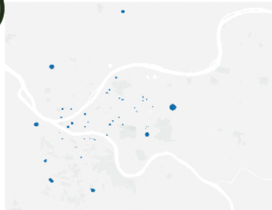
LOTS OF LOTS

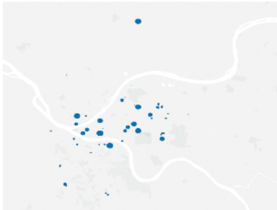
an investigation of the effect of business presence on parking utilization

team members: Colin Crawford, Catrina Raich, Laurel Purcell, Zoe Trexel, Lindsey Hartle, Brayden Simmons

What factors affect the patterns of parking usage in Pittsburgh? Is there enough parking available in each zone? Does business presence affect parking availibility in Pittsburgh?

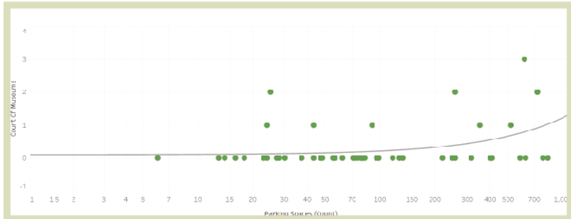


pittsburgh area restaurants



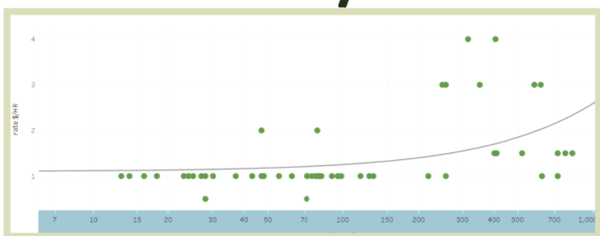
pittsburgh area parking

count of parking spaces vs. count of museums



This graph shows a moderate correlation between parking space count and museums. There is more available parking when there are more museums.

count of parking spaces vs. hourly rate



This graph shows a moderate correlation between parking space count and hourly rate. There is a generally higher parking rate in areas with more parking, or desirable areas.

Regression Statistics	
Multiple R	0.4426773573
R Square	0.1959632426
Adjusted R Square	0.1667255424
Standard Error	205.4814461
Observations	58

analysis sample

HYPOTHESES + ANALYSIS:

Null Hypothesis: Restaurants and museums have no effect on parking availability.
Significance Level (α) = 0.05
Museums:
P-Value = 0.108 --> > α, fail to reject null
Restaurants:
P-Value = 0.0006 --> < α, reject null

Null Hypothesis: Restaurant and museum presence (as a combination to represent business presence) and hourly rate have no effect on parking availability.
Significance Level (α) = 0.05
Sum of restaurants and museums:
P-Value = 0.75 --> > α, fail to reject null
Hourly rate:
P-Value = 0.0005 < α, reject null

DEFINITIONS:

geographical zone: parking zones, as created by the Pittsburgh Parking Authority, were used to determine counts of parking spaces in unique areas

desirable area: an area with prevelant attractions such as museums and restaurants

parking availability: combined number of parking lot spaces and street parking in a zone

CHALLENGES:

> rooting out irrelevant data & outliers
> converting "parking zones" into gps coordinates
> matching businesses with the "parking zones"
> creating accurate, clear visual representation of data

RESOURCES:

Western Pennsylvania Data Center: parking data as reported by the Pittsburgh Parking Authority, business license records, and the hourly rates of parking zones

City of Pittsburgh Website: municipal code for parking lots required per business

CONCLUSION

To determine the relationship between business presence and parking availability, we used a multiple variable regression test. Museum presence was shown to be a significant factor in parking availibilty, while restaurant presence had no significant effect. After running our second regression test, we found that hourly rate was a significant factor in parking availibilty while the combination of restaurant and museum presence had no significant effect. Thus, there are confounding variables that affect parking availability that we did not consider in our analysis. In order to predict parking availability in the city, it is necessary to explore additional variables.

Effects of Emissions on Public Health

Martina Tatalias, Marco Hebestreit, & Aletris Eckert
Bethel Park Team #1

Question:

Is there a correlation between asthma rates and air pollution in Pennsylvania counties?

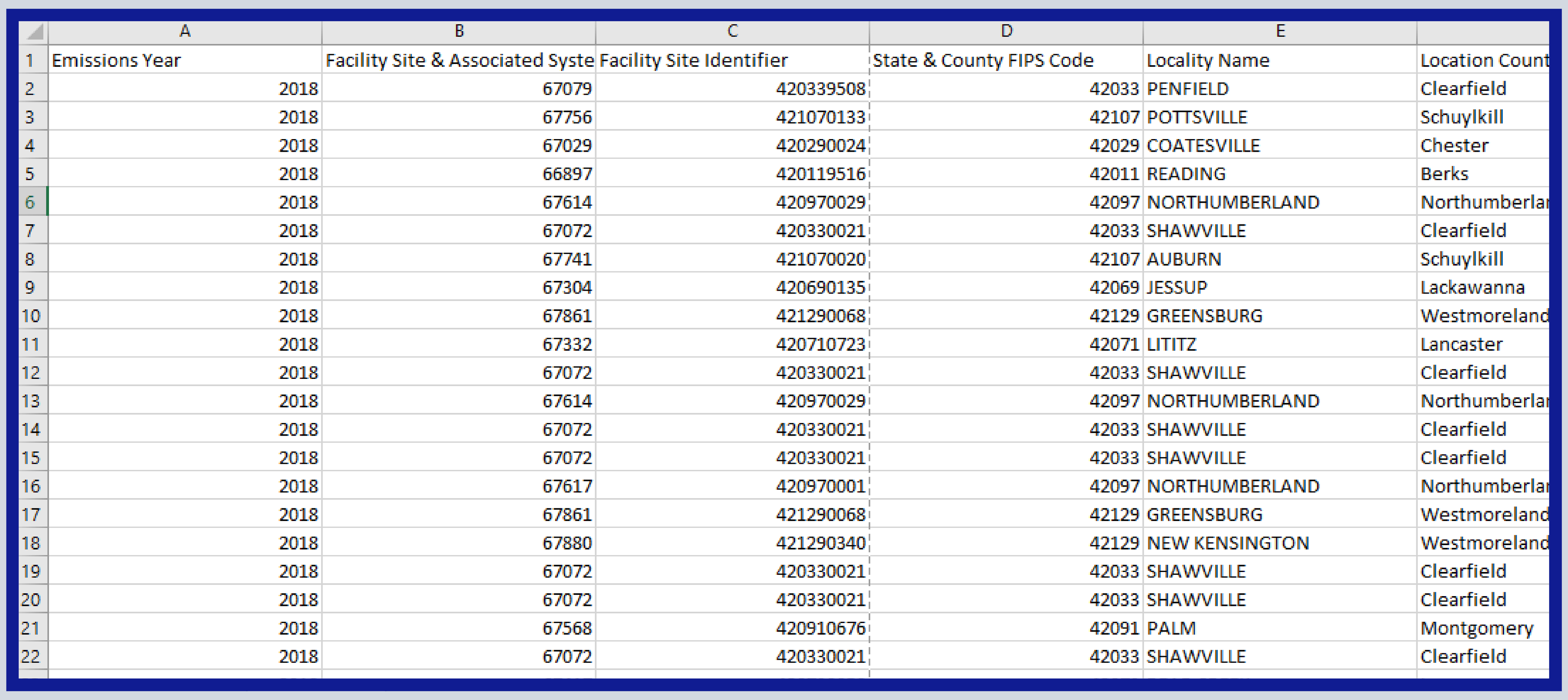
Challenges

We faced many challenges when finding data. For example, we had to find data sets that correspond to Pennsylvania counties by rate of emissions and asthma counts. Both of our data sets allowed us to analyze the correlation between these factors.

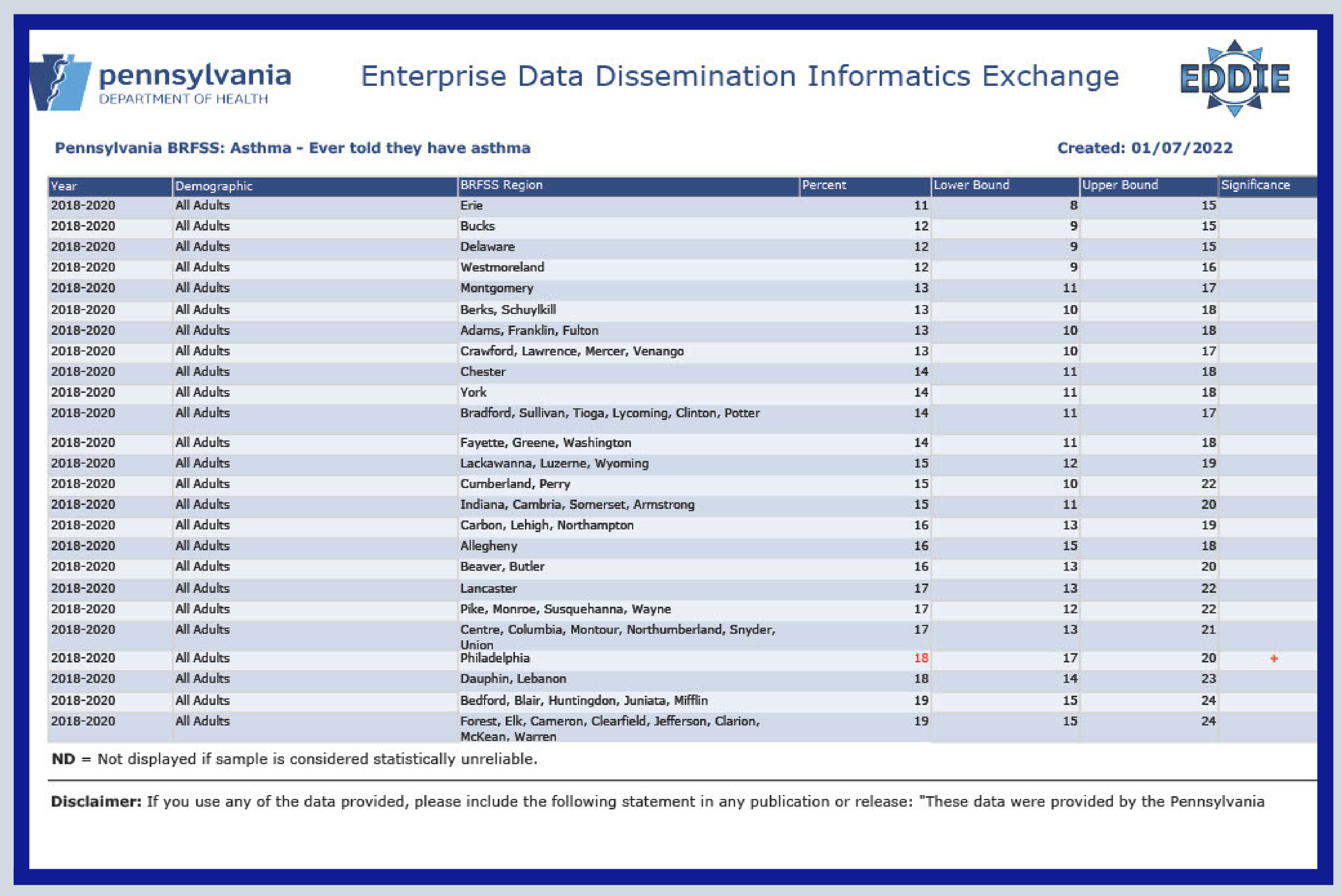
Data Sources

To find accurate data, we looked for the most recent data and made sure it was from a reliable source. We found data from Open Data Pennsylvania and the Pennsylvania Department of Health. One data set shows the total emissions rate of pollution by Pennsylvania county; this data set also shows the location, facility name, and emissions unit. The second data set shows the rate of people diagnosed with asthma by county.

Data Set Examples



	A	B	C	D	E	
1	Emissions Year	Facility Site & Associated Syste	Facility Site Identifier	State & County FIPS Code	Locality Name	Location Count
2	2018	67079	420339508	42033	PENFIELD	Clearfield
3	2018	67756	421070133	42107	POTTSVILLE	Schuylkill
4	2018	67029	420290024	42029	COATESVILLE	Chester
5	2018	66897	420119516	42011	READING	Berks
6	2018	67614	420970029	42097	NORTHUMBERLAND	Northumberlar
7	2018	67072	420330021	42033	SHAWVILLE	Clearfield
8	2018	67741	421070020	42107	AUBURN	Schuylkill
9	2018	67304	420690135	42069	JESSUP	Lackawanna
10	2018	67861	421290068	42129	GREENSBURG	Westmoreland
11	2018	67332	420710723	42071	LITITZ	Lancaster
12	2018	67072	420330021	42033	SHAWVILLE	Clearfield
13	2018	67614	420970029	42097	NORTHUMBERLAND	Northumberlar
14	2018	67072	420330021	42033	SHAWVILLE	Clearfield
15	2018	67072	420330021	42033	SHAWVILLE	Clearfield
16	2018	67617	420970001	42097	NORTHUMBERLAND	Northumberlar
17	2018	67861	421290068	42129	GREENSBURG	Westmoreland
18	2018	67880	421290340	42129	NEW KENSINGTON	Westmoreland
19	2018	67072	420330021	42033	SHAWVILLE	Clearfield
20	2018	67072	420330021	42033	SHAWVILLE	Clearfield
21	2018	67568	420910676	42091	PALM	Montgomery
22	2018	67072	420330021	42033	SHAWVILLE	Clearfield



pennsylvania DEPARTMENT OF HEALTH — Enterprise Data Dissemination Informatics Exchange — EDDIE

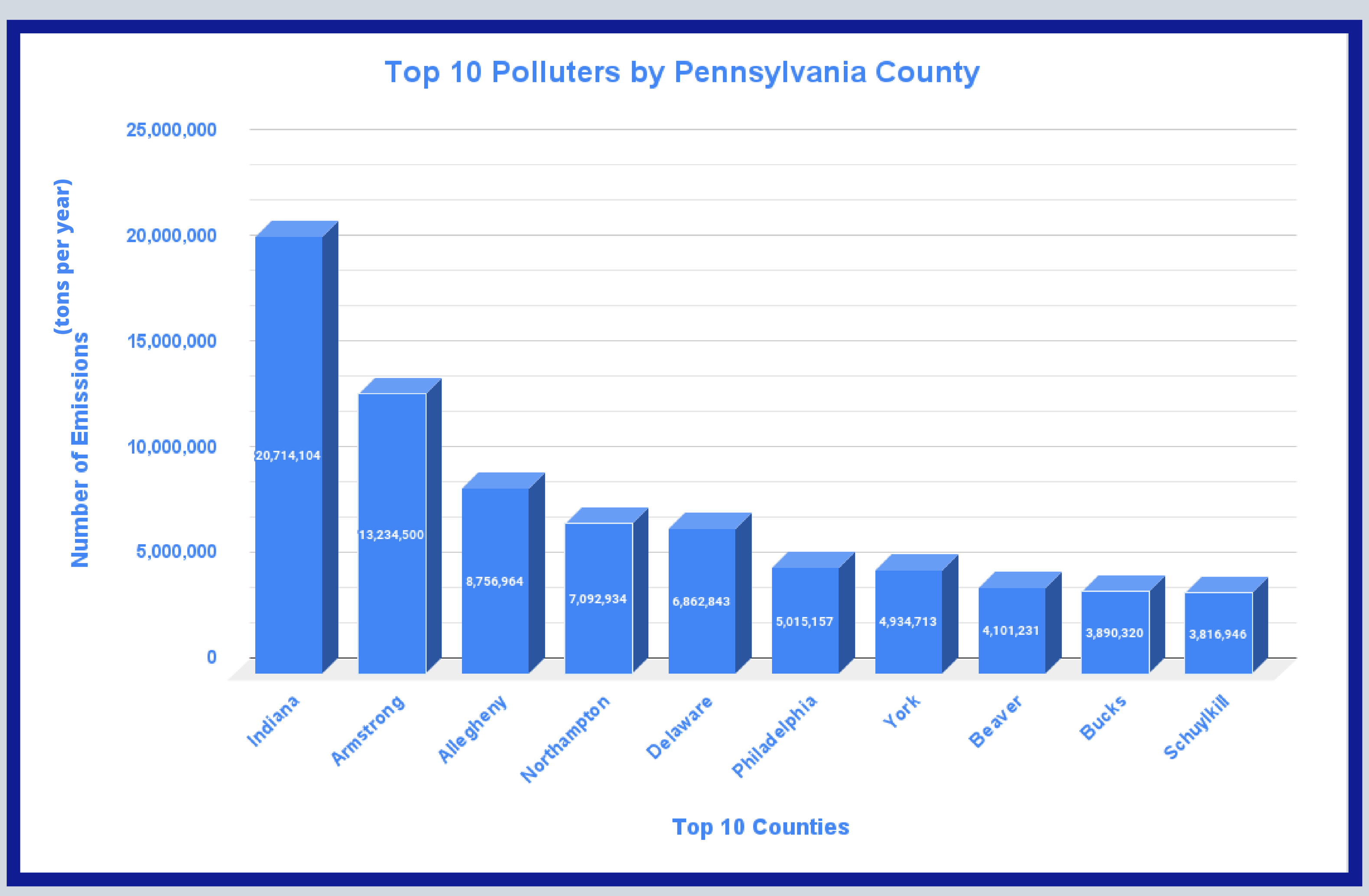
Pennsylvania BRFSS: Asthma - Ever told they have asthma — **Created: 01/07/2022**

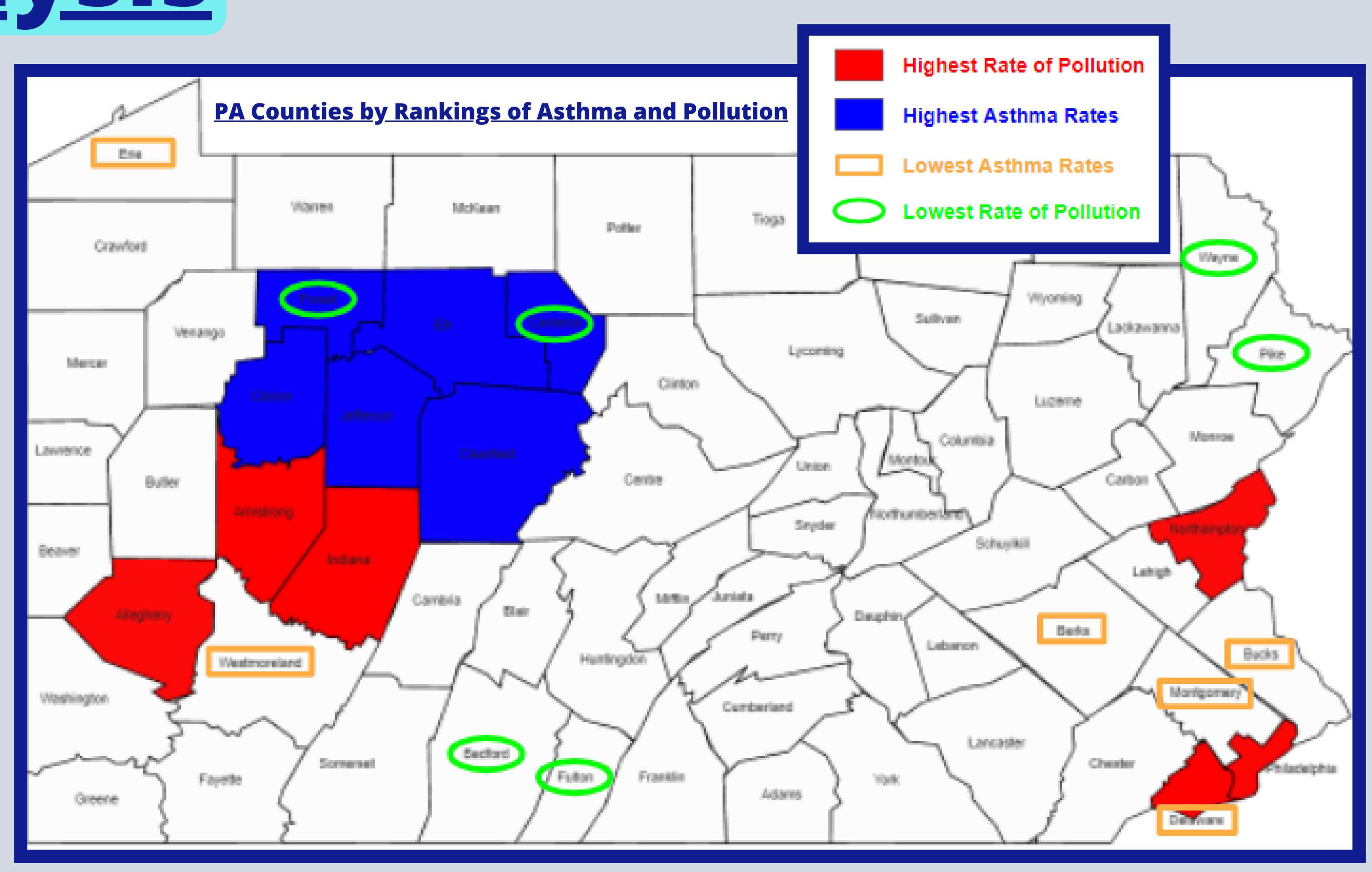
Year	Demographic	BRFSS Region	Percent	Lower Bound	Upper Bound	Significance
2018-2020	All Adults	Erie	11	8	15	
2018-2020	All Adults	Bucks	12	9	15	
2018-2020	All Adults	Delaware	12	9	15	
2018-2020	All Adults	Westmoreland	12	9	16	
2018-2020	All Adults	Montgomery	13	11	17	
2018-2020	All Adults	Berks, Schuylkill	13	10	18	
2018-2020	All Adults	Adams, Franklin, Fulton	13	10	18	
2018-2020	All Adults	Crawford, Lawrence, Mercer, Venango	13	10	17	
2018-2020	All Adults	Chester	14	11	18	
2018-2020	All Adults	York	14	11	18	
2018-2020	All Adults	Bradford, Sullivan, Tioga, Lycoming, Clinton, Potter	14	11	17	
2018-2020	All Adults	Fayette, Greene, Washington	14	11	18	
2018-2020	All Adults	Lackawanna, Luzerne, Wyoming	15	12	19	
2018-2020	All Adults	Cumberland, Perry	15	10	22	
2018-2020	All Adults	Indiana, Cambria, Somerset, Armstrong	15	11	20	
2018-2020	All Adults	Carbon, Lehigh, Northampton	16	13	19	
2018-2020	All Adults	Allegheny	16	15	18	
2018-2020	All Adults	Beaver, Butler	16	13	20	
2018-2020	All Adults	Lancaster	17	13	22	
2018-2020	All Adults	Pike, Monroe, Susquehanna, Wayne	17	12	22	
2018-2020	All Adults	Centre, Columbia, Montour, Northumberland, Snyder, Union	17	13	21	
2018-2020	All Adults	Philadelphia	18	17	20	+
2018-2020	All Adults	Dauphin, Lebanon	18	14	23	
2018-2020	All Adults	Bedford, Blair, Huntingdon, Juniata, Mifflin	19	15	24	
2018-2020	All Adults	Forest, Elk, Cameron, Clearfield, Jefferson, Clarion, McKean, Warren	19	15	24	

ND = Not displayed if sample is considered statistically unreliable.

Disclaimer: If you use any of the data provided, please include the following statement in any publication or release: "These data were provided by the Pennsylvania

Analysis





Conclusion

There is a strong positive correlation between carbon emissions and asthma rates in surrounding Pennsylvania counties. The counties with the highest pollution rates border those with the highest asthma rates. The data suggests that wind and weather play a role into asthma rates, and this can be further analyzed with weather data in the area.

Potential Policy

These data sets and correlations prove our hypothesis regarding emissions in counties and asthma rates of surrounding counties. In the future, we would like to implement programs for less vehicular traffic through the use of HOV lanes and a switch to clean energy in place of coal powerplants. We can create further regulations for pollution standards and install smart air monitors to monitor air pollution in areas where high pollution is predicted.

EFFECT OF COMMUNITY WEALTH ON TEST SCORES

SOURCES

- SAT Scores - www.education.pa.gov/K-12/Assessment%20and%20Accountability/SAT-ACT/Pages/default.aspx
- School District Expenditures - www.education.pa.gov/Teachers%20-%20Administrators/School%20Finances/Finances/AFR%20Data%20Summary/Pages/AFR-Data-Summary-Level.aspx
- School District Student Population: https://www.education.pa.gov/DataAndReporting/Enrollment/Pages/PublicSchEnrReports.aspx
- Median Household Income per District - https://censusreporter.org/tables/B19013/

Does the economic standing of a district affect students' SAT scores?

- We analyzed the median income of districts vs SAT scores in addition to school expenditures per student vs SAT scores.

Carlynton High School

Owen Schriver, Audrey Robb, Michael Kozy, and Azjia Gardner

INTRODUCTION

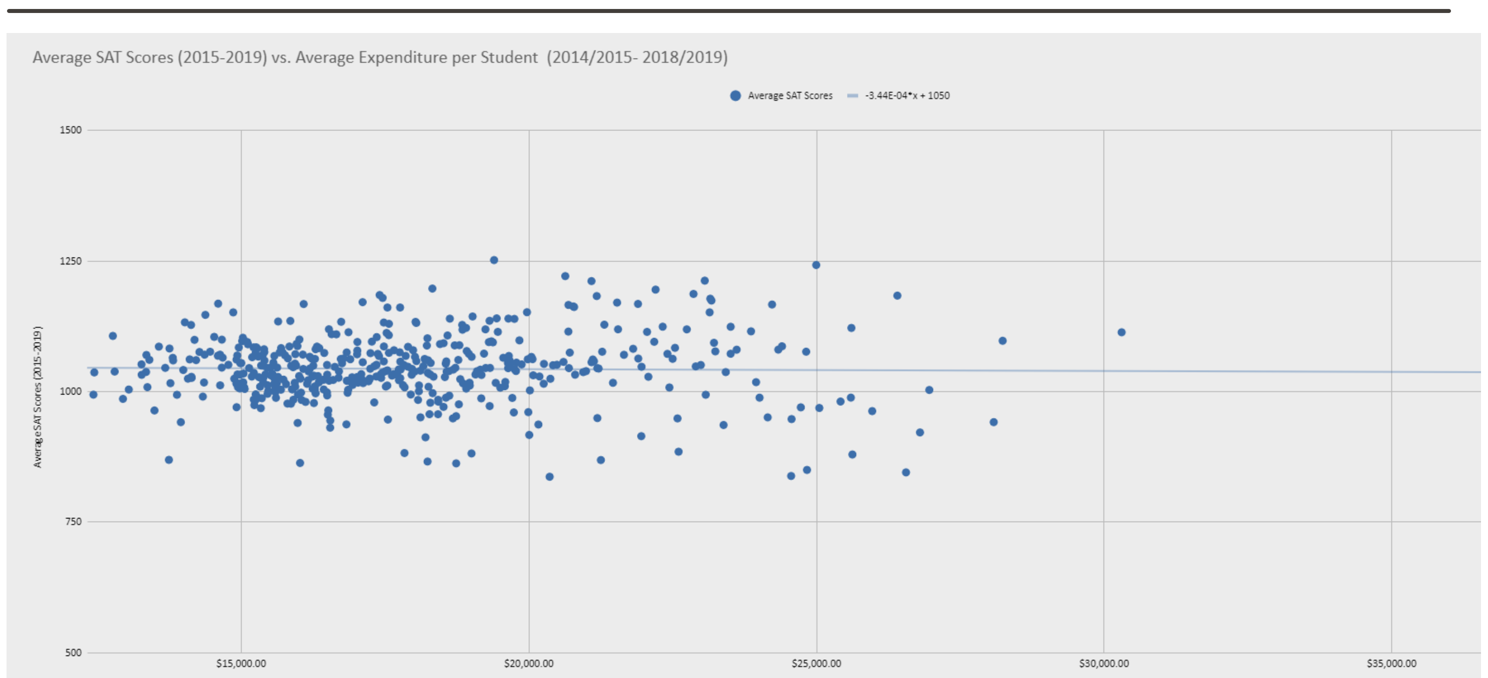
- We wanted to see how the different aspects of a district's wealth affect students, specifically their test scores
- We found reliable data sets for school districts
- We cleaned and organized the data to match only public school districts with our specific data collection areas.
- We did a correlation test, T-test, and P-test to see if there was a correlation and if the relationships were significant

Data

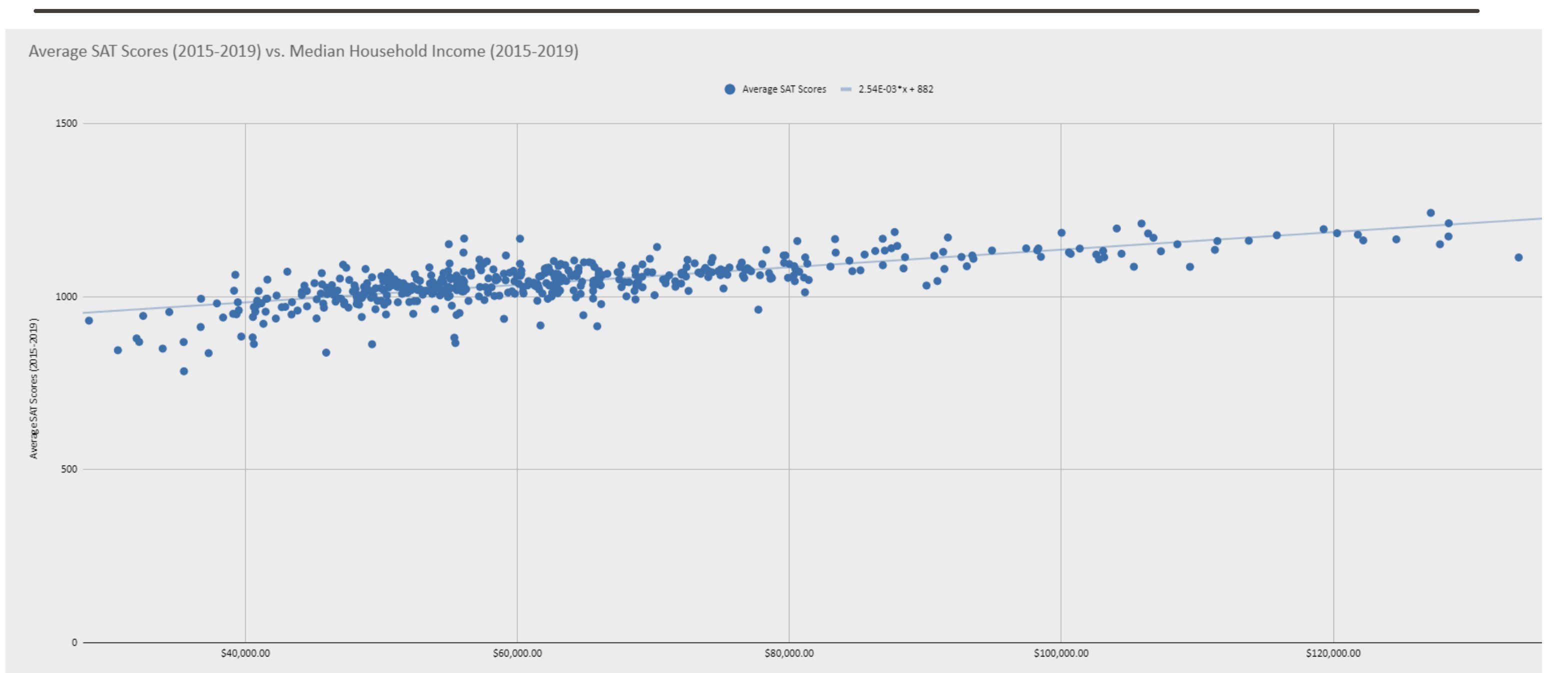
AUN	District Name	LEA Type	Average Expenditure/ Student (2014/2015-2018/2019)	Median Household Income (2015-2020) in 2019 inflation adjusted dollars	Average SAT Scores
123,460,302.00	Abington SD	SD	$18,226.36	$93,041.00	1088
101,260,303.00	Albert Gallatin Area SD	SD	$15,339.04	$45,761.00	968
127,040,503.00	Aliquippa SD	SD	$25,623.09	$31,985.00	880
103,020,603.00	Allegheny Valley SD	SD	$22,988.77	$55,791.00	1051
106,160,303.00	Allegheny-Clarion Valley SD	SD	$22,076.90	$51,172.00	1028
121,390,302.00	Allentown City SD	SD	$17,837.26	$40,520.00	882
108,070,502.00	Altoona Area SD	SD	$13,268.63	$44,328.00	1032
127,040,703.00	Ambridge Area SD	SD	$19,070.68	$60,742.00	1033
113,380,303.00	Annville-Cleona SD	SD	$15,692.23	$62,135.00	1083
114,060,503.00	Antietam SD	SD	$19,710.74	$56,393.00	988
128,030,603.00	Apollo-Ridge SD	SD	$18,557.10	$49,972.00	988
128,030,852.00	Armstrong SD	SD	$18,611.63	$50,478.00	1039
117,080,503.00	Athens Area SD	SD	$17,736.94	$50,554.00	1042
101,630,504.00	Avella Area SD	SD	$21,461.31	$68,603.00	1017

Data was gathered and compiled from multiple sources regarding the following: median household income per school district, school districts' SAT scores, students per district, and district expenditures over a five-year time period (2015-2019). We used a total of 480 PA school Districts.

SAT Scores vs Expenditure per Student



SAT Scores vs Median Household Income



CHALLENGES

- We had to manually separate charter schools and cyber schools from public schools to keep the data consistent
- We struggled to find data that was comparable in all categories because data sets often included different schools or were formatted differently
- Some districts has multiple schools, some of which didn't report SAT scores, or didn't have any students who tested, which we had to exclude from our data

Tests for District Expenditure/ Student:	
corelation (r) :	-0.0169
t stat	25.47301236
p-value	0.00E+00
critical value:	2.626
α	0.01
	*two tail test

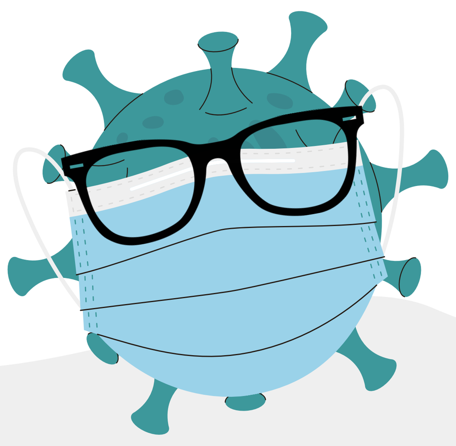
Test for Median Income of District:	
correlation (r):	0.7588
t stat:	-0.3690946933
p- value	0.7122205802
critical value:	2.626
α	0.01
	*two tail test

ANALYSIS AND CONCLUSIONS

Our data showed that there is no correlation between school district expenditures per student and SAT scores. However, it also showed that there is a statistically significant positive correlation between the median household income in a school district and the district's SAT scores. The equation defining the relationship between income (x) and SAT scores (y) is y = .00254x + 882. This data suggests that higher community wealth leads to higher test scores, while school district wealth has no effect on test scores. From the slope of our data, we found that for every increase of $20,000 in average household income, there is about a 50 point increase in average SAT score.

FURTHER QUESTIONS

- Do higher education levels in wealthier districts impact test scores?
- Do cultural expectations in communities impact their students' test scores?
- Does the income itself cause the increase in scores, or is it due to related factors, such as more accessible tutoring?



Forecasting COVID-19 Infections & Vaccinations in PA

Central Dauphin HS - Tiffany Chen, Matthew Bretz, & Jacob Miller

Research Question

How many Pennsylvanians will get infected with COVID-19?

Initial Hypotheses

H_0 : Data matches prediction model.
H_a : Data differs as of March 5th.

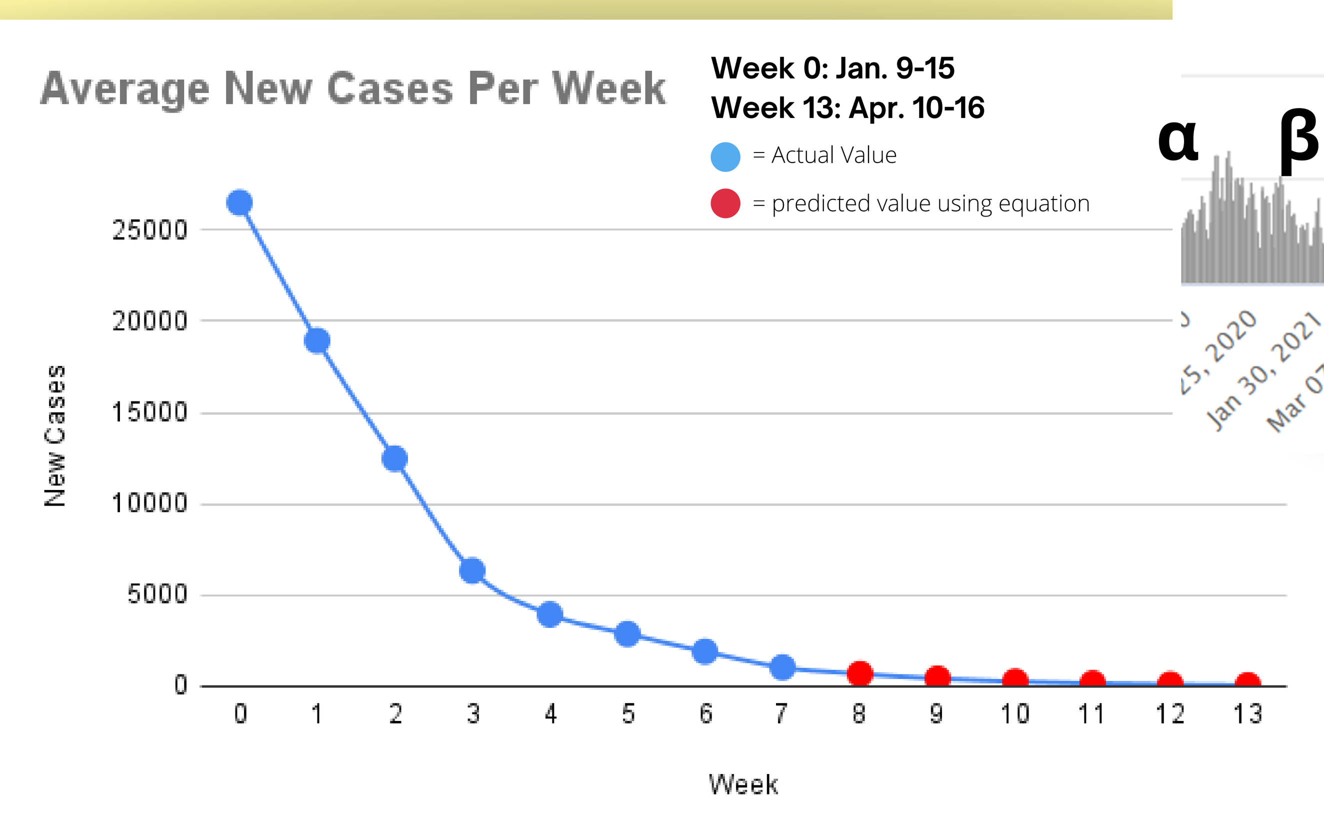
KEY

- **Unvaccinated** (Unvax): 0-1 doses
- **Vaccinated** (Vax): 2 doses
- **Boosted**: 3+ doses

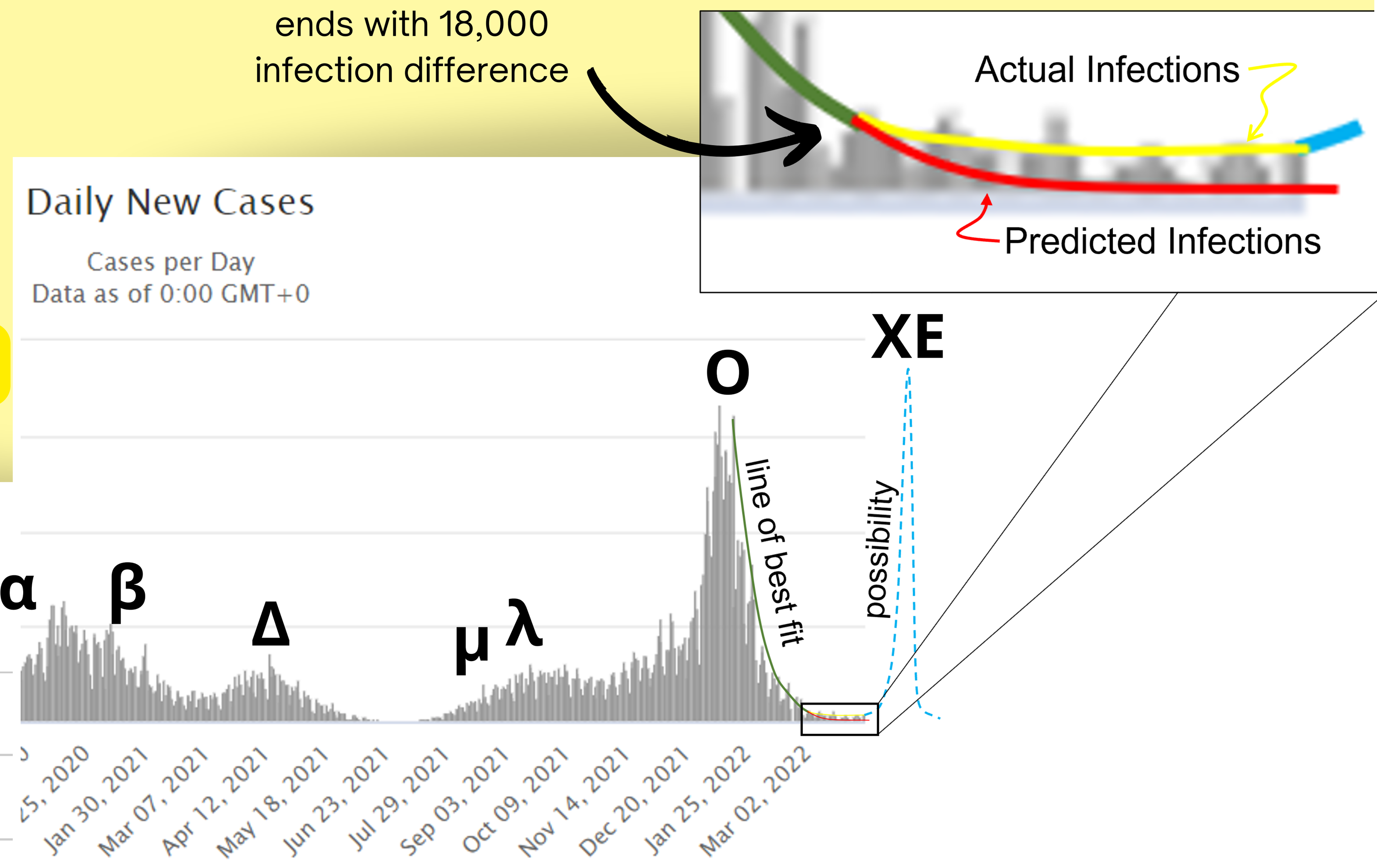
Method to Forecast Infections

In Minitab, a logistic nonlinear regression model was used to create a prediction equation for the number of new cases per week.

ln(AvgWeeklyCases) = 10.24 - .4639*Week



March 5th departure, ends with 18,000 infection difference



Null hypothesis	H_0: $p_1 - p_2 = 0.001157$	
Alternative hypothesis	H_1: $p_1 - p_2 \neq 0.001157$	
Method	**Z-Value**	**P-Value**
Normal approximation	2.47	0.013

NB: As our data did not pass the Anderson-Darling Test for Normality, even after applying a transform, we had to switch from a 2-sample t-test to a **2-proportions test** to meet the assumptions:

1. Binomial Distribution: Infected, Not Infected ✓
2. np & nq >=5 ✓

Possible cause of model departure on March 5th:

Change in behaviors: Less masking, less vaccinations, and less boosting observed.

https://centerfordignity.com/state-by-state-school-mask-mandates/
https://www.npr.org/sections/health-shots/2022/02/25/1082249002/cdc-says-americans-can-now-go-unmasked-in-many-parts-of-the-country
https://www.inquirer.com/news/philadelphia/philadelphia-vaccination-rate-children-covid-adults-20220309.html
https://www.cnbc.com/2022/02/24/bill-gates-australia-covid-blueprint-could-help-prevent-next-pandemic.html

To investigate these causes, we modeled the different subgroups (Unvax, Vax, & Boosted) using a Markov chain:

Method to Forecast Vaccinations

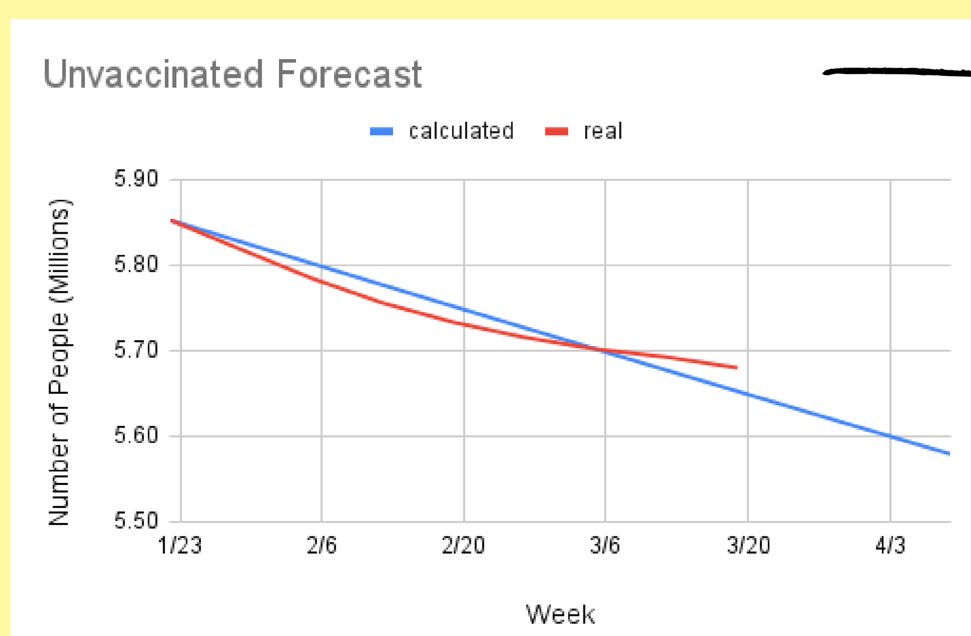
A Markov chain was used to create vaccination status matrix. An exponential decay equation to create matrix values. We chose a starting population vector (Jan. 23), then used matrix multiplication to predict the number of unvax, vax, and boosted for each week.

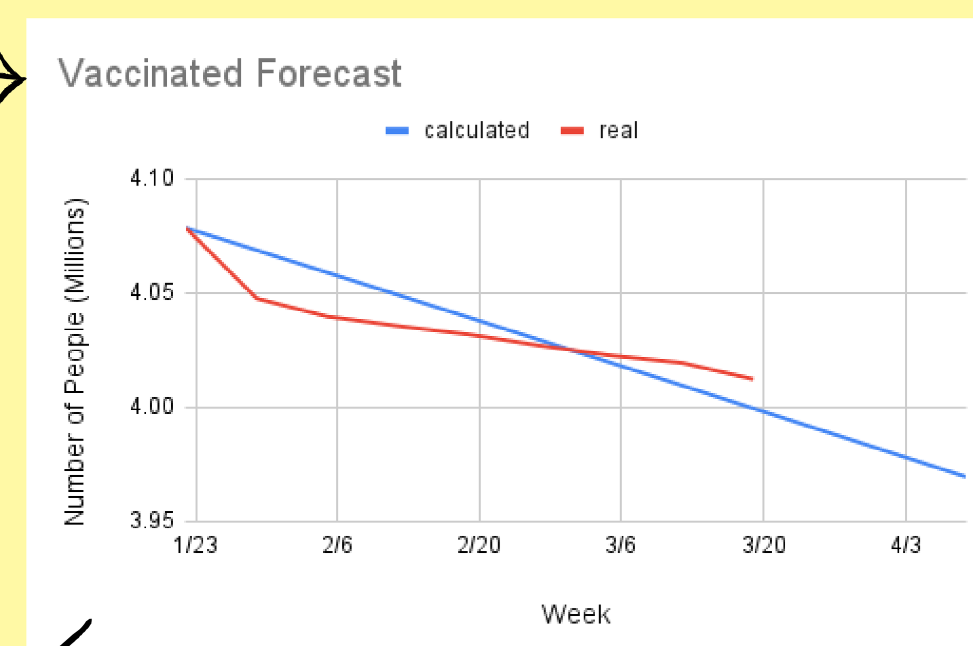
Vax Status Matrix

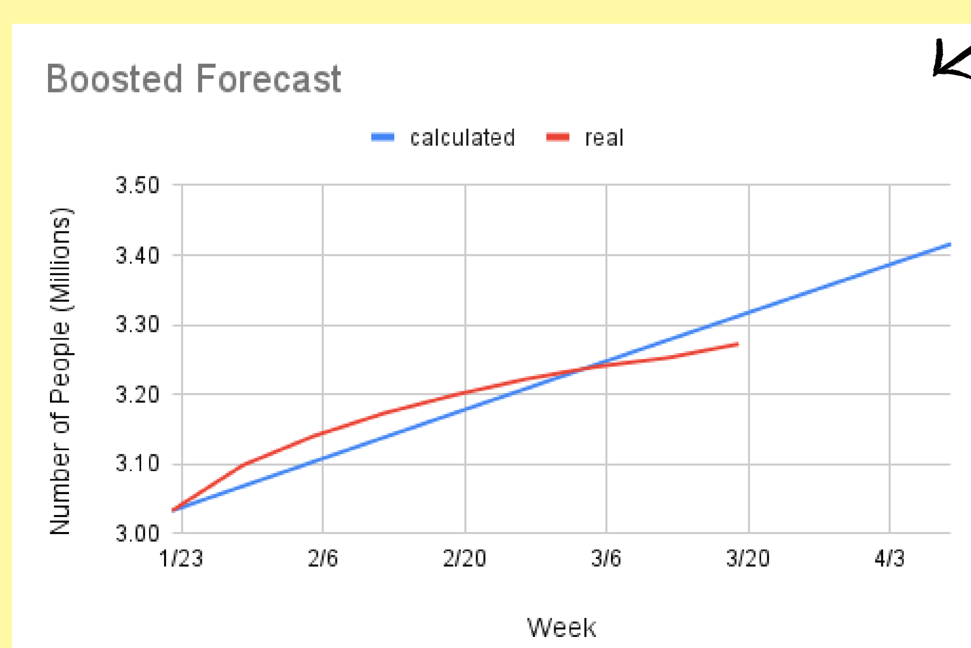
	unvax	vax	boosted
unvax	0.995654	0	0
vax	0.004346	0.991359	0
boosted	0	0.008641	1

Starting Population Vector

Unvax	5,852,447
Vax	4,078,499
Boosted	3,033,110







These graphs display our predicted vs. real populations. Note the inflection point at March 5.

Although our model fits the data better than Minitab's model, we are still investigating potential improvements.

Challenges

- **Testing**: The 2-proportions test is less robust than the 2-sample t-test because it is limited to a single equation/standard deviation.
- **Accuracy**: Vax status forecasts had limited accuracy due to the suspected change in behavior beginning March 5th.
- **Validity of data**: This data represents a minimum number of infected as reported by doctors and hospitals for those feeling unwell enough to seek medical attention. Home tests and untested infections are not reported but certainly abound.
- **False positives**: due to the probabilities involved, Bayes' Rule clearly shows the probability of a false positive is extremely low.

Conclusion

The data is strong enough to reject our null hypothesis with an alpha level of 0.02. The data clearly shows a departure from predicted infections. As of April 5th, 2022, with 98% confidence, we can safely say there have been at least 15,000 infections more than expected since March 5th, 2022.

Next Steps

1. Investigate Leslie and Perron-Frobenius Methods to better model the matrix.
2. Switch to natural base ***e*** and use a Vandermonde determinant to create a quadratic equation and find a steady state vector.
3. Test Vax and Boosted variables for Normality to do a 2-sample t-test
4. Investigate regression modelling and correlation analyses for both forecasts.

Databases Used:

www.health.pa.gov/topics/disease/coronavirus/Vaccine/Pages/Vaccine.aspx
www.worldometers.info/coronavirus/usa/pennsylvania/#graph-cases-daily

The Concealed Correlation with Drug Use

Greensburg Salem High School
Lea Kasmer, Mackenzie Halfhill, Adam Wilson, Alaina Blend, and Paige Gaughan

Is there a correlation between unemployment rates and drug overdoses?

Background

When the Covid-19 pandemic hit, the whole world came to a stop; millions were left without jobs, resulting in skyrocketing unemployment rates. With this rather current event, research has found that unemployment takes drastic mental tolls. Statistically speaking, people are likely to turn to drugs when experiencing such mental stress. Thus, analyzing this data may provide a possible correlation between drug overdoses and unemployment rates.

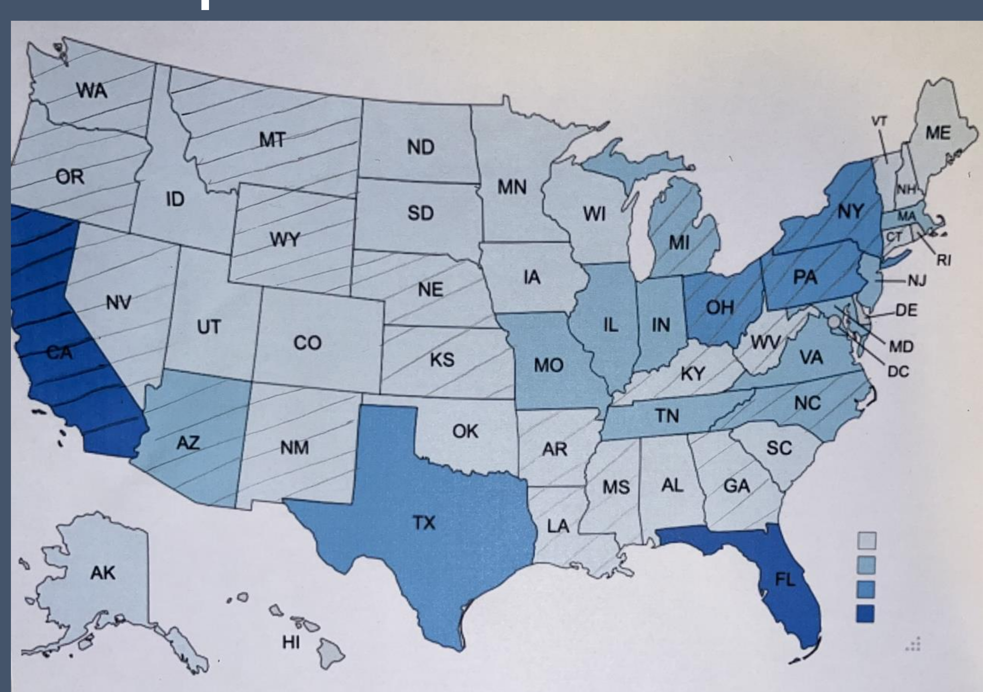
Purpose

Drug overdoses have been an epidemic for decades and has only showed signs of increasing. As this issue continues to threaten the lives of Americans its important to analyze possible correlations. The recent spike in unemployment rates creates the perfect opportunity to analyze this data and find a possible correlation.

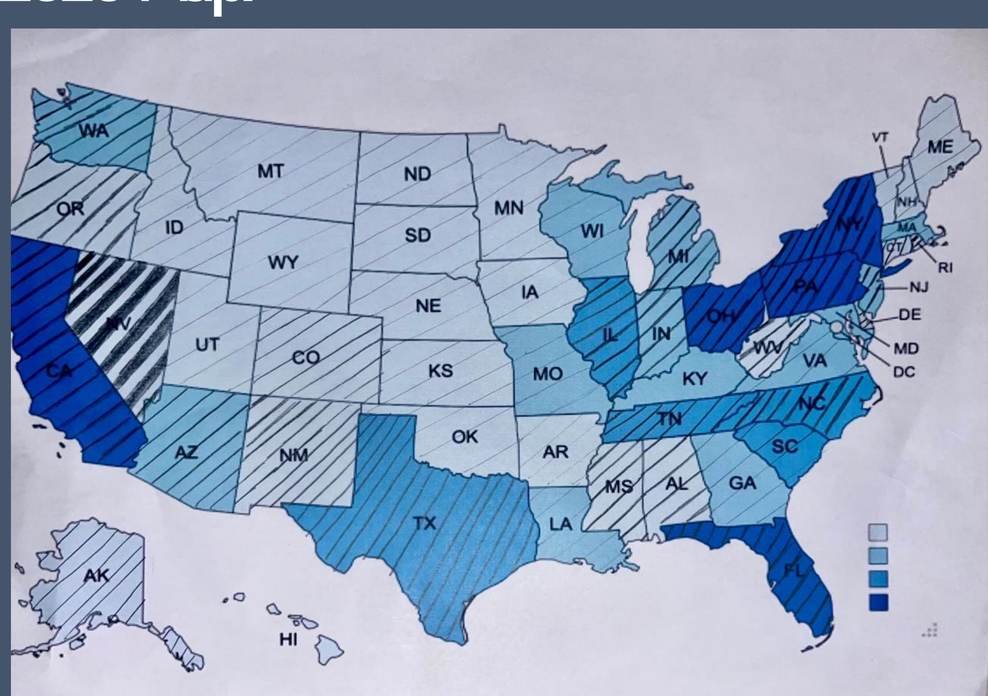
Hypothesis

There is a direct correlation between drug overdoses and unemployment rates across the United States.

2019 Map:

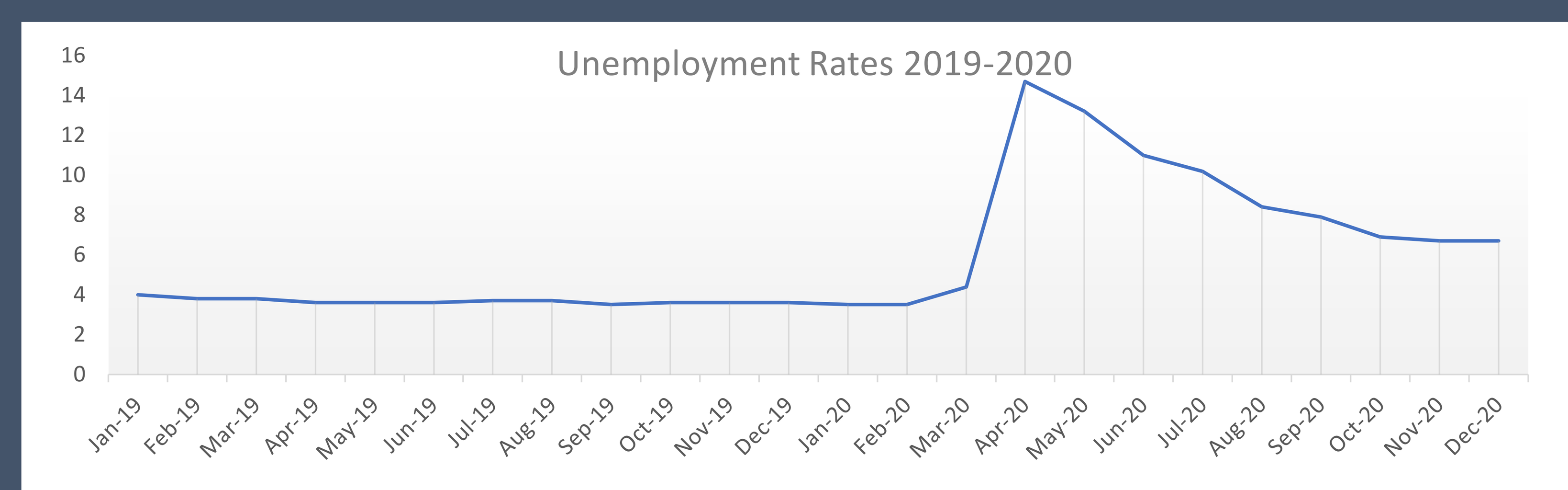


2020 Map:

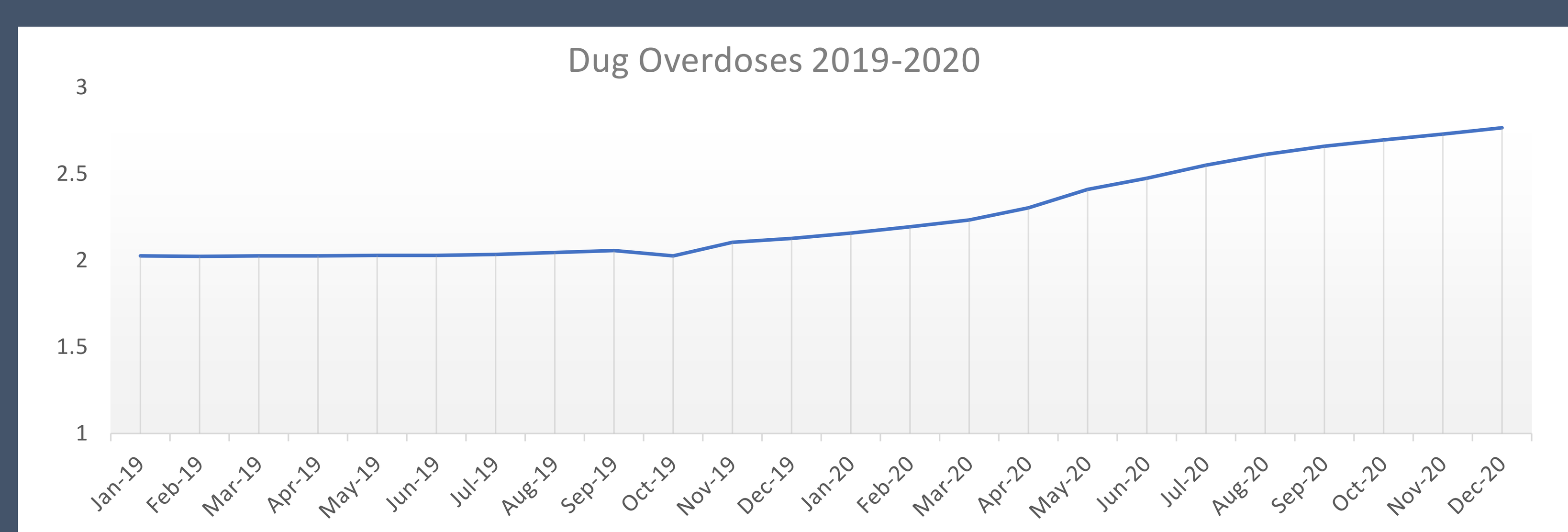


Legend
Overdoses by people:
Light Blue- 0-1500
Medium Blue- 1500-3000
Darker Blue- 3000-4500
Darkest Blue- 4500+

Unemployment Rates:
No lines- 0-3.5%
Thin Lines- 3.5-7%
Thicker Lines- 7-10.5%
Thickest Lines- 10.5%+



*percent of overdoses is multiplied by 1000 for graphing purposes



A Closer Look: North Dakota v. West Virginia

Most of the data so far has revolved around Covid-19. However, there could be other possible factors of the pandemic that drive Americans to drug use. So, to get more accurate data, more research to compare two individual states was gathered, averaging their overdose and unemployment rates over a 20-year time period. (1999-2019) The covid-19 pandemic was purposely left out to see more typical unemployment rates. The data sets were averaged and scaled to fit population statistics. By comparing individual states, it reinforces the correlation between unemployment and drug deaths, especially over a long period of time.

North Dakota (1999-2019)

Average unemployment rate: 2.8%
Average overdoses per year: 5.6 deaths per 100,000

West Virginia (1999-2019)

Average unemployment rate: 5.7%
Average overdoses per year: 30.8 deaths per 100,000

Analysis

The two timelines on the left are both from January 2019 through December 2020. Both unemployment rates and drug overdose rates see a significant increase around March of 2020. Turning to the "Closer Look" section, the data shows that North Dakota has CONSISTENTLY had a lower unemployment rate and a low number of drug deaths each year. Meanwhile, West Virginia has had a higher average unemployment rate and a significantly higher number of drug deaths per year. The two maps depict unemployment rates compared to drug overdoses, by state. The correlation is not evident considering the states with the worst drug overdoses do not have the highest unemployment. However, the map of 2020 shows many more regions with dark blue and dense stripes, meaning, over time, as unemployment got worse, drug overdoses did as well.

Challenges

- Data was often too general, too specific, or with incompatible units, so calculations were performed to tailor data to the project.
- 2020 statistics were harder to find because the data was not completely available, considering it is fairly recent

Data Sources

- CDC
- Bureau of Labor Statistics

Either data was taken directly, or calculations were performed.

Recommendations

When unemployment skyrocketed from 2019 to 2020, South Dakota and New Hampshire were the only two states without drastic overdoses. Why? Well, both had programs to help the rising unemployment rates; New Hampshire's Employment Security stationed 12 offices to help people find jobs, and South Dakota gave laid-off workers an extra $300 weekly. Because of these preventative measures, NH and SD were the only states with significantly lower overdoses. After finding the correlation between unemployment rates and overdoses, we recommend that other states implement similar programs to help manage these issues too.

Conclusion

After deep analysis it can be determined that the original hypothesis is somewhat correct. Over any given time period, fluctuation in unemployment correlates with fluctuation in drug overdoses. Statistically, as one increases the other appears to as well. When looking state by state though, the percentage of unemployed citizens and the number of drug overdoses seem to stand apart from one another and no direct pattern is indicated.

Availability of Inclusive Parks & ADA Accessibility Within the City of Pittsburgh

Hampton High School

Addison Gindlesperger, Kiana Kazemi, Lindsay Liebro, Eileen Lin, Becky Zhou, Gary Farrell, & Aaron Peng

Introduction

We noticed in our own area, our playgrounds are not user-friendly for people with mobility issues, even though some of them do meet ADA requirements. This observation led us to question how accessible other playgrounds are throughout the city of Pittsburgh, and what can be done to improve accessibility. We believe that lower income neighborhoods and those with higher populations of children lack playgrounds, especially those that are ADA compliant. The factors we will be considering in our research are median income, population of children (under 18 years old), and ADA compliance of Pittsburgh playgrounds. After examining the ADA requirements for accessibility in playgrounds (summarized below), we arrived at our research question:

To what extent is there a correlation between neighborhood demographics, median income, and accessibility within the playgrounds maintained by the City of Pittsburgh?

A major part of childhood is playing outside with your friends at parks and playgrounds. With new developments and advancements in technology and accessibility for children, upgraded playgrounds are a must in our world. However, many playgrounds in cities are not fully equipped to fulfill every requirement for every child, and some children in the area do not even have access to a nearby playground. We believe that lower income neighborhoods, notably those with a higher percentage of children, lack playgrounds - especially the ones equipped with ADA accessible equipment. Our goal is to look at the city of Pittsburgh neighborhoods and playgrounds to identify the correlation between neighborhood demographics, median income, and accessibility within the playgrounds maintained by the city of Pittsburgh. In addition, we want to provide feedback and improvements that will make all of the playgrounds available and accessible for children of all needs in the city of Pittsburgh. Many parks and recreation areas are receiving millions of dollars in grants, with many planning on using these funds to update and improve playgrounds with ADA accessibility (Himler, 2022). In recent years, people's interest in the relationship between green space exposure and children's mental health has risen, and as well as the association between green space exposure and children's behavioral and emotional difficulties (Vanaken, Danckaerts 2018). With the increased funds and knowledge, we believe parks can take recommendations and improve their current offerings for all children, including those with disabilities..

Methodology

Our chosen data sets included:

- Playground Equipment. -(from WPRDC) to determine ADA Accessible equipment per playground
- Age and Sex in Pittsburgh, Allegheny County, Pennsylvania- to determine the population under 18 in each neighborhood
- Household Income in Pittsburgh, Pennsylvania- to determine the median income in each neighborhood

After gathering our data sets and combining them into one spreadsheet, we began our analysis. We used descriptive statistics to learn more about our sample population. This included finding the percentage of playgrounds in the city of Pittsburgh that are ADA accessible, the average median income for the neighborhoods included in the data set, and the percent of the population that in each neighborhood under the age of 18. To fully answer our research question, we used inferential statistics, specifically regression analysis to look for any relationship between ADA accessibility and the median income or percent of residents under 18. Our final step was to make recommendations in solving this problem in Pittsburgh based on the analysis of our findings.

Limitations

The first limitation of this project was the data set itself. While the data from the City of Pittsburgh did include great detail about the parks in terms of playground accessibility (number of pieces of equipment, type of playground surface) and their locations, it did not include any data about the neighborhoods around the parks. Thus, we had to analyze 3 different datasets for (equipment, median income and population under 18) to have a thorough understanding of the answer to our research question.

For a small number of neighborhoods, the data on median income or the percent of the population under 18 was not included in the larger data sets for these variables. Since we had to find this information from alternate sources, there is always a possibility of inaccurate reporting. In addition, the data set included the playgrounds in every single neighborhood in the Pittsburgh area, which meant the data set had to be cleaned to remove variables not relevant to our research question..

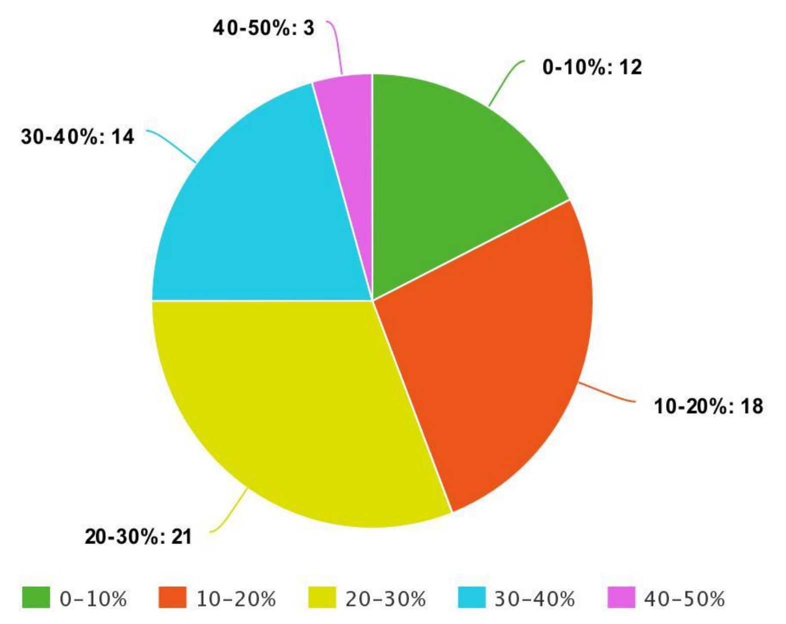
Our research question covered too many aspects of the issue, and it was revised multiple times, which reduced the amount of time we had to conduct our data analysis. We realize that our analysis may have been stronger if we had better focused our research question earlier in the project timeline.

Results

Within the City of Pittsburgh, we analyzed 68 neighborhoods, which had a total of 123 playgrounds and a total of 445 pieces of playground equipment. Looking at the percentage of ADA accessible equipment in neighborhoods, we see there are no neighborhoods with greater than 50% accessibility with 20-30% accessibility being most common (Figure 1). Additionally, only 21.35% of equipment across all parks are ADA accessible, and the only type of ADA accessible equipment is swings (Figure 2). Tables 1a and 1b highlight the stark differences in ratios between parks to kids between neighborhoods. Looking at income across the 68 analyzed neighborhoods, the average median income was $45,872.99, with 54.83% being above the average income and 45.17% below the average income. The highest median income was $120,504 (Squirrel Hill North) and the lowest median income was $18,910 (Crawford-Roberts). Throughout the 68 analyzed neighborhoods, the mean percentage of the population under the age of 18 is 18.7%. The highest extreme is 53.4% of the population below the age of 18 is Manchester. The lowest extreme is 0% of the population below 18, in the neighborhood of West End which has a population of 254. Finally, we ran a regression analysis for ADA accessibility and median income, and ADA accessibility and population under 18. Running this analysis, we found that there was no statistically significant relationship between ADA accessibility and median income (P=0.7111) and ADA accessibility and population under 18 (P=0.8880).

Figure 1.

Counts of Neighborhoods within Ranges of ADA Accessibility Percentage



Note. This pie chart depicts how many neighborhoods in the City of Pittsburgh have a certain percentage of ADA accessible equipment in their playgrounds.

Table 1a.

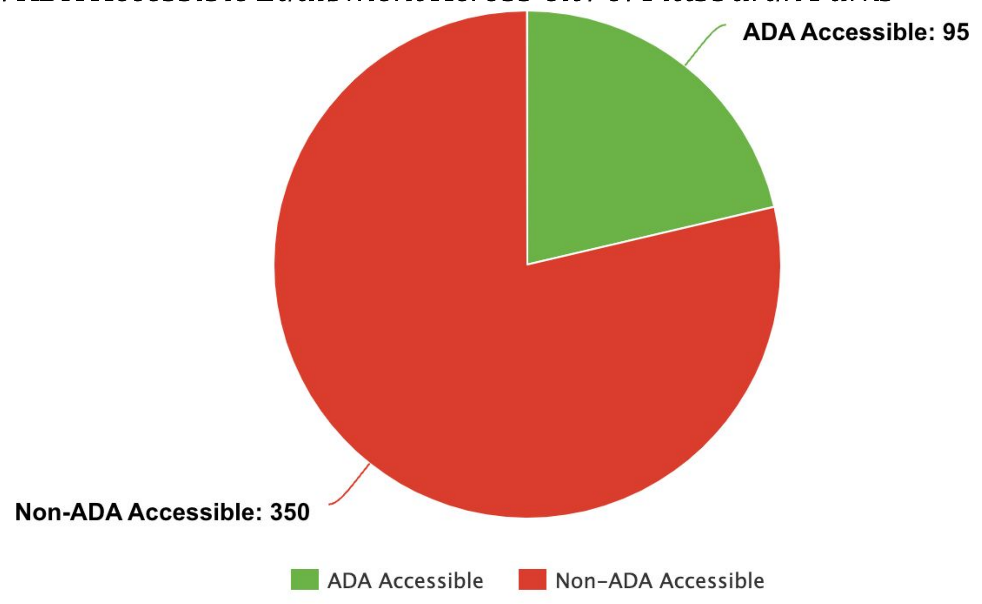
Ratios of Pittsburgh Parks to Persons Under 18 by Neighborhood (Highest Ratios)

Neighborhood	Ratio of Parks to Persons Under 18
Larimer	1 : 50.9
Bedford Dwellings	1 : 48.29
Homewood West	1 : 48.1
Homewood North	1 : 35.1
Allentown	1 : 30
Manchester	1 : 26.7
Spring Hill-City View	1 : 26
Point Breeze North	1 : 24.4
Westwood	1 : 22.8
Bon Air	1 : 21.9

Note. This table depicts the top ten neighborhoods with the highest parks to persons under 18 ratio.

Figure 2.

Number of ADA Accessible Equipment Across City of Pittsburgh Parks



Note. This pie chart depicts the number of ADA versus non-ADA accessible equipment in parks throughout the City of Pittsburgh.

Table 1b.

Ratios of Pittsburgh Parks to Persons Under 18 by Neighborhood (Lowest Ratios)

Neighborhood	Ratio of Parks to Persons Under 18
South Side Slopes	1 : 1.65
South Side Flats	1 : 1.75
Bluff	1 : 1.8
Squirrel Hill South	1 : 1.93
Central Oakland	1 : 2.1
South Oakland	1 : 2.67
Strip District	1 : 3
Beechview	1 : 3.32
Mount Washington	1 : 3.9
Bloomfield	1 : 4.5

Note. This table depicts ten neighborhoods with the lowest parks to persons under 18 ratio.

Conclusions & Recommendations

When looking at the connection between median income and accessibility, there is no significant relationship. However, it can be inferred that there is a connection within how much taxes/money is being used to build playgrounds in lower income neighborhoods. In our analysis of the data selected, we chose to focus on median income, accessibility of ADA accessible, and the percentage of population under 18 in the select 68 neighborhoods in Pittsburgh. There were increasingly more playgrounds without ADA accessible equipment compared to those with. Another significant finding was that the only equipment identified as ADA accessible in all of the parks was swings. While our goal is to increase ADA accessible equipment and playgrounds in Pittsburgh, but rather make areas of joy and play more user friendly for our younger population.

Based on our analysis, and research question, we would make these recommendations to the City of Pittsburgh. Increase the variability of ADA accessible equipment in the playgrounds. Playworld, an ADA accessible playground equipment organization, provides different types of ADA accessible equipment that makes parks user friendly as well. We also recommend that the city use budget funds not just for repairs of sidewalk gaps and ADA ramps, but adding more equipment where there are more kids for one playground. According to the new budget and five year plan created for the city, the Department of Mobility and Infrastructure will install several significant sidewalk gap projects and continue to update ADA ramps to modern standards. In this plan, the budget for Mobility and Infrastructure increased 11.4%. Using this larger budget to improve the playgrounds in the city of Pittsburgh and making them more user friendly to the population. A final recommendation, would be to partner with the Pittsburgh Penguins, Pittsburgh Pirates, and Pittsburgh Steelers to help build or improve ADA accessibility in parks around the city. With their help, the city could build parks, baseball fields, football field, and possibly even deck hockey arenas that are both ADA accessible and user friendly.

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
- Looking at future research questions related to ADA accessible equipment
- Determine what meets ADA accessible standards verus what is user friendly for the community

References

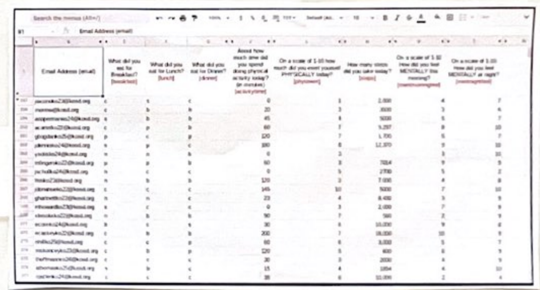
- *Accessible and inclusive playgrounds: where all kids can play.* (2021, August 27). Pittsburgh Parent. https://www.pittsburghparent.com/accessible-and-inclusive-playgrounds-where-all-kids-can-play/
- Accessible Playground Equipment. (n.d.). Retrieved from https://playworld.com/accessible-playgrounds#gref
- Himler, T. L. A. J. (2022). *Recreation, conservation projects in A-K Valley, Westmoreland County to get state grants.* TribLIVE.com. http://triblive.com/local/valley-news-dispatch/recreation-conservation-projects-in-a-k-valley-westmoreland-county-to-get-ion-in-grants/
- *Playground Equipment - Playground Equipment - WPRDC.* (n.d.). Data.wprdc.org. Retrieved April 5, 2022, from https://data.wprdc.org/dataset/playground-equipment/resource/e39ef76e-0a11-47c8-a86f-a37f55db7a2b
- Peduto. (2021, November 8). City of Pittsburgh Budget. Pittsburgh PA Document. Retrieved April 7, 2022, from https://apps.pittsburghpa.gov/redtail/images/16355_2022_Operating_Budget_-_November.pdf
- Stanton-Chapman, T. L., & Schmidt, E. L. (2019). Building Playgrounds for Children of All Abilities: Legal Requirements and Professional Recommendations. *Early Childhood Education Journal.* https://doi.org/10.1007/s10643-019-00947-3
- *The Demographic Statistical Atlas of the United States - Statistical Atlas. (2020). Statisticalatlas.com. https://statisticalatlas.com/place/Pennsylvania/Pittsburgh/Household-Income*
- *The Demographic Statistical Atlas of the United States - Statistical Atlas.* (n.d.). Statisticalatlas.com. Retrieved April 5, 2022, from https://statisticalatlas.com/county-subdivision/Pennsylvania/Allegheny-County/Pittsburgh/Age-and-Sex#figure/neighborhood
- U. S. Access Board. (2005). Accessible Play Areas. Retrieved January 17, 2022, from https://www.access-board.gov/files/ada/guides/play-areas.pdf
- Vanaken, G.-J., & Danckaerts, M. (2018). Impact of Green Space Exposure on Children's and Adolescents' Mental Health: A Systematic Review. *International Journal of Environmental Research and Public Health, 15*(12), 2668. https://doi.org/10.3390/ijerph15122668

moodMent()

Keystone Oaks High School
Greyson Barsotti, Robin Cesario, and Abriana Lyda

Research question: Does a particular diet and an increase in exercise have an association with improved mental health?

Dataset: we collected our own data by sending a Google Form to students and teachers in our High School. We asked questions about what they had to eat, the number of steps taken, and their mood. We had an incentive to fill out the form by picking a random person every week to receive an Amazon gift card.



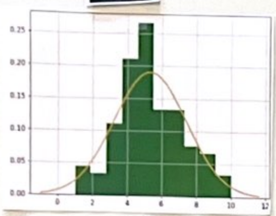
Hypothesis: We hypothesize that eating a balanced breakfast has a positive correlation with your mental health and amount of physical exertion through the day.

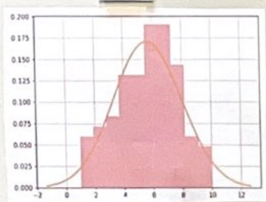
What Food Group They Ate For Breakfast

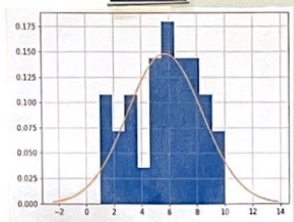
Nothing | Sugar | Balanced

Physical Exertion

Whole Day



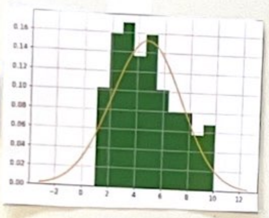


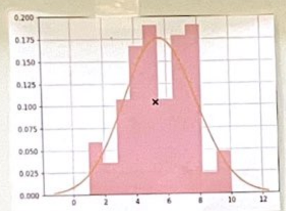


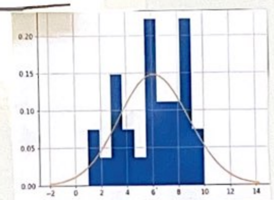
These graphs show levels of physical exertion depending on the meal they ate for breakfast. We see that both the people who ate nothing or sugar have a bimodal distrbution, except nothing has a slight right skew while sugar has a slight left skew. This tells us that people who ate sugar are more likely to physically exert themselves more than those who ate nothing. We also see that people who ate a balanced breakfast had a graph that was bimodal and left skewed, meaning they were likely to feel really good or really bad, but tended to feel better. Due to the median of balanced and sugar being higher than the people who ate nothing, we can come to the conclusion that by eating sugar or a balanced meal you are more likely to raise your level of physical exertion compared to eating nothing for the day.

Mental Feeling

Morning

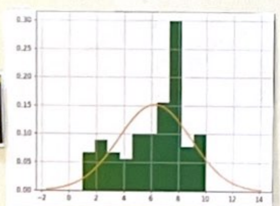


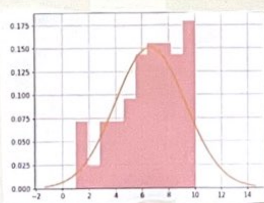


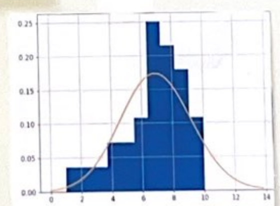


These graphs show mental feeling in the morning based on the meal they ate for breakfast. The graph of nothing shows a right skew meaning on average they felt worse after eating breakfast. Sugar has a bimodal shape meaning they were likely to feel pretty good or pretty bad after breakfast. The people who ate a balanced breakfast were all over the place even though the graph had a left skew. This means they are likely to feel a wide range of ways, but tend to feel better than they did worse. Looking at the medians, balanced had the highest average feeling while nothing had the lowest average feeling. We can conclude that by eating a balanced breakfast, you are more likely to feel better throughout the day.

Evening







These graphs show the mental feeling at night based on what they ate for breakfast. In the graph of people who ate nothing, we see that it is left skewed. This means that most people had a higher chance to feel at around an 8/10 while others felt a wide range of ways on average. For the people who ate sugar or a balanced breakfast, we see a large left skew, meaning that they, on average, felt much better at night. If you look at the medians, they are all very close together, but balanced is the highest. This means that while eating a balanced breakfast tends to make you feel better at night, breakfast typically didn't have much of an effect on your feeling at night.

Summary and Recommendation: In all of the graphs, we see that eating a balanced breakfast compared to nothing or a sugary breakfast demonstrated, on average, a higher median physical exertion and mental feeling throughout the day. Based on these results, we recommend that you eat a breakfast with a wide range of food groups in order to increase your likelihood of maximizing your physical exertion levels and your mental health during the day.

Challenges: During our process we ran into many challenges. Our first issue was cleaning the data. Many people put commas in their step count meaning the code treated them as strings instead of integers. We used the median step count to fill all numerical values that were left blank. Another challenge we faced was what exactly to do with the dataset. We ended up making many scatterplots, box plots, and histograms to represent the data, then manually compared them.

Doom Scrolling

Montour High School

Maddie Rimbey, Evan Witcop, Jordyn Seibel, Suzy Safko, Aidan Ferry, Allison Schindehette

Question

Does Screen Time Increase Rates of Domestic Violence?

Introduction

Throughout the years the usage of portable technology has increased in all age groups. With this increase of portable devices, more and more people are interacting with their screens on a daily basis. It is understandable that the threat of violence can be a driving force of fear in our society and at the same time introduction of portable technology, results in people spending more time interacting with their screens and less time interacting with people offline.

Hypothesis: Increased usage of portable technology across all age groups has led to a decrease in crime rates due to an increased amount of time spent using personal technology.

Challenges

- The original idea was to focus on domestic teen violence but enough domestic violence among teens data with corresponding screen usage for all states could not be found .
- Absence of information by demographic, therefore switched to state and gender.
- Multiyear data for screen time by state was unavailable.
- Several team members dropped out during the planning stages and team cohesion was low, at first, but as the year went along began working well together.
- Considered changing the topic halfway through.

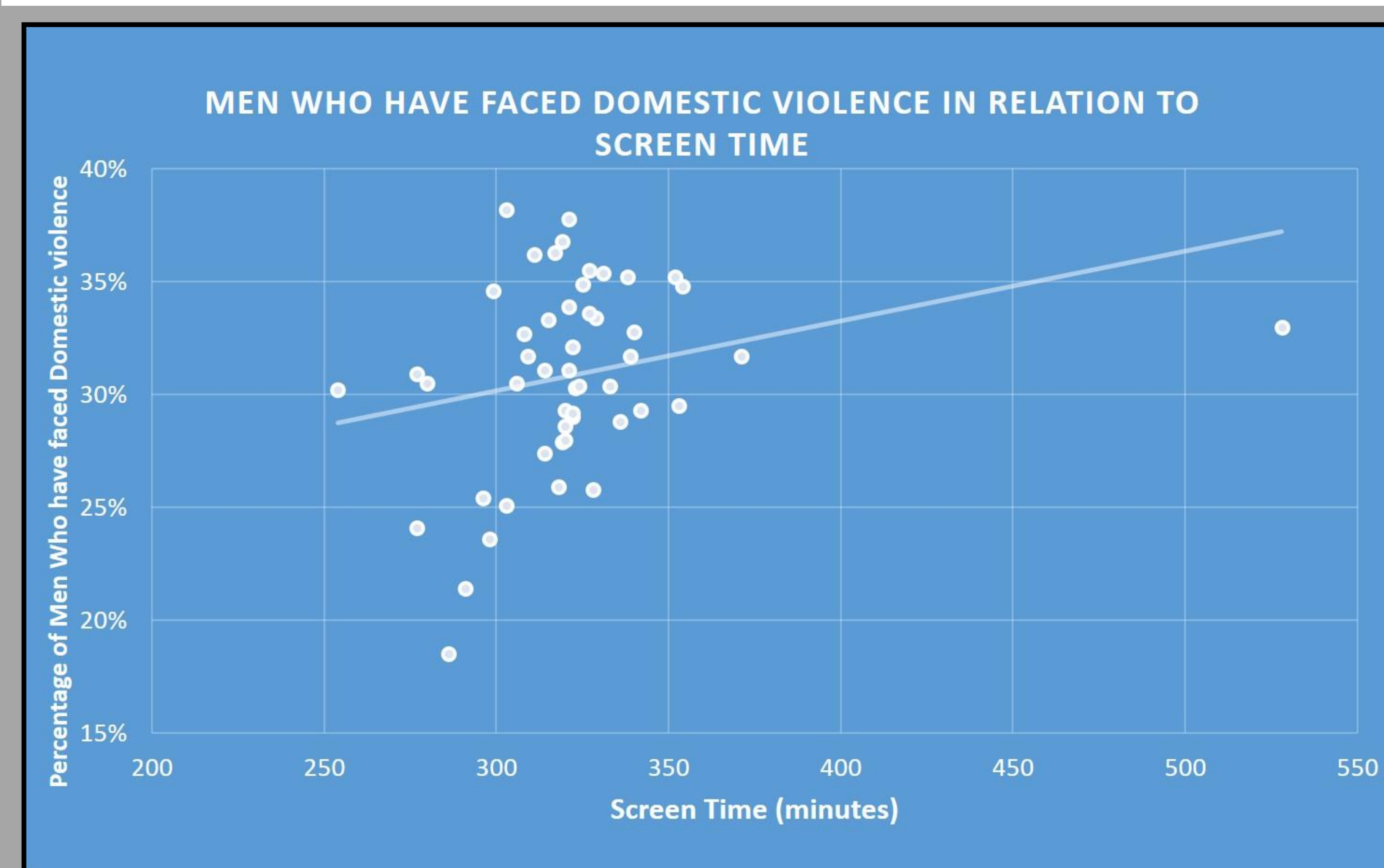
Data Set

- The data used in this case study was from 2017.
- On average 37% of women faced domestic violence and 31% of men. People spent 321.1 minutes on mobile devices.
- We obtained 2017 time spent on smartphones in minutes by state.
- With the help of Simple Texting we were able to find the time spent on mobile devices during COVID in minutes by state but this data was not used.
- Using the database built by National Coalition Against Domestic Violence we were able to find all the data about percentages of domestic abuse by state.

Sources

https://elitecontentmarketer.com/screen-time-statistics/
https://go.verizon.com/resources/teens-with-the-most-screen-time-per-state/
https://simpletexting.com/screen-time-survey/

Visuals



Findings

The correlation coefficient, which measures the strength of the relationship between screen time and domestic violence, identifies a relationship, though rather weak, with men at (0.1945) and women at (0.2764). In addition, when examining the graphs of the relationship, we can identify a clear cluster sitting around the averages for men who faced domestic violence at an average of (31%) and women at (37%) with a corresponding screen time average of approximately 325 minutes. Upon further investigation, one can see that the trend points to a prediction relationship revealing that as screen time increases, so does the rate of domestic violence for men and women. Finally, it appears that there is a stronger relationship for women when considering domestic violence and screen time.

Conclusion

Since the invention of portable technology, screen time has drastically changed the way we live our daily lives. In this project, our hypothesis was centered around the possibility that increased screen time, which has become a popular interest/hobby, could result in an overall reduction in the amount of domestic violence. To examine this hypothesis, we first looked at the amount of screen time, in minutes, per state as well as the percentages of men and women per state who faced domestic violence. Looking at our data, the connection between domestic violence and screen time is evident, but seemingly weak. The data revealed that there was a stronger relationship for the women and screen time and violence compared to the men. However, our case study solely focused on 2017. We would recommend, in the future, to look at data for women's screen time and domestic violence rates from multiple years and compare the data using a time study. Considering the law of large numbers, it would be possible to contextualize our findings within a broader scope of information and see an overall trend. A further study could also include the separation of crimes committed by men vs. crimes committed by women, due to men having committed more crimes statistically. All things considered, our data shows that there is a link, while small, between screen time and domestic violence, and this connection could be further explored by looking at a larger set of data spanning multiple years.

Analyzing the Legacy of Redlining in Pittsburgh

How has redlining in the 1940s affected demographics, housing, finances, and education in modern-day Pittsburgh?

Norwin High School

Aaron Berger • Dmitri Berger • Rex Wu

Background

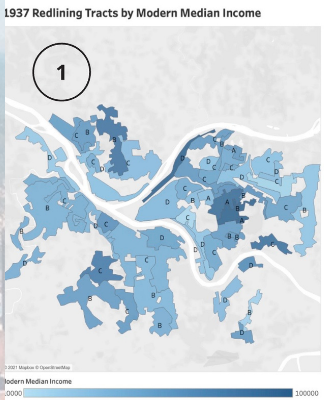
- Redlining was a grading technique used by banks during the early-mid 20th century.
- It was used to identify which areas would be favorable for loans, and which would not; we refer to areas that were redlined as 'redlining tracts'.
- Crucially, racial and ethnic background was a key consideration, with bankers giving lower 'grades' to areas with minority groups.

Hypothesis

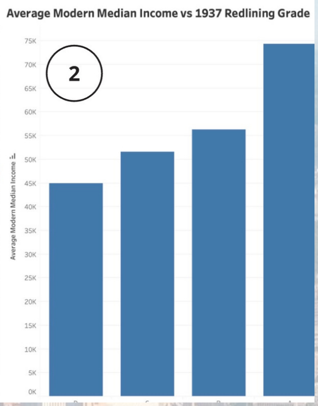
- We expect to find that the majority of formerly heavily redlined areas will be worse off in housing, finances, and education than areas that were not as severely target
- This hypothesis is partially based on our previous project, in which we found that redlining grades correlated with modern median income (MMI).

Dataset Descriptions

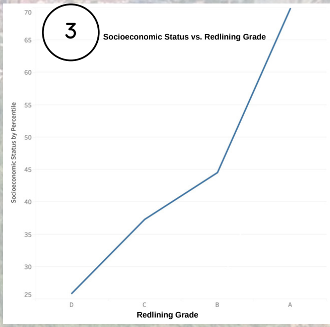
- Census Tract Data:An interactive map of census tracts, including Pittsburgh, with data on demographics, finances, housing, and educational attainment by for each tract. It was the primary source for our data.
- 2010 Census Trap Map: An index of maps of census tracts in Pittsburgh from the 2010 census. We used it to compare redlining tracts with census tracts, and to that extent to create spatial graphics.
- Mapping Inequality Redlining Map: A 1937 map of Pittsburgh, divided into redlining tracts by bankers. The tracts were color coded by the redlining 'grade' A through D, where A was 'most desirable' for loaning and D was 'least desirable'. The bankers determined the grade of an area based on factors like industry, age, wealth, and demographics, discriminating against ethnic or religious minorities and against immigrants. We used it to identify how certain areas were graded, and to match those grades onto modern census tracts.



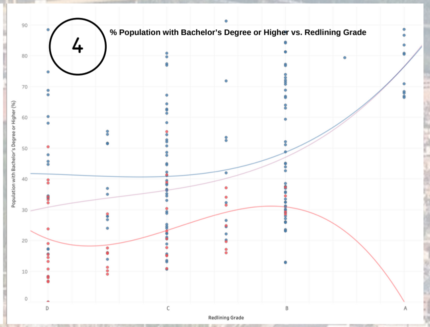
Graphic 1 depicts shows a map of Pittsburgh organized as it was by the bankers, in redlining tracts. Each tract has a grade from 1937 and a color, with darker blue representing a higher MMI. Most of these dark blue tracts are A and B tracts, while the tracts with lower grades have lower MMI.



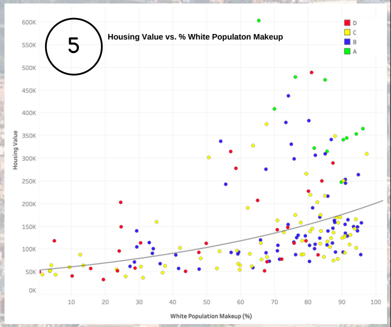
Graphic 2 depicts shows the average MMI of the redlining tracts, organized by grade There is a clear upward trend, with A-grade tracts possessing a significantly higher average MMI than C-grade and D-grade tracts.



Graphic 3 is a multivariate factor analysis of the redlining grade that the area a census tract encompasses was assigned and socioeconomic status as a percentile, which is a general measure of MMI, housing value, and academic achievement, all of which are generally very correlated with one another. The graph shows how inhabitants of B-grade and A-grade census tracts on average have a higher socioeconomic status than inhabitants of C-grade and D-grade census tracts.



Graphic 4 is a multivariate analysis of the redlining grade and average academic achievement of a census tract that, given the aforementioned information, predicts that likelihood of the district being majority-white or not. On the graph there are three trend lines for the data points; the red one shows the general educational achievement of non-majority-white census tracts, while the blue one shows the same for majority-white census tracts. Notably, educational achievement in majority-white tracts is considerably higher.The central purple line is the average between these two, and predicts the likelihood of a district being majority white. The farther above the line, the more likely the district is majority-white, and vise versa.



Graphic 5 is a multivariate analysis of housing value and demographic makeup. The color of each point indicates what grade the region that's the tract enoompasses was given in 1937. Notably, most of the red D-grade tracts and yellow C-grade tracts are confined to the lower half of the graph, indicating that on average property values in these areas are lower. Additionally, most of the B-grade tracts are kept to the right side of the graph, indicating low minority populations. Almost all of the A-grade tracts are in the top right; they are majority-white and possess considerable property value on average.

Challenges

- Finding organized, usable data from the 1930s and 40s on Pittsburgh and redlining in Pittsburgh for comparison with modern data
- Comparing data of modern tracts to data from the 1900s for differently-shaped tracts, and converting data—particularly modern median income—between the differently-shaped tracts.
- Manually taking redlining grades from redlining tracts and assigning them to modern census tracts within their region, and making mixed grades if needed.
- Creating a web crawler to automatically access sort through, compile, and organize the datac we needed census tract by census tract, and identifying a consistent method to extract data from our sources.

Summary

We found that there is a clear, distinct correlation between redlining grades and finances, housing, demographics, and education. Redlining in the 1940s devalued areas and economically disadvantaged inhabitants of those areas.

- Graphic 1 shows how for the most part only redlining tracts with higher grades had a higher corresponding MMI.
- Graphic 2 shows that on average redlining tracts with higher grades had a higher corresponding MMI.
- Graphic 3 shows that in general, inhabitants of census tracts in areas graded poorly during redlining see a lower socioeconomic status than areas graded favorably.
- Graphic 4 shows how in general tracts with minority populations saw lower redlining grades and lower educational achievement.
- Graphic 5 shows how generally D and C-grade census tracts have lower housing value and larger minority populations.

Reccomendations

Redlining, historically, has been shown to significantly disadvantage minority populations by devaluing areas with such populations, economically discriminating against them. The impacts of this are still visible today. We recommend that areas that had been redlined unfavorably in the past receive grants and funding to repair the damage that has built up over time, and economically revitalize the area.

Sources

Census Tract Data: https://censusreporter.org/profiles/14000US42003462100-census-tract-4621-allegheny-pa/
Census Map Index: https://www2.census.gov/geo/maps/dc10map/tract/st42_pa/c42003_allegheny/
Mapping Inequality Redliing Map: https://dsl.richmond.edu/panorama/redlining/#loc=12/40.442/-80.041&city=pittsburgh-pa

A Bikeride A Day Keeps the Doctor Away: An Analysis of the Effect of Bikeshare Ridership on City Health Metrics

By Oakland Catholic's Data Jam Team: Maura Schorr and Róisín Tsang

The Question

What are the potential health impacts of city bikeshare systems on a given city's population?

Null Hypothesis

There is no relationship between bikeshare ridership and obesity levels of a given city.

Alternate Hypothesis

We hypothesize that there is a positive linear relationship between bikeshare ridership and obesity levels of a given city. [given a higher obesity score indicates lower obesity rates]

Figure 1: Healthy Ride Bikes in Pittsburgh U Pitt Parking, Transportation and Services

Pivot table:

Row Labels	Sum of Number of trips controled for population (2020)	Sum of Number of trips controled for population (2021)	Sum of number of trips controled for population (2019)
Austin	0.3074	0.3062	0.3297
Boston	9.081637305	12.89878738	11.1230539
Chattanooga	1.3574	0.6153	1.1137
Chicago	3.7219	5.9286	4.1227
Columbus	0.1536	0.1373	0.1177
Jersey City	3.396089575	6.352440938	4.133120647
Los Angeles	0.1783	0.1764	0.2394
Minneapolis	1.433818502	1.730715379	2.324937086
New York	6.585716574	9.146739905	6.976218596
Philidelphia	1.474982183	1.650597925	1.413667066
Pittsburgh	0.790422186	0.403533672	0.83873374
Portland	0.421257833	0	1.456922037
San Fransico	14.3451	13.5147	17.1119
Washington D.C.	9.307272187	11.55849147	14.60606197
Grand Total	**52.55489634**	**64.41980667**	**65.90781504**

Table 1. : Pivot Table of Cities and the Sum Number of Trips Controlled for Population in 2019, 2020 and 2021

Results

The Logarithmic regression shows that as bike ridership increases, obesity levels rapidly decline, and eventually plateau. This would lead us to believe that a proportion past 6 of the population is riding bikes, there is not much of an effect on the overall Obesity levels. The R-squared value for our model was 0.2755, therefore, about 28% of the variation in obesity levels can be predicted by bike ridership as shown by our logarithmic model. We were able to establish an association, but because of a lack of statistical testing we can not yet say that there is causal relationship.

Challenges

Our raw datasets were difficult to understand how they were arranged straight from the databases, but we tidied and organized to make it more understandable. We originally planned on analyzing our data using a linear regression analysis, but after looking closer we found that may not accurately represent our data. We confirmed this by doing a test for homoscedasticity. There is also the issue the health metrics, and obesity in particular, can be influenced by many things, such a income and genetics, and this makes it difficult to prove that bike riding impacts health metrics of a city.

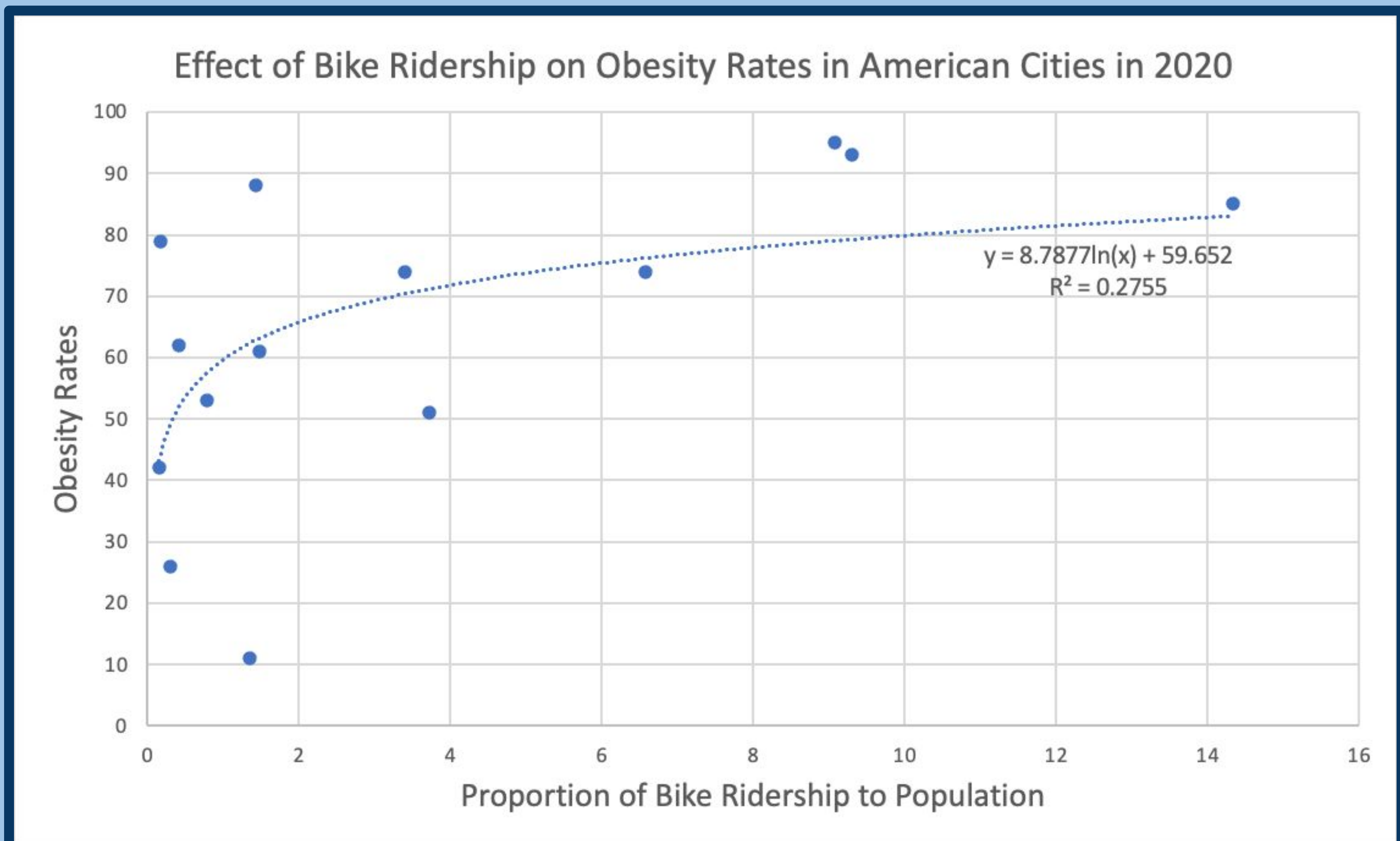


Figure 2: A scatterplot of bike ridership adjusted for population compared to obesity score, and a logarithmic regression analysis with an R^2 value of 0.2755

Recommendations

We believe there is a relationship between bike ridership and obesity levels. Cities who are struggling with high obesity levels and low bike ridership may want to promote bike ridership as a way to tackle obesity levels. Cities who have a population proportion past 6 riding bikes may want to direct funds towards other obesity lowering programs.

Datasets

Bikeshare ridership by city: https://data.bts.gov/Research-and-Statistics/Docked-Bikeshare-Ridership/6cfa-ipzd/data

Most Overweight and Obese American Cities: https://wallethub.com/edu/fattest-cities-in-america/10532

Examining pH Data in the San Luis Rey River within the Pala Native American Reservation and Beyond

Maniya Zwicker and Amara Sanchez, Pala Youth Center

Problem: Our project addressed our question of "How do pH levels differ within various spots in the San Luis Rey River"

Why is it important: We feel that this question is important because water samples measuring specifics such as pH levels can be indicators of environmental stress and we are concerned about the water and land within our reservation.

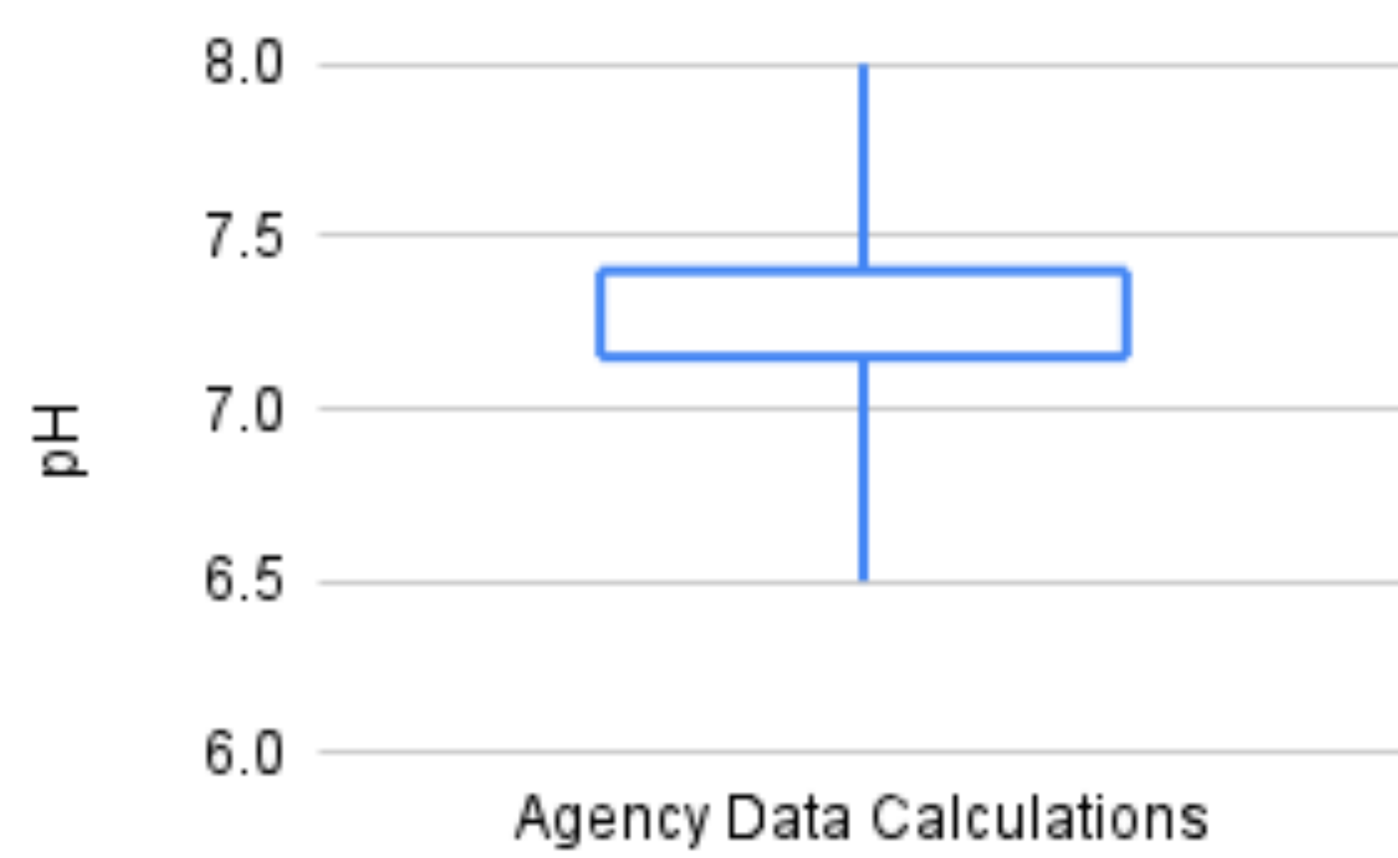
Hypothesis: We believe that the pH levels of the San Luis Rey River water on Pala land are not as healthy as others areas of the River.

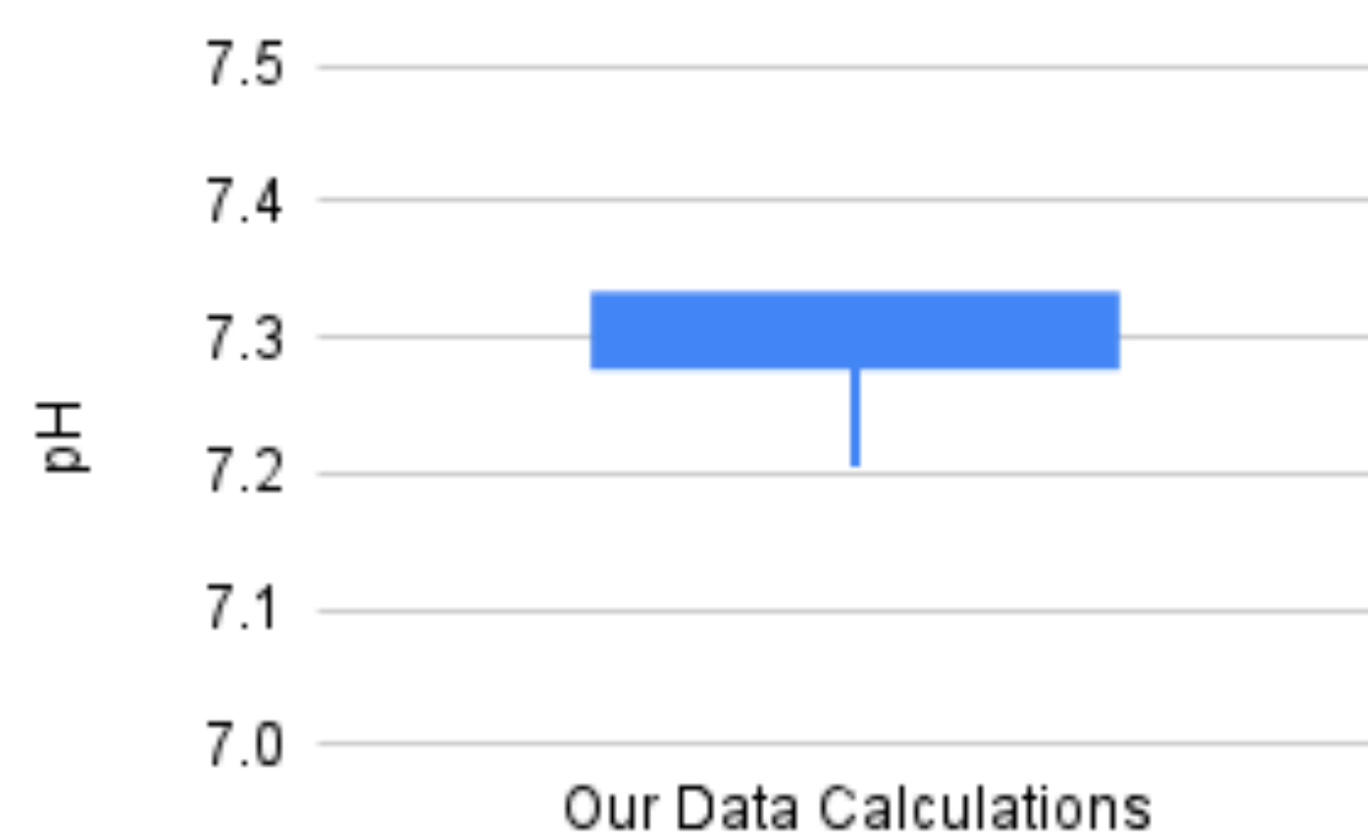
Dataset use: California Natural Resources Agency (CNRA) Data Set

Analysis: We compared local pH readings from the San Luis Ray River on the Pala Native American Indian reservation with those acquired by the CNRA. We found that our local pH samples (three) were close to neutral at 7.23, 7.33, and 7.33. Meanwhile, the CNRA dataset encompassed 101 pH readings from various stations in the San Luis Rey River; these ranged from 7.1-8.6.

Conclusion: The water at the site where we tested is currently healthy; however, we did note (see box charts) that our locally collected samples were more acidic than the CNRA data.

Future Work: We would now like to measure additional attributes at our reservation's site, such as minerals and temperature - to compare those with the CNRA data set at other points in the river.





2010-2016 NJ CRIME RATE COMPARISON
HOW PASSAIC COUNTY CRIME RATES CHANGED

Nicole Llerena, Joshua Sanchez, Jacqueline Flores, Yoselin Felix & Sheyla Vera-Portilla
Passaic Academy for Science and Engineering
Mentor: Nathaniel McDowell
University of Pittsburgh

PROJECT DEVELOPMENT:

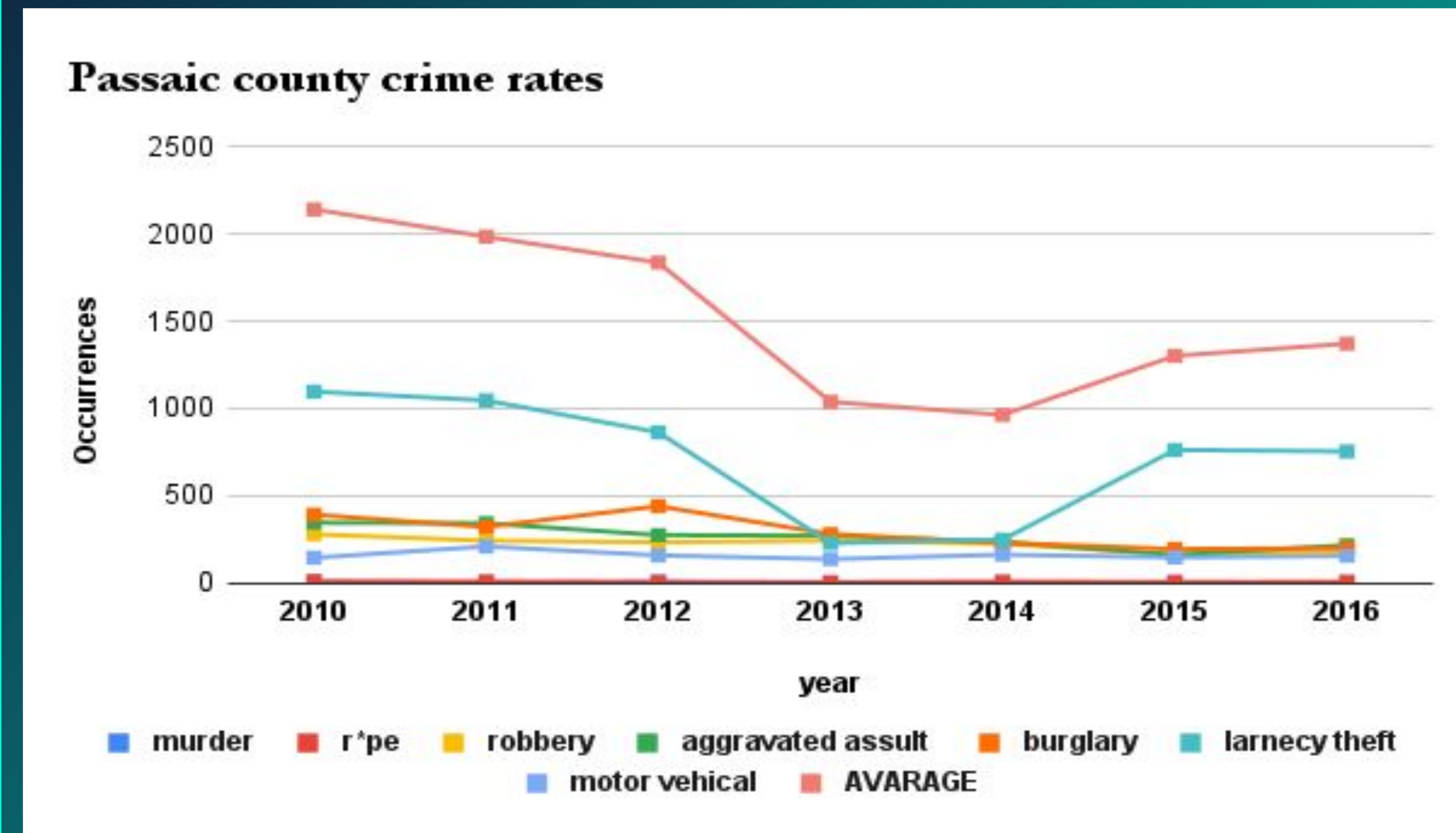
★ Our hypothesis when we started the project was to see that the crime rates were increasing in 2010 and later decreasing in 2016. Our project evolved as we connected it to our question "How Passaic County Crime Rates Change Over Time" We picked this project because it was very interesting to look into. We found it interesting because we wanted to see the development of hate crimes throughout the years and how badly it had became.

DATA:

★ The dataset we used was Google spreadsheet as it was easy to find. No, the data didn't need to be cleaned up. Some challenges we faced were not being able to find the correct website top use to find our data. We split the work so each of us can work with the data, making the graphs, and coming up with what can represent the graph. We made it equal so everyone could participate in the coming of crime rates increasing, decreasing or staying the same. We made a graph to show what crimes increased throughout the 6 years. All the crimes except rape has increased or decreased, since rape stayed constant over the course of 6 years. The limits with our project was the data that was given to us/we found was hard to work with because some of the data didn't work in our favor. Some confounding factors that are embedded in the data that we can't remove are the years and the average, these two cannot be removed because it represents the time it changed and how it was affected.

INTERPRETATIONS & RECOMMENDATIONS:

★ We can make recommendations based on our results is that even though it may seem or you might think that since the year goes by the average will go down but according to our data you can see that the data at some point the crime rates were high but in the years 2011-2014 it was starting to go down but in 2015-2016 it increased again by a bit. This shows that the crime rates will go low at some point but rise back up by a bit after a while.



★ The graph being presented above shows how crime rates decreased,increased, and stayed constant over the years 2010-2016.

REFERENCES:

★ https://docs.google.com/spreadsheets/d/10dF8P_dLoPfu8A8cmIzSUy_pS8tCF Yh6/edit?usp=sharing&ouid=112255094067407695336&rtpof=true&sd=true
★ https://www.nj.gov/lps/njsp/ucr/uniform-crime-reports.shtml
★ https://docs.google.com/document/d/1uzSWMPpS3Le2O89_J1QwNeZYUo_ou MpGlY8nKsEV6OQ/edit?usp=sharing

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

Air Quality Due to the Number of Trees

Ryan Balerio, Lori Perez, Kendrick Herrera, Jordana Del Angel
Passaic Academy for Science and Engineering
Mentors: Anthony Lucchitti
University of Pittsburgh

Research Question

How does the number of trees affect the air quality?

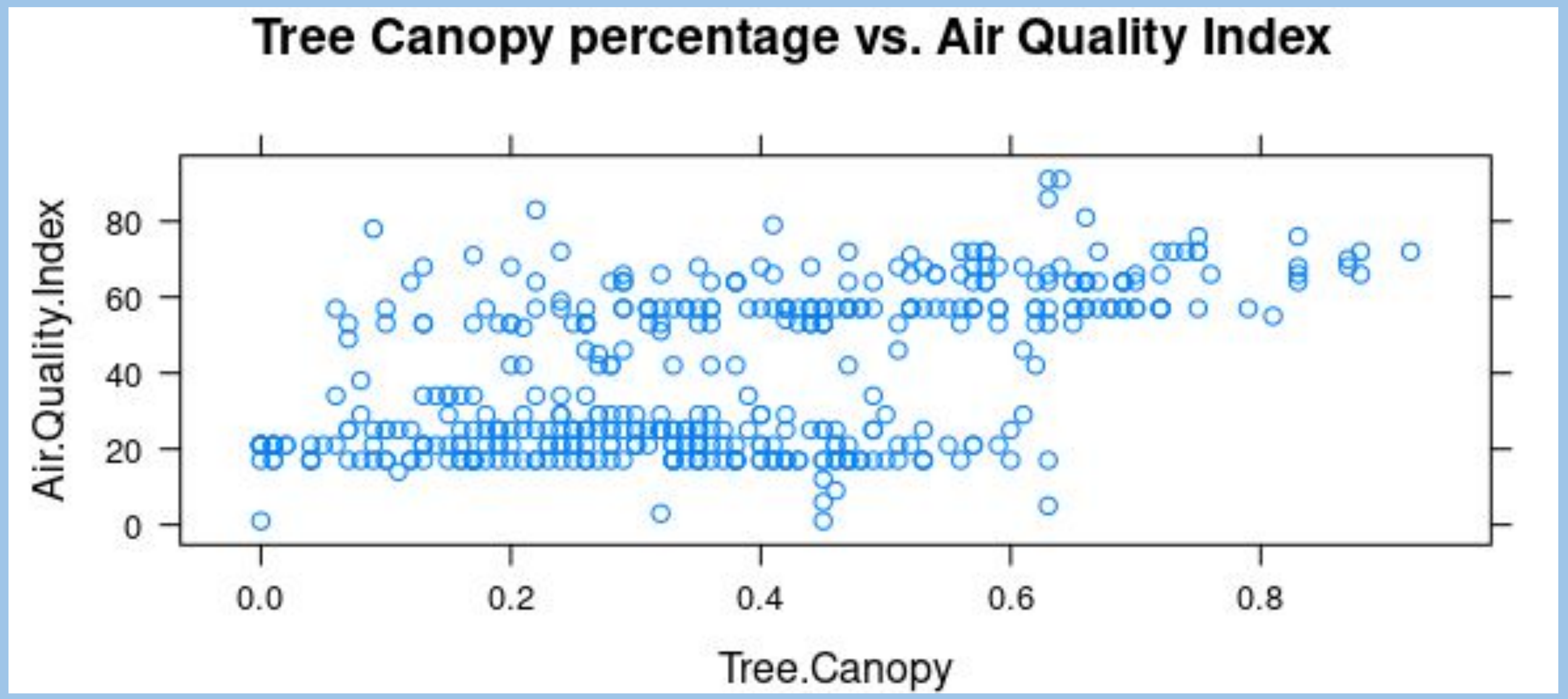
Project Development

- We chose this topic because we wanted to focus on something more environmental, so that it could raise awareness to people about this topic of trees and air.
- The hypothesis was that the air quality would be better because there are more trees since they are part of the green areas that purify the air.
- We find it interesting how important green areas are for improving air quality and yet we don't take them as important as we should.

Data

The datasets we used are about Tree Canopy per county, and Air Quality per county from two different websites one shows the air quality, and the other one contains the tree canopy and thanks to Anthony helping to separate the data. It was easy to find the data but we needed to filter the data only focusing on the tree canopy and air quality in NJ. Some challenges we found in the way of using the data was to separate NJ data from the other states. We limited our data to focus on NJ data for the Tree Canopy and Air Quality..

Findings



This graph shows the Air Quality due to the number of trees in that county but there are two horizontal lines that also show that the Air Quality does not depend that much on the number of trees.

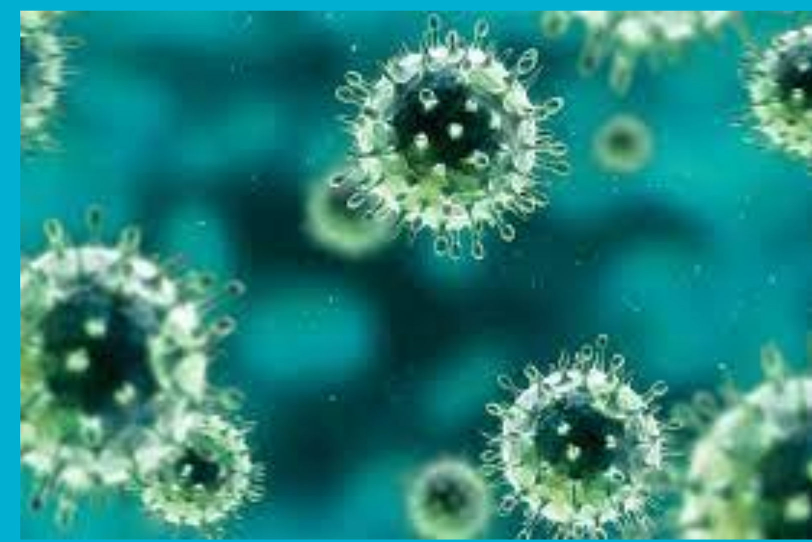
Interpretations and Recommendations

- The better the Air Quality the more trees there are.
- Some recommendations could be that there are reserved areas where trees cannot be cut.
- Apply a fine in case these areas are not respected.
- Further research into the variables affecting air quality

References

- https://www.iqair.com/us/usa/new-jersey
- https://docs.google.com/spreadsheets/d/1eIPv66aqI1SWXFvzlSCSVduka0naw_VYab-xnDiZKMc/edit#gid=0
- https://www.nrs.fs.fed.us/data/urban/state/?state=NJ

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



COVID and Income

Jorge, Alexandre, Cristian, Javier & Emmanuel

Passaic Academy for Science and Engineering

Mentor: Jackson Filosa

University of Pittsburgh

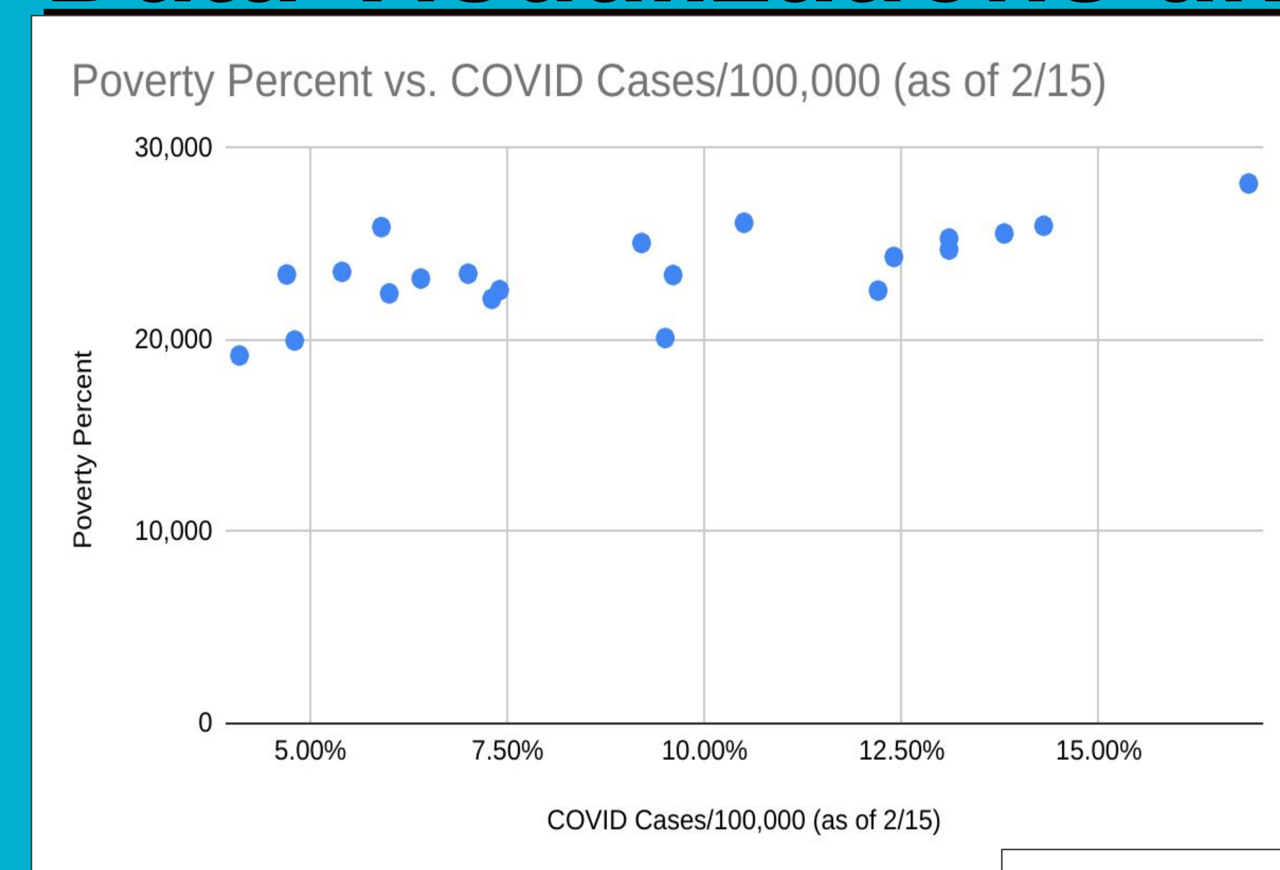
Research Goals:

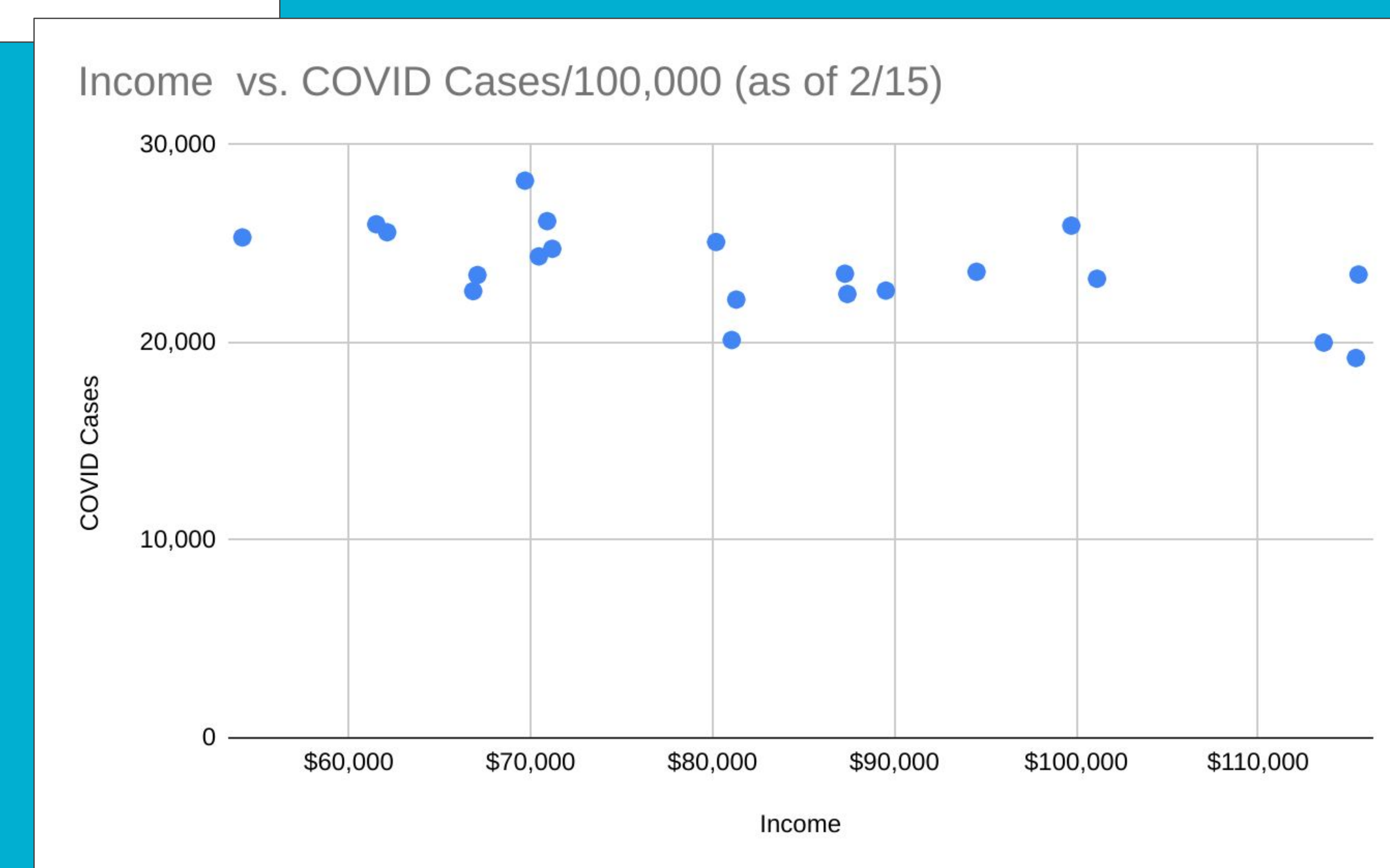
- Hypothesis: There is a negative correlation between Covid Cases and Income, and we predict that higher income families are less likely to get Covid
- Project Goals: To examine COVID cases in New Jersey, see which counties were affected most and how income affected them

Data

- Income and Poverty percentage data from the U.S Census
- Covid Data from the NY Times
- Did the data need to be cleaned up?
  - We had a very hard time finding data, and had to change our research question in order to find reliable data sets
- Considerations of the limits/scope of the data you are using?
  - Yeah it was used because it was better to use proportions to make it easier to collect data and get the coloration an estimate of the data.
- Are there confounding factors that could be in the data that you can't remove?
  - There is no factors that can not be removed that were implanted because everything like the covid and the pevrty and income rely on each other and other factors.

Data Visualizations and Findings





References

New Jersey Median Income and Poverty Percent: https://www.census.gov/quickfacts/

New Jersey COVID Cases by County: https://www.nytimes.com/interactive/2021/us/new-jersey-covid-cases.html

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

Can Federal Funding Flip-Flop the Fossil Fuel Fixation?

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

Has an increase in renewable energy production had a significant impact on the production of energy using fossil fuels?

What steps can our government take to further decrease our reliance on fossil fuels?

Definitions

Renewable Energy → Energy production that doesn't consume the source including wind, solar, hydroelectric, and geothermal power.

Clean Energy → Energy produced that doesn't generate greenhouse gas emissions.

Nuclear Power → Energy produced through nuclear fission; not defined as "Renewable Energy" under US legislation, but is "Clean"

Fossil Fuel → Energy produced by burning products such as oil and coal that formed from dead plants and animals. Energy production in this way leads to global warming and climate change.

Resources

Our main source of data was the **United States Energy Information Administration (EIA)**, which allowed us to collect data regarding energy production from renewable energy, nuclear energy, and fossil fuels.

Additional supplemental data was gathered from a multitude of reliable and academic sources such as **Statista** and **OurWorldInData,** regarding the specifics of the United State's energy investments, energy consumption, and renewable energy generation.

Data

Year	Fossil Fuels	Nuclear Power	Renewable	Total (Quadrillion BTUs)
2005	54.995	8.161	6.221	69.377
2006	55.877	8.215	6.587	70.678
2007	56.369	8.459	6.511	71.338
2008	57.527	8.426	7.192	73.146
2009	56.612	8.355	7.626	72.593
2010	58.159	8.434	8.315	74.909
2011	60.529	8.269	9.31	78.108
2012	62.296	8.062	8.896	79.254
2013	64.184	8.244	9.438	81.866
2014	69.622	8.338	9.798	87.757
2015	70.19	8.337	9.768	88.295
2016	65.43	8.427	10.48	84.337
2017	68.447	8.419	11.263	88.129
2018	75.758	8.438	11.584	95.78
2019	81.354	8.425	11.632	101.437
2020	75.734	8.251	11.687	95.627
2021	77.268	8.129	12.317	97.714

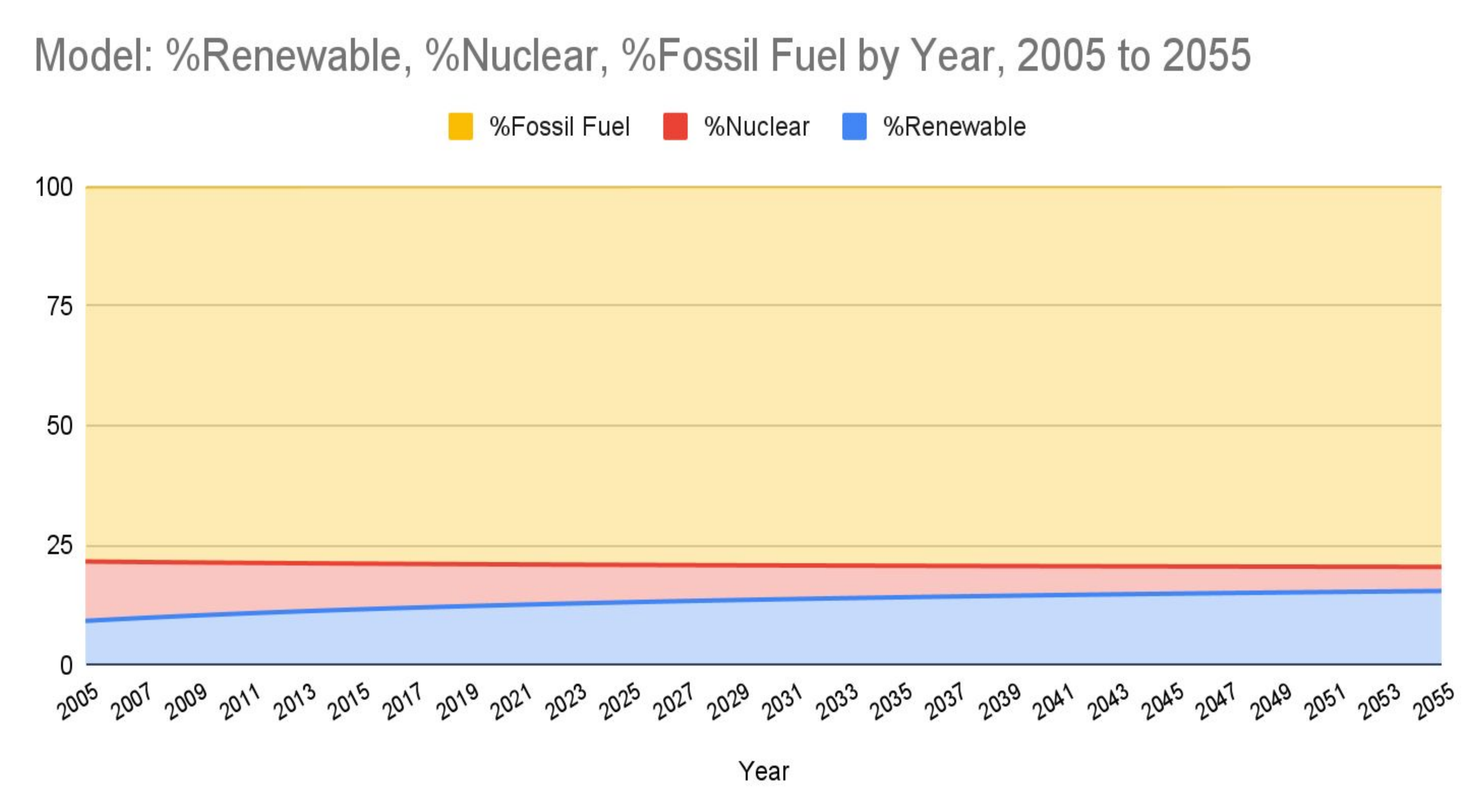
Sample data of energy produced from different sectors

%Change in Renewable Energy Production	%Energy from Fossil Fuels	%Energy from Nuclear
	79.26978682	11.76326448
5.883298505	79.05854721	11.62313591
-1.153787764	79.01679329	11.85763548
10.45922285	78.64681596	11.5194269
6.034482759	77.98548069	11.50937418
9.034880671	77.63953597	11.25899425
11.96632592	77.49398269	10.58662365
-4.446831364	78.60297272	10.17235723
6.092625899	78.40128991	10.07011458
3.814367451	79.33498182	9.501236369
-0.3061849357	79.49487513	9.442210771
7.289107289	77.58160712	9.992055681
7.471374046	77.66682931	9.553041564
2.850039954	79.09584464	8.809772395
0.4143646409	80.20150438	8.305647841
0.4728335626	79.19729783	8.628316271

Percent fossil fuel and renewable energy production have increased

Percent of energy from nuclear fission has decreased

Results



Model for total US energy production:

- Total Energy = 1.99 * X + 67
- Percentage Renewable Energy Generation by Year:

(0.392 * X + 6.19) / (1.99 * X + 67) * 100%

- Percentage Nuclear Energy Generation by Year:

(1.22e-3 * X + 8.31) / (1.99 * X + 67) * 100%

- Percentage Fossil Fuel Energy Generation by Year:

(1.6 * X + 52.5) / (1.99 * X + 67) * 100%

Where X is the number of years since 2005.

As X → Infinity,
%Renewable → 19.70%
%Nuclear → 0.06%
%Fossil Fuel → 80.40%

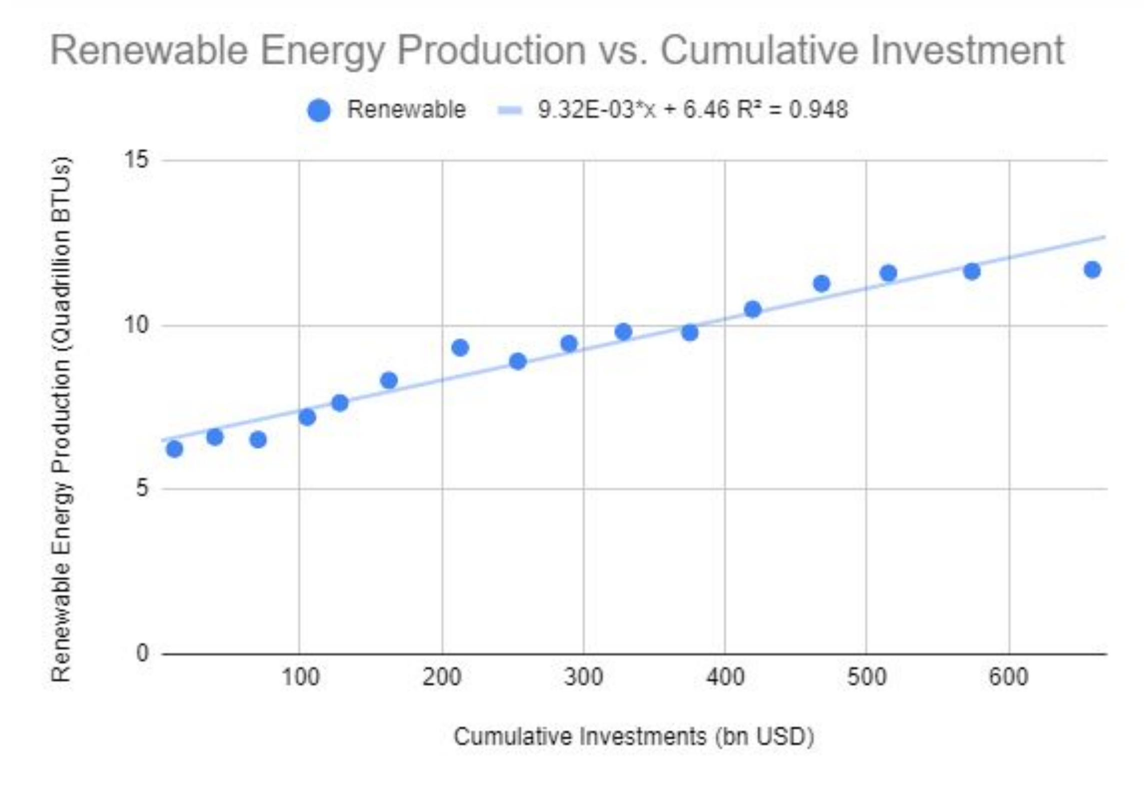
- With current trends, increasing renewable energy production will never cause a significant decrease in fossil fuel usage
- Fossil fuels are a major cause in detrimental climate change
  - ➢ Unless something is done, climate change will continue unhindered

Conclusion

The data collected was the energy productions in various sectors every year in the United States. By extrapolating current trends, it was determined that if there are no significant changes in government policies, energy production from fossil fuels would not decrease past 80%. One proposed solution to this problem is increased investment in renewable energy. However, it was found that the percentage of total energy from renewable sources would level off at about 20% of the total energy. Overall, while increasing investment in renewable energy would be a step in the right direction towards a "clean" future, there are too many confounding variables for it to be considered an absolute solution.

Extension

Reliance on fossil fuels is not decreasing despite increased government investment in renewables, what level of investments would be necessary to decrease this reliance?



Using this graph and looking at total energy productions, an extrapolation model can be used to find a model for the percentage of the total energy production that is from renewable sources as the cumulative investments increase.

- Total Energy vs. Cumulative Investments
  - Total Energy = 0.0485 * Cumulative + 68.1
- Renewable Energy vs. Cumulative Investments
  - Renewable Energy = 9.32e-3 * Cumulative + 6.46
- %Renewable Energy Production per Cumulative Investments
  - (9.32e-3 * Cumulative + 6.46)/(0.0485 * Cumulative + 68.1) * 100%
  - Levels off at about 20% as Cumulative Investments approach infinity

Challenges

- ➔ Choosing an interesting and complex topic relating to current events
- ➔ Learning methods of analyzing data
- ➔ Finding and collecting time series data for our topic
- ➔ Changing the research question based on the availability of data

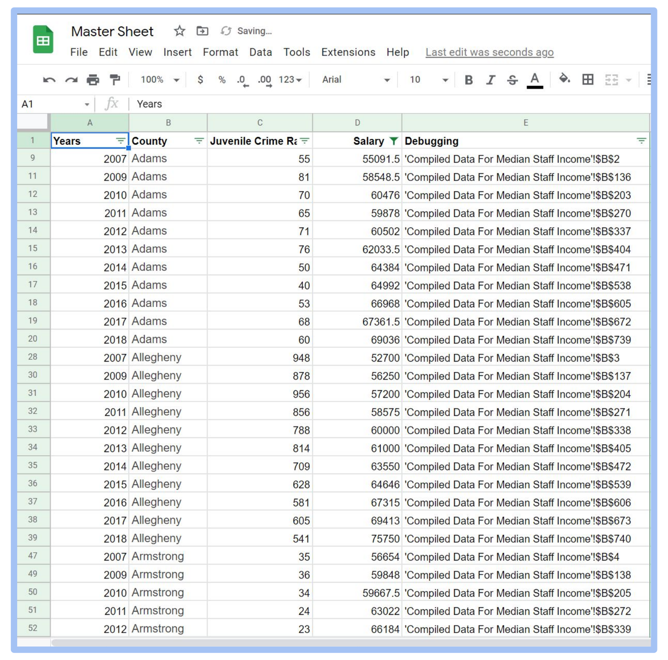
WHAT IS THE RELATIONSHIP BETWEEN SCHOOL STAFF SALARY AND JUVENILE CRIME RATES

Jeffrey Yan, Rudra Thakkar, Dhruv Thakkar, Caroline Madden, Serena Carnahan, Alex George, Carly Beninati
from Plum Senior High School

Key

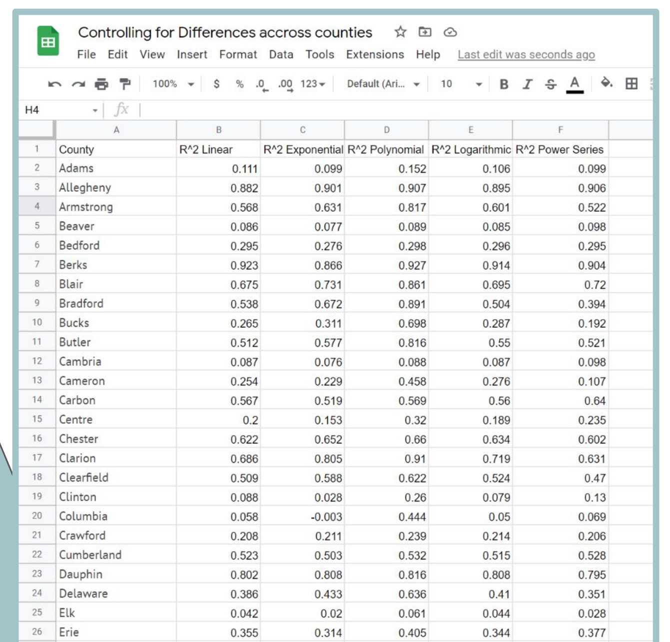
- **Years :** We had access to fourteen years of teacher salaries from 2007-2020, but a few of the years were omitted due to lacking county information. We also had 19 years of juvenile crime data from 2000-2018.
- **County :** The County refers to the county in which the crime was committed, and also the county of the school district in which the teacher teaches.
- **Juvenile Crime :** Juvenile crime refers to the amount of convicted juvenile crime cases within a county during a specific year.
- **Salary :** The salary refers to the median annual salary of all school staff within a specific county during a specific year.
- **R^2 :** The R^2 value refers to the how well the model fits the data.

Data Sets



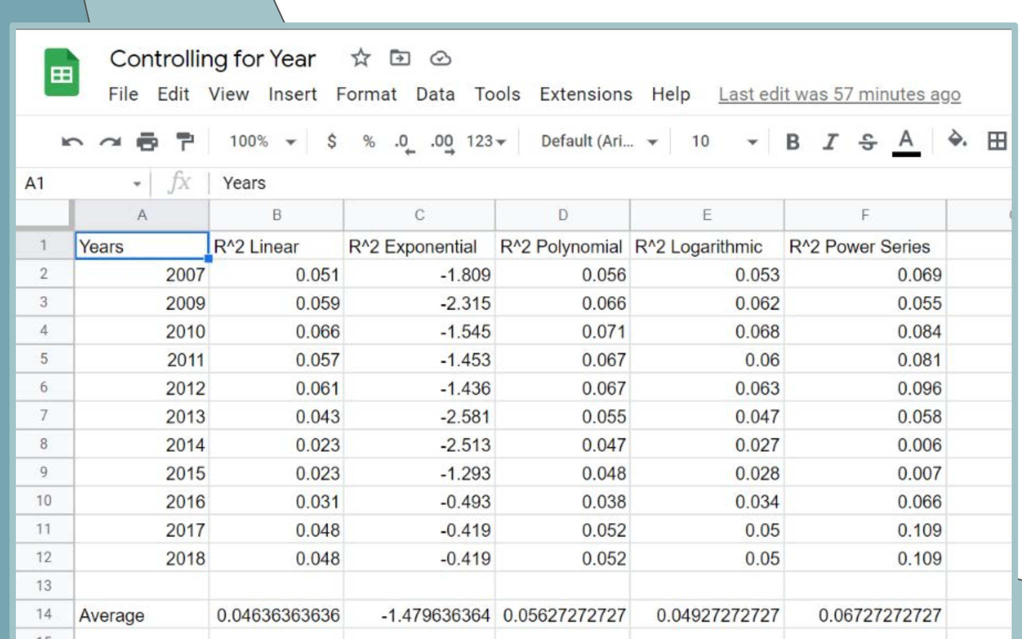
"Master Sheet"

- The "Compiled Data For Median Staff Income" sheet includes the summarized relevant data of staff members from every county and the years that I was able to process properly.
- The "Combination of Crime and Median Staff Income Data" sheet combines the juvenile crime data and staff data so that they can be compared.
- The "Pruned Combo of Crime and Median Staff Income Data" Sheet filters out the years that are missing the relevant staff salary data.



"Controlling for Differences Across Counties"

- The "R^2" Sheet contains various R^2 values for various relationships for multiple years of data.
- The "Master Sheet" data contains the combined data of juvenile crime and staff salary medians copied from the "Master Sheet" spreadsheet .
- The remaining sheets each filter all of the data from the Master Sheet so that it only contains the data for a particular county. Then we plot the data on a graph and find the R^2 values for various relationships.



Controlling for Year

Years	R^2 Linear	R^2 Exponential	R^2 Polynomial	R^2 Logarithmic	R^2 Power Series
2007	0.051	-1.809	0.056	0.053	0.069
2009	0.059	-2.315	0.066	0.062	0.055
2010	0.066	-1.545	0.071	0.068	0.084
2011	0.057	-1.453	0.067	0.06	0.081
2012	0.061	-1.436	0.067	0.063	0.096
2013	0.043	-2.581	0.055	0.047	0.058
2014	0.023	-2.513	0.047	0.027	0.006
2015	0.023	-1.293	0.048	0.028	0.007
2016	0.031	-0.493	0.038	0.034	0.066
2017	0.048	-0.419	0.052	0.05	0.109
2018	0.048	-0.419	0.052	0.05	0.109
Average	0.04636363636	-1.479636364	0.05627272727	0.04927272727	0.06727272727

"Controlling for Year"

- The "R^2" Sheet contains various R^2 values for various relationships for multiple years of data.
- The "Master Sheet" data contains the combined data of juvenile crime and staff salary medians copied from the "Master Sheet" spreadsheet .
- The remaining sheets each filter all of the data from the Master Sheet so that it only contains the data for a particular year. Then we plot the data on a graph and find the R^2 values for various relationships.

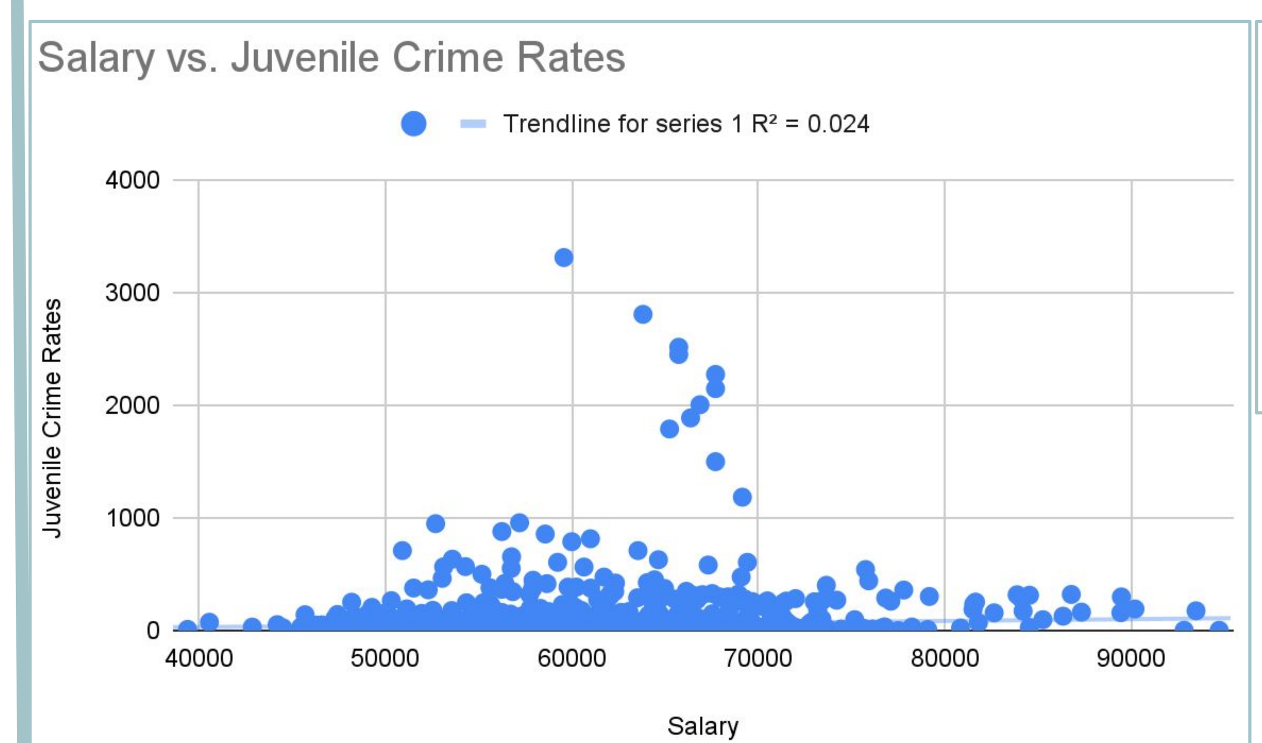
Challenges/Design Process

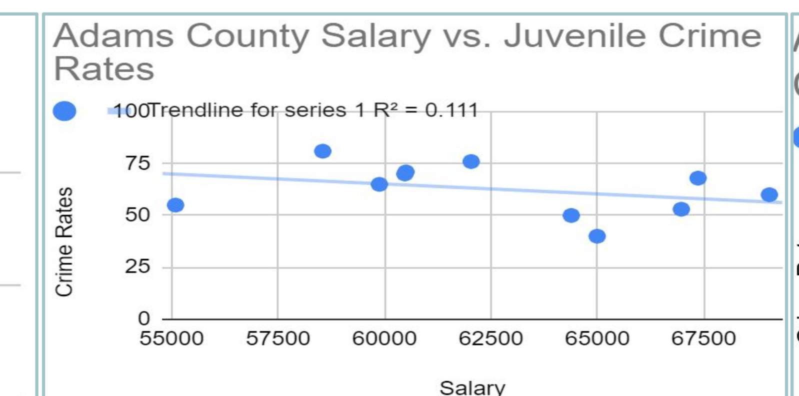
One of the first significant challenges came from processing the school staff data. The original staff data consisted of multiple datasheets for each year that contained various pieces of information concerning every school staff member in PA. We extracted data concerning the staff member's name or public ID, salary, and county. However, since there were approximately two hundred thousand staff members, copying and processing the data continued to crash and freeze google sheets. In the end, we copied every year's data into its own spreadsheet. Furthermore, the data from the government was not uniform between years, which required the usage of various cell functions to standardize the data, since manual processing would be too time-intensive. After standardizing the data, we found that some staff was repeated likely due to having multiple positions, thus we removed these repeats. Then we used a pivot table to find the median salary of staff in every county for that specific year. All of our processed staff data should be in this folder.

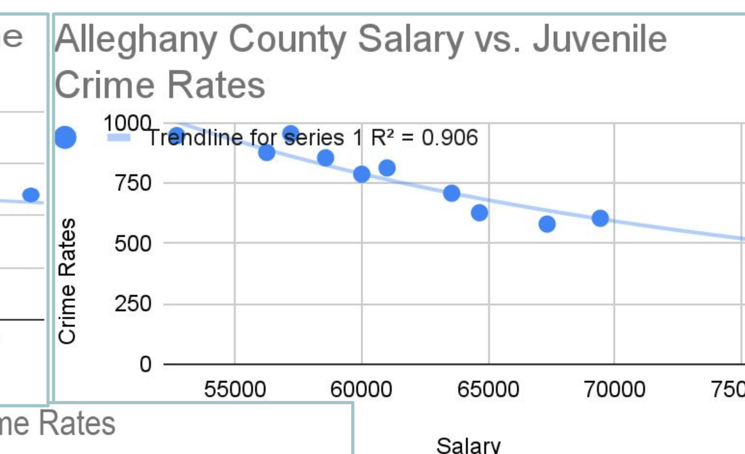
Processing juvenile crime rates also possessed its own difficulties. We were unable to find a spreadsheet detailing the amount of crime in each county. Additionally, after contacting various government institutions, it appeared that they did not have one either. Therefore we had to copy the data manually from a government website into a spreadsheet.

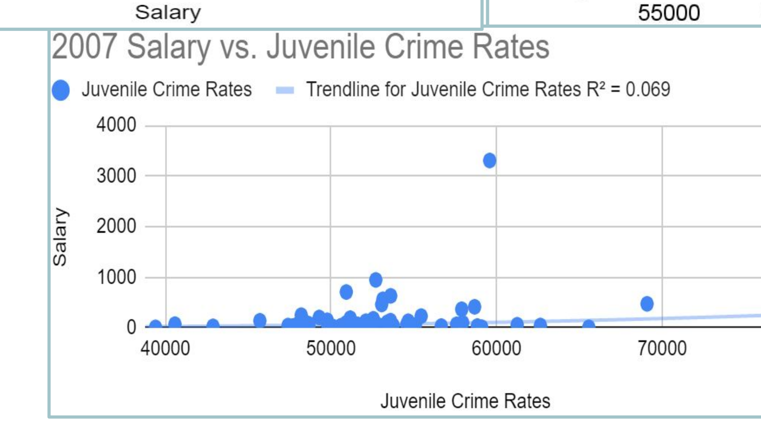
With the data of each section processed, we needed to combine it to find any consistent relationship between the salary and juvenile crime rates in PA. Thus we copied all of the pivot tables of the staff data into a spreadsheet and copied the juvenile crime rates into a separate sheet within the spreadsheet. However, due to the formatting of the data concerning the order of the years, the data for salary and juvenile crime was not directly compatible. Hence we used cell functions to reorder the data so that they would be compatible. After cleaning up and graphing the data, we found an extremely small R^2 value of 0.2 for linear relationships.

Due to the rather shockingly low R^2 value, we also tried to control for differences across years and counties to see if we could find a correlation after controlling for these confounding factors. The main challenge with removing the confounding factors was the amount of manual labor required to repeatedly filter for each year or county. In the end, we found that the strength of the correlation varied from county to county. Meanwhile, there was not a single year where there was an especially significant correlation.









Summary and Conclusion

No correlation between staff salary and juvenile crime rates is present across the entire state and time period from which the data is drawn. Our highest R^2 value was 0.028 for a second degree polynomial relationship (we did not sample other degrees of polynomial), which means that only 2.8% of variation in juvenile crime rates can be explained by staff salary. We recognized that there could be various confounding factors from regional differences and differences between each year. Thus we also analyzed the data by controlling for differences across counties by only graphing data from one county as it changes over time. We found that some counties had extremely high R^2 values, like Alleghany County, (0.907) and other counties with very low R^2 values, like Adams County (0.152). The inconsistencies between counties may be explained by the median not being a good general representation for all school staff within the county, since neighborhoods can be significantly different from data espoused by the median. Thus variations within these neighborhoods may not be reflected by differences in the median as time passes. A way to rectify the issue, would be to find the average of staff salaries within the county (but then outliers may be an issue), or by breaking down the area that is being analyzed. This means that we would try to find the median or average of the staff within a school district and compare it to juvenile crime within the school district. We analyzed the data by controlling for differences across different years by only graphing data from a single year as it changed from county to county and the R^2 values were very low for each year. For example, the highest R^2 value for 2007 was only the power series R^2 value of 0.69. This lent credence to the idea that there may be significant confounding factors from regional differences, but not necessarily differences across time.

What is the correlation between attendance of a sporting event and the success of the team?

Maura Marston, Megan Marston, Lauren Price, Amelia Faust, Alex Woltjen, and Matthew Diss

Plum Senior High School

50.79% of Big Ten Conference football games were won by the home team in 2021

BIG

University of Michigan

The Ohio State University

Michigan State University

Won **100%** of their games in the 2021 season

What We Learned

The group found that there was is a slight correlation between a team's home win percentage and attendance. However after further analysis it was determined that total expenses has a greater correlation on a teams at home win success. The group also discovered that there are many other variables that play a factor in the outcome of the game including talent of the players which cannot be quantified for our analysis.

Background

Home field advantage is a widely debated argument in the sports world. Even if someone is not interested in sports, they have probably heard of this term before. With all of the group members being athletes, a special and common interest was developed in this topic.

Method

The group observed each football team in the Big Ten Conference total football expenses, average attendance per home game, and average at-home win percentage. A team's total expenses includes medical, competition guarantees, recruiting, travel, facilities use, equipment, faculty compensation, student aid, revenues, corporate sponsorship/advertising/licensing, donor contributions, competition guarantees, NCAA and conference distributions, media rights, post-season football, ticket sales, institutional and government support, student fees, total academic spending (university-wide), total football spending, athletics related debt, annual debt service, leases, and rental fees on athletic facilities. All the expenses data was obtained from "Big Ten Financial Data". No financial data was recorded on Northwestern, so the team was dismissed from expenses calculations. The average attendance per home game was obtained from "NCAA Football Attendance" and spans from 2010-2019. Lastly, the at-home win percentage was obtained from "ESPN Big Ten Football Standings" from 2010-2019, focusing solely on each team's at-home record. Since the University of Maryland, University of Nebraska, and Rutgers University joined the conference after 2010, the average of their at-home win percentage was calculated from the year the team joined the Big Ten to 2019. A multiple regression analysis was performed after two linear regression models displayed a correlation and warrant further observation.

Challenges Faced

The biggest challenge the group faced throughout the process was determining if the size of attendance was the leading factor in the success of the team. After completing a multiple regression analysis the group was able to incorporate other factors such as revenue into our analysis. Additionally some of our analysis could not be performed using Google Sheets causing us to use other programs such as MiniTab and Social Science Statistics.

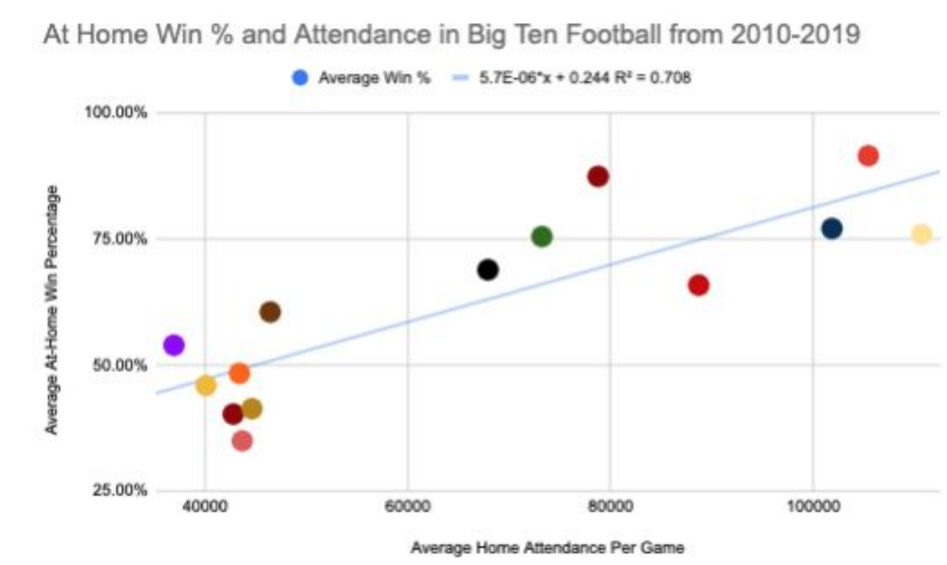
Data Points

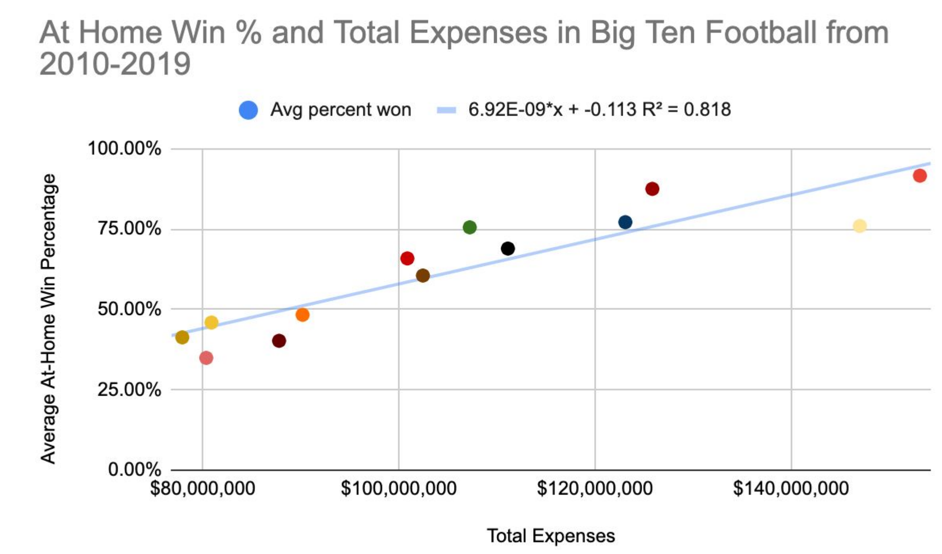
	Average Home Attendance (2010-2019)	Average Percent Won (2010-2019)	Total Expenses (2010-2019)
Illinois	43,379	48.33%	$90,170,927
Indiana	42,765	40.24%	$87,776,216
Iowa	67,902	68.93%	$111,110,520
Maryland	40,081	45.92%	$80,878,740
Michigan	110,718	75.95%	$146,993,441
Michigan State	73,224	75.54%	$107,218,708
Minnesota	46,416	60.54%	$102,443,293
Nebraska	88,701	65.87%	$100,856,633
Northwestern	36,913	53.93%	Not Available
Ohio State	105,442	91.61%	$153,127,530
Penn State	101,853	77.14%	$123,084,461
Purdue	44,608	41.31%	$77,902,376
Rutgers	43,648	34.92%	$80,349,009
Wisconsin	78,774	87.50%	$125,838,837

Colors shown on the above table correspond with the colors of the data points on the scatterplots.

Results

The first graph compares the average home attendance per game to the average home wins from 2010-2019. With an R-Squared value of .708, the data suggests about 70% of the at-home win percentage is explained by the average attendance. However, the slope was close to zero, meaning that the correlation is most likely smaller than 70.8%.





The second graph displays each Big Ten football team's average home wins and total expenses from 2010-2019. The calculated R-Squared value in the second graph is greater than the first, suggesting that 81.8% of the at-home wins is explained by the expenses of the football program.

Since both the above graphs demonstrated a high R-Squared value, multiple regression analysis was performed to further observe how attendance and expenses are related to a team's at-home success. The analysis demonstrated that attendance (X1) does not affect the prediction of the team's win percentage, since the coefficient is 0.

$$\hat{y} = 0X_1 + 0.00558X_2 - 0.0634$$

The Correlation Between Income and Covid-19 Vaccination Rates

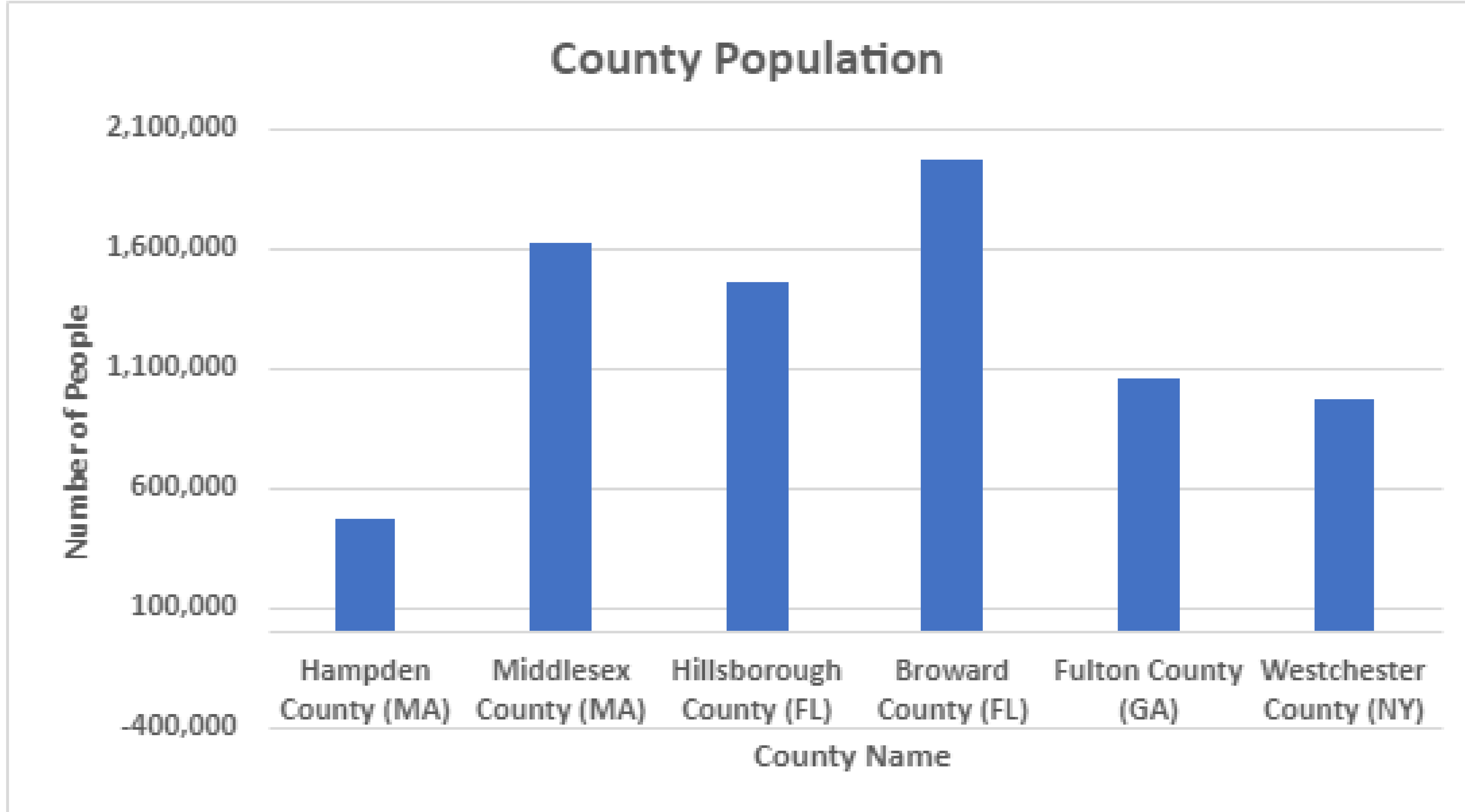
By Esteban Guzman, Faduma Haji, and Keagan Sicotte from the High School of Science and Technology

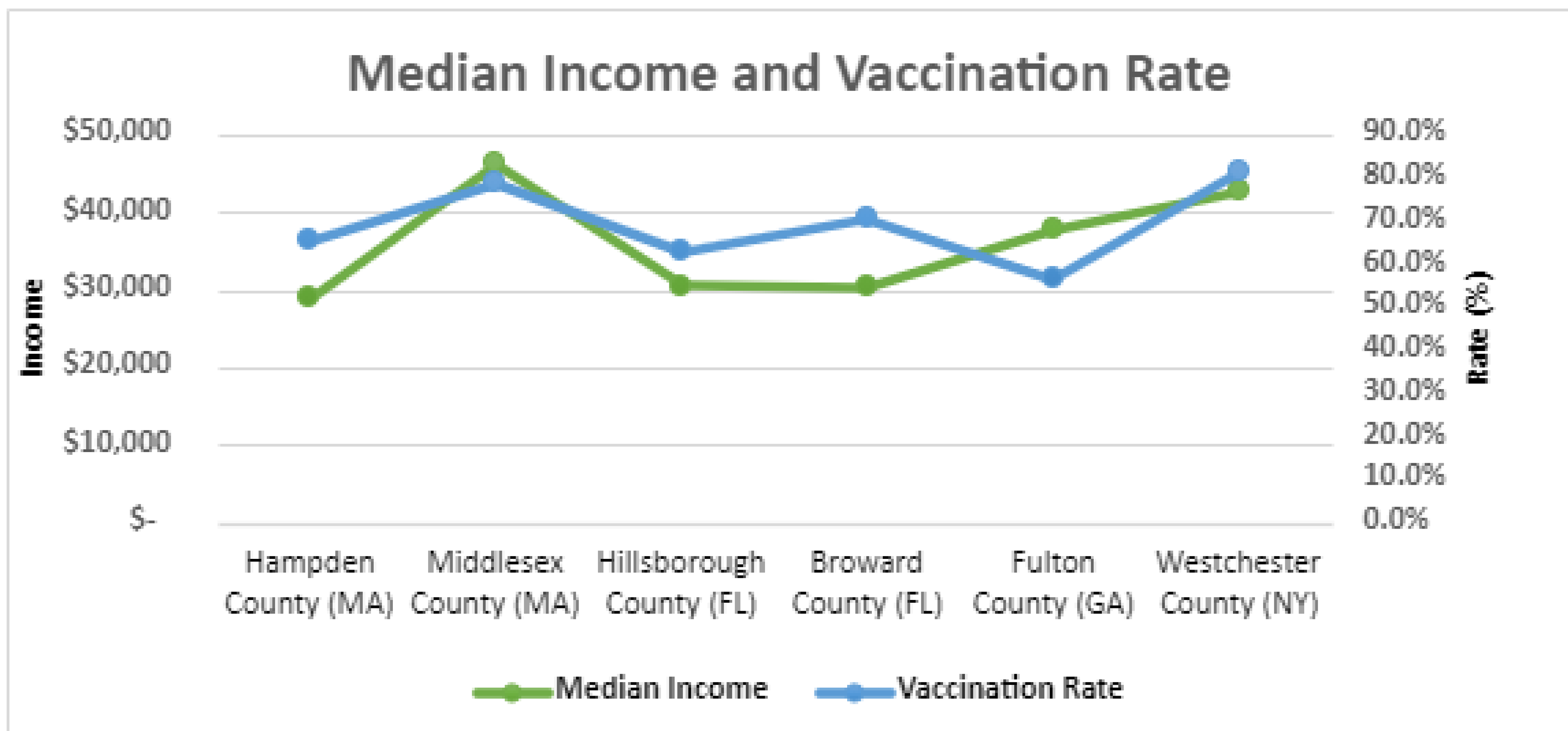
Hypothesis- It appears that highly urbanized cities have lower vaccination rates than more affluent cities.

Research Question- Does the amount of money a person earns affect the rates of vaccination? The Covid-19 pandemic has exposed the many inequities within our communities and country. Income is an important factor that contributes to those numerous inequities. In our research, we will explore the parameters of income, pertaining to vaccination rates.

Data Sources-

- COVID-19 by County (CDC)
- Florida Counties by Population (Florida-demographics.com).
- Georgia Coronavirus Map and Case Count - The New York Times (Nytimes.com)
- Massachusetts COVID-19 vaccination data and updates (Mass.gov)
- Median Income Data | U.S Census Bureau (Census.gov)
- United States (US) COVID-19 vaccine tracker (Mayoclinic.org)





Data Sets-

-Covid-19 by County (CDC.com)
-Population (State-demographics.com)
-US Census Bureau Quick Facts:(Enter County Name)

Results and Discussion- After compiling and organizing our data, we have come to a consensus that income does correlate with the Covid-19 vaccination rates. When the median income increased, the vaccination rates did so as well. There were a few exceptions due to the stark economic and geographical differences between the Northeast and the Southern part of the country.

A few questions that came up during our discussion included:
What other factors can we consider within our research?
Will there be a similar difference in hospitalizations?
Does the race demographics have some influence on the vaccination rates?
Does educational background alter vaccine literacy?
These are questions that we will explain more in our slides.

Conclusion- While conducting our research we faced the problem of finding a county that matched the Middlesex county within Massachusetts. While filtering through counties within Florida, we decided to observe other states in the Northeast of the US and discovered Westchester, NY. This county had a similar median income to Middlesex, MA. We also switched from analyzing cites to counties since, we found data sets mostly representing counties. Finally, we feel that there should be a call to action in legislation, prioritizing the literacy of vaccines. So, this will allow individuals from urbanized cities to make more informed decisions about their health. Thus, this legislation will aid in the concerns and hesitations within highly urbanized cities throughout our country.

Diversity and Student Achievements in Public High Schools

What is the correlation between socioeconomic diversity and student achievement in Pittsburgh Metropolitan Area public high schools?

Abigail Feinstein, Arina Sokolova, Jonah Rosenberg, Varun Bhat, Thomas O'Brien, Theo Rothstein, Oliver Brewer, Anya Zivanov

Allderdice Team 1

Definitions

Took AP Exam: % of students in the school who took at least 1 AP exam
Passed AP Exam: % of students in the school who passed at least 1 AP exam
Keystone Math/Reading: % of students who received proficient or above on the math or reading sections of the Pennsylvania Keystone Exam
Keystone Combined: % of students who received an overall score of proficient or above on the Pennsylvania Keystone Exam
Minority Enrollment: % of students at the school who are not white
% Below Poverty Line: % of students living with household income below the nationally-recognized poverty line

Resources

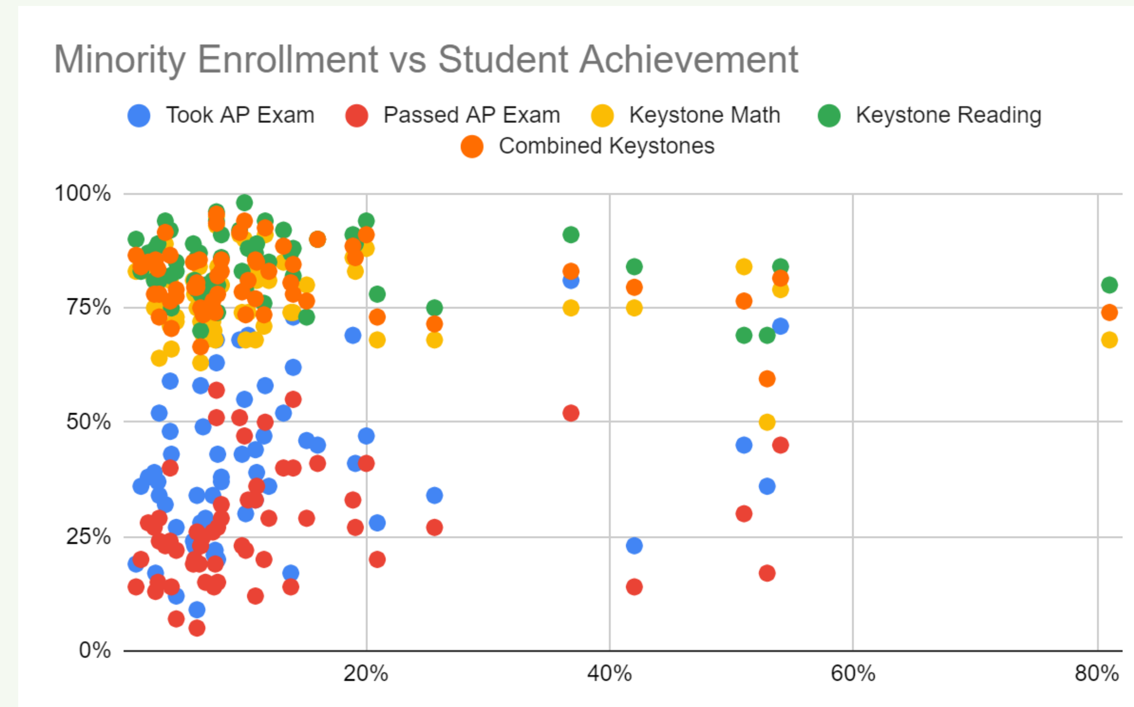
We used the **US News and World Report** website to access school profiles and demographic information about each school. This site provided data on Minority Enrollment, Keystone scores, and AP exams.

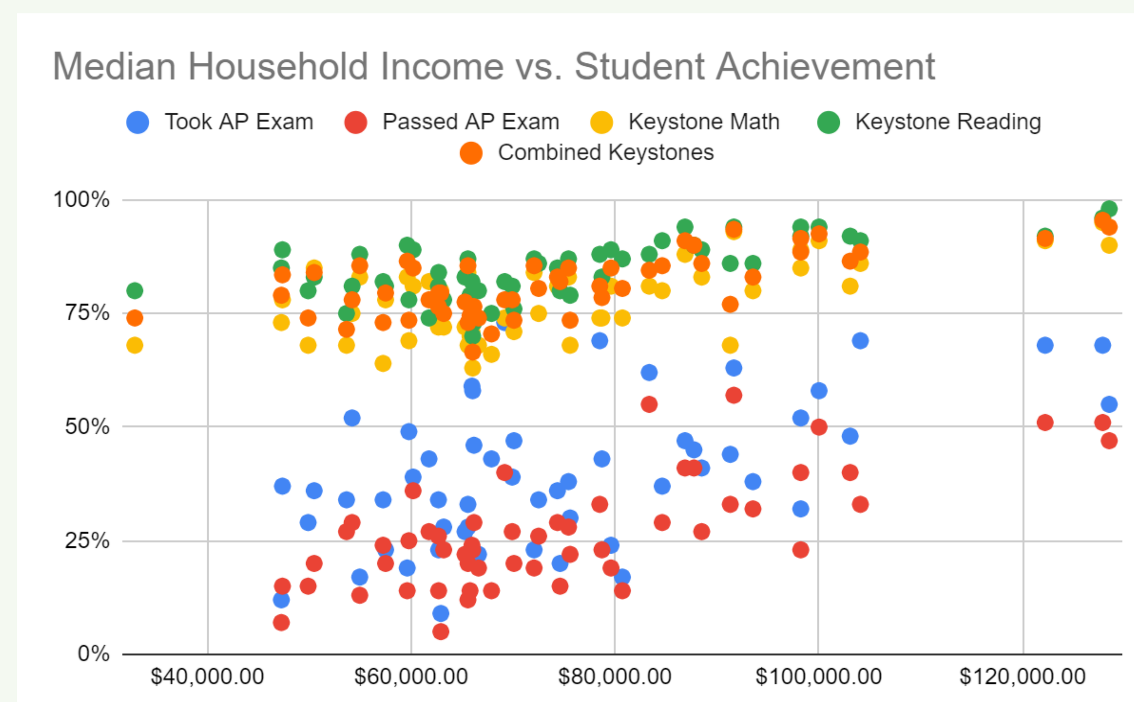
We used **Census Reporter** to access data regarding these various school districts such as the Median Household Income, Graduation Rate, and the Poverty Rate.

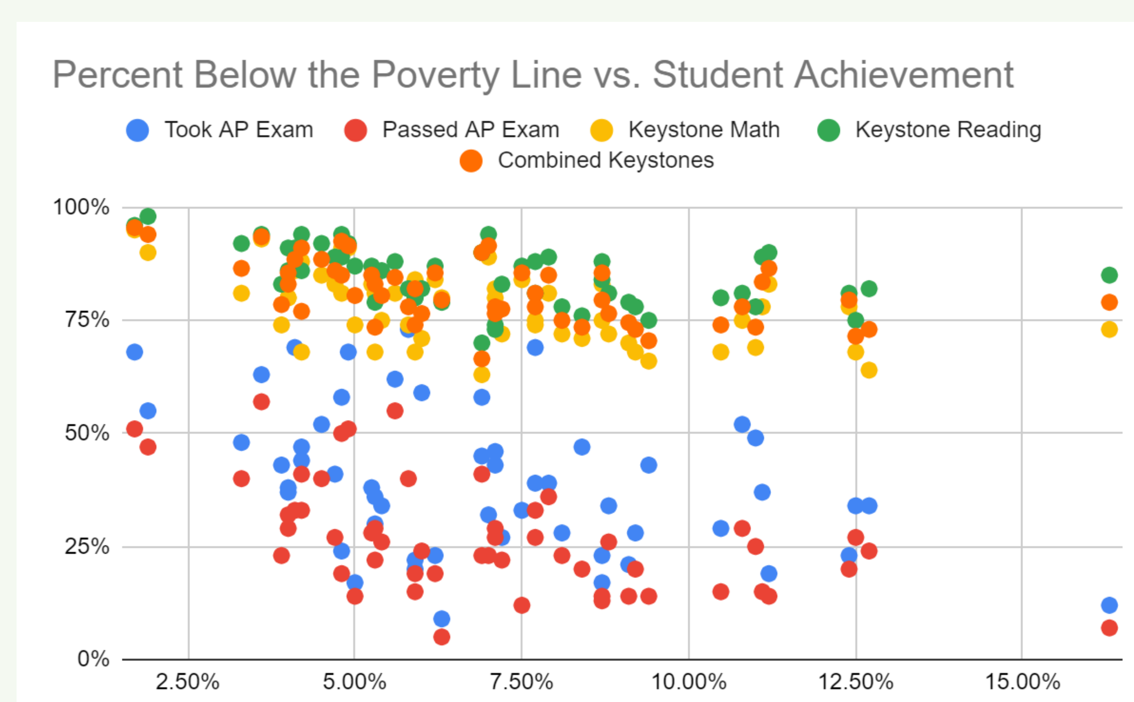
Challenges

The first challenge we faced during this process was identifying proper resources that had data specified to school districts and not counties.

Another challenge we faced was identifying what schools to review. Some of these schools didn't have recent data, and many Pittsburgh Public Schools were in areas that had students commute from all over Pittsburgh. Therefore, those schools were invalidated because of the over-saturated data values.







Results & Analysis

There are strong negative correlations between the median household income of a school district and the achievement of students, especially when measured in the percentage of passed AP exams and combined passage of keystone exams. In all categories of the minority enrollment graphs, our research found weak or no correlation. In the percent below the poverty line, there is a relatively strong correlation, although not as strong as with minority enrollment graphs.

	%Minority	MedIncome	%Below Poverty
Took 1 exam	0.2603259254	0.5388084667	-0.4262235089
Passed 1 exam	0.2054677226	0.707413699	-0.5585875878
Grad rate	-0.2162351065	0.5494784231	-0.6101474896
Math Keystone	-0.2042121604	0.6223380643	-0.5086211659
Reading Keystone	-0.2317312636	0.6700316976	-0.4805041692
Combined Keystone	-0.2340873074	0.6858036015	-0.527681449

Correlation Matrix

Conclusion & Recommendations

Further research needs to be done to see whether racial diversity plays a role in student achievement. From our data, we found there to be no effect. However, there is a strong correlation between income-based measures of diversity and student achievement, specifically test scores. We recommend that organizations like the College Board and the Pennsylvania Department of Education work to make their tests more equitable to people of all socioeconomic backgrounds, and they should so work to allow test prep access to more people, which have proven successfulness.

Taylor Allderdice High School

Median Income and Safety Measures in Pittsburgh Neighborhoods

Gabby Berger, Ezra Dubowitz, Liam Friedlander, Jennifer Lin, Bennett Matthews, Rachel Moreci, Audrey Olmsted

Definitions

Allderdice Safety Score: Score from 0-100 defining the safety of any given neighborhood, calculated by multiplying walk score by .15, multiplying the number of drivers at or below speed limit (derived from percent over) by .15, the number of crosswalks by 0.35, and the number of intersections by 0.35.

Walk Score: Score assigned to various census tracts throughout the city of Pittsburgh. Points are awarded based on distance to amenities and pedestrian friendliness. Pedestrian friendliness is determined through population density and road metrics, such as block length and intersection density. The score ranges from 0-100, as defined by walkscore.com.

Percent over Speed Limit: Percentages of vehicles that were speeding in an area, as provided by City of Pittsburgh Department of Mobility and Infrastructure (DOMI).

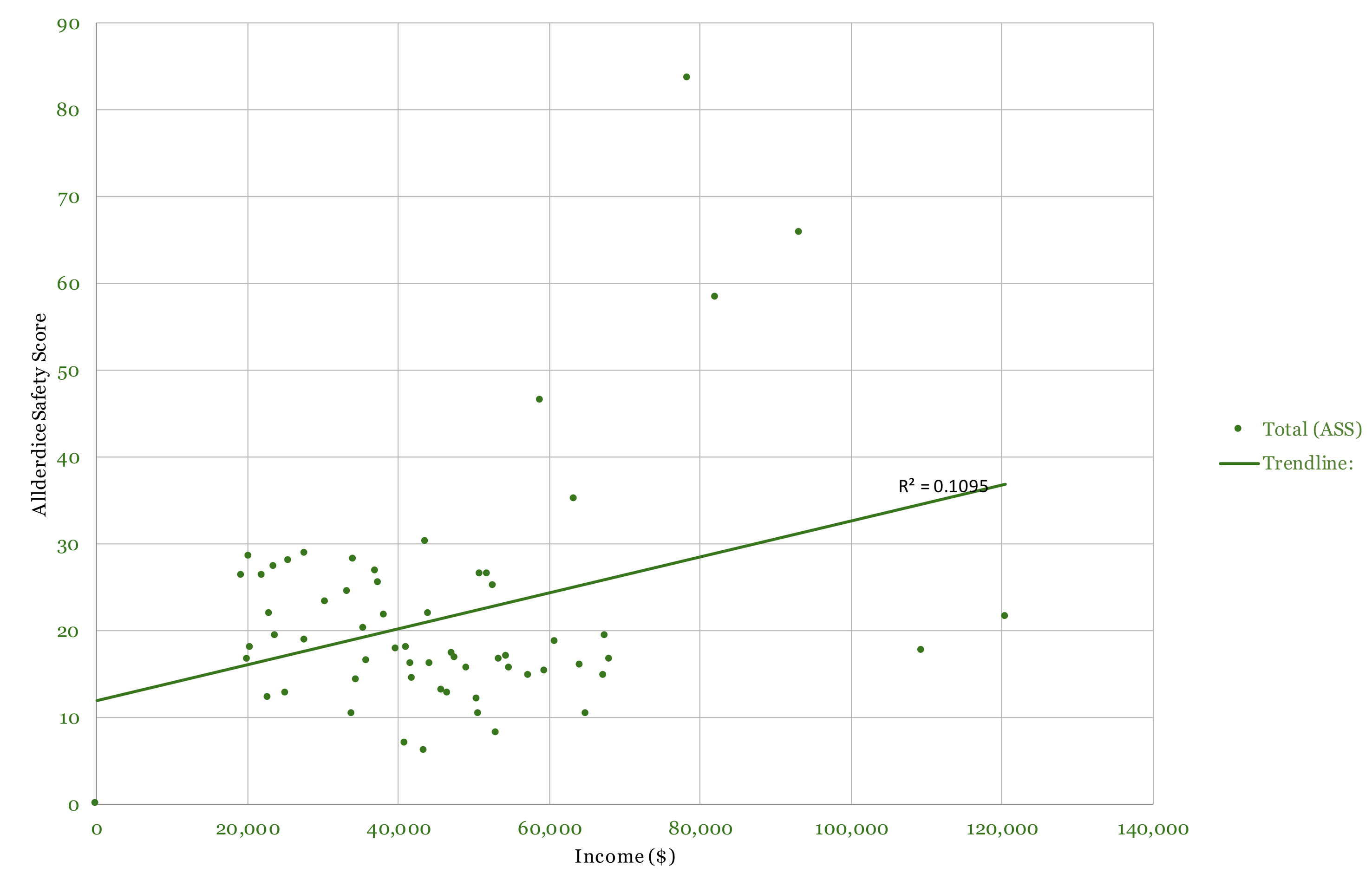
Number of Crosswalks: Amount of every type of crosswalk in each neighborhood.

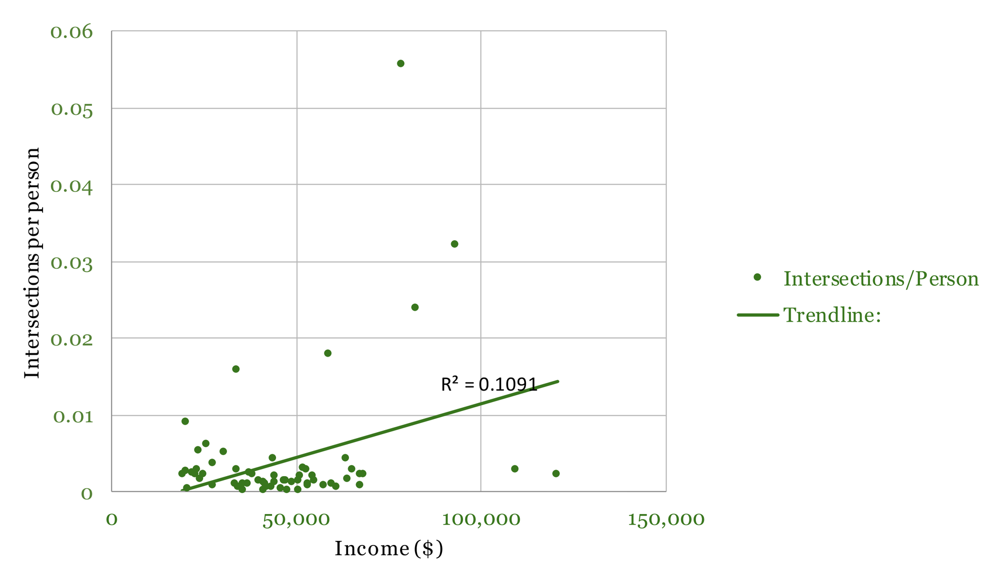
Number of Signalized Intersections: List of intersections with signals and which neighborhood the intersection is in.

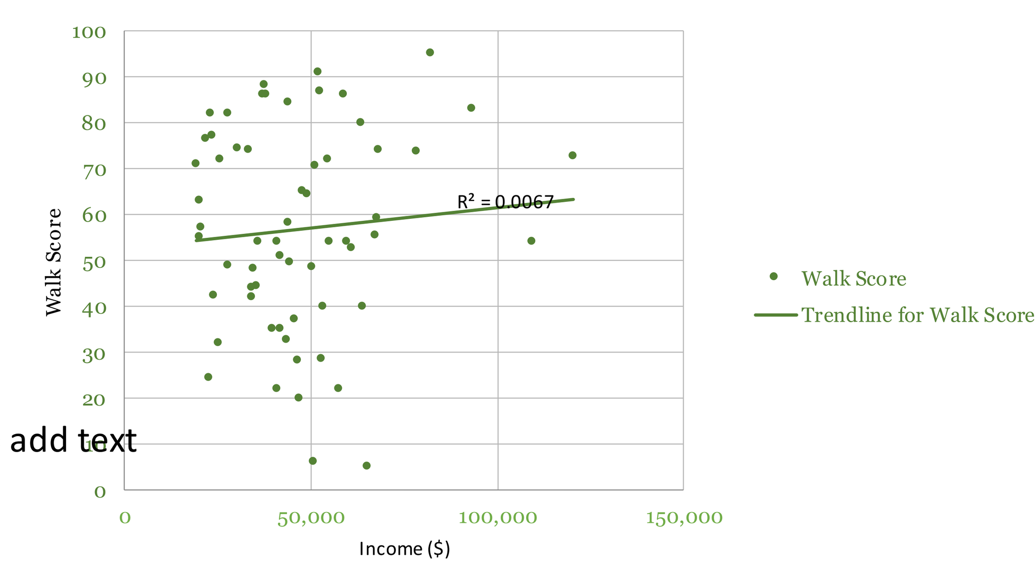
Square Mileage: Area of each neighborhood in square miles.

Population Density: People per square mile of every neighborhood.

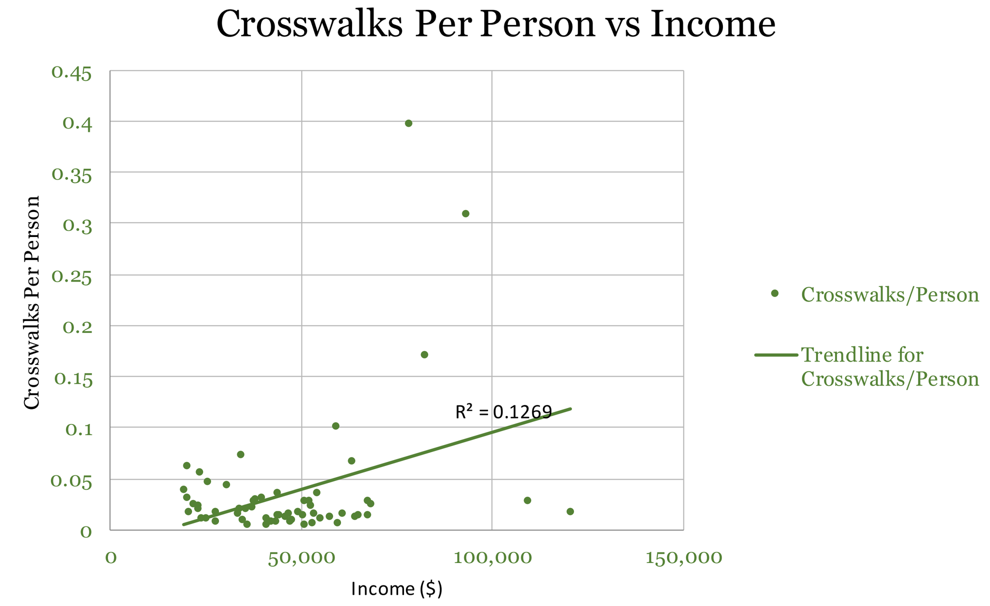
Graphs and Results

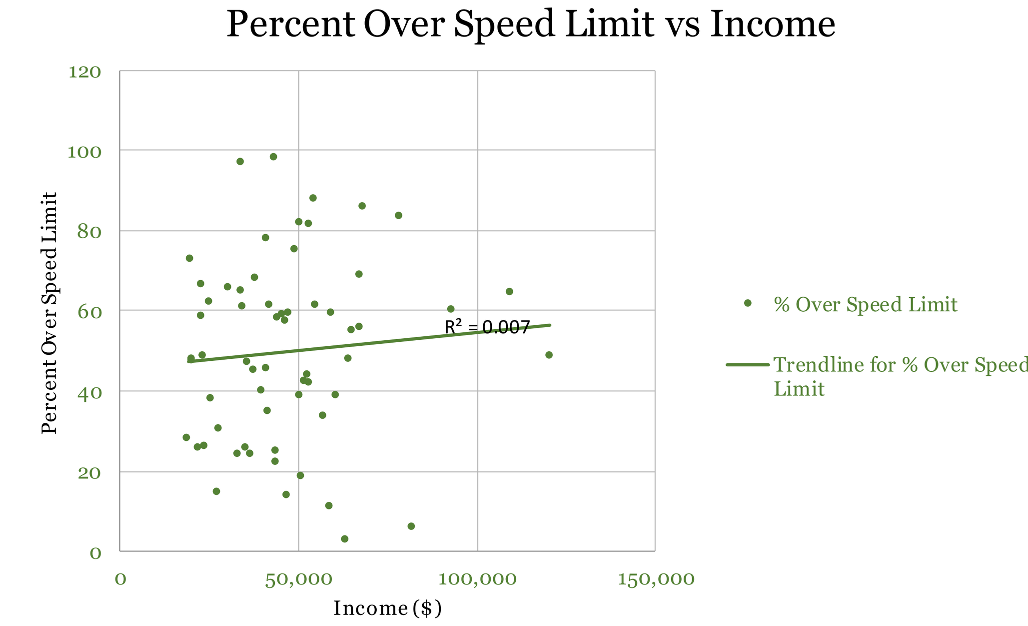


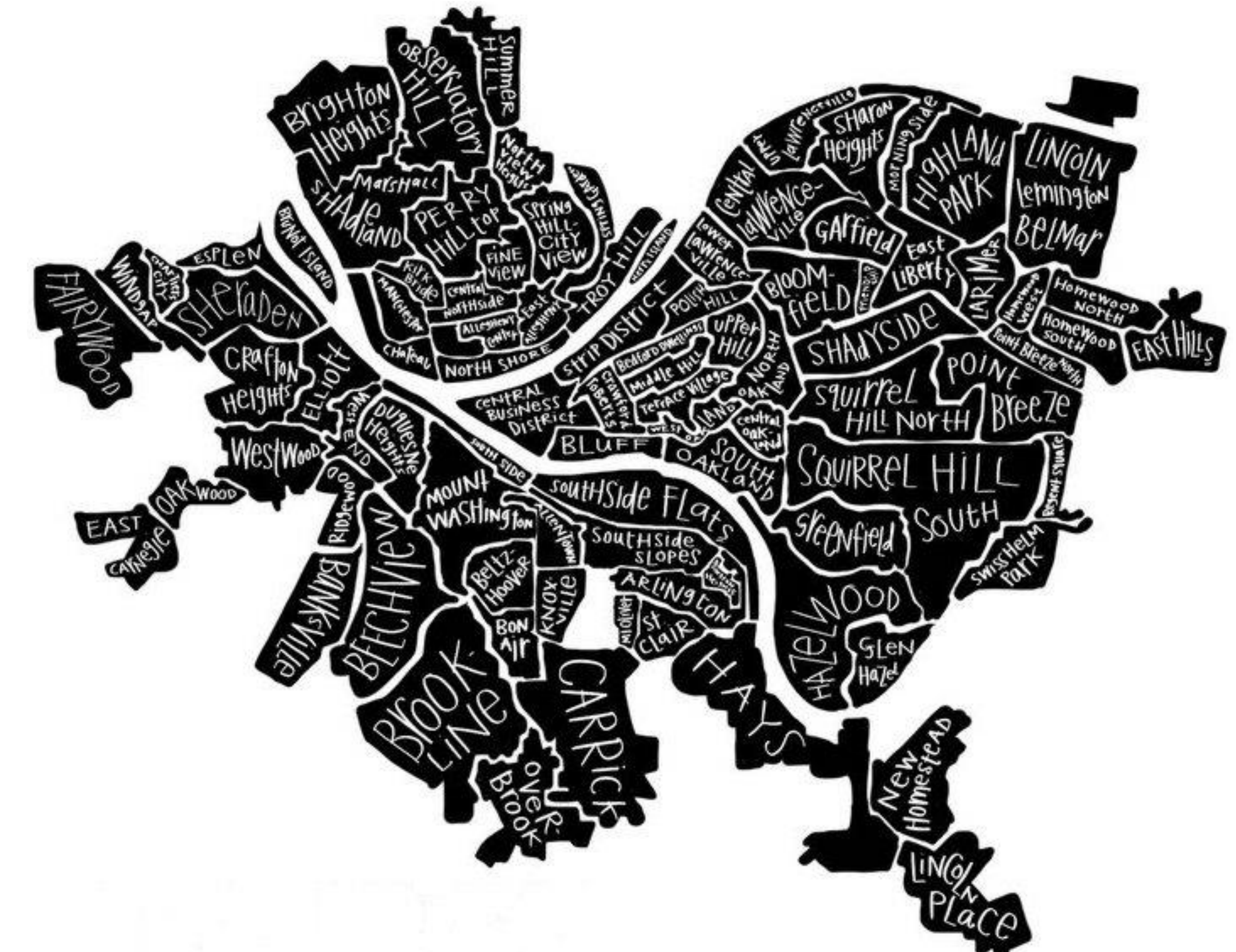




Click to add text







Challenges

- Lack of data points for some neighborhoods
- Uncertainty of exact level of median income
- Possibly outdated data sets
- Factors chosen were not comprehensive of every safety aspect
- Scoring based on multiple sets of data; possible inconsistencies

Top 3 AS Scores

Neighborhood	AS Score
North Shore	83.5
Strip District	65.8
Central Business District	58.4

Bottom 3 AS Scores

Neighborhood	AS Score
Crafton Heights	6
Chartiers City	7
Stanton Heights	8.1

Resources

Our primary resource was the Western Pennsylvania Regional Data Center website, which allowed us to retrieve demographic data and the data that we required in order to quantify safety within a neighborhood. We used this resource, along with some others to answer the question *How Does Median Home Income in a Neighborhood Affect Transportation Safety?*
We also created our own quantifying scale, the Allderdice Safety Score (AS score), which helped us to numerically analyze safety in a neighborhood through an exploration of various components of safety in a neighborhood including walk score, percent over speed limit, and crosswalks and intersections per person.

Conclusion

From our results, we found that there is a very minimal correlation between our Allderdice Safety Score and income in a neighborhood. The factors that had the strongest correlation with income of a neighborhood were intersections per person and crosswalks per person. This suggests that these two factors are more correlated to the income than the others. Percent over speed limit and walk score compared to income showed minimal correlation suggesting the income of a neighborhood does not correlate to these factors. Although the Safety Score did not show a strong correlation to median income, there still may be relationships between traffic safety and the income of a neighborhood, as the factors chosen were intended to represent the safety of a given neighborhood, not describe exactly how safe a place was for travelers.

KNOWING THE SCORE!

Does the Pittsburgh Steelers' performance correlate with crime rates?

UPPER ST. CLAIR HIGH SCHOOL

RITVIK SHAH, SCOTT CHEUNG, NIVEDHA SURESH, BEAR BOTTONARI, ROHAN MEHTA

BACKGROUND

For years, fans bases have burned couches and flipped cars to celebrate the success of their favorite teams. After the Eagles won the Superbowl a few years ago, Philadelphia observed a temporary increase in crime rate. The Steelers are an integral part of life in Pittsburgh. We want to identify whether a significant relationship exists between Steelers' performance and crime rates.

HYPOTHESIS

We believe that through our study, we will find a **positive correlation** between Steelers' performance and crime rates; if the Steelers play well, there will be an increase in crime.

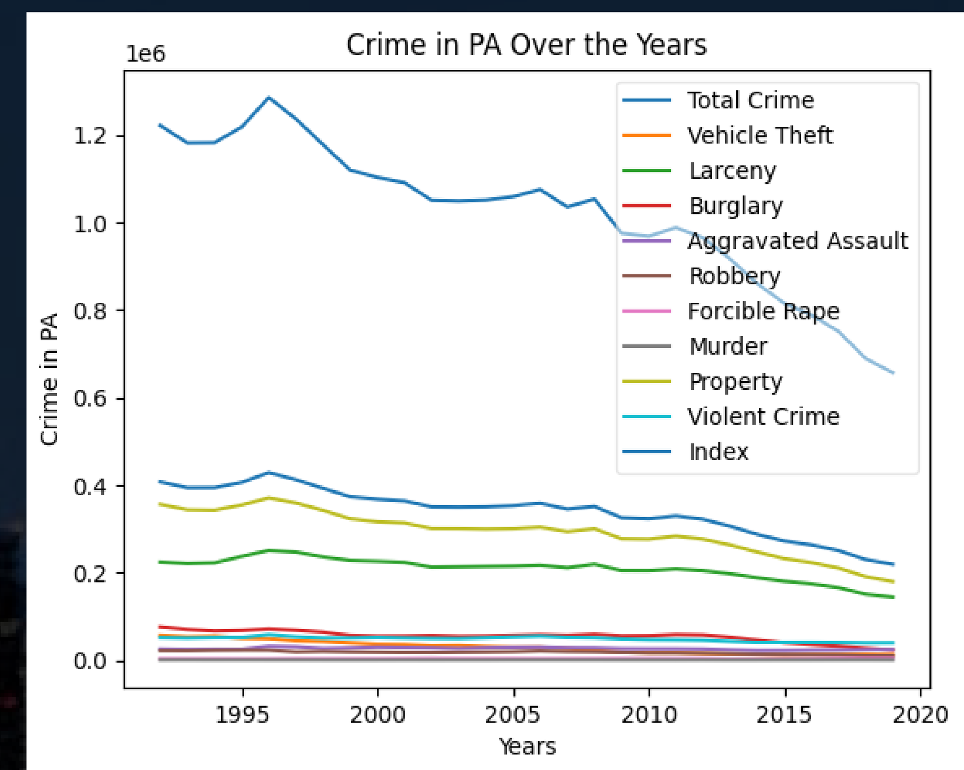
DATASETS

The main source for our Steelers' performance data was **Pro Football Reference**. Our crime data came from the **Western Pennsylvania Regional Data Center** for reports on non-traffic citations, arrests, and blotter data. We also obtained crime rates data from the **Disaster Center**, which provided different rates of a variety of crimes across multiple decades. Lastly, we used **Allegheny County Analytics** to procure data on motor vehicle theft in Pittsburgh.

CHALLENGES

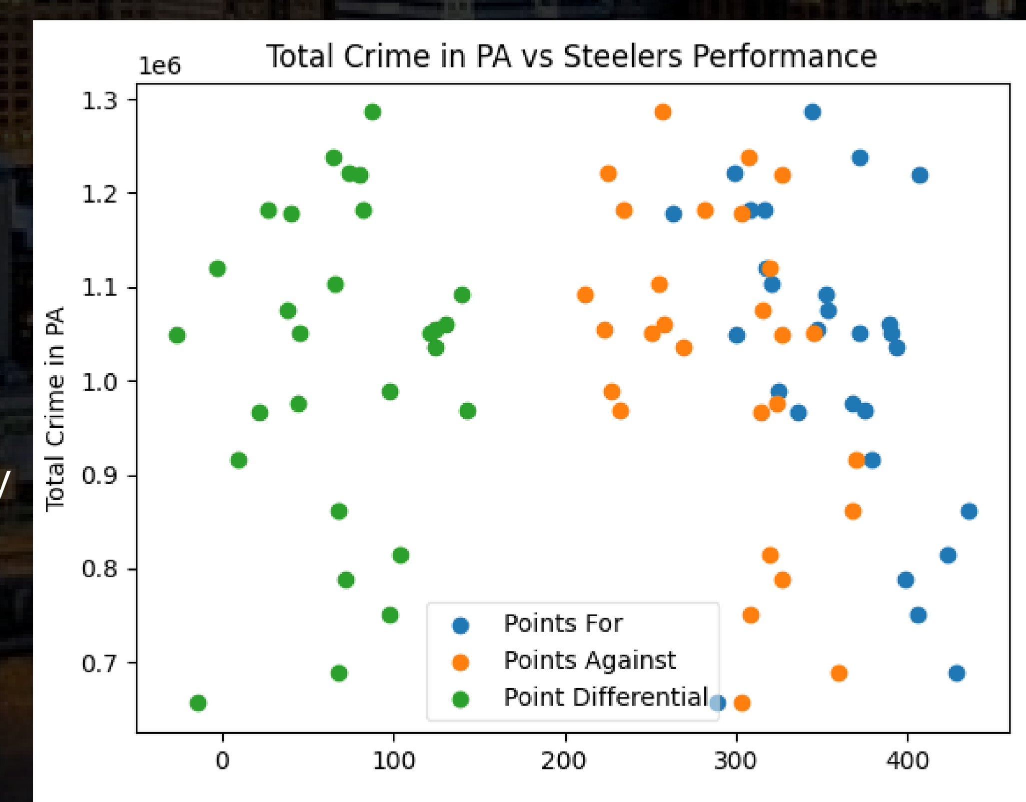
Data about crime in Pittsburgh since 1992 is not readily available. Half our datasets were from 2016-2019. This caused some of our initial graphs to be unrealiable as they only had 4 yearly data points, which shows current trends instead of historical trends. Another challenge was trying to factor in losing seasons with crime. Since the Steelers have not had a losing season since 2003, it was difficult to develop an understanding of crime during a Steelers' losing season. Another challenge we faced is evidenced by the results of a machine learning model that we developed with logistic regression to predict the Steelers' success from crime rate. The model predicted the Steelers making the playoffs with a 64% accuracy regardless of the crime level. We believe this to be attributed to the lack data readily available, which made it more difficult for our model to build a clear cut threshold for crime affecting the Steelers' success.

RESULTS

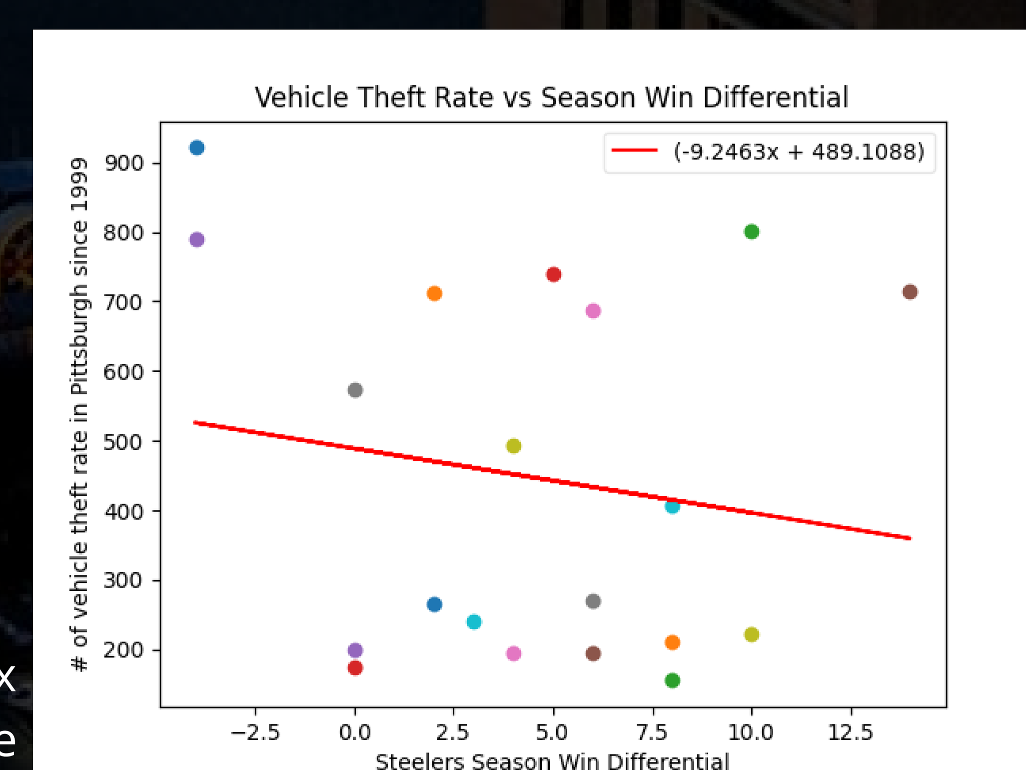


This depicts the trends of a variety of crimes across the past three decades in PA. Despite some peaks, crimes have generally declined since 1990. Due to limited data availability, we were unable to show the complete picture of crime in Pittsburgh.

This graph represents the comparison of the total crime in PA (millions) against the Steelers' Performance. The graph shows that in most situations where the **Points For** is above 400, there is less crime in PA. With the **Points Against**, we can see that more points given up by the defense correlated to less crime. Games that are both close and high scoring could be engaging people's attention; that in turn, could affect overall crime.



Since there is a correlation between Steelers' success and property-related crime, we further investigated vehicle theft rates for Pittsburgh. As we can see by the slope of -9.2463, the more wins than losses the Steelers have, the more the vehicle theft rate decreases. However, when looking at the total crime in Pittsburgh, we witnessed the Simpson's Paradox as there tended to be an increase in crime as the Steelers did better.



SUMMARY/CONCLUSION

While total crime in PA and Pittsburgh shows some correlation to the Steelers' success, we found relatively consistent evidence that property-related crime is negatively correlated with the Steelers' success. **Correlation does not mean causation.** We believe that the main cause for the decrease in property-related crimes could be attributed to people being more likely to be at home to watch to the game, which means that there is less opportunity for crime to occur.
In order to understand our data and develop our graphs, we used Python with Matplotlib, Pandas, Numpy, and Sklearn. We experimented with machine learning to determine if crime effects Steelers' success.

RECOMMENDATIONS

- Using data to predict the Steelers' success, police officers can prepare to handle specific types of crimes. For example, they can prepare to deal with more vehicle thefts and property-related crime when the Steelers have a bad season.
- Local Police Stations can hire sports statisticians to analyze the Steelers' success and its effect on crime.