the natural disasters onto one graph.
- Analyzing the data, as the different time frames affected our time series analysis

There were several challenges we faced, but the biggest challenge was finding data for a large time span. Some of the data was limited to the United States and to the technology available in the past years. Because of this, it was difficult to determine the time frame to address. Additionally, contracting datasets with a yearly frequency instead of monthly was a challenge.



Correlations

	Global Temperature Anomalies
US Tornadoes	0.673
US Hurricanes	0.412
Earthquakes Magnitude 5+	0.303
US Wildfire Acreage	0.573

The final part of our investigation involved running all of our data through a regression analysis to look at the r-squared values and p-values to determine if we would reject our null hypothesis. Our null hypothesis is that the coefficient for global temperature anomalies is equal to zero, meaning there would be no effect on the increase in temperature anomalies to the frequency of a particular natural disaster. According to our regression equations, the coefficients did not equal zero and the p-values for each equation were small enough to reject the null hypothesis. This means that the global temperature anomalies do affect the frequency of the four natural disasters studied and that as the temperature anomalies increase, the frequencies of the natural disasters also increase by some degree.

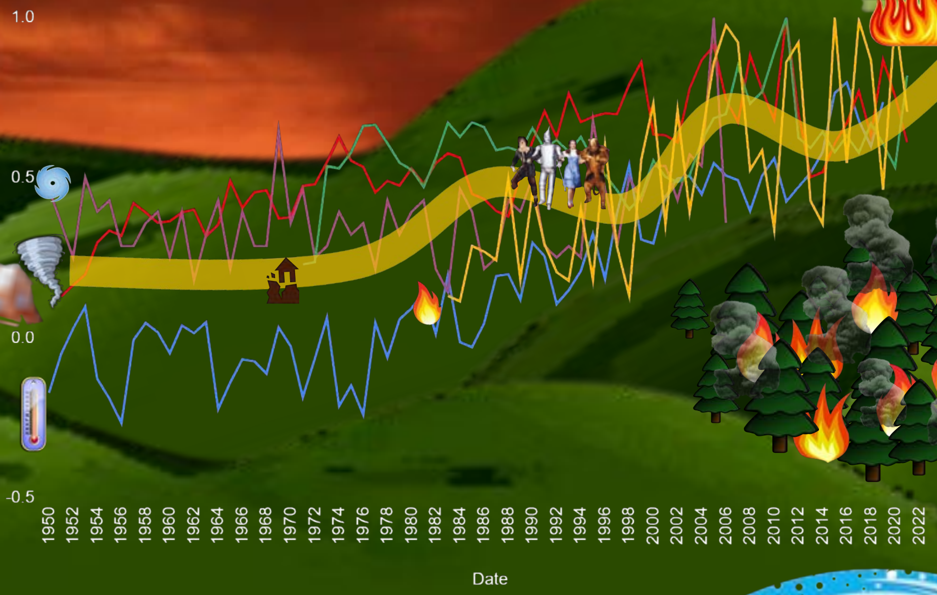
R-sq values:			
Tornadoes	Hurricanes	Earthquakes	Wildfires
45.6%	16.97%	9.19%	32.89%

Regression Equations:
US Tornadoes = 0.47 + 0.45 Global Temp. Anomalies
US Hurricanes = 0.36 + 0.27 Global Temp. Anomalies
Earthquakes Magnitude 5+ = 0.5535 + 0.1486 Global Temp. Anomalies
US Wildfire Acreage = 0.2138 + 0.741 Global Temp. Anomalies

Ho: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$

Global Temperature Anomaly p-values:
US Tornadoes: 0.000
US Hurricanes: 0.001
Earthquakes Magnitude 5+: 0.034
US Wildfire Acreage: 0.000

$p < 0.05$ → **Reject Ho**



Conclusion and Recommendation:

According to our research, the majority of weather-related disasters are likely to become more frequent by the end of the century. There will be an increase in the frequency of tornadoes, hurricanes, and wildfires all of which are affected by an increase in the global temperature anomaly by some degree. The world needs to get ready for this shift as natural disasters increase in frequency. Governments should make investments in robust infrastructure that can endure growing threats, such as rising sea levels and stronger winds. To cut expenses in the future, it will be crucial to update zoning regulations and building requirements to take climate change into consideration. Currently, nations should save so they can prepare for an increase in government spending to help their economies when natural disasters related to climate change strike.

Next Steps

1. Create equations that represent the correlation between temp. anomalies and the natural disasters.
2. Identify the additional correlation with rising sea levels and glacial change
3. Use the regression equation and other predictive analytics to predict the number of future disasters
4. Account for daily or seasonal temperature anomalies to more accurately analyse the effect of temp. anomalies on natural disaster frequency

Pittsburgh Regional Transit - Vehicle Age and Why it Matters

Team Member Names: Sruthi Yerram, Sravya Cavuturu, Sai Sri Sowgandika Ramadugu, Krishna Autade, Sai Krishna Ramadugu,

Mentors: Phil Light Sam Winward

Central Question

How does Pittsburgh Regional Transit (PRT) manage its aging fleet of buses?

Why is it Important?

PRT gets federal funding to buy vehicles.

The FTA mandates these vehicles be driven for at least 12 years and/or 500k miles before being retired or send to scrap.

For that reason, it's important these vehicles continue to accrue miles even if they are less desirable to drive than newer vehicles.

Hypothesis

We expect that newer buses will accrue more miles than older buses due to them being more reliable and desirable.

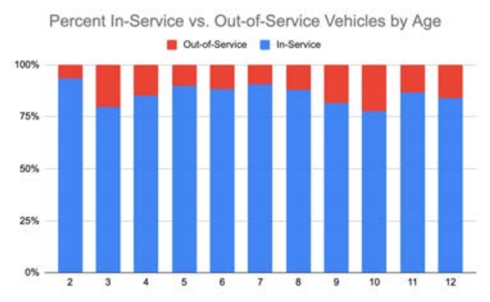
Datasets

1. AVL (automatic vehicle locator) trip-level PRT data from March 2023. Showing the vehicle used on each trip. Over 200k records.
2. PRT scheduled service data. Total trip distance and garage associated with all scheduled trips. About 20k unique records for the current March 23 schedule.
3. PRT fleet data. A smaller dataset showing the year/make/model etc. of all buses in PRT possession.

Methodology

- Combined datasets. Used vehicle ID to get vehicle age and trip ID to join distance and garage.
- Pivot tables. Summarized trip-level data across dates by age and garage.
- Weighted average. Calculated avg age weighted by distance traveled.

Results

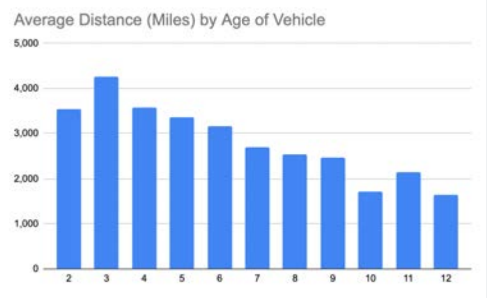


OOS vehicles are those that did not run a single trip in March 23.

Avg Age vs. Weighted Avg Age by Garage

	Ross	Collier	Mifflin	Liberty	Total
Avg Age	6.95	6.93	7.19	7.12	7.07
Avg Age by Use	6.69	5.94	6.21	6.50	6.33
Difference	**-0.26**	**-0.99**	**-0.98**	**-0.62**	**-0.74**

PRT has four bus garages each with their own fleet of vehicles.



Total March 23 miles traveled per vehicle of each age.

Bus age does not appear to be a significant predictor of out of-service.

Likely because reasons for long term out-of-service, like crashes, do not discriminate by age.

Avg vehicle age is approximately 7 years for all garages.

However, avg age by use varies. Collier and Mifflin have the largest discrepancy with vehicles used being almost a year younger on average.

Newer vehicles tend to be driven more!

Conclusion

- We found that vehicle usage drops as vehicle age increases.
- Some garages have a bigger difference between vehicle avg age and avg age used.

There could be a couple reasons this is happening:

1. Older vehicles may be more likely to incur mechanical issues that take them off the road for some period of time (but not the entire month).
2. Operators or supervisors may prefer newer vehicles because they are more reliable, comfortable, etc.

Discussion

Newer vehicles are preferable to both operators and riders.

However, it is important older vehicles continue to accrue miles so they can be retired, and new vehicles can be purchased.

Further investigation is needed into the reliability of buses as they age. And why such a difference in usage exists across garages.

Redlining: A Broader Perspective

How has redlining in the 20th century affected socioeconomic status in the modern day? How have other socioeconomic influences compounded with those of redlining?

Norwin High School

Aaron Berger • Dmitri Berger • Adam Guskiewicz • Suhana Navalgund • Simone Pal • Rex Wu

Background

- Redlining was a grading technique used by banks during the early-mid 20th century. 'HOLC Grades', named for the Homeowner's Loan Corporation that oversaw and distributed redlining maps, were assigned to urban areas.
- It was used to identify which areas would be favorable for loans, and which would not; we refer to areas that were redlined as 'redlining tracts,' and their grades as 'redlining grades.'
- Loan eligibility was judged based on factors such as potential and present industrial development, location, history, etc.
- Crucially, racial and ethnic background was a key consideration, with bankers giving lower 'grades' to areas with minorities and other 'undesirable' groups.

Hypothesis

1. We hypothesize that the majority of formerly unfavorably redlined areas will be worse off in socioeconomic measures than areas that were not as severely targeted.
2. We hypothesize that, disproportionately, areas unfavorably redlined in the past will consist of high minority population makeups.
3. We hypothesize more damaging influences–namely low geographic mobility–to have a negative impact on areas
4. We hypothesize areas more unfavorably redlined in the past to demonstrate greater potential for and occurrence of gentrification, compared to their favorably redlined counterparts

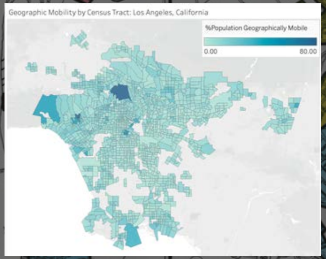
- These hypotheses are in part drawn from our previous projects, in which we found redlining to have a damaging socioeconomic influence in Pittsburgh.

Hypothesis Testing

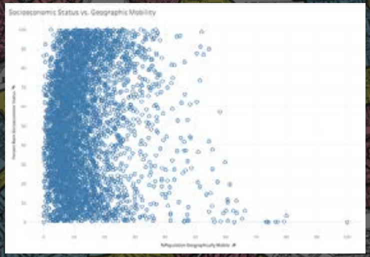
1.
   - Null Hypothesis: On average, census tracts with a lower redlining grade do not necessarily have a lower modern socioeconomic status.
   - Alternative Hypothesis: On average, census tracts with a lower redlining grade do have a lower modern socioeconomic status.
2.
   - Null Hypothesis: On average, census tracts with a lower redlining grade do not necessarily have a disproportionately modern minority population makeup.
   - Alternative Hypothesis: On average, census tracts with a lower redlining grade do have a disproportionately large modern minority population makeup.
3.
   - Null Hypothesis: On average, census tracts that demonstrate low geographic mobility do not necessarily have lower modern socioeconomic status.
   - Alternative Hypothesis: On average, census tracts that demonstrate low geographic mobility do have lower modern socioeconomic status.
4.
   - Null Hypothesis: On average, census tracts with a lower redlining grade do not necessarily demonstrate greater potential for and occurrence of gentrification.
   - Alternative Hypothesis: On average, census tracts with a lower redlining grade do demonstrate greater potential for and occurrence of gentrification.

Methodology

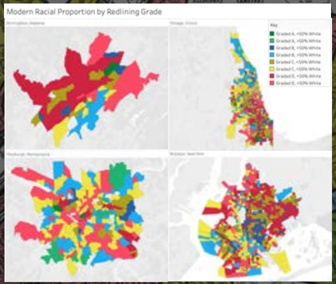
- We used Python software, written in the Spyder 5 IDE, to web-scrape data.census.gov, downloading almost a dozen data tables for each of nearly 16,000 census tracts across 200 cities throughout the United States, compiling and converting information into usable form for statistical analysis and visualization.
- We used software to scrape the geometries of over 200 redlining maps from the Mapping Inequality Project, and similarly gathered the geometry of census tracts of 37 different US states from geospatial file repositories from census.gov, correcting or removing inconsistencies in the maps, using Google Earth Pro.
- In order to assign a redlining grade, originally provided to an arbitrary area in the 1930s and 40s, to a modern census tract, we used geospatial visualization software, specifically Google Earth Pro, and programming modules designed to work with geometries to assign a grade between 0 and 100 to the tract, based on how much of each grade of redlining tract it intersected.
- We used software to import our data into Tableau in a usable format that could render geospatial visualizations.
- We used Tableau to create geospatial and statistical representations and analyses of our data, incorporating multivariate analysis techniques into our process.
- To be concise, this poster will display geospatial visualizations of the most prominent cities in the data.


Geographic Mobility by Census Tract: Los Angeles, California
%Population Geographically Mobile
0.00
80.00

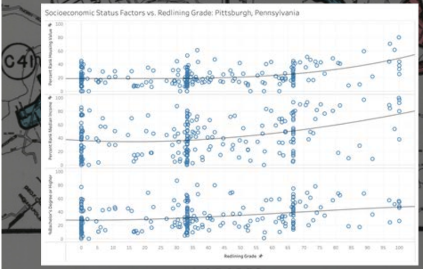
This geospatial visualization displays a map of the city of Los Angeles, the second most populous city in the United States. It is subdivided by census tracts. The census tracts are shaded by the percentage of their population that is geographically mobile–that is, the percentage of each tract's population that moved into the tract from another location. Specifically, the different categories of such movement have been compounded into one factor, Geographic Mobility, in a simple factor analysis. This factor encapsulates the occurrences and to a small extent circumstances of mobility, which can then be correlated with other measures to determine a relation. It is not a measure for ease of movement. Notably, the city demonstrates very little geographic mobility overall.



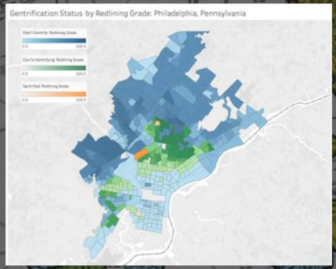
This statistical visualization is a comparison of socioeconomic status, as a function of geographic mobility. It contains close to 5,000 census tracts from the most prominent urban occurrences of redlining. Demonstrably clear is that, disregarding a few outlying groups of census tracts, the scatter plot lacks a clear correlation between socioeconomic status and geographic mobility. The overwhelming majority of census tracts–each one represented by a single point on the plot–are on the lower end of the spectrum of geographic mobility, and further demonstrate no clear pattern in this regard. That is to say, socioeconomic status in a census tract seems to be largely independent of geographic mobility. As such, it can be taken that there is at best a minimal or negligible correlation between socioeconomic status and geographic mobility, as displayed by this large sampling of census tracts. Other influences upon socioeconomic status must be taken into account.



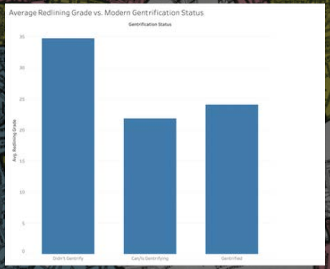
The visualization depicts geospatial maps of the cities of Birmingham, Chicago, Pittsburgh, and Brooklyn. They have been organized by redlining grade, such that modern census tracts corresponding to an aggregate area in the 1930s and 40s display a redlining grade, with 'A' being in green, 'B' being in blue, 'C' being in yellow, and 'D' being in red. Brighter hues indicate that the census tract has a majority-white population. Notably, the green 'A' tracts and especially the blue 'B' tracts on the maps–the more favorable redlining grades–are on average more likely to be a majority white tract. Similarly, the less favorable grades–the yellow 'C' and red 'D'–are more likely to be dimmer than their favored counterparts, in such a case indicating that the population of the census tract is a majority of minority groups. These four cities more broadly represent a common pattern in this relation, indicating a strong correlation.



This statistical analysis displays the relation between the past redlining grade of census tracts and key facets of socioeconomic status, namely educational attainment, median household income, and critically important to the potential influences of redlining, median housing value. The census tracts of subject are those of Pittsburgh. Polynomial regression trend-lines have been added to more effectively display the relationship. As clearly evident, all three categories of socioeconomic status demonstrate some increase in tandem with the increase of redlining grade–that is, a higher redlining grade is indicated to correspond to a higher socioeconomic status. Notably, however, more dramatic influences by redlining grade are seen on median household income and median housing value than that of educational attainment. This discrepancy is deserving of further investigation and analysis. Such inconsistency notwithstanding, the data clearly provides an example wherein a lower redlining grade has a negative influence on socioeconomic status, and vise versa. It can thus be supported, as this example pattern is observed throughout much of the accumulated data in similar fashion, that at least a non-negligible correlation exists between redlining grade and socioeconomic status.


Gentrification Status by Redlining Grade: Philadelphia, Pennsylvania
Didn't Gentrify: Redlining Grade
0.0
100.0

The visualization depicts a geospatial map of Philadelphia, Pennsylvania. It is shaded based on the calculated redlining grade of each census tract, determined through compounding the grades of all parts of redlining tracts within the census tracts, where '0' corresponds to a tract encompassing purely areas with a redlining grade of 'D', and '100' corresponding to a tract encompassing purely areas with a redlining grade of 'A'. The gentrification statuses of the tracts were divided into three categories: 'Did not gentrify,' 'Can Gentrify/Is Gentrifying,' and 'Gentrified.' This is measured over a decade long span, from 2010 to 2019 (whereafter census tracts would change due to the decennial census). Tracts eligible to gentrify are in short determined to be of acceptably low overall socioeconomic status at the start of the decade; tracks that are considered to have gentrified demonstrate significant increases in said status by the end of the decade. It should be noted that tracts that are 'eligible' to gentrify may be in the process of gentrifying. Notably, most of the tracts that are 'eligible' to gentrify or have gentrified are of a light shade, indicating a past unfavorable redlining grade. This map is exemplary of a broader pattern observed in our analysis, wherein tracts within areas more poorly graded in the past demonstrate gentrification or the potential for it than their counterparts, thus possibly recovering from the more damaging effects of redlining when they were first graded. This indicates a non-negligible correlation between redlining grade and potential for and occurrence of gentrification.



This visualization demonstrates the relationship between the redlining grade of a modern census tract and its potential for and occurrence of gentrification in the same 5000 modern census tracts shown in previous visualizations. The average redlining grade of census tracts across urban centers is measured on the y-axis, with the three categories of gentrification activity represented by the bars on the x-axis. As exemplified in the adjacent geospatial visualization, tracts that can gentrify or have gentrified are more often than not tracts with a relatively lower redlining grade. This is further demonstrated by this bar-graph analysis of urban census tracts. The average past redlining grade of tracts that did not demonstrate gentrification is significantly higher than those that did. A noteworthy feature of this analysis is that the composition of tracts that are eligible to gentrify or have not yet 'concluded gentrifying,' in a manner of speaking, possess an even lower average historical redlining grade than tracts that have gentrified. Further analysis would be required to determine the source, if any, of this discrepancy. Regardless, the visualization clearly indicates a strong correlation between past redlining grade in modern census tracts and modern gentrification patterns.

Challenges

- Establishing boundaries to test and analyze hypotheses covering multiple influences
- Scraping significant amounts of data for thousands of different census tracts, across hundreds of cities, over almost a dozen years, in an optimal timeframe
- Assigning a redlining grade, taken from arbitrarily divided redlining tracts from the 1930s and 40s, to modern census tracts of different geometry
- Working with two different census maps from two different decades.
- Identifying, correcting & removing inaccuracies in redlining and census data
- Creating programs to optimize and automate different aspects of our approach
- Creating, compiling, and analyzing multiple compounded socioeconomic measures and influences

Summary

We found that while all of our null hypotheses could be rejected with a high degree of confidence, as the p-values of all four hypothesis tests were well below our significance level of 5.00%, we also found that not all of our alternative hypotheses could be readily accepted as predicting a clear correlation.

1. As is evident in our analysis of the relationship between facets of socioeconomic status and historical redlining grade, our findings support our more broadly predicted alternative hypothesis as to the nature of that relationship. To at least a non-negligible extent, past redlining grade appears to affect modern socioeconomic status such that an area encompassed by a modern census tract and once favorably graded during the redlining era is more likely than not to display a higher socioeconomic status than its less-favorably graded counterparts. This supports a present, albeit not necessarily strong correlation between socioeconomic status and past redlining grade.
2. As exemplified by the visualized analysis, our findings indicated a strong support of our alternative hypothesis; our data indicates that on average, census tracts encompassing areas less-favorably graded under redlining were more likely to consist of disproportionately large minority populations, demonstrating a significant correlation between past redlining grade and demographic makeup in modern census tracts.
3. As displayed by the provided data, there exists little to no correlation between socioeconomic status and geographic mobility in a census tract. This contradicts our alternative hypothesis, as we had predicted a direct relationship between the two. Noteworthy is that this potentially eliminates geographic mobility as an interfering influence on other analyses of socioeconomic status. More investigation would be required to determine this nature.
4. As demonstrated by our findings on the occurrence of gentrification, our data strongly supports our alternative hypothesis that on average tracts with less favorable redlining grades are more likely to gentrify. This underscores a significant correlation between past redlining grade and modern gentrification in modern census tracts, as well as weakening the direct correlation between redlining grade and modern socioeconomic status.

Next Steps

- Automate the identification of discrepancies in data so as to avoid manual corrections
- Explore different definitions of gentrification in order to access a broader swath of potential analyses and correlations
- Analyze population density to more accurately characterize aggregated geometries
- Analyze differences in urban areas that might lead to inconsistent correlations and patterns between cities

Dataset Descriptions

- Census Tract Data: A website of the United States Census Bureau. Contains dozens of census tables for census tracts all throughout the United States, with data compiled through the decennial census and organizations such as the American Community Survey.
- Census Trap Maps: A repository of the United States Census Bureau's geospatial maps of census tracts throughout the United States, organized by state and containing tracts from the 2010 and 2020 census.
- Mapping Inequality Project: An interactive website published by the University of Richmond and associated groups storing, organizing, and digitizing redlining maps and data from the 1930s and 40s. Besides containing a user interface for viewing those maps, it also contains shape-file data for the geometries of redlining tracts across redlined cities in the United States.

Resources

Census Tract Data: https://data.census.gov/tables
2020-Present Census Tract Maps: https://www2.census.gov/geo/tiger/TIGER2022/TRACT
2010-2019 Census Tract Maps: https://www2.census.gov/geo/tiger/TIGER2019/TRACT/
Mapping Inequality Redlining Maps: https://dsl.richmond.edu/panorama/redlining
Programming and Web-Scraping: https://www.spyder-ide.org
Gentrification Identification: https://www.governing.com/archive/gentrification-report-methodology.html
Geospatial Visualization and Manipulation: https://www.google.com/earth/versions/#download-pro
Statistical Analysis and Visuals: https://public.tableau.com/app/discover

Factors Affecting Air Quality In NY State

Passaic Academy for Science and Engineering
By: Dhruv, Roberto, Andrea and Brenda
Mentors: Connor Woods Caldwell University

Project Development

As we embarked on our project, we recognized the challenge ahead of us, to uncover the hidden variables that shape air quality. Our initial exploration revealed a numerous amount of factors that contribute to AQI. However, we choose, after much research, to experiment with different topics until we discovered the perfect subject. We chose to look into the often overlooked aspects of air quality, recognizing the potential cause of high AQI or maybe not. We firmly believe that air pollution demands greater attention, as it has an impact on our health and the well being of our planet.

Statistical Question

How does the population and amount of factories contribute to air quality in NY?

Data

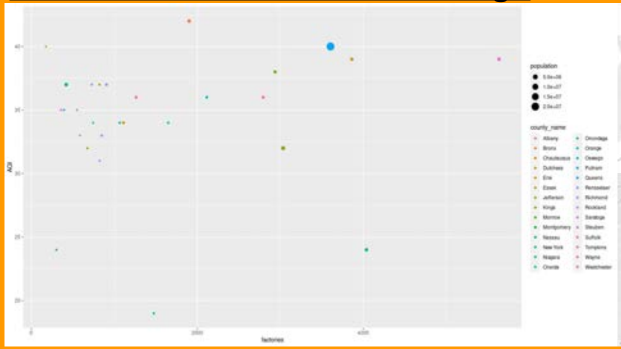
- In total three datasets we used: one for factories, one for population and one for air quality.
- Finding the data was tricky at first but we eventually did find the data we needed through thorough research
- The data needed to go through extensive cleaning and filtering
- We filtered the location of these variables since many had other locations
- We didn't focus on simplifying the project because we wanted to have a more informative research
- When filtering the data, we had realized in multiple instances that the data wasn't useful so we had to look for other datasets during our research. The datasets that we were looking for were simple so we could join them together into a meta data.

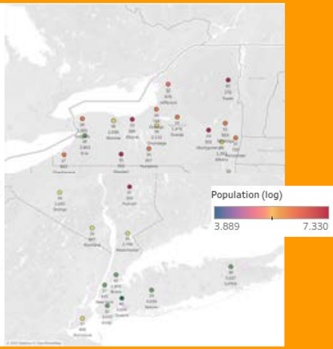
Feedback and Possible Future Research

- We could have added a bit more graphs with grouping to show how the range of AQI is affected by certain variables.
- Things we could look at moving forward are :
- How might the gender of the majority of the population affect the AQI?
- How does the size of the factories affect the pollution?
- Does the region of factories have an affect on the AQI?

Data Visualizations and Findings

Counties with different colors depending on the population along with AQI and number of factories





Factories and AQI with dots for each county and size depending on the population

Interpretations and Recommendations

- We observed that the air quality index (AQI) levels in the corresponding counties were surprisingly decent compared to the population
- A good AQI score is 0-50 and a moderate score is 51-100
- Given the size and population density of a state like New York, one might expect to see a significant leve of pollution, however, the data says otherwise
- The research allows us to see how certain unnoticed variables in a state affect the air quality
- The graph that was generated helps in seeing how states with high amounts of factories seem to have higher AQI but we also saw how there was cluster of points at a good AQI close to lower amount of factories
- The graphs are not as we expected and to our surprise, the graphs with the data collected, shock us as there were more counties who had a greater numbers of factories with better AQI than with less
- The main variable affected the AQI is the pollution since there is higher AQI as the dot points get bigger with the population
- Based on our results, the population and factories do affect the AQI by quite a bit as there is a increase in AQI when there is an increase in population and factories

References

- "Annual Estimates of the Resident Population for Selected Age Groups by Sex for New York: April 1, 2020 to July 1, 2021 (SC-EST2021-AGESEX-36)." Index of /Programs-Surveys/Popest/Tables/2020-2021/State/Detail, U.S. Census Bureau, Population Division, June 2022, https://www2.census.gov/programs-surveys/popest/tables/2020-2021/state/detail/.
- "Facilities Licensed by the Department of Motor Vehicles (DMV): State of New York." Facilities Licensed by the Department of Motor Vehicles (DMV) | State of New York, NY Open Data, 3 Apr. 2023, https://data.ny.gov/Transportation/Facilities-Licensed-by-the-Department-of-Motor-Veh/nhjr-rpi2/explore/query/SELECT%0A%20%20%60facility%60%2C%0A%20%20%60facility_name%60%2C%0A%20%20%60facility_name_overflow%60%2C%0A%20%20%60facility_street%60%2C%0A%20%20%60facility_city%60%2C%0A%20%20%60facility_state%60%2C%0A%20%20%60facility_zip_code%60%2C%0A%20%20%60facility_county%60%2C%0A%20%20%60owner_name%60%2C%0A%20%20%60owner_name_overflow%60%2C%0A%20%20%60business_type%60%2C%0A%20%20%60origional_issuance_date%60%2C%0A%20%20%60last_renewal_date%60%2C%0A%20%20%60expiration_date%60%2C%0A%20%20%60georeference%60/page/filter.
- "AirData Website File ." EPA, Environmental Protection Agency, https://aqs.epa.gov/aqsweb/airdata/download_files.html.

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

Correlation between income & crime in NJ counties within minority groups.

Project Development

Our project focused mainly on crime and income, we picked this project because we were interested to know if crime and income were correlated with each other. At the start of this project, we developed a hypothesis that stated that a higher percentage of minority groups that have a lower total median income are more likely to commit violent crimes compared to non-violent crimes. However, as our project evolved in the sense that we realize how the income and crime in some counties were relevant while in others it wasn't.

Data

- We used multiple data set's from Census, Federal Bureau of Justice and Statista
- We had to clean our data several times due to variable names, invalid values, unknown values and unnecessary data
- In the start there were two data sets that had to be combined one was a data set which included the income in the 6 counties and data set of crime rates in these counties
- We looked at 6 counties with different levels of income and minority group's
- We looked at 6 counties as looking at all county's would have been very time consuming due to the data set's being unorganized.

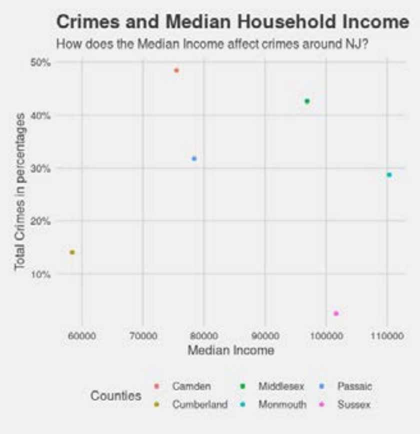
Interpretations and Recommendations

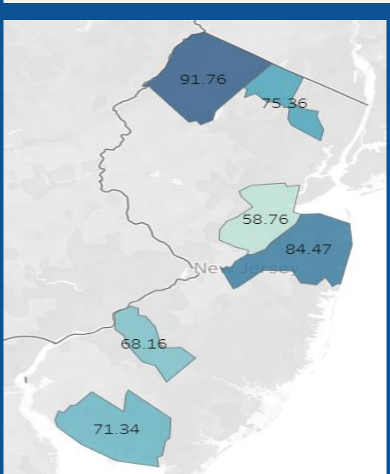
- A recommendation we can make from this project is to prioritize having police in low- income areas to deter violent crime.
- Someone could use these results to predict where it is most likely to find a form of violent crime and prevent anything tragic from happening in these areas

Jit Patel, Xzavier Aguilar, Jafet Fernandez, Juan Vega
Passaic Academy for Science and Engineering

Mentors: Connor Woods
Caldwell University

Data Visualizations and Findings





White population in percentages

From these graphs, we see how places with a higher median income (above $85,000) have a lower violent crime rates compared to the counties with lower median income which have higher violent crime rates. From the other graph, we see how the white population does have an impact on crime rates. For example, looking at Sussex County we see how the majority of the population is white which is around 91.76% and the violent crime rate is 2% and the median income is $103,000. This shows the impact of the white population on violent crime rates and having higher income results in less violent crimes.

References

- "Pittsburgh Data Works." Pittsburgh DataWorks, https://www.pghdataworks.org/resources.
- "Uniform Crime Reporting." Current Crime Data | UCR | New Jersey State Police, https://nj.gov/njsp/ucr/current-crime-data1.shtml?agree=0.
- "U.S. Census Bureau Quickfacts: New Jersey." United States Census Bureau, https://www.census.gov/quickfacts/fact/table/NJ/SBO001217.

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

Close Calls with High Cholesterol

How are behavioral factors like smoking rates and walk score and environmental factors like convenience, supermarket and fast food locations related to hyperlipidemia within Allegheny County?

Definitions

Walk Score - A score from 0 - 100 based on walkability of ZIP Code through analysis of hundreds of walking routes to amenities

Hyperlipidemia - High Blood Cholesterol

CT - Census Tract

Our Process

Brainstorming

Created Google doc of topics of interest. Each member contributed ideas. As a team, we analyzed importance + pros/cons of each and asked whether it was a worthy topic to invest time in

Data Gathering + Resources

Used mainly WPRDC datasets. Gathered datasets containing hyperlipidemia rates per ZIP Code/Census Tract. Data was gathered from Gateway Health Plan, Highmark Health, and UPMC. Separate datasets found for locations of stores and rates for smoking + walk. Chose to categorize fast food into 4 separate types: Breafast, Takeout, Dollar Menu, and No Dollar Menu.

Asking a Question + Basic Plan

Crafted both null and alternative hypotheses. Created an analysis plan. Simple organizing and linear regression were done in Excel. R Studio was used to study 2 multi-linear regression models. Took into account p values of each factor to determine statistical significance. Linear regression is displayed in scatter plots, tables of p values also created as visuals.

Data Cleaning + Organizing

Import Datasets ➡ Excel, manually cleaned and checked for invalid data. Used vlookup() in Excel to match each factor's data points with hyperlipidemia rates based upon ZIP or CT. Created Pivot tables and graphs to visualize our analysis.

Analysis

Used Tableau to create a heat map of hyperlipidemia rates. Create scatterplots of each dependent variable with independent variable (linear regression). Examined r^2 value. Then we used R-studio for multi-regression. Chose to create 2 models, categorized as behavioral factors and environmental factors. Smoking and walk score were in behavioral model, others were in environmental model. Looked for p-values <0.05 to show statiscal significance.

Hypothesis

Null: There should be no significant correlation between store locations, smoking, obesity rates, and the rate of hyperlipidemia.

Alternative: There should be a significant correlation between store locations, smoking, obesity rates, and the rate of hyperlipidemia.

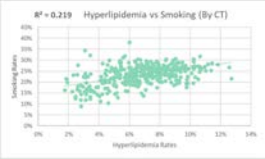


Figure 1: Example of a scatterplot of smoking and hyperlipidemia in terms of CT

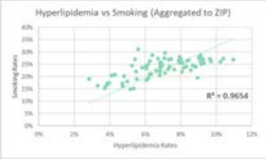


Figure 2: Example of a scatterplot of smoking and hyperlipidemia in terms of ZIP

Hyperlipidemia Rates Heat Map

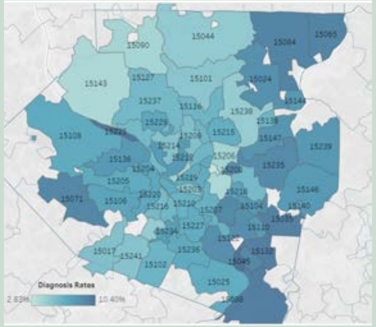
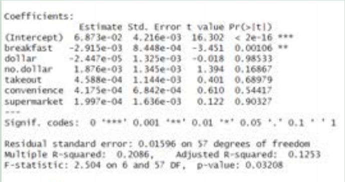


Figure 3: Heat map describing the rates of hyperlipidemia in the specified ZIP codes

Multi-Linear Regression Analysis

Environmental

Factor	P-Value	Reject or Accept Null Hypthesis
Fast Food: Dollar Menu	0.985333	Accept
Fast Food No Dollar Menu	0.16867	Accept
Fast Food: Takeout	0.68979	Accept
Fast Food: Breakfast	0.00106	Reject
Convenience Stores	0.54417	Accept
Supermarkets	0.90327	Accept



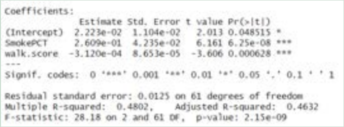
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.873e-02  4.216e-03  16.302  < 2e-16 ***
breakfast   -2.915e-03  8.448e-04  -3.451  0.00106 **
dollar      -2.447e-05  1.325e-03  -0.018  0.98533
no.dollar    1.876e-03  1.345e-03   1.394  0.16867
takeout      4.588e-04  1.144e-03   0.401  0.68979
convenience  4.175e-04  6.842e-04   0.610  0.54417
supermarket  1.997e-04  1.636e-03   0.122  0.90327
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01596 on 57 degrees of freedom
Multiple R-squared:  0.2086,	Adjusted R-squared:  0.1253
F-statistic: 2.504 on 6 and 57 DF,  p-value: 0.03208
```

Figure 4: Multi-linear regression model of environmental factors and hyperlipidemia

Behavioral

Factor	P-Value	Reject or Accept Hypothesis
Smoking	<2e-16	Reject
Walk Score	5.99e-7	Reject



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.223e-02  1.104e-02   2.013 0.048515 *
SmokePCT     2.609e-01  4.235e-02   6.161 6.25e-08 ***
walk.score  -3.120e-04  8.653e-05  -3.606 0.000628 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0125 on 61 degrees of freedom
Multiple R-squared:  0.4802,	Adjusted R-squared:  0.4632
F-statistic: 28.18 on 2 and 61 DF,  p-value: 2.15e-09
```

Figure 5: Multi-linear regression model of behavioral factors and hyperlipidemia

Conclusions

Our study provided lots of useful insight on chronic illness. Firstly, our p-values and results indicated that fast food has little correlation to hyperlipidemia, even showing a negative relationship, which was different from the general public consensus. It is interesting to note the outlying statistical significance of breakfast stores. Our project instead shows that factors like smoking and exercise have the greatest significance. These factors mainly deal with the behavioral habits we have and can improve upon, and not our surrounding environment, which are often times out of our control. In conclusion, our personal habits like smoking and exercise affect our health more than the influence of the surrounding environment, and changing those habits can lead to a decrease in chronic illness. Exposing the public to the harms of hyperlipidemia is the best path forward!

North Allegheny Team 2

Andrew Li, Ethan Hu, Lucas Pu, Matthew Guo

Observations

- **Walk Score, Smoking, and Breakfast** are considered statistically significant towards hyperlipidemia rates according to p-values
- Greatest significance on hyperlipidemia rates (descending):
  1. Behavioral: Smoking (P-value: **6.25e-08,** T-value : 6.161)
  2. Behavioral: Walk Score (0.000628, -3.606)
  3. Environmental: Breakfast (0.00106, -3.451)
- Most Surprising: General decrease in hyperlipidemia rates even as fast food store numbers increase.
- Weak correlation between fast foods stores and hyperlipidemia rates even though we originally believed it to be most influential factor
- **East & South West Allegheny County** have a greater prevalence of hyperlipidemia
- There are areas where 1 of every 10 people have hyperlipidemia
- 3 of our examined factors reject the null hypothesis, other 5 accept null hypothesis

Challenges

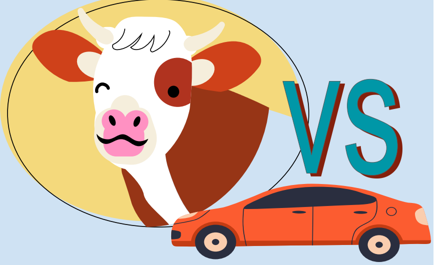
1. **Converting ZIP and CT**: Walk Score + Smoking by CT, others by ZIP. Found UPS dataset converting ZIP to the CTs.
2. **Learning R-Studio**: Previous experience in Java, C++, Python, but no experience in non-object-oriented programming like R. There was a learning curve, but guidance from our mentor helped us succeed.
3. **Creating the Best Analysis Plan**: We had to make sure that our plan created the most accurate picture of what the data can tell us. Examined both linear regression and multilinear-regression.
   a. **Smoking Data**: Not scaled to represent differing populations, only a %. In order to convert all to ZIP, needed to make assumptions. Led to idea of using two models.
4. **Understanding Statistical Figures**: Terms like p-value, multilinear-regression, t-value and others were new to most of our group. Teammate in Statistics and Mentor provided guidance

Diving Deeper

- Examining the Heat Map and data, ZIP codes with higher hyperlipidemia rates generally have fewer supermarkets
- **Reasons for breakfast significance + decline in rates**
  - More breakfast places = fewer places for other fast food types (generally more unhealthy)
  - Breakfast consumption reduces craving and hunger for unhealthier fast food during the rest of the day
- **Solution Proposal**: Increase Breakfast stores and Supermarkets in East and South West Allegheny County as both are significant towards lowering rates. Promote safety in streets of South East Allegheny as it is less safe than other regions, which in turn promotes increased walking + exercise and a greater walk score. Inform the public about the dangers of smoking and nicotine addiction, especially to the youth through the education system.

Cows vs Cars!

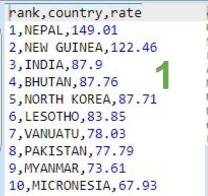
Do higher levels of greenhouse gases lead to health issues and more fatalities? If so, which type of greenhouse gas pollution is more harmful to public health: methane from **cows** or carbon emissions from **cars?**



Initial Hypothesis:
There is a relationship between the number of cows or cars in a country and the death rate.

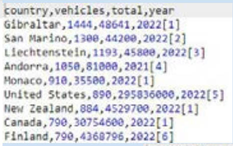
Data Sets Used
1. Death rate by country
2. Number of vehicles per country
3. Number of cattle per country

Examples of the top 10 values in each data set->



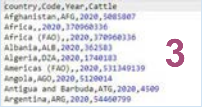
```
rank,country,rate
1,NEPAL,149.01
2,NEW GUINEA,122.46
3,INDIA,87.9
4,BHUTAN,87.76
5,NORTH KOREA,87.71
6,LESOTHO,83.85
7,VANUATU,78.03
8,PAKISTAN,77.79
9,MYANMAR,73.61
10,MICRONESIA,67.93
```

1



```
country,vehicles,total,year
Gibraltar,1444,48641,2022[1]
San Marino,1300,44200,2022[2]
Liechtenstein,1193,45800,2022[3]
Andorra,1050,81000,2021[4]
Monaco,910,35500,2022[1]
United States,890,295836000,2022[5]
New Zealand,884,4529700,2022[1]
Canada,790,30754600,2022[1]
Finland,790,4368796,2022[6]
```

2

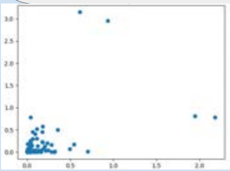


```
country,Code,Year,Cattle
Afghanistan,AFG,2020,5085807
Africa,,2020,370960336
Africa (FAO),,2020,370960336
Albania,ALB,2020,362583
Algeria,DZA,2020,1740183
Americas (FAO),,2020,531349139
Angola,AGO,2020,5120014
Antigua and Barbuda,ATG,2020,4509
Argentina,ARG,2020,54460799
```

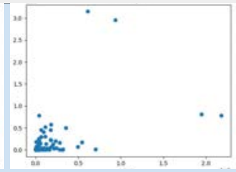
3

By: Tina Tran, Jenna Scholl, Ava Stoker-Jakab,
Dom Feden-Kist, and Naudia Booker
Keystone Oaks High School

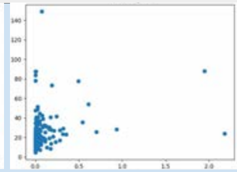
Analysis: We used the computer programs of Python and R. We did a 1 variable analysis with summary statistics such as mean, standard deviation, and a 5 number summary for each variable. We constructed box plots and histograms for each variable.
We conducted a 2 variable analysis with correlation plots and scatter plots. We used formal hypothesis testing (Linear Regression T Tests) to determine if there was a significant linear relationship between cars and deaths, cows and deaths, and cows and cars.
We found that the p value for relationship between cars and death rate was 0.69, which is greater than 0.05, so no relationship.
The P value of cows and death rate was 0.006, which is less than 0.05, indicating a relationship.
Challenges: We encountered missing data and removed them from the model. We had to clean the data, removing commas and quotations. We had to merge 3 different data frames. We acknowledge that factors such as covid 19, alternate types of pollution, and lack of medical resources available to certain countries could impact the data.



X Axis: Cars
Y Axis: death rate
P value:

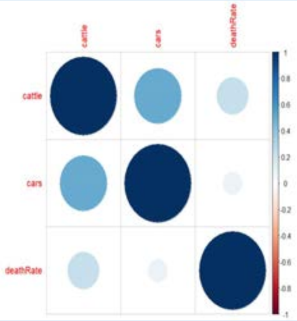


X Axis:
Y Axis:
P value:



X Axis:
Y Axis:

Cattle and cars have a moderately positive correlation
Cattle and death rate have a moderately positive correlation
Death rate and cars have a weak positive correlation



Results: There is convincing statistical evidence to suggest that there is a relationship between number of cows in a particular country and the death rate in that country. We found no relationship between cars and death rate.