How does local crime intensity correlate with factors that demonstrate the quality of life in Pittsburgh neighborhoods? What can these correlations tell us about an area?

Avonworth High School

Luc Azen ~ Charlie Bozada ~ Alivia Wright ~ Addison Dexter ~ Catrina Raich ~ Laurel Purcell ~ Mike Frank

RESOURCES

Western Pennsylvania Regional Data Center

United States Census Bureau

Crime by area Data Sets

COVID-19 rates Data Sets

DEFINITIONS

Regression is the measure of relation between mean value of one variable and corresponding values of a second variable

A residual is a data point showing measure between mean value and the corresponding value

A tract is a way of arranging data values in a broader form; tracts include larger sectioning than neighborhoods

A heatmap is a data visualization technique using a range of colors to define intensity

ASSUMPTIONS

Local crime intensity correlates with other factors defining quality of life, such as median income, unemployment rates, and COVID-19 rates.

Lower median income areas have more crimes committed.

CHALLENGES

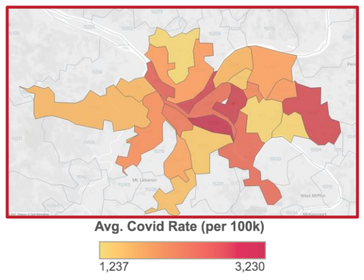
Relevant and recent sources

Correlation / not establishing causation

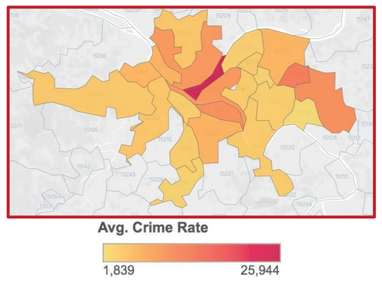
Dataset compatibility

Programming

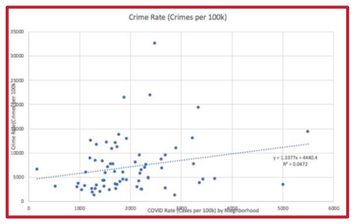
SUMMARY OF RESULTS



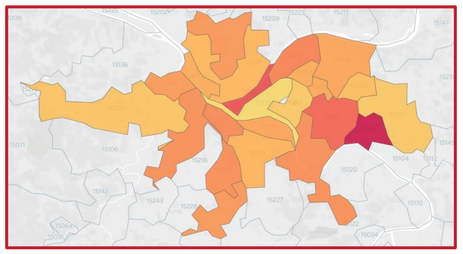
This heatmap shows the rate of COVID-19 infections by ZIP code. Darker shades correspond with higher COVID-19 rates.

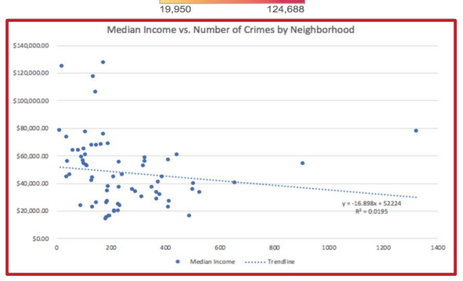


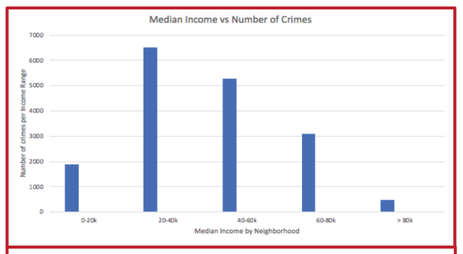
This heatmap shows the crime rates for each ZIP code in the Pittsburgh area. Darker shades correspond with higher crime rates.



This data plot shows the crime rates on the y-axis, and COVID-19 rates on the x-axis. The r value is about 0.217.

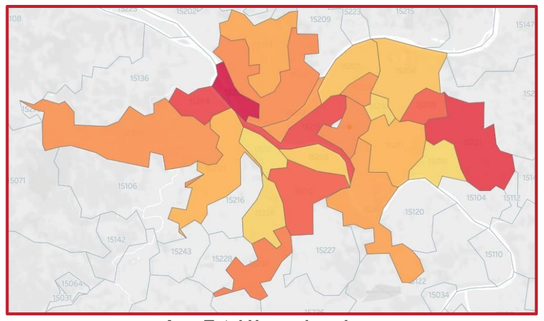


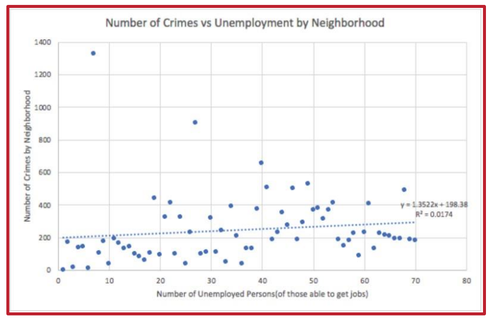




	0-20k	20-40k	40-60k	60-80k	> 80k	Totals
Crimes	1896	6510	5288	3082	471	17247
Neighborhoods	8	22	21	14	4	69
Percent of Crime by "Income Bucket"	11%	38%	31%	18%	3%	100%

This heat map shows median incomes of each Pittsburgh ZIP code. The darker shades correspond with higher income levels. The data plot shows the median income levels on the y-axis, and crime rates on the x-axis. The r value is about 0.1396. The bar graph shows the number of crimes for each income bucket, and the table shows clearer data values.





This heatmap shows the rates of unemployment by Pittsburgh ZIP codes. Darker shades correspond with higher COVID-19 rates. The data plot shows the number of unemployed people in each ZIP code on the x-axis, and the amount of crimes by each Pittsburgh neighborhood on the y-axis. The r value for this graph is about 0.1319.

CONCLUSION

COVID-19 rates and unemployment rates had relative regression values, so we are unable to state there is any relation present. In regards to crime and median income data, there is no disparity present regardless of median income, and a high regression value is present. About 50% of crimes happen in areas with a median income of 0-40k, and the other 50% happen in areas with a median income of 40k-80k. This was contrary to our expectations going into the analysis portion of the project. We felt that crime would be more prevalent in neighborhoods with high unemployment rates and low income median incomes, but this was not the result. The results found could not substantiate the argument that crime rates had any correlation to several factors relating to the quality of life.

Pittsburgh's Communities: Status of Pools Versus Crime

Analyzed By: Desmond Corrado, Martial Delrosario, Abigail McClain, & Reece Smith, of Carlynton Jr./Sr. High School

Research Question: What is the relationship between pool status in Pittsburgh neighborhoods and crime rate?

Obtaining the Data:

We obtained data from the Pittsburgh police blotter data. The data contains information about all police reports in the Pittsburgh area, including the date, time, address, and community. We also obtained the location of City Parks from the City website.

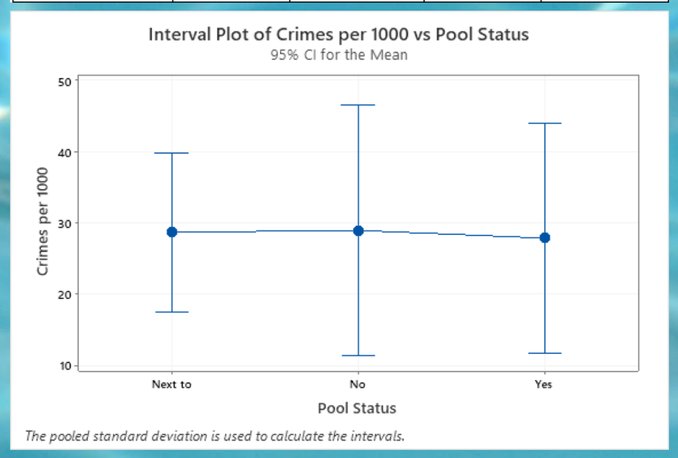
Challenges:

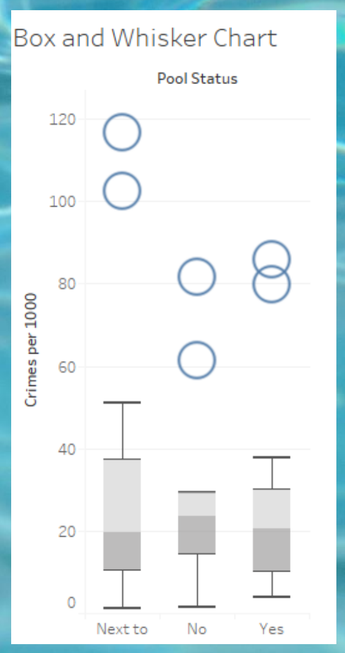
1. Police data spanned over several years
2. Police data consisted of all reports rather than solely criminal ones
3. Some neighborhoods of the Pittsburgh area aren't residential areas

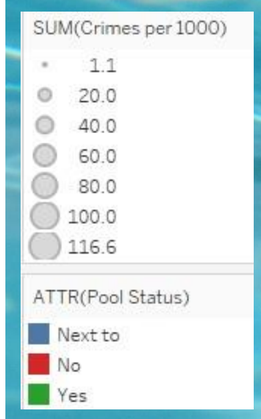
Solutions:

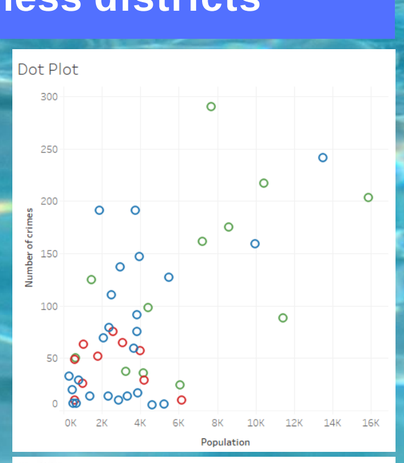
1. We analyzed only the data from the summer months of 2019
2. We excluded any non-criminal offense, as well as any crime committed only by adults
3. We excluded neighborhoods from the Point and others, which are primarily business districts

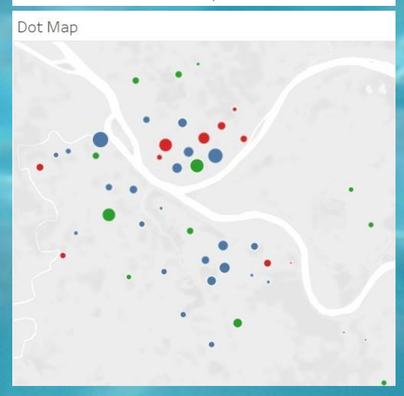
Neighborhood	Pool Status	Number of crimes	Population	Crimes Per 1000
Perry South	Next to	147	3949	37.2
Marshall-Shadeland	Next to	127	5480	23.2
Arlington	No	52	1785	29.1
Fairywood	No	26	1002	25.9
Perry North	Yes	98	4400	22.3
Carrick	Yes	290	7698	37.7











Summarization:

From the data, we can conclude that there is no significant difference between the status of pools in communities in conjunction to the crime rate. We reached this conclusion from multiple analyses, from an ANOVA test, the F-value was .01 and the P-value was .995, as well, the confidence intervals of the means of the different pool status and the box and whisker charts by pool status largely overlap.

Factors in Car Accidents Involving Drivers Age 16-18

Hampton High School

Addison Gindlesperger, Kiana Kazemi, Lindsay Liebro,
Eileen Lin, Abigail Pursh, Becky Zhou

Introduction

According to the CDC, teenage drivers are most at risk for motor vehicle accidents (Teen drivers, 2020), and this is true in Allegheny County. It is alarming that there are so many accidents, because drivers under 18 must have 6 months of practice driving with a permit. Teenage drivers can choose to take a driver's training course, but it is not mandatory. Because so many teenage drivers in Allegheny County are getting into accidents, it is important to understand what is causing these accidents. For this reason, our research question is:

- *How frequently are car accidents involving 16-18 year old drivers a result of the time of day, type of intersection, weather conditions, and road illumination in the city of Pittsburgh?*

This research question was chosen to investigate why so many teen drivers are involved in car accidents, with the goal of finding ways to prepare drivers to reduce accidents. We chose to use data from Allegheny County Crash Data: Cumulative Crash Data (2004-2019). We focused on data for the the time of day, type of intersection, weather conditions, and road illumination in the city of Pittsburgh that occurred in accidents involving teenage drivers. We chose to focus on these factors because they are things that a driver cannot control. Road conditions, weather, and lighting of roads cannot be changed by the driver. While drivers can control, within reason, the time of day that they choose to drive, they cannot control the amount of traffic on the road during certain times of the day. They may also have to drive at certain times of the day due to obligations or appointments. By determining how often these factors are causing accidents, parents and drivers' training instructors can focus their attention when teaching 16-18 year old drivers. This can help teen drivers to learn to be cautious of these conditions that frequently invoke crashes. With better awareness of and practice in handling these factors, they may avoid future collisions, saving money, legal trouble, and even lives.

Methodology

Data was entered into Excel and filtered by ages 16-18, with a total of 16,643 accidents. This represents the sample of our analysis (n = 16,643). It was then further filtered by 6 factors: weather conditions, the number of drivers aged 16-18 involved, road conditions, time of day, the number of accidents per year for the sample, and road illumination.

To analyze each individual factor, each team member determined frequencies for each factor. Trends in these data sets were observed and reported. In some instances, data for drivers age 16-18 was compared to rest of the drivers in the full data set to determine if trends were unique to the sample (16-18 year old drivers). Graphic representations of the data (pie chart, bar graph, frequency table) were used to show significant trends in these conditions.

Using these trends and additional sources of information, we determined changes that need to be made to driver's training programs to help reduce the number of accidents involving teenage drivers in Allegheny County.

Limitations

As this was a publicly available data set, we determined several challenges or limitations in our analysis. Some of the data entries were incomplete or not recorded, which was a challenge since the data was collected by responding officers or other third party reporters. In addition, drivers in the 21-49 year old age range were not included in this data set. As such, it was difficult to truly compare the trends for our sample (16-18 year old drivers) to the rest of the population of drivers.

There were some unexpected trends that arose in the analysis, which presented some challenges in making recommendations to solve the problem. For example, the most frequent weather condition in accidents involving teenage drivers was blowing sand, dirt, and soil. It was difficult to explain, from just this data set, why so many teenage drivers encountered this type of weather condition in Allegheny County.

To help address these challenges, we could use inferential statistics for a more detailed view of all of the factors that can cause accidents. Alternatively, we could still use the data from Allegheny County, but use other sources of data to fill in data for drivers age 21-49.

Results

Although there were 16,643 accidents involving drivers age 16-18, there were 16,931 drivers involved in these accidents age 16-18, which means some accidents involved more than 1teenage driver. Most accidents involved a single driver. Drivers in the 18 year old group were most frequently involved in accidents (*Table 1).* Data shows that the total number of accidents per year for drivers age 16-18 was highest in 2004 (n= 1,492, 8.96%) and lowest in 2019 (n= 775, 4.66%), indicating that the number of accidents has generally decreased over the years of data in this study *(Figure 1).* This could indicate that the teen safety driver campaign implemented in 2007 and the change to licensing requirements for drivers under the age of 18, passed in 2011, are helping (DMV, 2011).

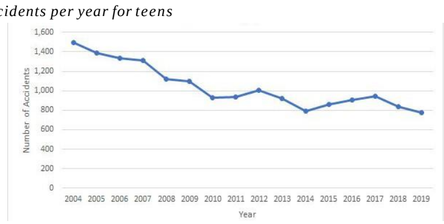
Data shows that Blowing Sand, Soil, and Dirt accounted for 73.56% (n= 12,243) of accidents. Since this study was done in Allegheny County, a place where such occurrences are not common, this was unexpected. Snow only accounted for 30% of accidents (n= 50) and sleet/hail accounted for 0.14% (n= 24). Typically, one would think that snow, sleet, and hail would be weather conditions that were frequently seen in conjunction with accidents. (*Figure 2).* Data also indicated that drivers age 16-18 most frequently have accidents on dry road conditions (n= 11080, 66.57%), which could indicate drivers are more confident and less focused in dry conditions. Standing/moving water on the road was least frequent (*Figure 3*). Most frequently, accidents occur between the hours of 2 pm to 6 pm, with 4pm being the most common hour (n=1,462, 8.79%) (*Figure 4).* It is possible this results from teenage drivers leaving school and going to activities or home. There were still a great deal of accidents following 11 PM (the assumed curfew for teenage drivers). This could potentially be 18 year old drivers, or 16 and 17 year old drivers breaking the law. There was an outlier at 7 AM of accidents (n= 861, 5.17%), which may be attributed to times when students drive to school. Data also indicated that Daylight is the most common illumination type (n= 10,455, 62.82%) for accidents involving drivers age 16-18, with Dark/Street Lights as the second most frequent (*Figure 5*). Based on these results, there are some areas in which drivers age 16-18 most likely need more training and practice before getting their license in order to reduce the number of accidents involving this age group of drivers.

Table 1.
Number of drivers age 16-18 involved in car accidents

Age	1 Driver	2 Drivers	3+ Drivers	Total
16	2,545 (15.03%)	31 (0.18%)	0	2,573 (15.20%)
17	6,356 (37.52%)	9 (0.05%)	18 (0.11%)	6,383 (37.70%)
18	7,839 (46.30%)	124 (0.73%)	9 (0.05%)	7,972 (47.09%)

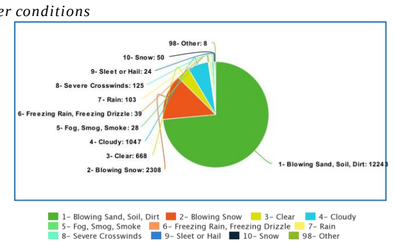
Note. The total number of accidents involving 16-18 year old drivers was n= 16,643, but there were 16,931 drivers in these accidents.

Figure 1.
Accidents per year for teens



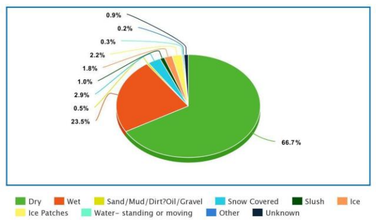
Note. This figure shows the general decrease in accidents for drivers age 16-18 from 2004-2019. When compared with the data for other aged drivers, this same trend was not observed.

Figure 2.
Weather conditions



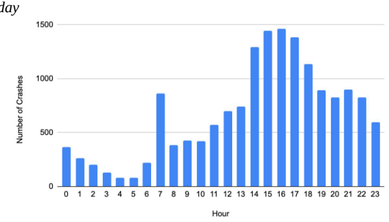
Note. Of the 16,643 accidents involving drivers age 16-18, for this factor only 16,619 accidents had this data reported.

Figure 3.
Road conditions



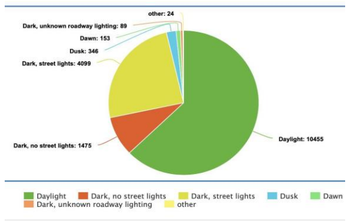
Note. There were 40 accidents in the data set that did not have this factor reported. However, frequencies were calculated with n= 16,643.

Figure 4.
Time of day



Note. In the data set, 427 accidents had the time of day recorded as "unknown". These are not shown in the figure, but they accounted for 2.57% of accidents from the 16,643 reported.

Figure 5.
Road illumination



Note. Of the 16,643 accidents of drivers 16-18, 26 accidents for this data set do not have data reported for this factor.

Conclusions & Recommendations

Current driver's education programs do require instructors to give classroom (theoretical) training on hazardous conditions and their effects, along with the challenges of night driving and appropriate responses (Pennsylvania Department of Education, 2003). While many schools do not offer this course now, private driver training programs are also an option. These programs must offer classroom training on managing adverse conditions (Driving School Association of the Americas, 2017). There is not, however, a requirement of certain types of adverse conditions that must be simulated in road practice. Considering this, and the observations from our data analysis, we recommend that driver training programs include:

- Students practice more during times of heavier traffic conditions
- Increased number of on the road hours with a certified driving instructor to improve proper attitude, focus, and safety in favorable conditions
- Require more on the road experience with different weather and road conditions

We also recommend an increase in the number of public education campaigns to raise awareness of driver safety, targeted at teenage drivers, such as National Teen Driver Safety Week in 2007 (Announcement: Teen Driver, 2007). The data showed a drop in accidents among teens in 2008, leading us to believe that a public awareness campaign could help address this problem.

It would also be beneficial to require a driver training program/education to qualify for driver's license. We suggest that Allegheny County try to create/provide incentive for 18 year old drivers, who can bypass the permit process, to complete a driver's education program.

Our final recommendation is for Allegheny County to develop a simulation to practice driving in different road conditions, lighting, and weather conditions to practice responses in accident situations in a safe environment. We believe all of these recommendations could help reduce the number of accidents involving drivers aged 16-18 and address this problem in Allegheny County and the city of Pittsburgh.

References

Allegheny County Cumulative Crash Data (2004-2019). (2020). Pennsylvania Regional Data Center. https://data.wprdc.org/datastore/dump/2c13021f-74a9-4289-a1e5-fe0472c89881

Announcement: National Teen Driver Safety Week --- October 18--24, 2009. (n.d.). Retrieved March 6, 2021, from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5840a3.htm

Driving School Association of the Americas (2017). *Beginning driver education and training: Curriculum content standards*. https://dsaa.org/files/galleries/DSAA_Standards_Revision_January_2017.pdf

DMV. (2011). *New Teen Driver Law*. PennDOT Driver & Vehicle Services. https://www.dmv.pa.gov/ONLINE-SERVICES/Pages/New-Teen-Driver-Law.aspx

Pennsylvania Department of Education. (2003). *Content and performance expectations for driver education*. https://www.education.pa.gov/Documents/Teachers-Administrators/Curriculum/Driver%20and%20Safety%20Education/Content%20and%20Performance%20Expectations%20for%20Driver%20Education.pdf

Teen drivers: Get the facts. (2020). Retrieved April 08, 2021, from https://www.cdc.gov/transportationsafety/teen_drivers/teendrivers_factsheet.html

“Tweet”le Dee or “Tweet”le Dumped?

AJ Manges, Cade Moffatt, Josh Knapp, Grayden Jackson, Cole Marshall,
Brock Martindale, Jake Bonato, Ronan Dicsare, and Nathan Knause

Does an increase in social media users have an effect on marriage, birth, and divorce rates?

Social media platforms have become a common and preferred method for people to communicate. By using technology to talk to one another, what effect is this having on interpersonal relationships? Could social media usage disrupt or enhance modern-day relationships as it relates to birthrates, marriage rates, and divorce rates? The number of social media users is at an all time high. People young and old document their lives through snapshots and selfies. Social gatherings, social achievements, and interpersonal relationships tend to be the focus of these posts. As a result of interacting online there could be affects on real life relationships. In our project we tried to find correlations between the number of users and marriage, divorce, and birth rates. Are we in danger of only interacting online? Are we more worried about likes on posts, as opposed to finding a partner to spend their lives together? Is the need for online interactions ruining current marriages or pushing couples to avoid becoming parents? Maybe the opposite is true. We have tried to see if specific social media platforms contribute to changes in marriage, divorce, and birth rates.

Hypothesis

Increased usage on social media platforms correlates with a decline in the number of marriages, an increase in divorces and a decrease in birth rates between 2008-2018.

Analysis

Results

r values:

Twitter v. Birthrates: -.784 Twitter v. Divorce: -.893 Twitter v. Marriage: -.167
Snapchat v. Birthrates: -.941 Snapchat v. Divorce; -.935 Snapchat v. Marriage: -.499
Instagram v. Birthrates: -.905 Instagram v. Divorce: -.991 Instagram v. Marriage: -.425
Facebook v. Birthrates:-.735 Facebook v. Divorce: -.843 Facebook v. Marriage: -.170

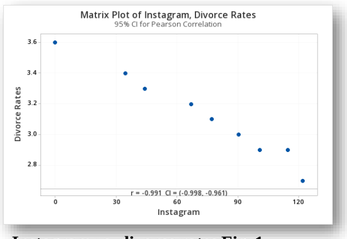
- All social media platforms show a strong correlation with divorce rates, as the number of users increased, divorces decreased. Instagram appears to be the greatest predictor of divorce rates.
- The two variables that had the weakest relationship were Twitter and marriage rates
- The two variables with the strongest relationship were Snapchat and marriage rates with a correlation coefficient of -.499 (Fig. 5)
- The two variables with the strongest relationship were Snapchat and birth rates with the highest correlation coefficient of -0.941. (Fig. 4)
- The two variables with the strongest relationship were Instagram and divorce rates with a correlation coefficient of -0.991. (Fig. 1)
- Multiple regression analysis of all platforms gave p values of:
  - .363 – birth rates
  - .980 – marriage
  - .329 divorce rates (Fig. 6)

- Using Analysis Of Variance, we see that the social media platforms are the best predictors for divorce rates

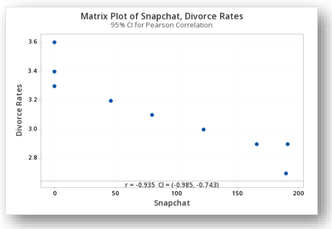
General Conclusions Based on r value:

Birth rates and social media have strong negative correlation. As social media increases, birthrates decrease. Divorce rates and social media have strong negative correlation. As social media increases, divorce rates decrease. Marriage rates and social media have no correlation. The amount of social media users has no effect on marriage rates.

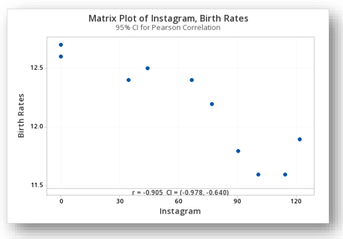
Visuals



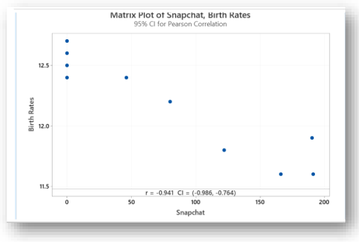
Instagram vs. divorce rates Fig. 1



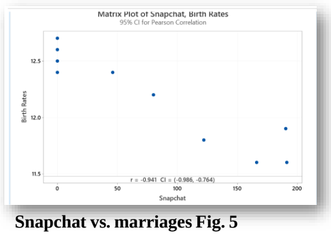
Snapchat vs divorce rates Fig. 2



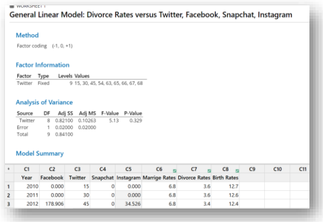
Instagram vs. birth rates Fig. 3



Snapchat vs. birth rates Fig. 4



Snapchat vs. marriages Fig. 5



Minitab Presentation and breakdown of P value for multiple regression of divorce rates vs. all social media. Fig. 6

Observations

- People are connecting, talking, and dating all through social media around the world 24 hours a day and 7 days a week. This leads to an affect on how many people are having children, getting married, or getting divorced.
- Instagram, Snapchat, Facebook, and Twitter have the strongest correlation with declining birth and divorce rates among people in the United States. We can see that these applications have an impact on people not having children and not getting divorced.
- Instagram, Snapchat, Facebook, and Twitter have no correlation with declining marriages rates among people in the United States.
- Thousands of people are spending more time on Instagram, Snapchat, Facebook, and Twitter than ever before, so people are not interacting face to face as much as they used to.

Challenges

- Trouble finding data for Facebook, Twitter, and Snapchat when they were not publicly traded.
- We had to adjust our populations to US adults only to match our percentages.
- We had trouble finding data isolated to the US
- There was a lot of international data and data that grouped the US and Canada
- We were struggling to figure out how we wanted to have a longer stream of data with social media only being available for a limited numbers of year
- Only a certain number of students had computers that could run Minitab allowing for less people to work on scatter plots finding a constant stream of CDC data that wasn’t divided and had gap years

Sources

- Statista.com
- Pewresarch.org
- Tradingeconomics.com
- Businessofapps.com
- Datacenter.kidscount.org
- Finance.yahoo.com
- Kff.org
- Cdc.gov

Conclusion

Our initial hypothesis was centered around social media and the effects it would have on all aspects of relationships. We tried to find correlations between the amount of US users on social media and marriage, divorce, and birth rates. First, we used the amount of users on Snapchat, Facebook, Instagram, and Twitter; and found a high correlation between social media usage on all sights with birth rates and divorce rates using r values. On the other hand there was little to no correlation with social media usage to marriage rates. We also would recommend further analysis to find if other variables such as rising student loan debt, more people attending college, later marriages, and a mixture of other variables would be affecting marriages initially. However, our data suggests that social media has had a significant effect on the birth and divorce rates that stem from marriage. One of the only drawbacks from our initial hypothesis was that we failed to recognize that if marriages decrease, then divorces would decrease as well. Overall, our data shows that while there may be other variables, we can’t say that relationships and social media are a post hoc fallacy. In fact we can conclude that social media sites, specifically Snapchat, are a good predictor of divorce and birth rates.

What demographics in the Pittsburgh area can influence standardized test scores other than academic intelligence?

North Allegheny School District
Anjali Bandi, Alina Zaidi, Sonya Dhussa, Helen Mao

Hypothesis:

Our hypothesis is that a higher income will affect certain races and regions to influence a school's higher average SAT score than others. We also hypothesized that the percentage of boys in a school had a positive correlation with SAT score.

Challenges:

- Selecting data sets that were specific to the Pittsburgh region
- Pulling out data specific to some high schools
- Trying to find income averages in a particular school/region
- Combining and analyzing collinearity

Observations:

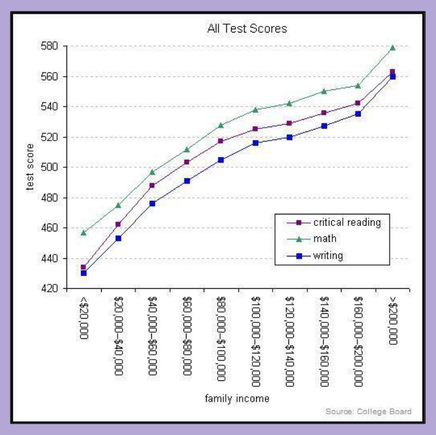
- The data shown are regional statistics and show a clear relationship between higher incomes and higher test scores.
- There are clear distinctions between areas that tend to have a higher median household income than others. This higher income is often used to its advantage, as it provides some students with more opportunities to learn and excel when it comes to standardized testing.
- White people tend to have the highest income in Pittsburgh, and the second highest income in all of Pennsylvania
- Areas with a higher median household income are more likely to have larger 504 plan percentages.

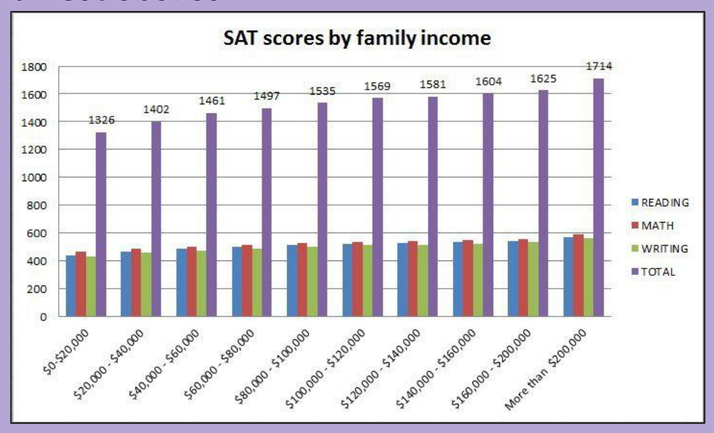
Sources:

- https://www.wsj.com/articles/many-more-students-especially-the-affluent-get-extra-time-to-take-the-sat-11558450347
- https://statisticalatlas.com/place/Pennsylvania/Pittsburgh/Household-Income

Visualizations

Income and Test Scores



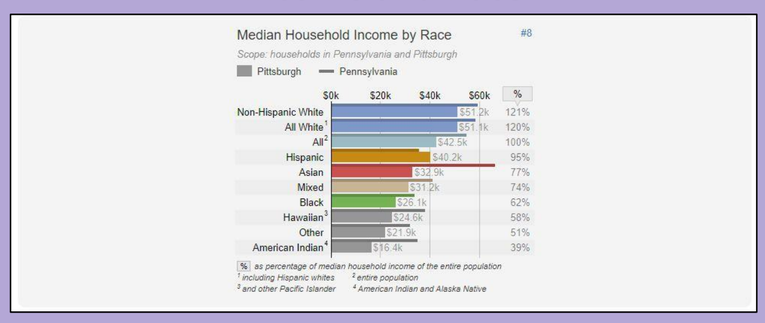


Correlation Matrix

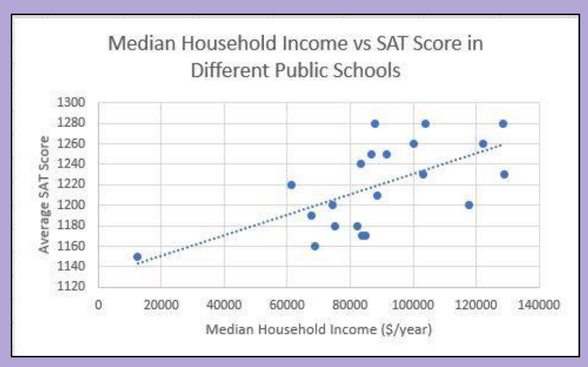
Correlations

	Median Household Income ($/year	Average SAT
Average SAT	0.624	
% boys	-0.242	0.094
Asian	0.299	0.647
Black2	-0.693	-0.474
White	0.618	0.321

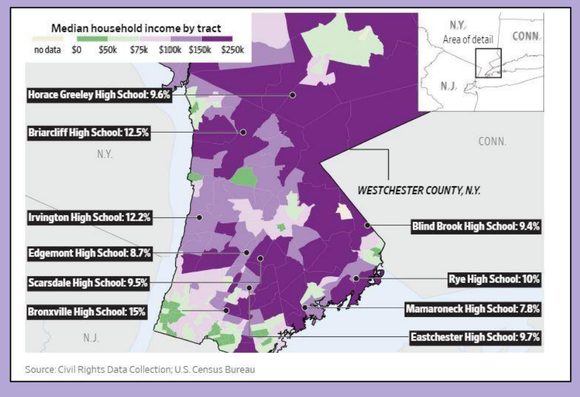
Income and Race

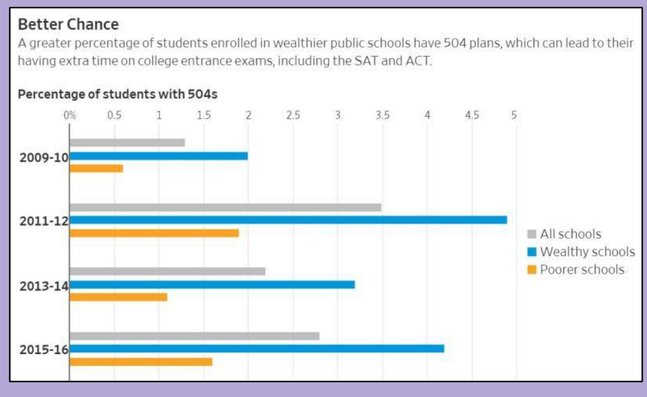


Region of Residence



Connection to Disabilities





- Our hypothesis about higher SAT scores from higher incomes and certain races was correct
- However, our hypothesis about the percentage of boys causing a higher SAT score was not supported strongly because of such a weak correlation

Summary:

- It is clear that there is a strong correlation between income and academic test scores
- However, income plays a role in many other demographic factors like race, region, and school district

Analyzing the Legacy of Redlining in Pittsburgh

Is there a correlation between redlining loan grades and modern median income?

Norwin High School

Arnav Bedekar • Aaron Berger • Dmitri Berger • Lydia Berger • Abrielle Brown • Simone Pal • Alexander Puskaric

Background:

- Redlining was used by banks in the early 20th century
- They distributed loans based on **racial and ethnic demographics**, with the intention to harm minority groups

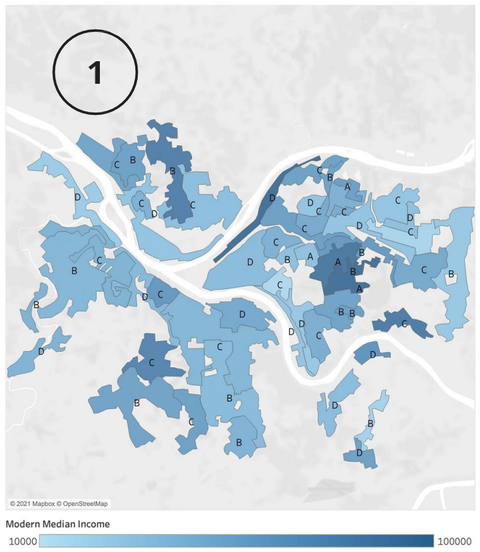
Dataset Descriptions:

Census Median Income Data: An expansive and interactive map of census tracts in Pittsburgh, with census data including **median income** by census tract.

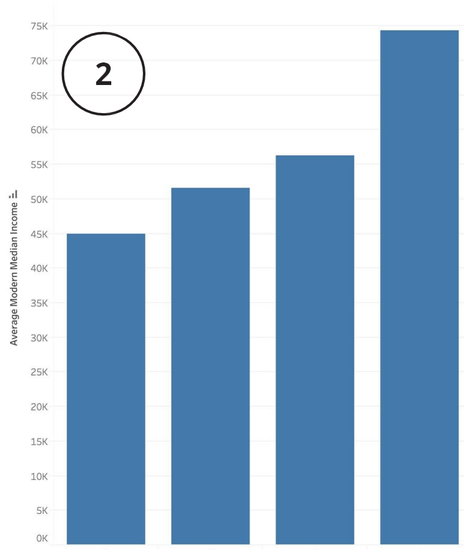
Census Tracts: Maps showing **Pittsburgh census tracts** from the 2010 census.

Redlining Data: A 1937 redlining map of Pittsburgh. It divides the city into redlining tracts, color-coded by redlining grades from A through D where A was 'most desirable' and D was 'least desirable'. Redlining loan grades were determined by the **demographics of a tract**.

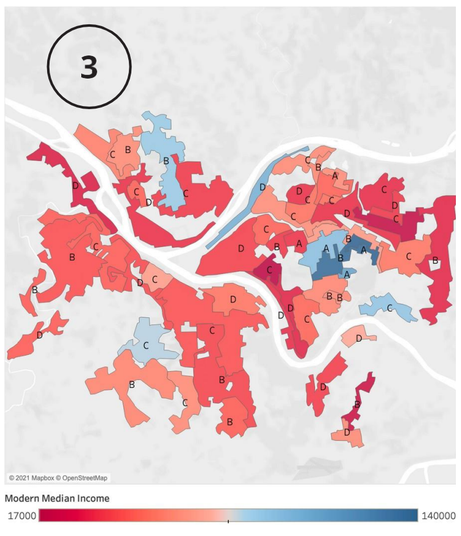
1937 Redlining Tracts by Modern Median Income



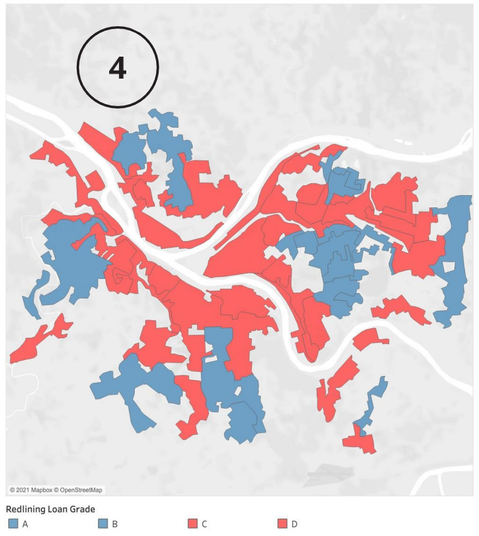
Average Modern Median Income vs 1937 Redlining Grade



Redlining Tracts by Modern Median Income Over $65,000



1937 Redlining Tracts in Category A-B vs Category C-D



Sources:

Census profile: Census Tract 305, Allegheny, PA. Census Reporter. (2019). https://censusreporter.org/profiles/14000US42003030500-census-tract-305-allegheny-pa/.

Census Tracts:
Index of /geo/maps/dc10map/tract/st42_pa/c42003_allegheny. (2011, March 10). https://www2.census.gov/geo/maps/dc10map/tract/st42_pa/c42003_allegheny/.

Redlining Data: Nelson, R. K., Winling, L. D., Marciano, R., & Connolly, N. (n.d.). Mapping Inequality. https://dsl.richmond.edu/panorama/redlining/#loc=11/40.442/-80.149&city=pittsburgh-pa.

Our Challenges:

- Adding geospatial census tract data to Tableau, because our instructions were outdated
- Creating our own conversions between MMI per census tract and MMI per redlining tract, since there is no standardized conversion
- We originally planned to incorporate median income data from 1937, but we couldn't find clear, usable data on it
- Manually processing and organizing census data on median income

Summary:

We found a **distinct correlation** between redlining tract grades and modern median income (MMI):

- Visualization 2 - as the redlining grade increased from D to A, the MMI rose significantly
- Visualization 3 - redlining tracts in grades A and B are more likely to have a corresponding MMI greater than $65,000.

We recommend that people living in low-income, formerly redlined areas be eligible for grants to help with housing affordability.

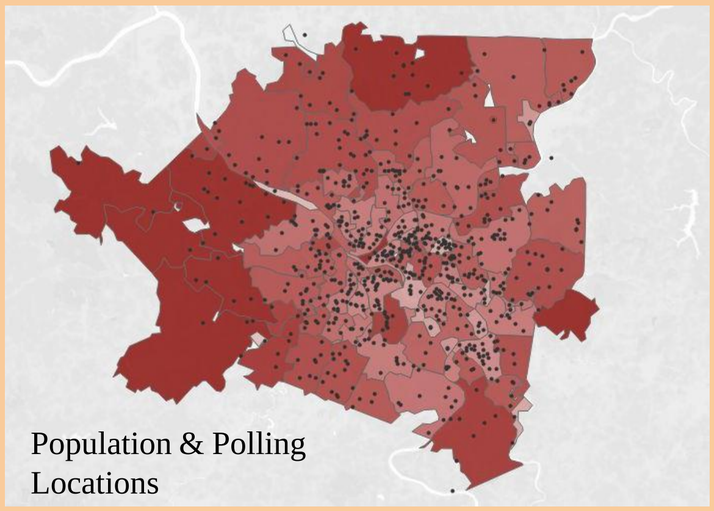
Polling Location Accessibility in Allegheny County by Median Income, Race, Population, and Political Party Demographics

Oakland Catholic High School
Olivia Marangoni, Angela McKinzie, Róisín Tsang, Yolanda Yang

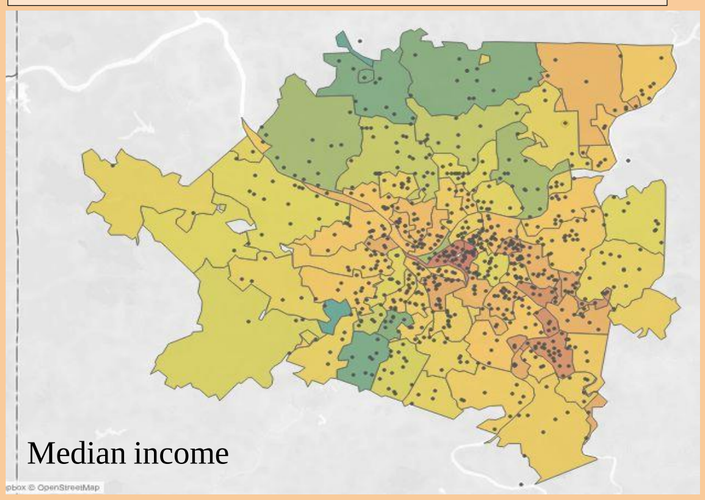
Problem

Does median household income, race, and political party have correlations with the number of polling locations per zip code? How do demographics correlate with voter-accessibility in Allegheny County?

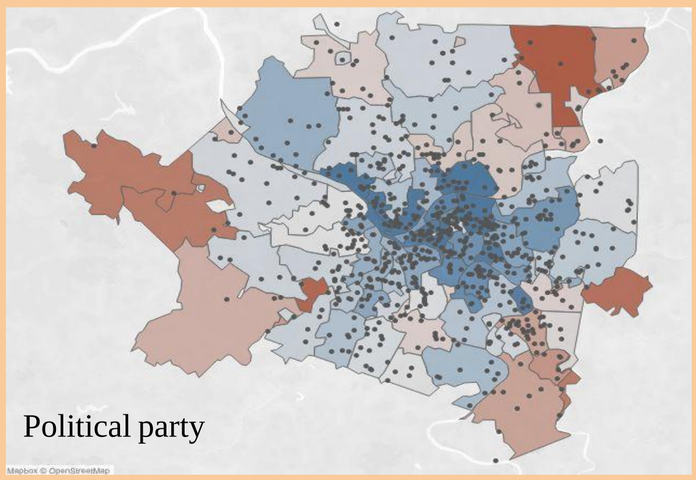
- With the demographic variables of race, political party, and income, we examined the possible relationship between them and the number of polling locations there are per zip code. We also used the population of each zip code as another variable to measure it.
- We measured voter-accessibility as: ease of eligible voters to vote, and how many polling locations are in a district compared to population of said district.



This map shows population divided by the number of polling locations. The darker red regions are areas where the deficit between population and polling locations is greater. As the red becomes lighter, it shows a smaller deficit between the two variables.



The color scale goes from red to green→ green being the wealthiest while red is the poorest

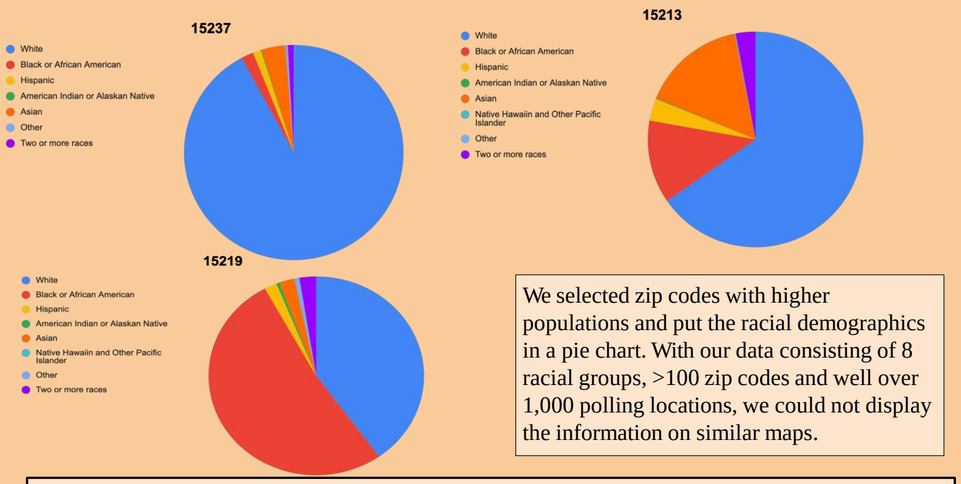


The color scale goes from red to blue→ if a region is more blue, then it leans Democratic and if it is more red, then it leans Republican

Importance

On a national scale, voter suppression represents a big issue that affects the notion of democracy in the U.S. Since the mid 2000's, more states have seen a trend towards more restrictive voter access policies, including photo verification, difficulty and repeals on voter registration, restricting convicted felons from participating in the voting process, and a limited number of polling locations[1]. We were also inspired by the 2020 election and talks of voter fraud and suppression in the U.S. during the pandemic.

Race and Select Zip CodePie Charts



We selected zip codes with higher populations and put the racial demographics in a pie chart. With our data consisting of 8 racial groups, >100 zip codes and well over 1,000 polling locations, we could not display the information on similar maps.

Challenges

→ The scale of the data was an issue because we wanted to use municipalities as the location, but some municipalities in Allegheny County had the two or more names for the same location. As a result, we switched to zip codes-based data
→ Demographic information specific to some zip codes was hard to find because some were P.O. boxes or the population was not large enough to collect data on
→ Finding data-by-zip code for political party was a problem because most political voting data went by district number
→ Displaying the demographics vs. polling locations proved difficult because political party and race had more than one category within the variable

Summary and Potential Actions

→ In conclusion, we found that there was a correlation between population and polling locations in Allegheny County. The outer regions of Allegheny County are less voter accessible and are harder to vote in, despite the fact that they are more populated. The prioritization of urban over suburban or rural regions limits the eligible voters in those regions, so there should be more polling locations placed there.
→ With the political party demographic, the location of polls does not favor one party over another. While the Democratic regions have "more" polling locations, it is not out of an intent to suppress the Republican vote; rather, this data simply reflects the how Allegheny County votes.
→ When looking at the race demographic, there was no consistent trend or correlation with polling locations. As a result of gentrification, minority groups were pushed closer to the city of Pittsburgh, which is where most polling locations are. The broad nature of our data does not look specifically at a possible correlation between race and polling locations

Datasets

Each variable required a separate dataset and source.
Population was found through US Census data.
Race was found through Zipdatamaps. Political party was found through an interactive map curated by the *New York Times*. Median income was found through Income by Zip Code. Polling locations was found through Western Pennsylvania Regional Data Center (WPRDC).

[1] https://abcnews.go.com/Politics/timeline-voter-suppression-us-civil-war-today/story?id=72248473 (ABC News)

Charge off-rates and Unemployment: The Stimulus Effect

Peters Township High School

Victor Yu, Maya Nagiub, Jackson Busche, James Wang, Sheng Wang, Larry Lu

How does unemployment and delinquency rates affect how much people default on their loans in the United States?

Definitions

Charge Off Rates -> rates that are used by banks to show how many loans are charged off, meaning they are not likely to be paid back.

Delinquency Rates -> rates that are used by banks to show when a loan is considered late or overdue

Unemployment Rates -> number of unemployed people (those out of work and seeking a job) divided by total labor force times 100

Resources

Our main sources of data was the **Federal Reserve**, which allowed us to retrieve data regarding the charge-off rates and the delinquency rates of commercial banks in the United States. The **U.S. Bureau of Labor Statistics** provided data regarding the civilian national unemployment.

When observing the effect of Covid-19 and the following stimulus check on our data sets, we collected additional data from the **FRED**, or Federal Reserve Economic Data.

Data

Year/Quarter	Unemployment	Delinquency	Charge off All
2000 Q4	3.9	1.97	2.63
2001 Q1	4.3	2.06	2.41
2001 Q2	5.7	2.16	2.6
2001 Q3	5	2.15	2.78

Sample data Charge-off, Unemployment, Delinquency by Quarter

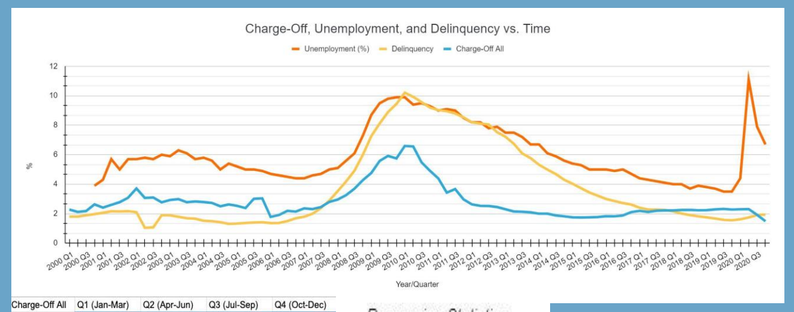
2008 Q4	7.3	6.00	4.28
2009 Q1	8.7	7.26	4.76
2009 Q2	9.5	8.12	5.58
2009 Q3	9.8	8.90	5.92

2008 Recession

2019 Q4	3.5	1.55	2.28
2020 Q1	4.4	1.62	2.3
2020 Q2	11.1	1.74	2.31
2020 Q3	7.9	1.92	1.92
2020 Q4	6.7	1.94	1.49

Coronavirus

Results



Charge-Off All	Q1 (Jan-Mar)	Q2 (Apr-Jun)	Q3 (Jul-Sep)	Q4 (Oct-Dec)
2000	2.28	2.12	2.18	2.63
2001	2.41	2.6	2.78	3.08
2002	3.71	3.07	3.1	2.77
2003	2.93	2.99	2.78	2.83
2004	2.79	2.73	2.5	2.63
2005	2.53	2.38	3.01	3.04
2006	1.78	1.91	2.19	2.14
2007	2.36	2.31	2.45	2.8
2008	2.96	3.26	3.7	4.28
2009	4.76	5.58	5.92	5.75
2010	6.6	6.56	5.48	4.91
2011	4.39	3.43	3.67	2.97
2012	2.63	2.53	2.52	2.45
2013	2.3	2.15	2.13	2.09
2014	2	2	1.88	1.82
2015	1.75	1.74	1.75	1.77
2016	1.83	1.82	1.88	2.11
2017	2.19	2.12	2.2	2.22
2018	2.22	2.25	2.26	2.23
2019	2.24	2.29	2.32	2.28
2020	2.3	2.31	1.92	1.49

Regression Statistics	
ultiple R	0.7595664236
Square	0.5769411518
djusted R Squa	0.5655071289
tandard Error	0.7503571184
bservations	77

	Coefficients	*Standard Error*	*t Stat*	*P-value*
tercept	-0.5414378436	0.4620315275	-1.171863415	0.2450110496
nemployment	0.6515589178	0.1195609263	5.449597437	0.00000063413
elinquency	-0.1253544795	0.07606646301	-1.647959884	0.1036007943

Unemployment spiked in 2020 while Charge Off Rates did not

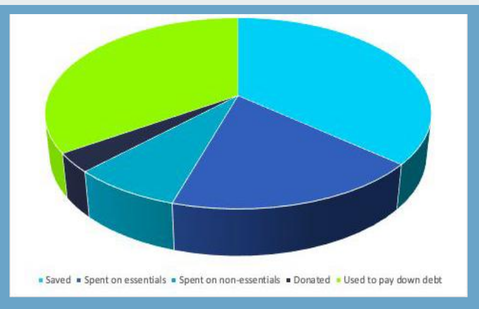
- Covid-19 Pandemic
- Data can be considered an outlier

❖ Null Hypothesis - Unemployment and delinquency rates have no effect on charge-off rates
❖ Significance Level (α) = 0.05
❖ Regression done without 2020 data which was an outlier
❖ Unemployment
  ➢ P-Value = 6.3e-7
  ➢ 6.3e-7 < 0.05
  ➢ Reject null
❖ Delinquency
  ➢ P-Value = 0.104
  ➢ 0.104 > 0.05
  ➢ Fail to reject null

Extension

Why did unemployment spike while charge off rates did not?

Stimulus



Regression Statistics	
Multiple R	0.7581881329
R Square	0.5748492448
Adjusted R Squa	0.5582849297
Standard Error	0.7491665935
Observations	81

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.45336676	0.3698148314	3.929985054	0.000184141959
Unemployment	0.318158997	0.07978511968	3.987698437	0.000150683254
Delinquency	0.1003369141	0.05244329684	1.913245737	0.0594340531
Transfer Payment	-0.00057284902	0.000134295643	-4.26558163	0.000056107472

❑ Regression shows significance of transfer payments (stimulus)
  ❑ P-value of transfer payments is 5.6e-5
  ❑ Far below significance level of 0.05
  ❑ Transfer payments are significant factor in charge-off rates especially when 2020 data is included

Challenges

➔ Finding late 2020 to early 2021 data sets
➔ Choosing an interesting and complex topic relating to current events
➔ Learning methods of analyzing data
➔ Coordinating and communicating remotely due to Covid-19 restrictions
➔ Maintaining focus on the main research question

Conclusion

The data collected was the quarterly nationwide charge-off rate in the US,unemployment rate, and delinquency rate. A multi-variable regression was performed to determine the relationship between the two explanatory variables (unemployment and delinquency) on one response variable (charge-off rates). The unemployment was shown to be a significant factor in the charge-off rates while the delinquency rate had no significant effect. There was an outlier in the unemployment and charge-off rate data for 2020, as the unemployment spiked while the charge-off rate did not. Another multi-variable regression showed that government transfers (stimulus) were a significant factor in the outlier. Overall, why unemployment is a significant factor in the charge-off rates there are other confounding variables that must be considered when trying to predict charge-off rates.

Fatal Crash Rates Compared to License Requirements

Alexandra George, Amelia Faust, Carly Beninati, Lauren Price, Maura Marston, Megan Marston
Plum Senior High School

What We Learned

No correlation is present between the amount of time required by state law to hold a permit and fatal crash rates among the teenage population. While some states in the Northeast require nine to twelve months of instructional driving with a driver's permit, the average crash rate percentage among these states did not outscore the competition. Massachusetts, New York, and Rhode Island displayed the lowest fatal car crash rates among teen drivers, and all these states require only six months of instructional driving. Fatal crash rates among the adult population were lowest in the same states (Massachusetts, New York, and Rhode Island), emphasizing that external factors majorly influence fatal crash rates compared to driver inexperience.

Background

Finding a correlation between driving requirements and teenage crash rates could save the lives of many young drivers. Government and road regulation committees would benefit from research to promote better, safer conditions for teenage drivers.

Method

By utilizing the Fatality and Injury Reporting System Tool by the National Highway Traffic Safety Administration, the group was able to gather fatal crash data by age, state, and year. We began by collecting the fatal crash data from 2005 to 2019 of a young driver (Ages 15-20). This group is the experimental group. To account for differences in state populations, the total fatal crashes each year per state was divided by the corresponding state population, and a fatal crash percentage was calculated for each state. This percentage allows for a more accurate comparison between the states. The group then observed fatal crash data of drivers ages 20-65. This group is the control group, and a fatal crash percentage was calculated per state for the control group as well. After all the data was collected, each group was ranked upon the average fatal crash percentage for the 14 year period. The state with the lowest fatal crash percentage was ranked first and is said to possess the "safest" driving conditions. When comparing the control and experimental group rankings, they were almost identical. The similarity suggests that fatal crash rates must be influenced by other variables other than driver inexperience.

Challenges Faced

Initially, we planned to analyze the crash rates of all states, and organize according to region (Northeast, Southeast, South, Midwest, etc.) in order to mitigate the effects of climate. However, we ultimately decided to only analyze the crash rates of the Northeast because we found that there was no correlation in that region, and extrapolated that there would be no correlation in any other region, and we did not believe it would be worth the time and effort to account for the different climates among regions. Furthermore, the Northeast is local to our team, and had the most states to compare between.

We also initially struggled with compiling our data, as we had to draw from multiple sources in order to find not only the fatal crash rates of each state, but also their overall population and climate region. This resulted in us having to find the data, and then create our own datasets in order to draw comparisons.

State Rankings

(Lowest Average % Fatal Crashes of State Population)

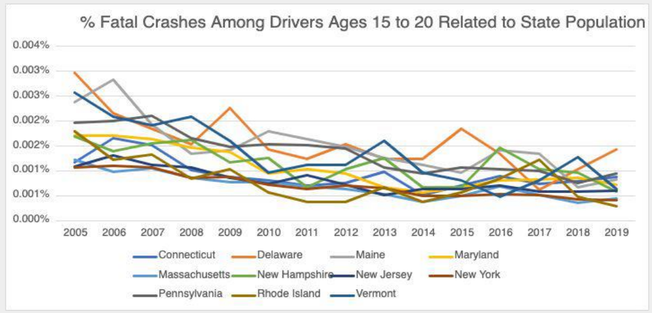
Ages 15-20 (Experimental)

1. Massachusetts
2. New York
3. Rhode Island
4. New Jersey
5. Connecticut
6. Maryland
7. New Hampshire
8. Vermont
9. Pennsylvania
10. Maine
11. Delaware

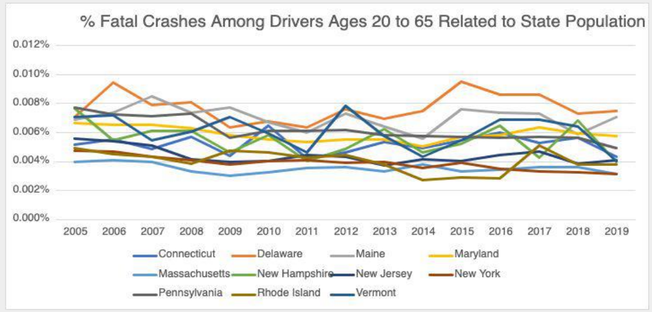
Ages 20-65 (Control)

1. Massachusetts
2. New York
3. Rhode Island
4. New Jersey
5. Connecticut
6. New Hampshire
7. Maryland
8. Vermont
9. Pennsylvania
10. Maine
11. Delaware

Results



The graph to the left displays fatal crash percentages among young drivers in the Northeast from 2005 to 2019. Although some fatal crash rate percentages fluctuated, each state's percentage trended downward and decreased throughout the time interval. The slight downward trend suggests that the teenage population is becoming better, safer drivers, which could be a result of new government driving regulations, better drivers education, or safer road conditions for teen drivers.



The graph displays fatal crash percentages among adult drivers in the Northeast from 2005 to 2019. Unlike the above graph pertaining to fatal crash percentage among teenage drivers, this graph does not depict a downward trend, and fatal crash percentages remained relatively constant throughout the fourteen year period. However, both graphs follow a similar order, meaning that the states possessing the highest fatal crash rate percentages among teenagers also possess the highest percentage among adults.

The table displays the amount of time required by law for a young driver to possess a drivers learning permit before acquiring a driver's license. Delaware and Vermont require the longest time (12 months). If the amount of time with the learners permit led to less teenage drivers experiencing fatal car crashes, these two states would display the lowest average fatal crash percentages. However, this is not the case as Vermont and Delaware possess a teen fatal crash percentage greater than the majority of the other states in the Northeast.

	Length Permit is Held		
	6 Months	9 Months	12 Months
Connecticut	X		
Delaware			X
Maine	X		
Maryland		X	
Massachusetts	X		
New Jersey	X		
New York	X		
Pennsylvania	X		
Rhode Island	X		
Vermont			X

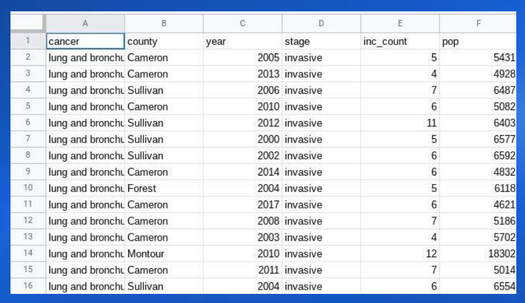
What is the relationship between airborne particulate matter concentration and lung cancer incidence?

Eric Kilduff, Rudra Thakkar, Lucas Wycich, Jeffrey Yan from Plum Senior High School

Key

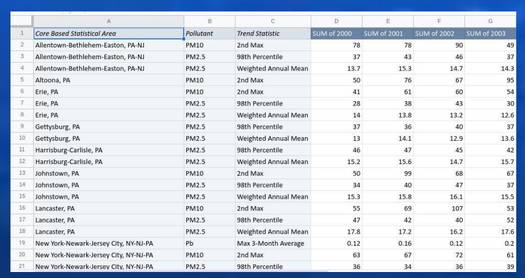
- **PM2.5** - Particulate matter that is 2.5 micrometers or less in diameter
- **PM10** - particulate matter that is 10 micrometers or less in diameter
- **Ppm** - parts per million of air particles in a sample
- **Ppb** - parts per billion of air particles in a sample

Data Sets



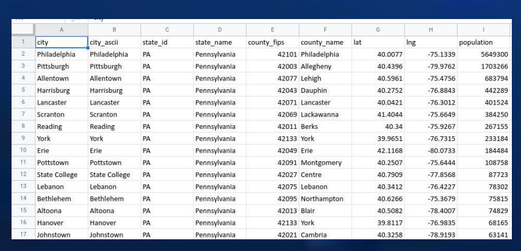
	A	B	C	D	E	F
1	cancer	county	year	stage	inc_count	pop
2	lung and bronch	Cameron	2005	invasive	5	5431
3	lung and bronch	Cameron	2013	invasive	4	4928
4	lung and bronch	Sullivan	2006	invasive	7	6487
5	lung and bronch	Cameron	2010	invasive	6	5082
6	lung and bronch	Sullivan	2012	invasive	11	6403
7	lung and bronch	Sullivan	2000	invasive	5	6577
8	lung and bronch	Sullivan	2002	invasive	6	6592
9	lung and bronch	Cameron	2014	invasive	6	4832
10	lung and bronch	Forest	2004	invasive	5	6118
11	lung and bronch	Cameron	2017	invasive	6	4621
12	lung and bronch	Cameron	2008	invasive	7	5186
13	lung and bronch	Cameron	2003	invasive	4	5702
14	lung and bronch	Montour	2010	invasive	12	18302
15	lung and bronch	Cameron	2011	invasive	7	5014
16	lung and bronch	Sullivan	2004	invasive	6	6554

- **"Cancer Incidences"**
- A dataset depicting various types of cancer and their respective incidence counts per year for each county. The incidence counts are the central point for analysis, which are geographically related by counties.
- Data was taken from the columns titled "inc_count", "county", "year", and "cancer".



	A	B	C	D	E	F	G
1	*Core Based Statistical Area*	*Pollutant*	*Trend Statistic*	SUM of 2000	SUM of 2001	SUM of 2002	SUM of 2003
2	Allentown-Bethlehem-Easton, PA-NJ	PM10	2nd Max	78	78	90	49
3	Allentown-Bethlehem-Easton, PA-NJ	PM2.5	98th Percentile	37	43	46	37
4	Allentown-Bethlehem-Easton, PA-NJ	PM2.5	Weighted Annual Mean	13.7	15.3	14.7	14.3
5	Altoona, PA	PM10	2nd Max	50	76	67	95
6	Erie, PA	PM10	2nd Max	41	61	60	54
7	Erie, PA	PM2.5	98th Percentile	28	38	43	30
8	Erie, PA	PM2.5	Weighted Annual Mean	14	13.8	13.2	12.6
9	Gettysburg, PA	PM2.5	98th Percentile	37	36	40	37
10	Gettysburg, PA	PM2.5	Weighted Annual Mean	13	14.1	12.9	13.6
11	Harrisburg-Carlisle, PA	PM2.5	98th Percentile	46	47	45	42
12	Harrisburg-Carlisle, PA	PM2.5	Weighted Annual Mean	15.2	15.6	14.7	15.7
13	Johnstown, PA	PM10	2nd Max	50	99	68	67
14	Johnstown, PA	PM2.5	98th Percentile	34	40	47	37
15	Johnstown, PA	PM2.5	Weighted Annual Mean	15.3	15.8	16.1	15.5
16	Lancaster, PA	PM10	2nd Max	55	69	107	53
17	Lancaster, PA	PM2.5	98th Percentile	47	42	40	52
18	Lancaster, PA	PM2.5	Weighted Annual Mean	17.8	17.2	16.2	17.6
19	New York-Newark-Jersey City, NY-NJ-PA	Pb	Max 3-Month Average	0.12	0.16	0.12	0.2
20	New York-Newark-Jersey City, NY-NJ-PA	PM10	2nd Max	63	67	72	61
21	New York-Newark-Jersey City, NY-NJ-PA	PM2.5	98th Percentile	36	34	36	39

- **"Air Quality"**
- A table displaying concentrations of particles in the air over a given area. There are typically multiple areas represented per row. The chemical concentrations of PM10 and PM2.5 are the independent variable in the analysis, graphed against the cancer incidences to uncover a possible correlation.
- Data was used from the columns titled "Core Based Statistical Area", "Pollutant", "Trend Statistic", and the rest of the columns to the right depicting the chemical concentrations in either ppm or ppb.



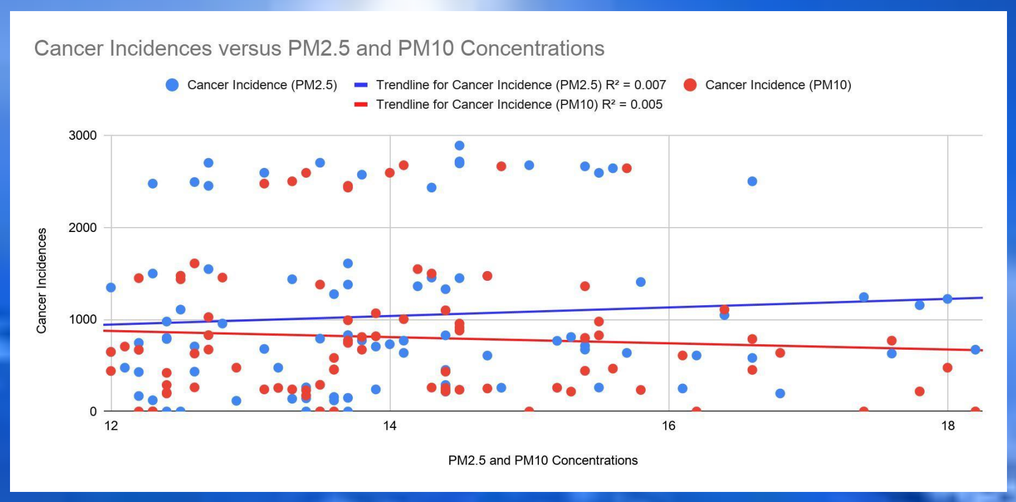
	A	B	C	D	E	F	G	H	I
1	city	city_ascii	state_id	state_name	county_fips	county_name	lat	lng	population
2	Philadelphia	Philadelphia	PA	Pennsylvania	42101	Philadelphia	40.0077	-75.1339	5649300
3	Pittsburgh	Pittsburgh	PA	Pennsylvania	42003	Allegheny	40.4396	-79.9762	1703266
4	Allentown	Allentown	PA	Pennsylvania	42077	Lehigh	40.5961	-75.4756	683794
5	Harrisburg	Harrisburg	PA	Pennsylvania	42043	Dauphin	40.2752	-76.8843	442289
6	Lancaster	Lancaster	PA	Pennsylvania	42071	Lancaster	40.0421	-76.3012	401524
7	Scranton	Scranton	PA	Pennsylvania	42069	Lackawanna	41.4044	-75.6649	384250
8	Reading	Reading	PA	Pennsylvania	42011	Berks	40.34	-75.9267	267155
9	York	York	PA	Pennsylvania	42133	York	39.9651	-76.7315	233184
10	Erie	Erie	PA	Pennsylvania	42049	Erie	42.1168	-80.0733	184484
11	Pottstown	Pottstown	PA	Pennsylvania	42091	Montgomery	40.2507	-75.6444	108758
12	State College	State College	PA	Pennsylvania	42027	Centre	40.7909	-77.8568	87723
13	Lebanon	Lebanon	PA	Pennsylvania	42075	Lebanon	40.3412	-76.4227	78302
14	Bethlehem	Bethlehem	PA	Pennsylvania	42095	Northampton	40.6266	-75.3679	75815
15	Altoona	Altoona	PA	Pennsylvania	42013	Blair	40.5082	-78.4007	74829
16	Hanover	Hanover	PA	Pennsylvania	42133	York	39.8117	-76.9835	68165
17	Johnstown	Johnstown	PA	Pennsylvania	42021	Cambria	40.3258	-78.9193	63141

- **"Translation"**
- An index of each Pennsylvania county and the respective city that the county falls under. Because Air Quality and Cancer Incidence were not covered by the same geographic representations, this third spreadsheet was a major factor is equating the two.
- Data was used from the columns titled "city" and "county_name".

Challenges

The first problem was deciphering the data. In the Air Quality dataset, the meaning of the "Pollutants" column was not apparent, because the airborne chemicals followed scientific classifications. Furthermore, the data did not define what amount of each chemical was considered toxic. In both cases, additional research was required to determine the definitions of each chemical, or in some cases an element, and how much of each chemical was considered toxic. From there, the data of two harmful chemicals were selected for analysis: PM2.5 and PM10. It was also discovered that chemical concentrations were measured in ppm and ppb.

Following that, there was a conflict in locationally relating the Cancer Incidence and Air Quality datasets. Ultimately, the lung cancer incidences had to be summed and displayed next to each area in the Air Quality dataset. However, Cancer Incidence was sorted by county, whereas Air Quality was sorted by "Core Based Statistical Area," which could be any combination of 1-5 cities. To bridge this discrepancy, a third "Translational" dataset containing each Pennsylvania county's respective city had to be used. Each county's city was identified, and the corresponding lung cancer incidence was displayed next to the correct city after a complex manipulation and reordering through Google Sheets.



Summary and Conclusion

There is no correlation between air particulate matter concentration and lung cancer incidence. Our highest R^2 value is .007, which means that only .7% of variation in lung cancer incidence can be explained by the linear relationship between lung cancer incidence and particulate matter concentration. This means that other factors, such as average age or lifestyle choices, are impacting lung cancer incidence. One potential solution is to divide the lung cancer incidence by the city population and graph the quotients against air quality. This would eliminate any errors resulting from differences in population size among different areas.