SUCCESS IN SMALL BUSINESS

Analyzing the factors that contribute to the success of small businesses in Pennsylvania

Avonworth High School
Data Jam 2018

Prosper Tjelmeland • Spencer Hall • Marion Haney • Benjamin Houser
Owen North • Richard Danylo • Katie Carlson • Alexie Auth • Lauren Pflueger

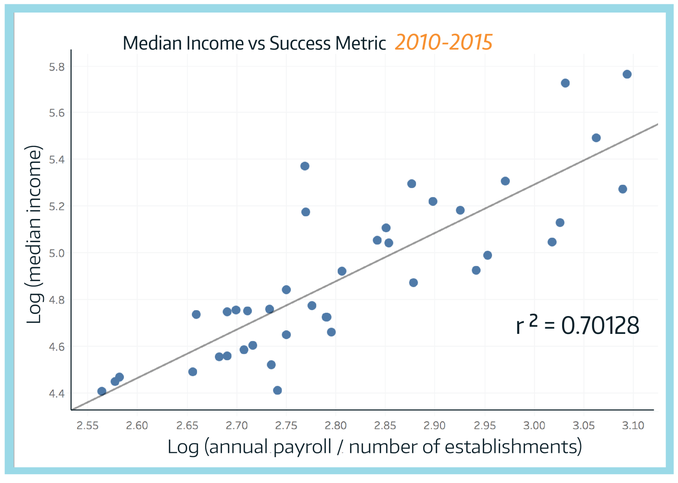
Research Question

Pennsylvania's economy has grown considerably in the past decade and the difficulty of entrepreneurship has greatly increased as a result. Thus, as a group, we believe it is critical to ascertain which factors can contribute to success in the current economic conditions. Sole proprietor entrepreneurship is struggling. **Why?**

How can we instruct business-owners so that they are more successful? Where is the best place to locate a business and why?

Our success metric calculates overall sucess of business. When we separate data based on the number of employees per establishment, we see that large businesses are growing and small businesses are stagnating.

Change in Income



Success Metric = Total Payroll / Number of Establishments

We standardized all of the data we collected by county, limiting ourselves to data sets organized by that category. In the long run however, this simplified comparison across datasets and creation of visual representations.

Immigration r = 0.2505

HS Dropout r = 0.104

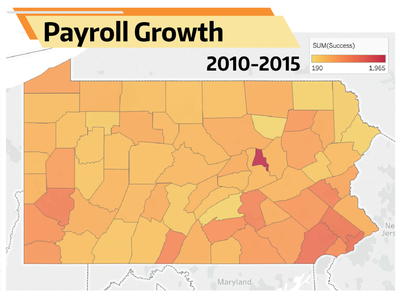
Poverty r = -0.1555

Fertility · Bankruptcy · Crime · Revenue

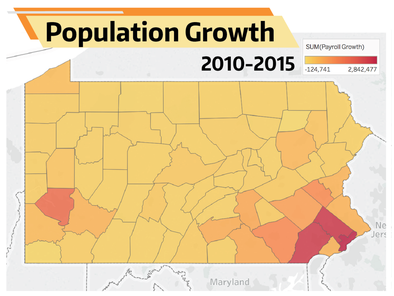
Our team struggled to find revenue data for small business, as well as consistent data on the factors listed above.

At the beginning of our research phase, our group was very interested in the demographics of each county, hoping to find an unexpected factor that had a major impact on small business. These factors however did not result in **strong** association.

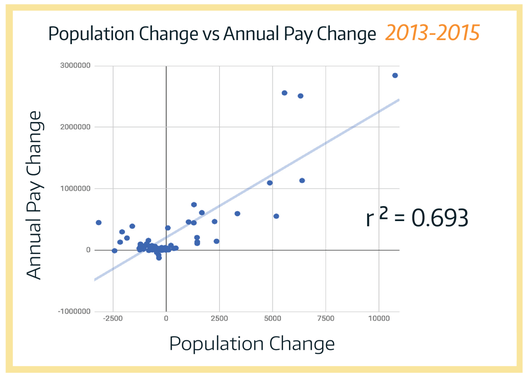
Summary of Results

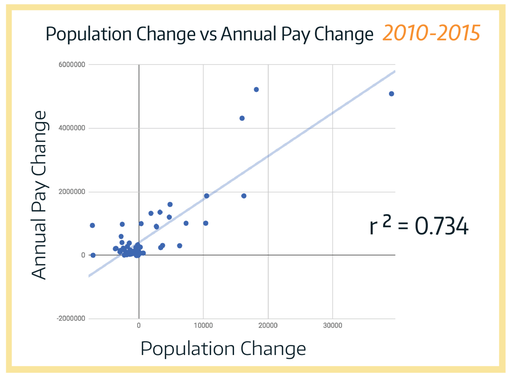


Heatmap ▼ Visual Display ▼ Counties with the Most Growth



We studied the effect of population on the success of a business. We found convincing correlations between population change and annual payroll. While we cannot assume that population change **caused** change in payroll, we can claim an **association** between the two measures. Counties that have a population increase tend to have increased payrolls, while counties with a decrease in population tend to have lower payroll increases.





Fulton County removed • lack of data

Allegheny County outlier • suffered a population decrease of 3,355 people over 2 years, but still had a major payroll increase.

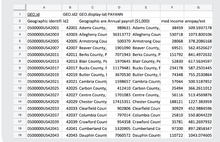
The importance of this graph is to show that the results of the 2013-2015 were not random. Increasing the range of years for the data increased our correlation, strengthening our argument that there is an association between change in population and change in payroll.

Resources

Most of our data came from an online database of information. We quickly learned how to interpret abbreviated categories in each column and organize data that was separated by two different demographics. The US Census Bureau was our primary guide to defining business terms, classification of industrial groupings, and finding reliable data sets.

United States Census Bureau

Sample Excel Spreadsheet



Categorization of Data

Challenges

- Years did not line up across datasets
- Finding revenue data for small business for success metric
- Narrowing down the list of factors to analyze one in depth
- Distinguishing small businesses from the data pool to isolate our analysis on sole proprietors

Conclusion

Most Influential Factor

Population Change

Factors that do not Influence Success

Birthrate • Poverty • HS Dropout • Immigration • Migration

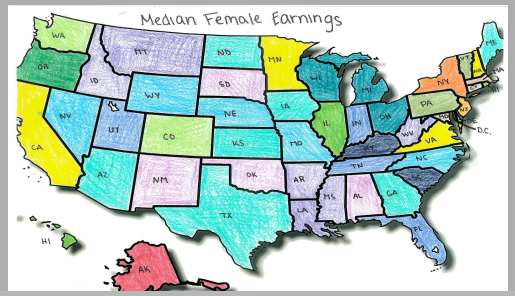
Birthrate, immigration rates, and migration rates are all subcategories; individually, they are not strong enough to effect the success of business. When all the **factors are combined** and labeled as change in population, there is an association. Therefore business-owners should seek out areas that are changing: new neighborhoods, construction, progress that encourages people to live there.

Does income affect abortion rates?

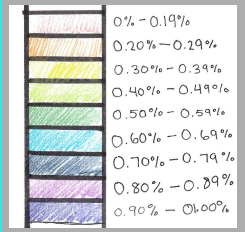
Background and purposes

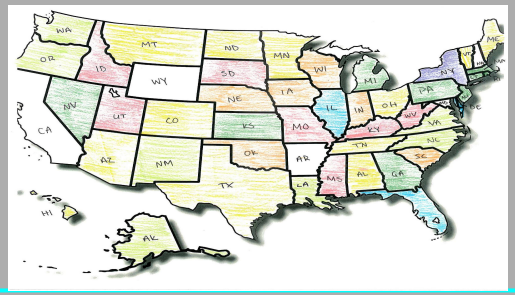
- Create awareness of repercussions of teen pregnancy
- Gives us information about abortion
- Information about Income
- Helps life get easier for some people.

Data and results



Abortion rate vs Population

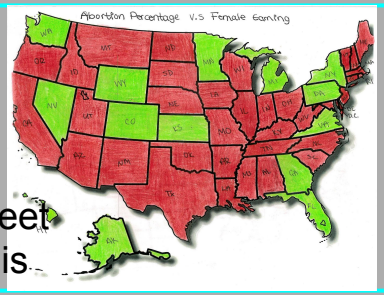




Conclusion

We've come to the conclusion that our hypothesis is wrong. There are more states with high income and high abortion rates and low incomes and low abortion rates.

Red- Did not meet our hypothesis
Green Did meet our hypothesis



Analysis

- 33.2% of all abortions in 2014 were reported by women between the ages of 20-24
- The second most frequent reason that abortion occurs is because parents can't afford the baby and people get pregnant too young like when they are teenagers.
- Abortion is most increasingly concentrated among low income women
- Poor women not trying to conceive are also three times more likely to get pregnant than their higher income counterparts (9% compared to 3%), and ultimately at 5 times more likely to give birth.
- Abortion helps unplanned pregnancies, but could also cause wombs to be destroyed.

Keystone Effectiveness: Hit or Miss?

Bethel Park High School - Team 1

Sabrina Tatalias, Regina Thase, Elise Bermudez,
Dolan Stapleton, Brett Marquardt & Lexi Schanck

Research Question

Are the Keystone Exams a good indicator of high school students' academic success?

We decided to research if the Keystone exams are a good indicator of student success versus alternative assessments, such as: the SAT and ACT tests.

Challenges

Since we only wanted to analyze data from Allegheny County, we had to manually enter all of the data after creating a spreadsheet with the school names. We also used data form our own school, which had to be anonymized before we could view it.

<u>Challenges Encountered:</u>

- Insufficient county data available (ACT)
- Unavailable composite Keystone score data sets
- Cleanse and pull out the Allegheny county school data
- We had to sort out some data by hand
- Many different data spreadsheets needing to be joined together

Data Sources

County level data was obtained from the Pennsylvania Department of Education's website. We also received **individual student data from the Bethel Park School District.** These current sources provided us with the most accurate information to help see if the Keystone exam is a good indicator of student success. The data was then analyzed in Excel and Google Sheets.

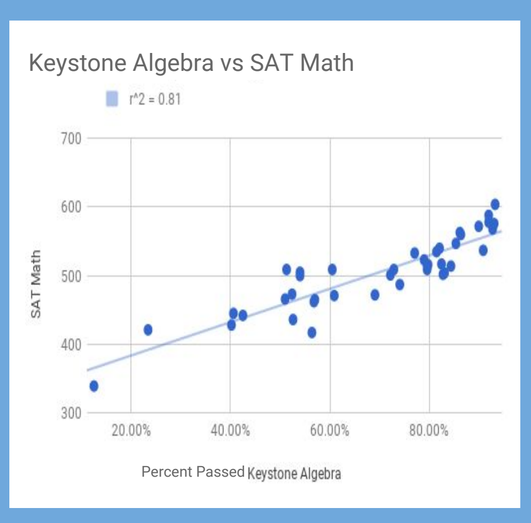
Data Set Examples

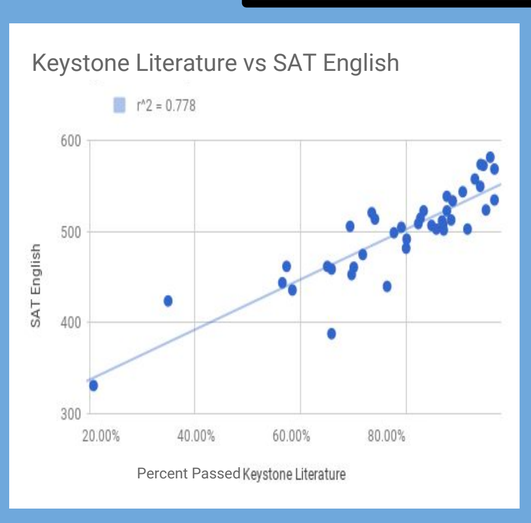
	A	B	C	D	E	F	G	H
1		School Name	Number of Graduates	% College Bound	Passed Algebra Keystone	Passed Biology Keystone	Passed Literature Keystone	SAT Number Tested
2	Allegheny Valley/Springdale	Springdale JSHS	72	69.44%	72.20%	73.20%	80%	50
3	Avonworth	Avonworth HS	93	87.10%	85.40%	87.10%	96.60%	77
4	Baldwin-Whitehall	Baldwin SHS	320	83.13%	83.10%	75.60%	85.60%	236
5	Bethel Park	Bethel Park HS	357	82.07%	77.10%	76%	88.70%	268
6	Brentwood Borough	Brentwood HS	82	82.93%	69.10%	69.40%	77.60%	45
7	Carlynton	Carlynton JSHS	95	72.63%	54%	72.30%	79%	60
8	Chartiers Valley	Chartiers Valley HS	262	85.50%	79.60%	88.40%	82.20%	186
9	Clairton City	Clairton MS/HS	40	60.00%	56.40%	41%	65.80%	19
10	Cornell	Cornell HS	44	63.64%	52.60%	52.70%	76.30%	27
11	Deer Lakes	Deer Lakes HS	169	87.57%	54%	70.30%	87%	130
12	East Allegheny	East Allegheny JSHS	113	73.45%	42.50%	53.60%	56.50%	75
13	Elizabeth Forward	Elizabeth Forward SHS	199	64.82%	74.10%	75.50%	79.90%	122
14	Fox Chapel	Fox Chapel Area HS	363	90.63%	93.10%	93.70%	96.60%	291
15	Gateway	Gateway SHS	296	78.04%	79.60%	55.50%	74%	204
16	Hampton	Hampton HS	261	88.51%	86.20%	83.90%	92.90%	228
17	Highlands	Highlands SHS	190	58.95%	60.90%	52.50%	65.80%	112
18	Keystone Oaks	Keystone Oaks HS	164	74.39%	81.50%	55.30%	83.20%	108
19	McKeesport Area	McKeesport Area HS	239	59.83%	40.20%	43.60%	58.40%	126
20	Montour	Montour HS	217	84.33%	84.40%	83.60%	86.90%	170
21	Moon Area	Moon Area SHS	289	91.00%	90.90%	90.50%	95%	246

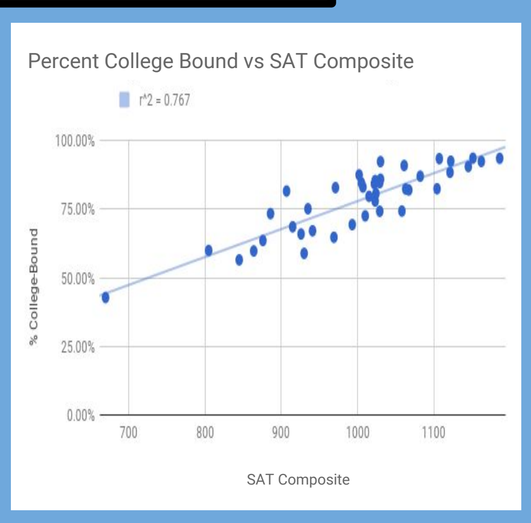
This is a screenshot of the data after it was put into Google Sheets

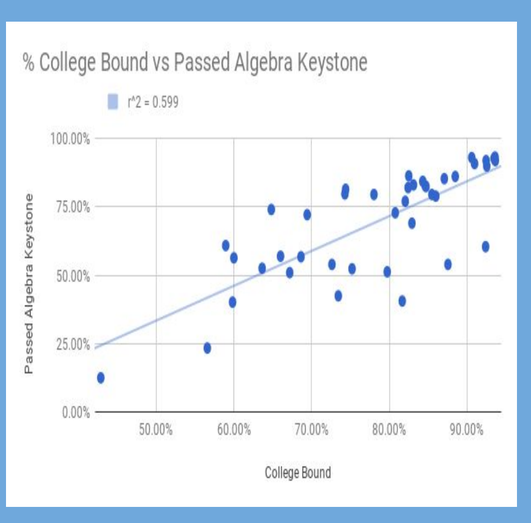
Visualizations

Allegheny County

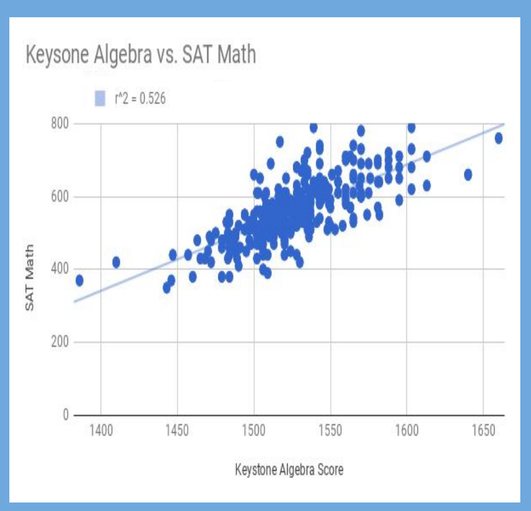


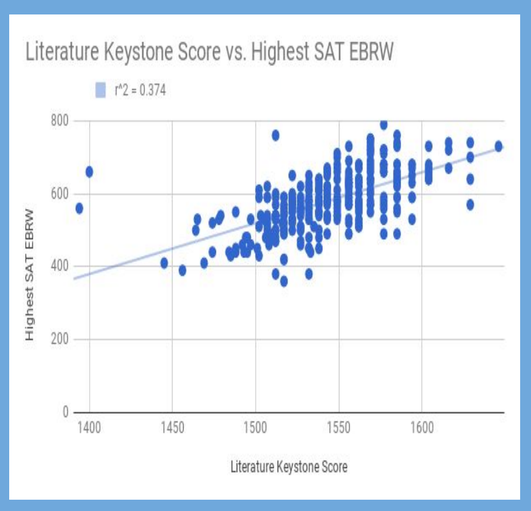


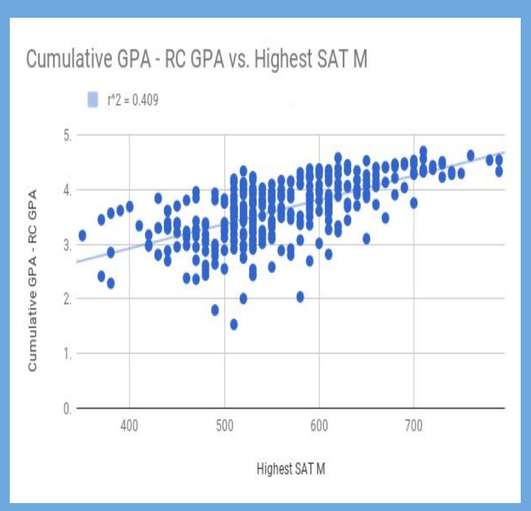


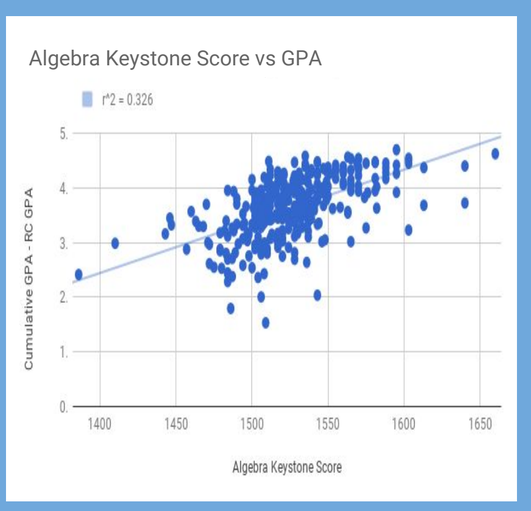


Bethel Park









Summary

- ★ There is a <u>strong</u> correlation between SAT **math** and Keystone **algebra** scores.
- ★ There is also a <u>strong</u> correlation between SAT **English** and Keystone **literature** scores.
- ★ Schools who had students that performed well on standardized testing <u>strongly</u> correlated with college bound students.
- ★ In general, **our school** data was similar to county data findings, but each school results differs.

Policy

We believe that it is redundant for students to take both the SATs and the Keystones since they produce such similar results. A good change in policy would be for the state to pay for students to take the SATs once and use that as a graduation benchmark in place of the Keystones. This would allow the state to save money, eliminate wasted school days dedicated to taking the Keystones, and students would not be burdened by the stress of additional, redundant high-stakes standardized testing.

The Effects of Teen Sleep Habits

Bethel Park High School – Team 2

Alyssa George, Alaina Cerro, Maria Schiller, and Sean Conroy

Research Question

How does the amount of sleep teenagers get affect their lives (health, academic performance, emotions, etc.)?

Delaying the start time for high school students has been a hot topic locally. We also recognized students are lacking sleep and theorized the possible results affecting many aspects of their lives. We decided to conduct research on the effects of the amount sleep teenagers get as it relates to academic performance and health. We intended for this research study to help schools and students better understand the importance of sleep and to determine if a later start time for high school students is worth future exploration.

Challenges

We encountered some challenges while gathering our data. First, we began by comparing the data received from our own created survey with the public school data we researched online. We compiled information such as graduation rates, average SAT scores and average ACT scores from the public schools. However, a few of the schools that we researched did not post their scores and graduation rates. This was seen as the biggest challenge when collecting data from the other schools.The data from our school was easily collected and analyzed via Google Forms.

Data Sources

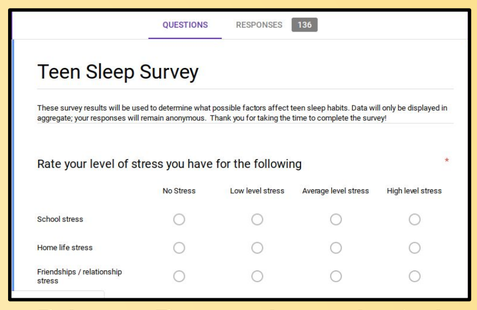
Secondary Data Sources: Our main focus of finding data was to research schools who have moved to a later start time. Once these schools were identified, we went to each school's website or a third-party source to find the most accurate data. We created a Google Sheet and recorded start time, graduation rate, and average ACT/SAT scores for each school. Also, we decided to use schools from different states to gather a variety of data to compare to our survey responses.

Primary Data Source: Next, a survey was sent to approximately 400 students our high school to question students on their sleep habits. In the survey, the students were asked questions such as how many hours are you exposed to screens per day, if they worked, played sports, etc. From the four hundred students we surveyed, around 320 completed the survey. Through Google Sheets ,we created charts to compare the data to the average data of the schools that have a later start time.

Data Set Examples

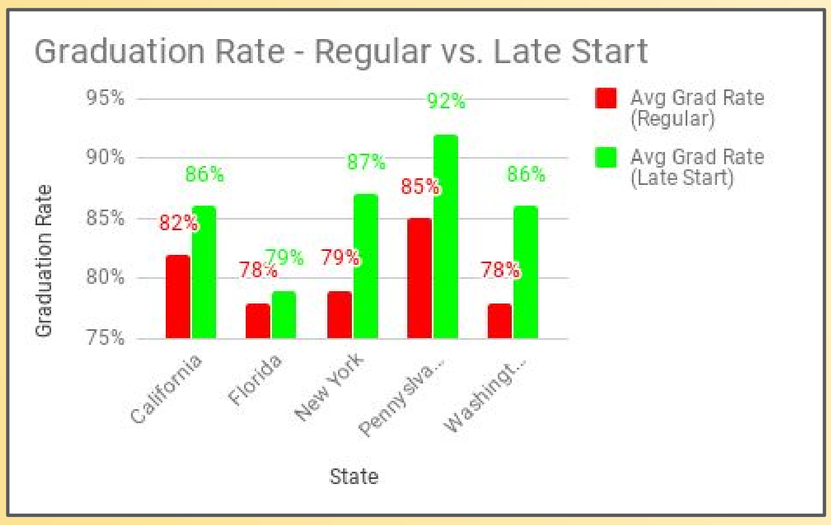
High School	State	Zip code	Start Time	Graduation Rate	Average SAT	Average ACT
Central High School	Alabama	36870	8.05	82%	960	22
Kenai Peninsula Borough High Sch	Alaska	99556	8	90%	1110	23
Mid-Peninsula High School	California	94025	9.3	79%	1260	30
Moorpark High School	California	93021	7.3	94%	1170	24
Novato Unified High School	California	94945	8	97%	1170	23
Palo Alto High School	California	94301	8.15	72%	1210	28
Milford County High School	Delaware	18337	8.25	96%	1160	24
Lake Wales High School	Florida	33853	8	78%	1050	22
Bell High School	Florida	32619	8.15	88%	1040	23
Glades High School	Florida	33065	8.2	91%	1100	23
Bradford High School	Florida	32091	9.1	77%	1040	23
Brevard Public School	Florida	32937	8.3	90%	1260	27
Desoto High School	Florida	34266	8.29	65%	1020	21
Suwannee High School	Florida	32064	8.1	67%	1160	24
Flagler High School	Florida	32164	8	77%	1120	24
Osceola High School	Florida	34741	8.2	83%	1050	22

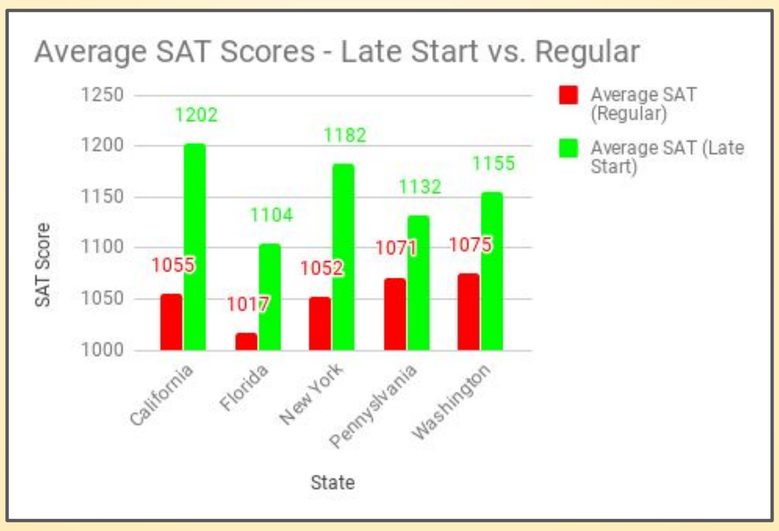
A collaborative spreadsheet was created in Google Docs

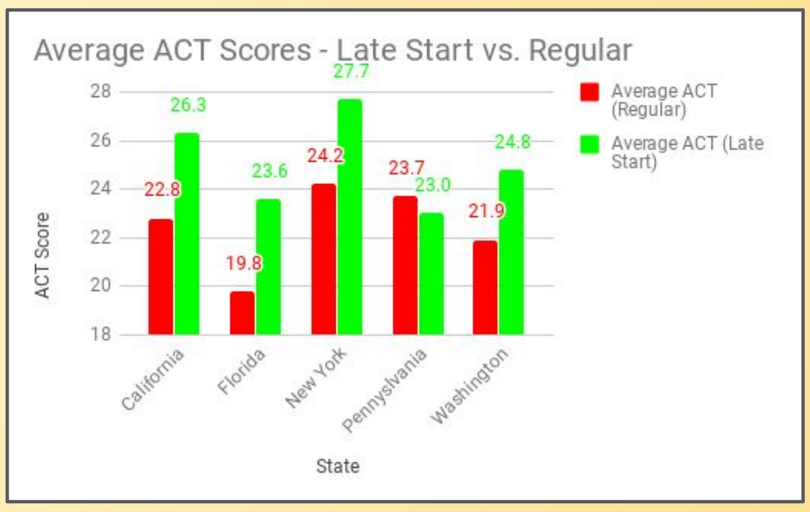


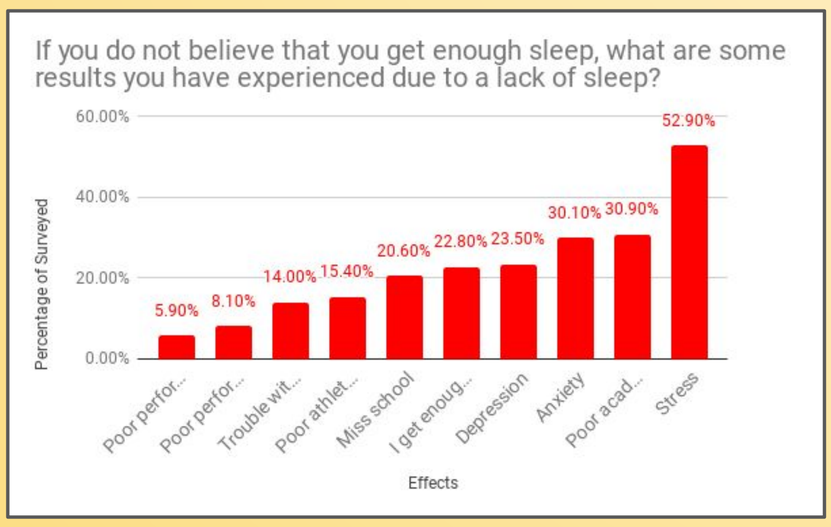
Primary Research conducted via Google Survey

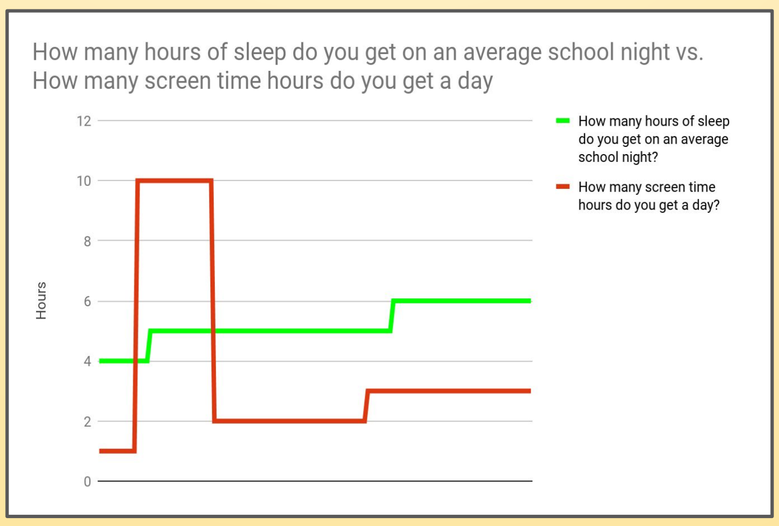
Visualizations

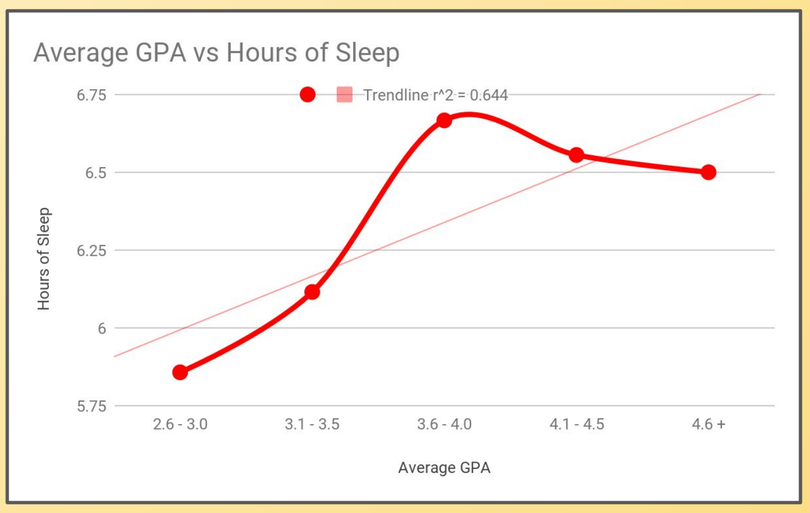












Summary

- There is a strong relationship between Average SAT Scores and Average ACT Scores in comparison to start time. The later start time is, the higher the standardized test scores are.
- Graduation rate and high school start time is positively correlated, on average 5.6% more students graduate high school with a later start time
- According to the National Sleep Foundation, teenagers ages 14-17 require 8-10 hours of sleep a night; however, on average only 15% of teens reported sleeping 8 ½ hours on a school night.
- Students surveyed have numerous side effects that affect academic performance such as stress (52.9%), poor academic performance (30.9%), and Anxiety (30.1%).
- On average, as sleep increased, the amount of screen time a day decreased.
- In order to maintain a 3.6 grade point average and above, students must achieve at least 6.5 hours of sleep a night. As the amount of sleep students get decreases, GPA decreases, creating a positive correlation.

Policy

- High schools should consider beginning classes at a later start time
  - North Allegheny, Fox Chapel, and Hampton are considering this issue currently in the school district. Quaker Valley now starts at a later time.
  - However, some of these considerations are crumbling due to outside issues such as bus and parent scheduling.
- It is clear through the data and visualizations that a later start time has a direct effect on test scores and academic performance as a whole.
  - In states such as California, Florida, New York, Pennsylvania, and Washington, the average ACT and SAT scores are much higher for schools with a later start time.
  - It was also found that GPA increases when one's amount of sleep increases.
- Considering later start times would reduce anxiety among teachers and students
  - It was found the leading cause of lack of sleep in students was stress
  - Also, later start times will provide teachers more time to prepare for classes, tutor students and get work done

The Effects of Food Deserts on Obesity Rates in Allegheny County

Introduction

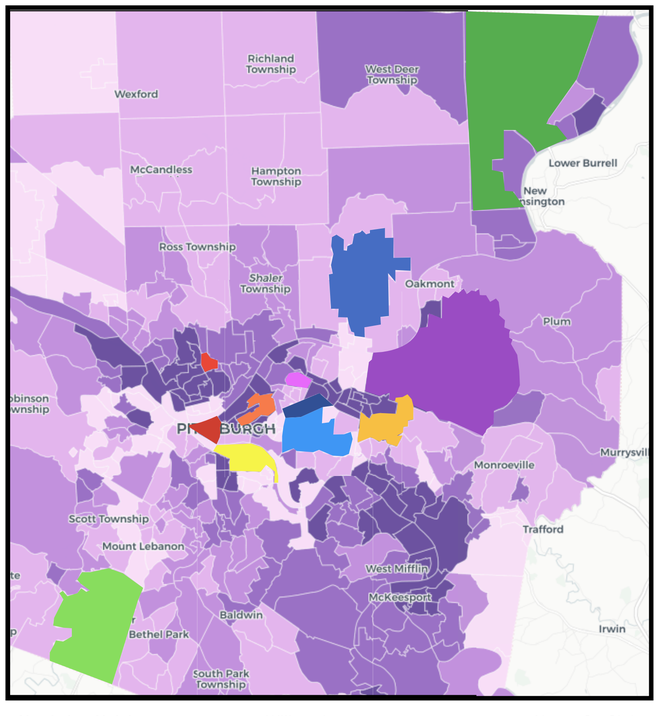
Health can be affected by a plethora of variables, including what food you put into your body. This is greatly influenced by environment and access to different kinds of foods. We used data about food deserts in Allegheny County recorded from the years 2006 to 2010.

Data sets used (wprdc.org):

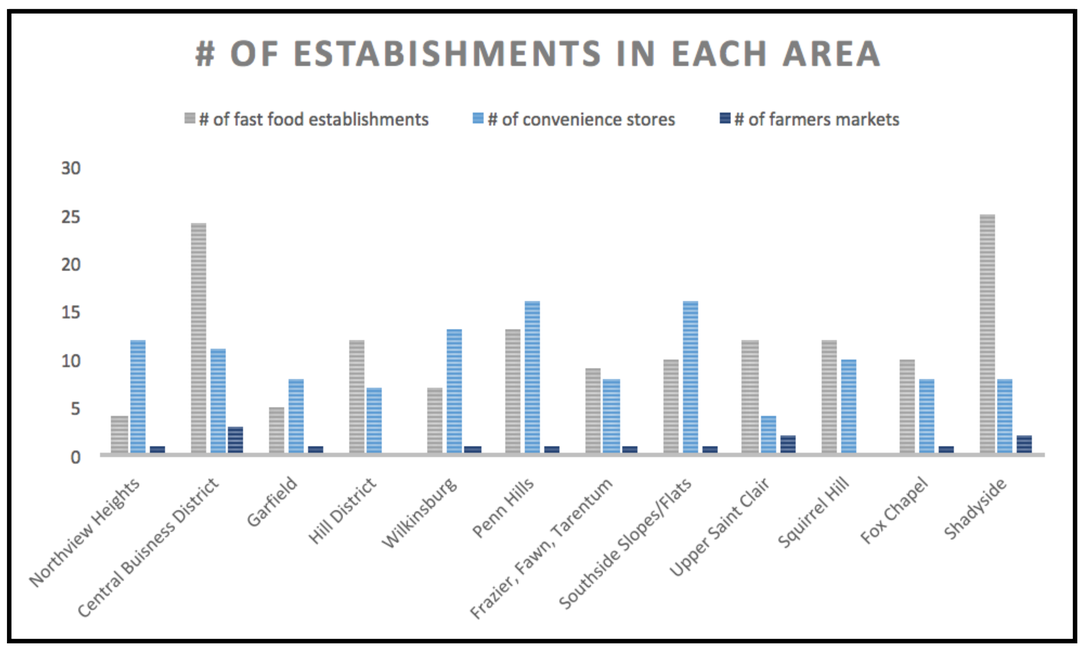
- Allegheny County Fast Food Establishments
- Allegheny County Obesity Rates:
- Allegheny County Supermarkets and Convenience Stores
- Allegheny County Farmers Markets Locations

Process

- Filter data sets (eliminate unnecessary columns, categorize, prioritize, time frame)
- Sort by neighborhood/zip codes, compile each set of data for specific areas
- Calculate amount of establishments in each area per capita (used population to determine how many there would be per 100,000 people)
- Make scatter plots to find possible correlation
- Form conclusion based off of scatter plots and consider outside factors

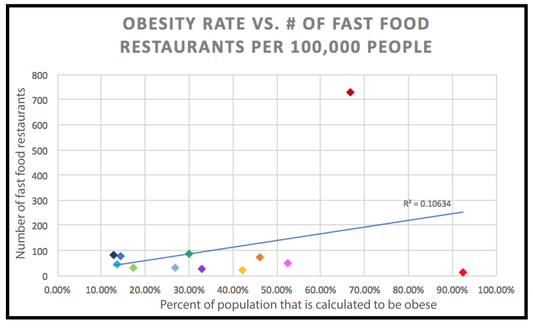


- Northview Heights
- Central Business District
- Garfield
- Hill District
- Wilkensburg
- Penn Hills
- Frazer, Fawn, and Tarrentum
- Southside Slopes/Flats
- Upper St. Clair
- Squirrel Hill
- Fox Chapel
- Shadyside



Examples of Fast food restaurants included in data

- Arby's
- Ben & Jerry's
- Bruegger's Bagels
- Chipotle
- Crazy Mocha
- Domino's Pizza
- Dunkin' Donuts
- Five Guys Burgers and Fries
- Jimmy John's
- McDonalds
- Panera
- Subway
- Starbucks
- Wendy's

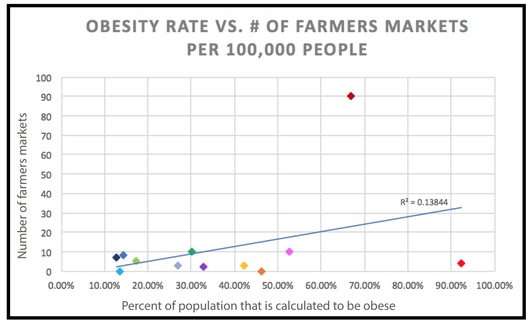


Examples of Convenience Stores included in data

- Beechview Farmer's Market
- East Liberty Farmer's Market
- JL Kennedy Farmer's Market
- Market Square Farmer's Market
- Mellon Square Farmer's Market
- Pennsylvania Market Building

Examples of Farmer's Markets included in data

- Bryant Street Market
- Forbes Avenue Market
- Garfield Hill Market
- Giant Eagle
- Kuhn's Market
- Kwik E Mart
- Le's Super Grocery Kart
- Sam's Market
- Sunoco
- Trader Joe's
- Walgreens
- Whole Foods



Expectations with scatter plots and r2 value:

- Positive correlation between fast food establishments and obesity rate in each area
- Obesity rates would be smaller in areas with more farmers markets
- Little to no correlation between amount convenience stores and obesity rate per area

Analysis

We found that there was little to no correlation between farmers markets, convenience stores, and obesity rates in Allegheny County. Our hypothesis was not supported by the data analyzed regarding the obesity rates we used for eleven areas in the county, and the amounts of fast food restaurants in each area. In order to have the most accurate analyzation, we used the population of each area from 2010 to find the number of each fast food restaurants if there were about 100,000 people living in each of the eleven places. Although this did clarify our results and make them more useful in analyzing the issue at hand, due to population size, we did not find the results we expected.

Conclusion

These results may not have aligned with our predictions for a number of reasons including how the original obesity data was collected by the Allegheny County Department of Health. They formulated the data points for each zipcode using a statistical model of a demographically similar area to Allegheny County, rather than the specific data of the region's populants. Statistical modeling was used rather than real data because of the Health Insurance Portability and Accountability Act (HIPAA), the privacy act which prevents doctors from sharing confidential medical information about their patients. Due to HIPAA, the obesity rates could have differed from the data that was calculated. Obesity is also a matter of other lifestyle habits such as exercise and income, and these factors could play a role in why the obesity rates did not differ much based off of if the area was a food desert.

Sources

data.wprdc.org
www.wesa.fm
Could Development Be Shaping Health in The Hill District?

By Sydne Ballengee, Sierra Brandegee, Katharine Ference, Samantha Honig, Kaia Iverson, Miranda Lightner, Carly McGuire, Renee Petersen, Amelia Rosenstock, Alanna Steals, Savanna Stein, Alison Taylor, Olivia Wilson, Abi Zimmerman
From The Ellis School

On the Relationship Between Distance to Pittsburgh and School Performance Metrics

Introduction:

Education is an integral part of our society and thus much research goes into what makes a school excellent. Typically school performance is thought of in terms of affluence or ethnicity, but we decided to think more geographically. It's common thought that inner-city schools are worse than suburban schools and so we wanted to test if that was true and if this proposed trend continued out into rural schools.

Our Prediction:

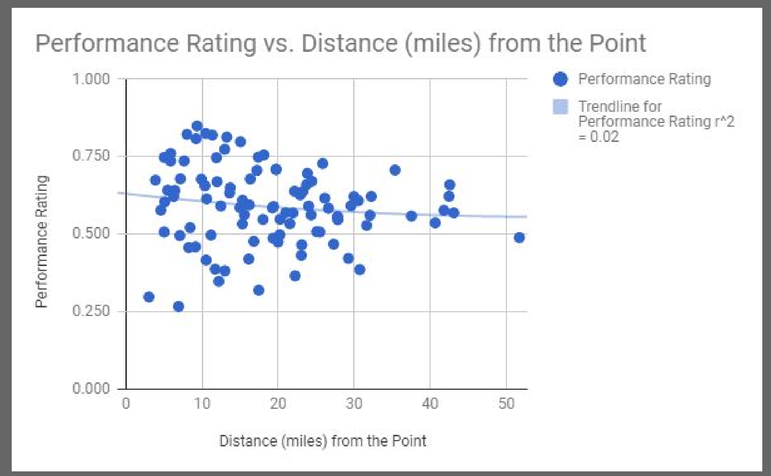
We expect schools close to Pittsburgh to perform below average, then for school performance to increase toward the suburbs and then taper off again with rural schools in adjacent counties.

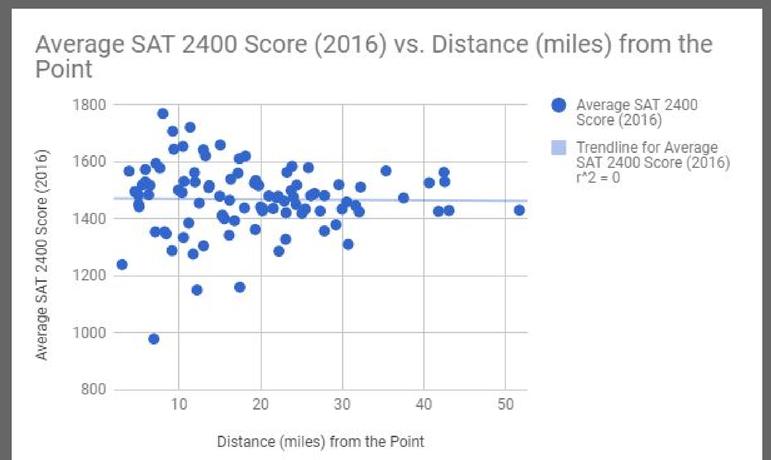
Procedure:

First, we had to catalog all the school districts in Allegheny, Butler, Beaver, Washington, and Westmoreland counties. Then we manually inputted data for SAT scores, Niche.com rankings, reading and math proficiency, expenditures/student, and other possible confounding factors. We obtained this data from the Pennsylvania Department of Education, NIche.com, and several school districts' websites. We then used an online map to calculate the distances from Point Park in Pittsburgh to the high schools of each district. We combined 4 of our metrics (average SAT score, Niche ranking, reading proficiency percent, and math proficiency percent) into an aggregate "Performance Rating" which we used to analyze our research question.

Data Sample: (only a small portion of spreadsheet shown)

	School District	Ring N	Distance (miles) fi	Performance Rati	Average SAT	Niche School	Niche score	Niche Teacher	Niche Teacher	Reading	Math f	N/S	Expenditure/Student	Dropout	Number of Years	LIF (%)
3	Sto-Rox	1	3.01	0.298	1240	C-	4	C-	4	21	13	North	$13,951.94	2.45%	12.05	79.1
4	Keystone Oaks	1	3.87	0.674	1566	A-	11	A	11	68	53	South	$22,106.26	0.78%	12.81	31.3
5	Carlynton	1	4.58	0.578	1496	B	8	B+	9	62	40	South	$22,106.26	1.41%	12.29	51.4
6	Northgate	1	5.02	0.507	1494	C+	6	B	8	55	36	North	$15,725.58	2.86%	14.37	54.6
7	Mt.Lebanon	1	5.04	0.748	1452	A	11	A+	12	81	66	South	$16,916.31	4.00%	13.98	10.7
8	Brentwood Borough	1	5.1	0.604	1443	B	8	B	8	69	46	South	$19,074.57	0.91%	13.42	42
9	Shaler	1	5.47	0.642	1519	B+	10	B+	9	68	42	North	$16,573.11	0.61%	13.62	38
10	Montour	1	5.86	0.736	1530	A-	11	A+	12	78	61	North	$21,137.47	0.78%	13.99	27.6
11	North Hills	1	5.87	0.761	1574	A	11	A+	12	81	66	North	$16,334.68	0.69%	15.23	24.8
12	Chartiers Valley	2	6.42	0.641	1518	B+	9	A	11	69	49	South	$16,985.36	0.64%	14.67	30.7
13	Wilkinsburg Borough	2	6.91	0.267	979	D+	3	C-	4	26	15	South	$24,244.46	3.47%	18.07	87.1
14	Baldwin Whitehall	2	6.26	0.621	1465	A-	8	A	11	69	51	South	$15,475.79	1.08%	13.77	42.2
15	Steel Valley	2	7.08	0.495	1355	B	8	B+	9	46	29	South	$17,864.29	0.00%	7.11	67.1
16	Avonworth	2	7.15	0.679	1526	B+	9	B	8	80	50	North	$16,477.35	0.29%	13.1	20.3
17	Bethel Park	2	7.64	0.736	1579	A	11	A+	12	78	59	South	$18,110.26	0.18%	15.04	14.4
18	Upper Saint Clair	2	8.01	0.822	1770	A+	12	A+	12	86	69	South	$18,287.96	0.00%	13.13	9.8
19	Woodland Hills	2	8.24	0.457	1355	B-	7	B	8	41	27	South	$16,933.76	2.86%	11.15	62.6
20	West Mifflin	2	8.43	0.521	1348	B-	7	B	8	58	36	South	$16,678.97	0.87%	13.58	54.7
21	Cornell	3	9.14	0.459	1289	C+	6	B+	9	52	28	North	$17,842.50	0.00%	11.29	100
22	Fox Chapel Area	3	9.22	0.808	1708	A+	12	A+	12	83	69	North	$20,292.72	5.00%	15.54	19.4
23	South Fayette	3	9.35	0.849	1645	A+	12	A+	12	91	80	South	$14,722.38	0.31%	9.67	12
24	South Park	3	9.91	0.677	1502	A-	10	A-	10	73	52	South	$15,252.86	0.32%	12.23	22.8
25	Duquesne City (K - 8)	3	9.96	#VALUE!	N/A		C-	4	B-	7	25	15	South	$22,747.64	NA	10.94	79.7
26	West Jefferson Hills	3	10.38	0.656	1493	B-	7	A-	10	84	58	South	$14,145.21	0.22%	11.31	17.1
27	Hampton Township	3	10.48	0.825	1665	A+	12	A+	12	88	73	North	$15,019.60	0.13%	13.75	11.8
28	Penn Hills	3	10.54	0.417	1335	C+	6	C	5	41	20	North	$19,169.74	1.36%	14.78	68.7
29	Riverview	3	10.6	0.614	1532	B	8	B+	9	73	42	North	$16,800.50	0.21%	13.54	46.3





Analysis:

- The top scatterplot to the right represents the culmination of our data. We plotted the Performance Rating vs. Distance to Pittsburgh for all the schools in the counties we studied. The plot clearly indicates no connection between the variables, and the r^2 value numerically demonstrates that fact.
- Instead of using a linear best-fit line, we tried to use a polynomial in order to better represent the increase-then-decrease we expected. The polynomial best fit line does not better represent the data, and contrary to our prediction is slightly concave up,
- It is somewhat interesting also that the data seem to approach the average as they get farther from the city, indicating that there is more variance in school performance in the 5-20 mile range. This points to a rather interesting conclusion: close access to the city allows a school to be exceptional, but general success is not contingent upon such access.
- Comparing our performance rating to just SAT scores, we found a very similar trend, which of course points to SAT scores meaning a higher rating, but the extreme data tended to pull toward the middle for the aggregate metric, meaning that extremely good SAT scores don't necessarily mean extremeness in other metrics.

Challenges:

- The data we needed was not easy to find, nor in an easy-to-use form from the sources.
- Additionally, confounding variables naturally plagued our analysis because school performance is the result of many different variables. Here are some variables that we anticipated:
  - Teacher Experience
  - Spending per student
  - Dropout Rate
  - Low-income Family
- Our results don't necessarily point themselves to a clear recommendation.

Further Interests:

- While inputting some data, we noticed that schools on the north side of the city tended to be above average schools, which raises the question whether the north side tends to be better than the south side of the city in terms of performance.
- We'd also like to look more into what other unconventional factors might be connected to school performance.

Recommendation:

- Our analysis indicates that parents looking for school districts should expect both the best and worst options between about 7 and 15 miles from Pittsburgh. However, they need not be afraid that rural schools are exceptionally bad nor that there are no good options very close to the city either, since there is no evident correlation between distance and academic quality.

References:

- "Explore Schools and Neighborhoods." Niche, www.niche.com/.
- "Performance Profile." Welcome to PA School Performance Profile, www.paschoolperformance.org/.

Names: Owen Chase, Jerry Chen, Ivan Voinov, Richard Yan, Gloria Ye, Andrew Zhang,

A Machine Learning Approach to Diagnosing Common Thorax Diseases

Fox Chapel Area High School

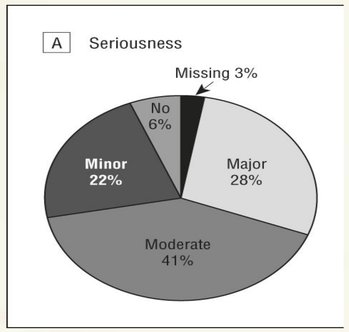
Justin Choo, Rajeev Godse, Arnav Gupta, Zachary Lakkis, Albert Liu, Andreas Paljug, Max Wolfendale

Research Question

Can we use a convolutional neural network to predict common thorax disease diagnoses from chest X-ray images with at least 50% percent accuracy?

Background

- According to Johns Hopkins University, "...diagnostic errors could easily be the biggest patient safety and medical malpractice problem in the United States."
- 87% of all cases have some diagnostic error, of which 28% were life-threatening or resulted in the patient's death or permanent disability.

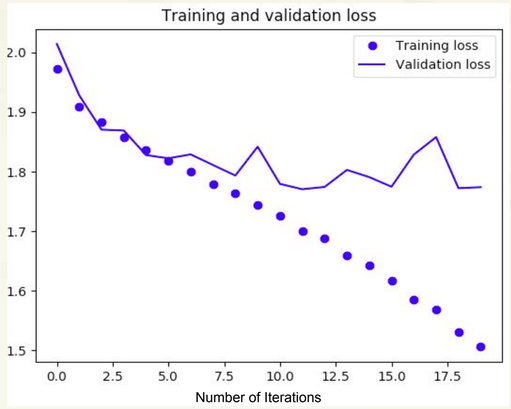


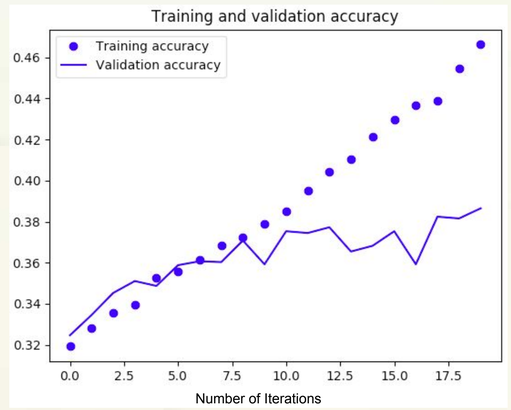
Dataset

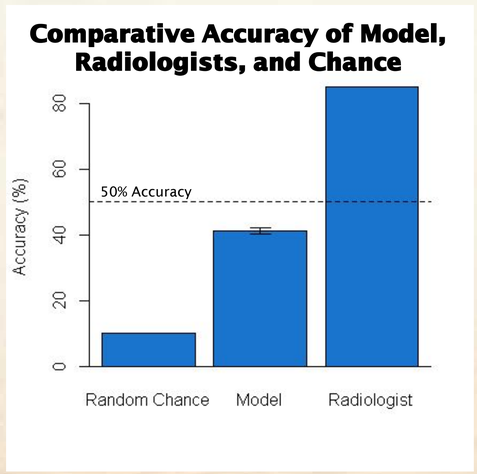
	Diagnosis	Visit#	ID	Age	Gender	View	Image Re:	Image Pixelation
2	Cardiomegaly	0	1	058	M	PA	7372818	0.143
3	Cardiomegaly Emphysema	1	1	058	M	PA	7897726	0.143
4	Cardiomegaly Effusion	2	1	058	M	PA	5120000	0.168
5	No_Finding	0	2	081	M	PA	5120000	0.171
6	Hernia	0	3	081	F	PA	7722762	0.143
7	Hernia	1	3	074	F	PA	5120000	0.168
8	Hernia	2	3	075	F	PA	5120000	0.168
9	Hernia Infiltration	3	3	076	F	PA	8069718	0.143

- Our Data is a set of 118,120 chest X-rays and their diagnoses provided by the National Institute of Health.
- This data shows the diagnoses of chest X-ray images, as well as the image size and resolution, view position, patient age and gender.
- We used R and Python to filter the data to only include single disease diagnoses and those with over 1000 occurrences.
- We finally converted the images into 128x128 matrices of grayscale pixel values to be inputted into a 3-layer Convolutional Neural Network to be classified into 1 of 10 possible disease labels.

Results







Analysis and Conclusion

Analysis:

- Training loss, a testing value that the model attempts to minimize over time, decreased over each successive iteration.
- Accuracy increased over each iteration. Final accuracy was 41.14%.
- Chi-squared analysis at $\alpha=.01$ showed that our accuracy was better than random chance ($p=2.1928E-2342$), but still worse than 50% ($p=2.9368E-70$).
- Our validation loss and accuracy stopped improving around 8/20 iterations, indicating that the model is overfitting the training data.

Conclusion:

- We were unable to attain a 50% accuracy with our 3-layer Convolutional Neural Network.
- We are significantly more accurate than random chance, showing that such an approach to diagnosis could have promise in the future.

Challenges in Gathering Data:

- The images had to be reduced in resolution by 64-fold in order to preserve computing time.
- The images did not have the same viewing frames and had slightly different angles of chests.
- Over 50% of the images were healthy, reducing the number of diseased images that went through the network.

Implications and Future Studies

- Machine learning is a novel approach to the diagnosis of diseases.
- This project demonstrates the potential for such an approach.
- In future work, we hope to use higher resolution images with similar view frames along with a deeper convolutional neural network.

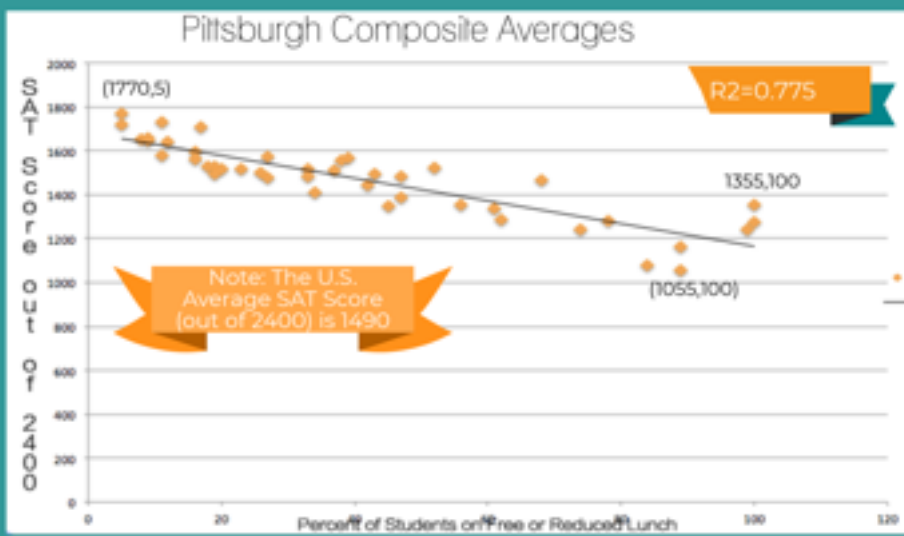
SATs & Me

Oakland Catholic Team 1

The Question:

Across the US, wealthier students rank higher on standardized tests. This "achievement gap" seems to be directly correlated to family income level. Are these trends as strong in the Pittsburgh area? What can Pittsburgh do to close this gap?

The Data:



Source: PA Dept of Education

This graph, which has a coefficient of correlation of .775, shows us that there is a strong correlation between the percentage of students on free or reduced Lunch and SAT scores of a school district. There no extreme outliers from this trend.

The Conclusion:

We can confidently conclude from the data, that Pittsburgh has a serious education gap. Higher income areas are scoring substantially higher than lower income areas (a difference of over 700 points between the highest income and lowest income). As a solution, we propose free sat prep classes at libraries and schools in lower income areas, summer programs for high school students that prevent learning loss over the summer, and the promotion of free resources such as Khan Academy.

Criminal Rates of Pittsburgh Investigation

Sylvia Li; Katie Henningsen; Grace Antonic; Elizabeth Gu; Eliane Rectenwald; Helen Tan; Jasmine Alston | Oakland Catholic High School Team 2

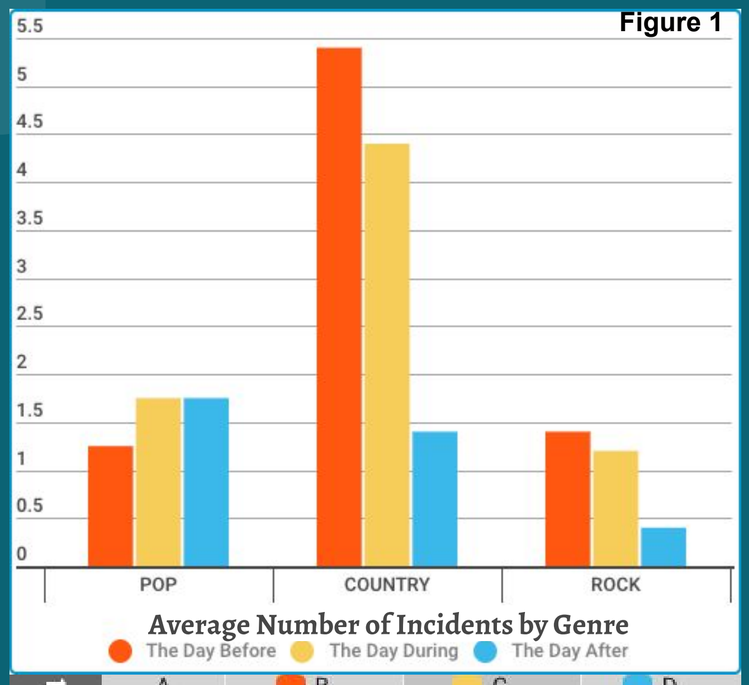
Introduction

Research question: *Do specific genres of music shows/concerts held at large scale venues have higher criminal rates over others?*

Importance: This project is important because it could show officials that more security is needed if we find any trends. We can suggest that local police and security take certain genres into consideration whenever an event is held so that people's safety is ensured. For instance, Kenny Chesney concerts in Pittsburgh are notorious for crime. In 2015, 17 minors were cited for underage drinking, and horrible environmental conditions were created by trash after his concert.

Data & Analysis

Figure 1

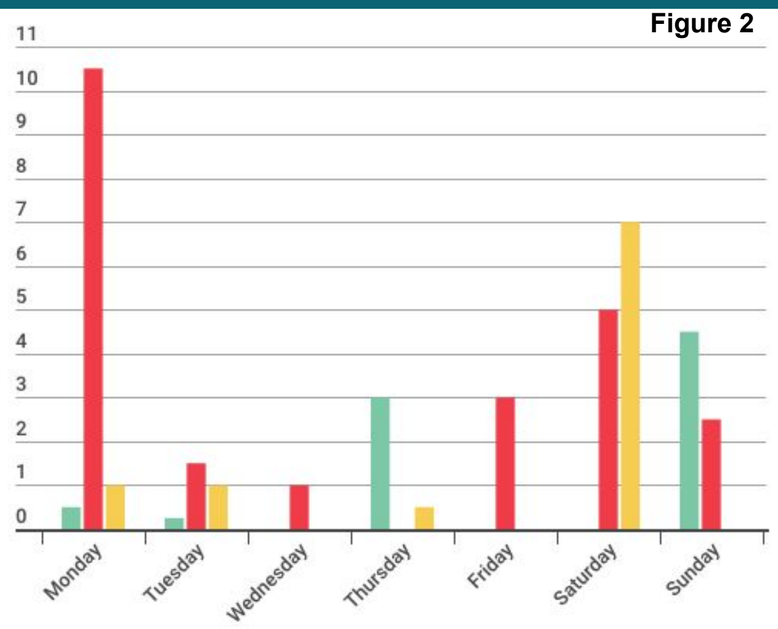


	A	B	C	D
1		The Day Before	The Day During	The Day After
2	POP	1.25 (5/4)	1.75 (7/4)	1.75 (7/4)
3	COUNTRY	5.4 (27/5)	4.4 (22/5)	1.4 (7/5)
4	ROCK	1.4 (7/5)	1.2 (6/5)	0.4 (2/5)

→ Figure 1:

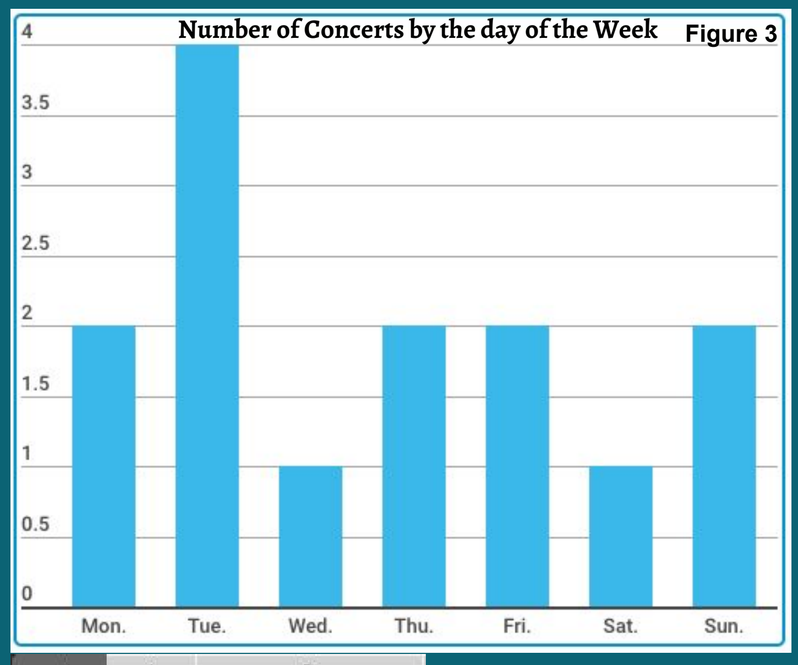
- Rather than making the graph show the total number, we made it to show the average number of incidents per concert.
- In general, crimes tend to take place on the day before and during the concerts actually starts. Only few crimes would take place on the day after, especially for the genre of rock concerts.
- Surprisingly, most of the crimes take place when there's Country Concerts instead of Pop or Rock concerts.

Figure 2



	A	B	C	D
1		The Day Before	The Day During	The Day After
2	Monday	0.5(1/2)	10.5(21/2)	1(2/2)
3	Tuesday	0.25(1/4)	1.5(6/4)	1(4/4)
4	Wednesday	0	1(1/1)	0
5	Thursday	3(6/2)	0	0.5(1/2)
6	Friday	0	3(6/2)	0
7	Saturday	0	5(5/1)	7(7/1)
8	Sunday	4.5(9/2)	2.5(5/2)	0

Figure 3



	A	B
1		Number of Concerts
2	Mon.	2
3	Tue.	4
4	Wed.	1
5	Thu.	2
6	Fri.	2
7	Sat.	1
8	Sun.	2

→ Figure 2 & Figure 3:

- While most of the concerts take place on Tuesday, most of the crimes take place on Monday, especially on the day during the concert.
- There were two concerts taken place on Monday, one was held by Lady Gaga (pop music) with the number of crime 6, and the other one was held by Steve Moakler (country music) with the number of crime 15 witch is much higher than that of Lady Gaga's.
- Wednesday and Saturday share the same number of concerts while Friday has even more concerts than Saturday has, yet there are much more crime incidents happening on Saturday than on Wednesday or Friday.

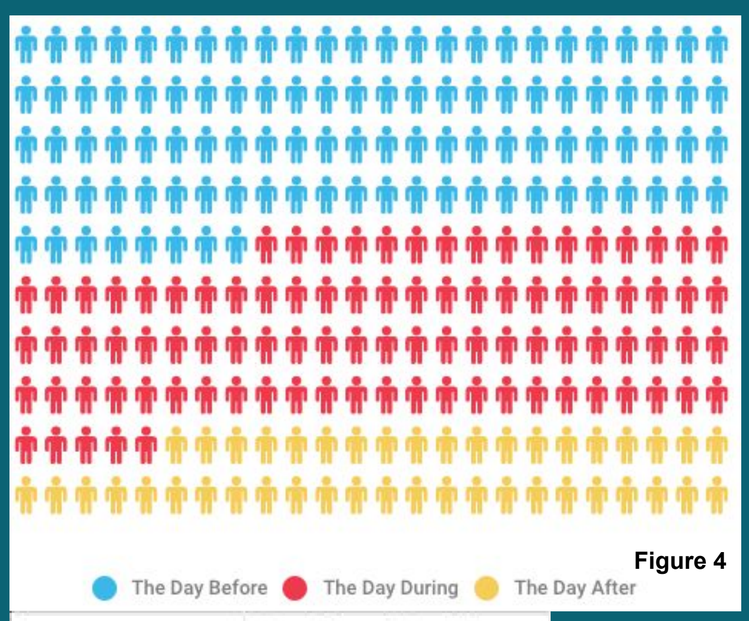
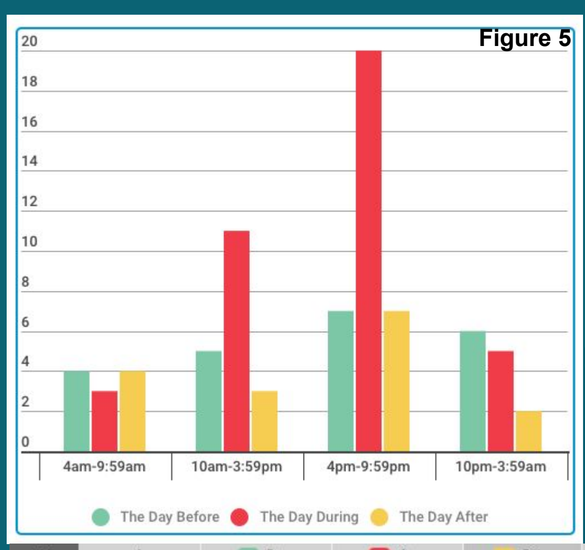


Figure 4

	Number of Incidents
The Day Before	39
The Day During	35
The Day After	16

→ Figure 4: Crimes often take place on the day before and during the concerts and about the same number.

Figure 5



	A	B	C	D
1		The Day Before	The Day During	The Day After
2	4am-9:59am	4	3	4
3	10am-3:59pm	5	11	3
4	4pm-9:59pm	7	20	7
5	10pm-3:59am	6	5	2

→ Figure 5 :

- Number of the incidents which take place at dusk and night(4pm to 9:59 pm) during the day of the concert been held is extremely high compare to others.
- Overall there are more crimes during the concerts.
- Number of crimes of the day before are slightly more than that of the day after.

Figure 6

	A	B	C	D	E	F
1	Location - signer	what type of music is it?	Date	Crime	counting	
2	Heinz Field/PPG paints - person who held it		month/day/year -time		Number of Incidents	Day of the Week
3	PPG Paints Arena -Lady Gaga	pop	11/19/2017 21:42:00	paraphemalia -use or possession	6	Sunday
4			21:37	3332 limitations on turning around		
5			17:35	stop signs and ield signs		
6			16:58	marijuana: possession small amount		
7			3:06	accidents incolving damage to unattended vehcle		
8			0:43	failure to appear/arrest on attachment order		
9		(the day during)	*11/20/2017 21:45:00*		6	Monday
10			*18:45*	*aggravated assult/ robbery*		
11			*10:00*	*possession of controlled substance*		
12			*10:00*	*possession of weapon on school property*		
13			*3:13*	*driving on right side of roadway/stop signs and yield signs*		
14			*1:15*	*rechlessy endangering another person*		
15			11/21/2017	possession of controlled substance	4	Tuesday
16			21:30	possession of small amount		
17			19:53	Theft from vehicle / possession of controled substance		
18			12:45	failure to appear/arrest on attachment order		
19	PPG Paints Arena -Bruce Springsteen	Rock	9/10/2016	/	0	Saturday
20			*9/11/2016 19:20:00*	*failure to comply with Megara Law*	1	Sunday
21			9/12/2016 8:50:00	Simple Assault -intent., know., Reckless	2	Monday
22			18:45	Paraphemalia -use or possession		

→ Figure 6: A sample of the data set that we are using.

Conclusion & Reflection

Conclusion:

In conclusion, the data shows that country concerts are more susceptible to crime and performances on and throughout the weekend shows higher criminality. These findings show that our suspicions based on previous knowledge about country concerts were proven correct. The data showcases the types of concerts and environments that authorities and venue managers should oversee with caution. Also seen in the graphs is the time most crimes occur within close proximity to these entertainment sites. This will give officials an inkling of alertness during such periods in which concerts will take place. Time of day also has a correlation with misdemeanors, for there is an obvious spike in crime during the evening performances. Overall, the data is practical to community leaders and venue managers for the supervision of what precautions should be taken to ensure a safe environment for attendees and populated communities.

Challenges Faced While Gathering Data:

- Unable to find the specific time when crimes were committed for a few of the data points , which leads to the consequence that the data could be misleading.
- Limited availability to find more concerts with all the data, time, location, where and when, the crimes were committed.
- Determining which data was in the area of the concert.

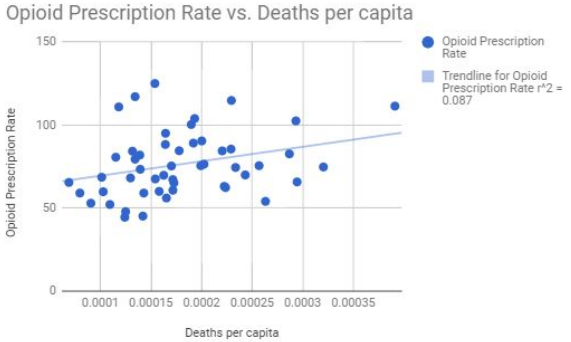
Refrences

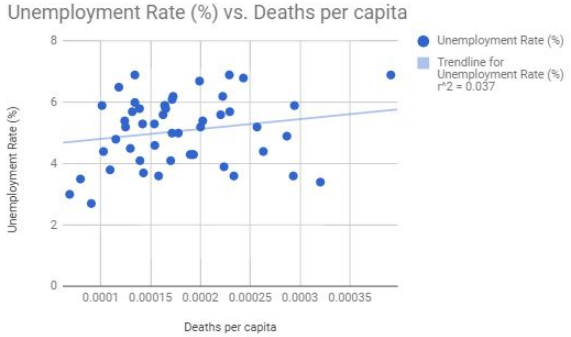
1. http://www.wtae.com/article/citations-issued-in-kenny-chesney-concert-aftermath-1/7471855
2. https://data.wprdc.org/dataset/arrest-data/resource/e03a89dd-134a-4ee8-a2bd-62c40aeebc6f
3. https://www.jambase.com/venue/stage-ae/past-shows?y=2016
4. http://www.ppgpaintsarena.com/events
5. http://www.post-gazette.com/ae/music/2016/12/15/Scott-Mervis-lists-Pittsburgh-s-best-pop-and-rock-concerts-in-2016/stories/201612150019

What Factors Affect Opioid Deaths?

Emily Veltri, Taylor Maida, Aubrey Lutz, David Gubinsky, Joey Black, Savannah Vetterly

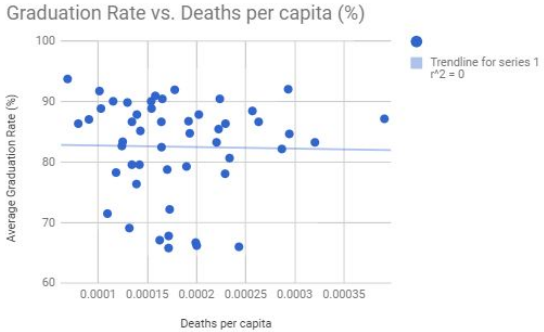
Introduction: As of recently, there has been a strike in opioid deaths, especially in Pennsylvania. We wanted to figure out what factors affected that increase, so we decided to look at unemployment rates, graduation rates, and prescriptions given out to patients. Additionally, we wanted to look at the r values for each of the factors with opioid deaths to see which has the highest correlation. We thought that through these observations we could come to a conclusion about which of these factors affect the opioid deaths, possibly leading to new solutions of this crisis.





Challenges: Some of the challenges we came across in this process were finding data sets that matched together. By this we mean that the data set we were using originally was by region within a single state, but we ended up not being able to use this information because it did not match up with the state by state data sets. Also, we struggled to find a correlation within the data, and this will be discussed in the conclusion.

Methods: We simply just looked up specific data sets on various factors we thought was interesting, such as graduation rates, prescriptions, and unemployment rates. All of the data we used was from the year 2015 in order to remain consistent and it was comparing state to state. With this information we were able to plug in the data in excel and construct graphs. With these graphs we were able to calculate the r and r squared values to find any type of correlation.



Conclusion/Analysis: Overall, we were not able to find any correlation between opioid deaths per capita and unemployment rates, graduation rate, and prescriptions. We calculated both r and r squared, which were extremely low. The r squared value for unemployment rates was .037, for graduation rates it was exactly 0, and for prescriptions it was about .087. We found that there is no correlation for any of the factors we decided to test.

Is there a correlation between the amount of fast food chains in a school district per capita and student test scores in Allegheny County?

(Propel Andrew St. High School: Da'Mani W., Shakeema A., Isaiah M., Ricky J., Nia M., Andre P., Destiny T., Mikayla M., Quincy D.)

INTRODUCTION:

We believe the above question is of importance because in many of the communities in which Propel draws it students, fast food chains are some of the only options for food. Many of the communities would be considered impoverished "food deserts" in which there are few if any locations in which to buy nutritious food items. Because of this, students and their families can be over-reliant on fast food. When researching issues pertaining to food deserts, we found articles in which there appears to be a connection between fast food consumption and student achievement. This inspired our group to push forward to see if there is a correlation here in Allegheny County regarding this issue.

METHODS:

With help from Pitt students and other resources, we analyzed various aspects of the data to see if there is a correlation for the question listed above. These tasks were "chunked" within our rather large team and put together when a group finished it's data gathering. The following are examples of these tasks:

1. Record the total number of students enrolled (K-12) in public school districts in Allegheny County.
2. Find PSSA scores for math, reading, etc. for 8th graders in public school districts in Allegheny County. Record the percentage "PROFICIENT" and "ADVANCED" for each.
3. Record and map the amount of fast food chains inside very public school district boundaries.

This data was compiled and put into an Excel file. Scatter-point charts/graphs were generated from this data. We entered data into the CARTO website to create the necessary maps for data gathering.

CHALLENGES:

-It's a complex question that required a lot of work in gathering the large amounts of data then making sense of it.

-What constitutes a true "fast food chain"? We decided on non-take-out, non-pizza, and non-coffee oriented establishments.

-The realization that here may be outliers because of areas like Monroeville, McKnight Rd., The Waterfront, etc.

-There are often multiple school districts within single zip codes making gathering data on fast food locations extremely difficult.

-We discovered the CARTO website for making maps from data too late.

-Lack of experience, time, and sometimes motivation, for our Data Jam team.

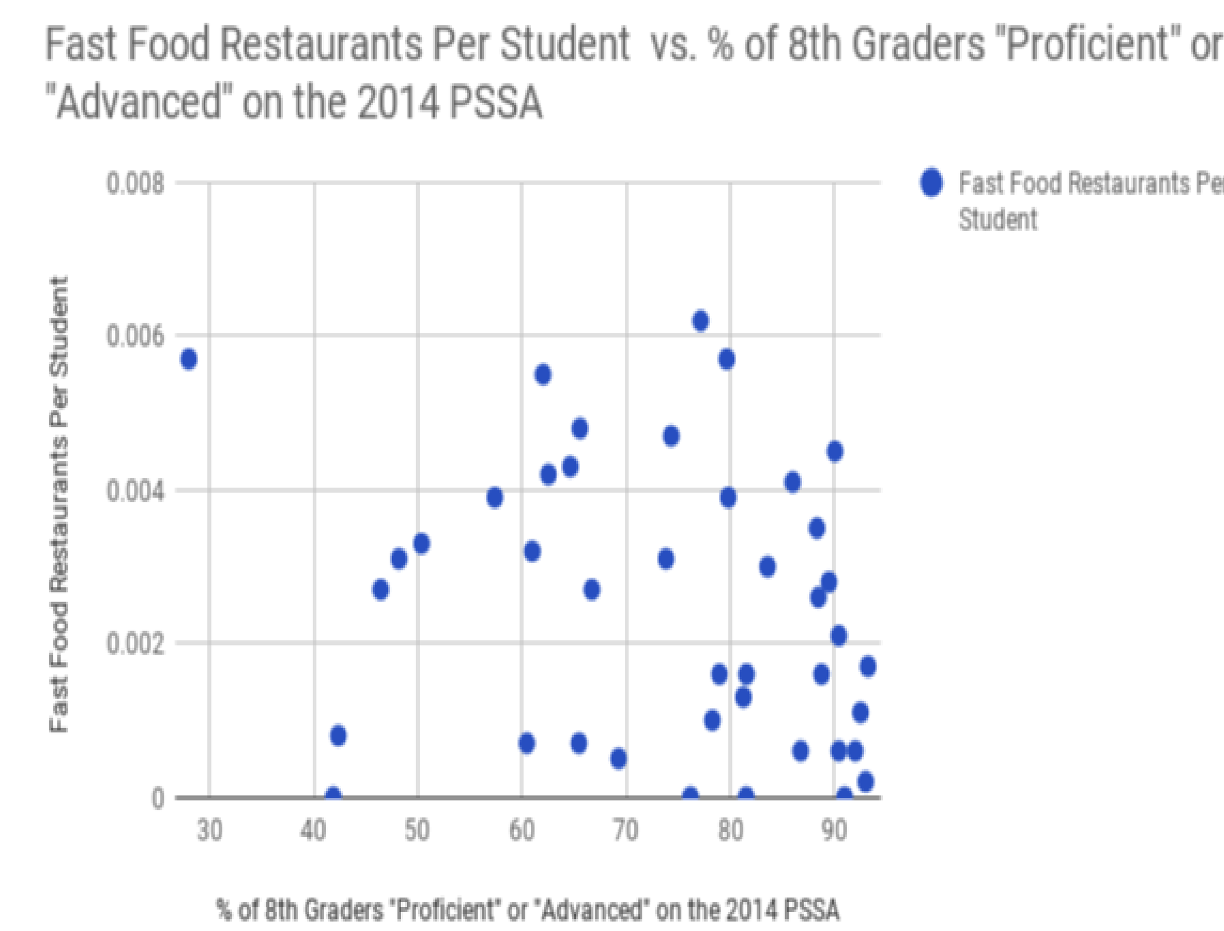
DATA SETS COMPILED:

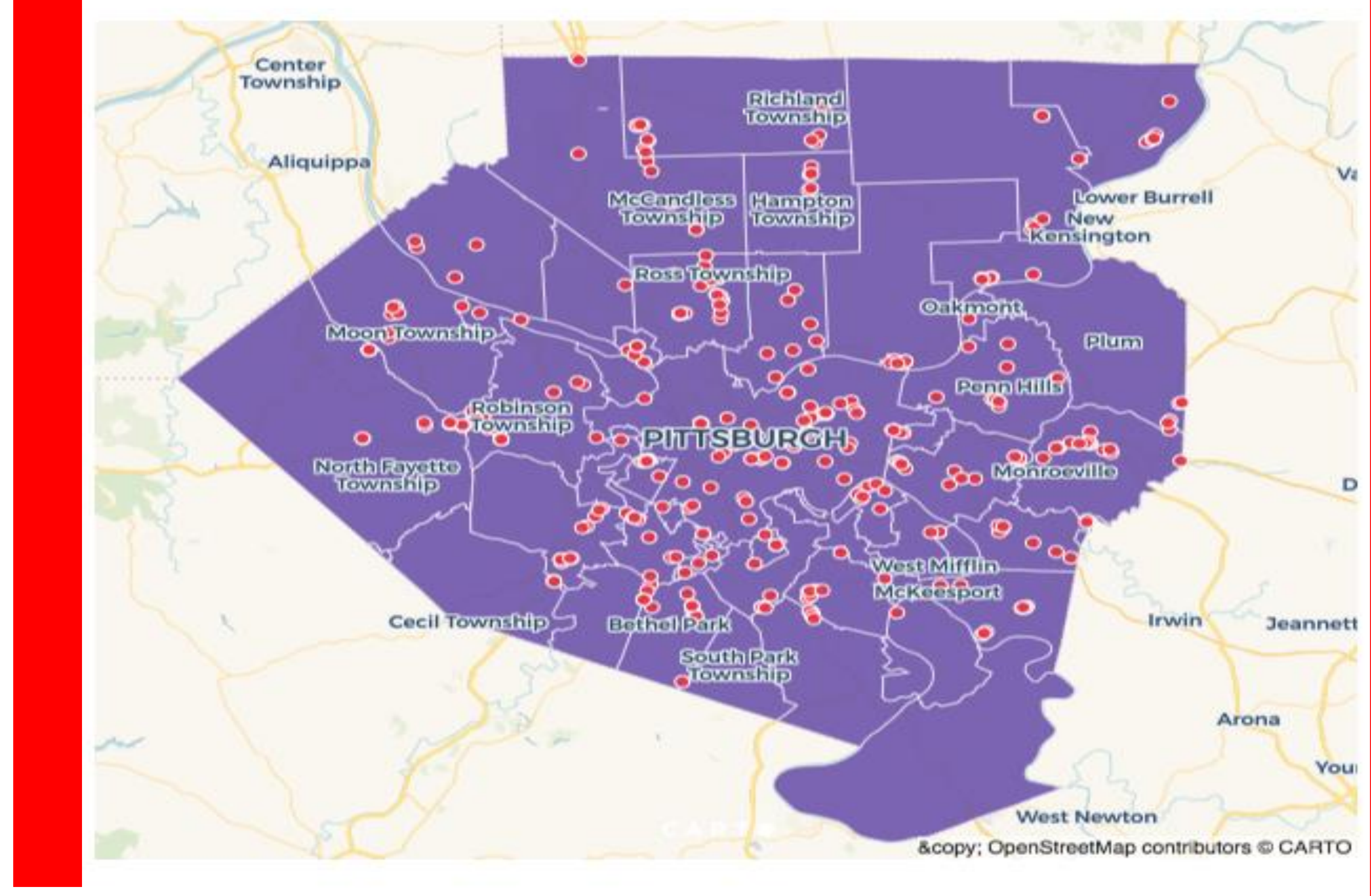
O'Neill Data Jam Complete Stastistics (Student)

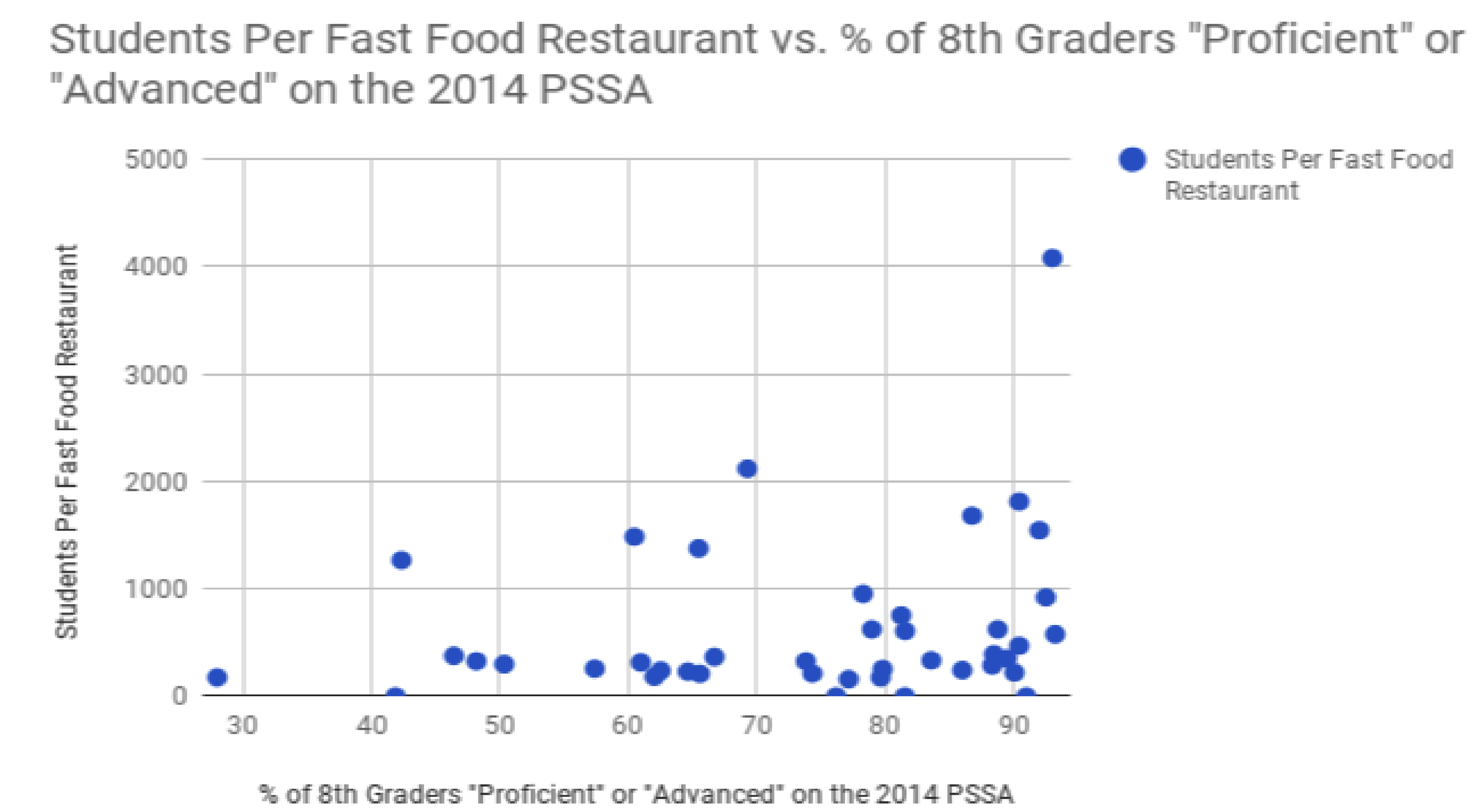
File Edit View Insert Format Data Tools Add-ons Help Last edit was yesterday at 8:03 PM

School District	Total Student Enrollment	% of 8th Graders "Proficient" or "Advanced" on the 2014 PSSA	Total Number of Fast Food Establishments	Students Per Fast Food Restaurant	Fast Food Restaurants Per Student
Allegheny Valley	961	77.15	6	160.16	0.0062
Avonworth	1,682	86.75	1	1,682	0.0006
Baldwin-Whitehall	4,240	69.275	2	2,120	0.0005
Bethel Park	4,293	88.45	11	390.27	0.0026
Brentwood	1,216	81.55	2	608	0.0016
Carlynton	1,378	65.475	1	1,378	0.0007
Chartiers Valley	3,296	79.8	13	253.54	0.0039
Clairton	809	41.875	0	0	0
Cornell	630	65.575	3	210	0.0048
Deer Lakes	1,960	73.825	6	326.67	0.0031
East Allegheny	1,673	62.525	7	239	0.0042

RESULTS:







Analysis/Conclusions:

From observing our data, there is little to no correlation between the amount fast food chains per capita in a school district and student test scores on the PSSA in Allegheny County. We concluded that there may be more overall economic development in areas with high PSSA achievement, including fast food chains, thus not making them "food deserts". However, some high achieving districts such as Upper St. Clair and Fox Chapel have one or zero locations in their districts proper. We also acknowledged that students and families can easily leave district boundaries to get fast food that is still close to home. Essentially, there is no clear data trend from analyzing the scatter point graphs.

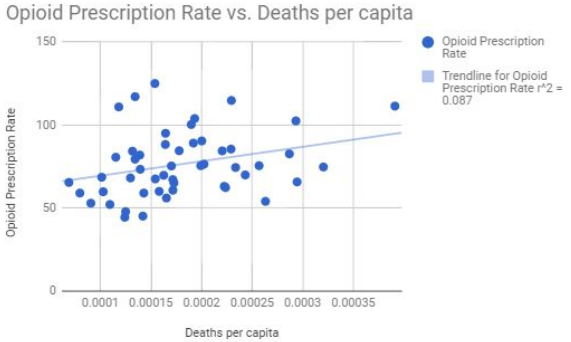
REFERENCES:

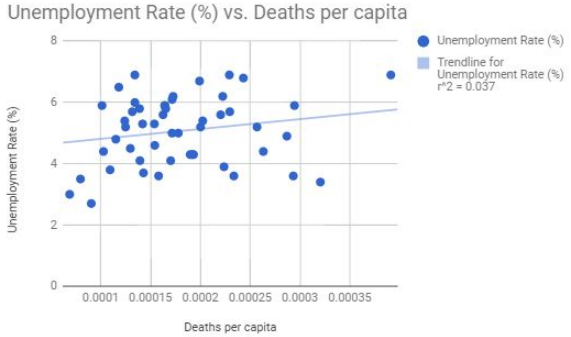
1) PA Dept. of Education: http://www.paschoolperformance.org/2/Districts
2) PA Dept. of Education: https://docs.google.com/a/propelschools.org/spreadsheets/d/1CfMZwNcNXG6Js0IRKJoICb42XuSAP7SgW8hF5YIEoZc/edit?usp=sharing
3) Western PA Regional Data Center: https://data.wprdc.org/dataset/allegheny-county-fast-food
4) Allegheny County GIS Open Data: http://openac-alcogis.opendata.arcgis.com/datasets/f5ac385118344bbf8f707cd88ba31a45_0/data

What Factors Affect Opioid Deaths?

Emily Veltri, Taylor Maida, Aubrey Lutz, David Gubinsky, Joey Black, Savannah Vetterly

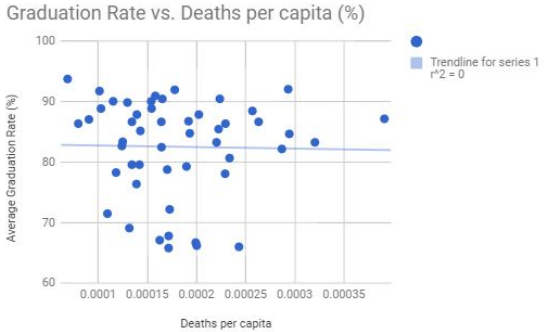
Introduction: As of recently, there has been a strike in opioid deaths, especially in Pennsylvania. We wanted to figure out what factors affected that increase, so we decided to look at unemployment rates, graduation rates, and prescriptions given out to patients. Additionally, we wanted to look at the r values for each of the factors with opioid deaths to see which has the highest correlation. We thought that through these observations we could come to a conclusion about which of these factors affect the opioid deaths, possibly leading to new solutions of this crisis.





Challenges: Some of the challenges we came across in this process were finding data sets that matched together. By this we mean that the data set we were using originally was by region within a single state, but we ended up not being able to use this information because it did not match up with the state by state data sets. Also, we struggled to find a correlation within the data, and this will be discussed in the conclusion.

Methods: We simply just looked up specific data sets on various factors we thought was interesting, such as graduation rates, prescriptions, and unemployment rates. All of the data we used was from the year 2015 in order to remain consistent and it was comparing state to state. With this information we were able to plug in the data in excel and construct graphs. With these graphs we were able to calculate the r and r squared values to find any type of correlation.



Conclusion/Analysis: Overall, we were not able to find any correlation between opioid deaths per capita and unemployment rates, graduation rate, and prescriptions. We calculated both r and r squared, which were extremely low. The r squared value for unemployment rates was .037, for graduation rates it was exactly 0, and for prescriptions it was about .087. We found that there is no correlation for any of the factors we decided to test.

Pittsburgh Allderdice HS
Data Jam Competition
February 27, 2018

Food Safety in Allegheny County

Benjamin Cummings
William Johnson
Kazuma Parkinson
Ashwin Srinivasan
Elijah Zingarelli

Background

- Food facilities such as restaurants, supermarkets, etc. can pose **serious health risks** to consumers through mishandling of items during transportation, storage, and preparation.
- The Allegheny County Health Department's Food Safety Program routinely conducts **in-person inspections** of over 9,000 food facilities in the region in order to ensure compliance with Pennsylvania Article III regulations.
- Establishments found to have **critical violations** can be cited with a warning to consumers, fined thousands per violation, and ultimately closed if necessary.

Source: ACHD

Purpose

We aim to determine whether the **wealth** (i.e. median income) of a food facility's neighborhood is associated with the results of its **food safety inspection**. We hypothesize that areas with a higher income will have, in general, safer food facilities.

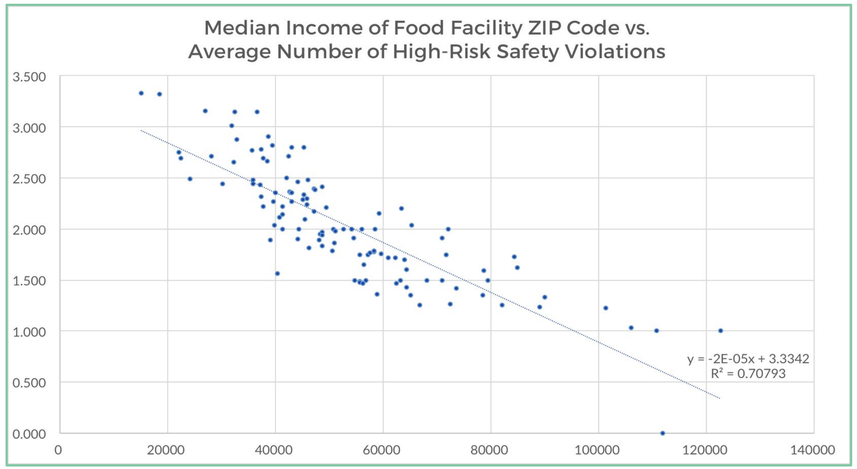
Data

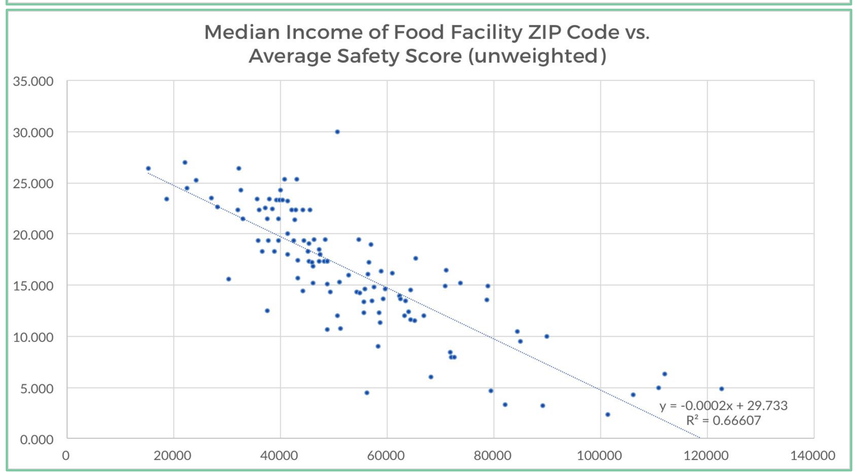
Violation Entry	Example
Facility Name	McDonald's
Facility Type	Chain Restaurant
Address	2925 Freeport Road, 15238
Inspection Date	10/24/2016
Violation Type	Failure to maintain temperature
Violation Risk	Low

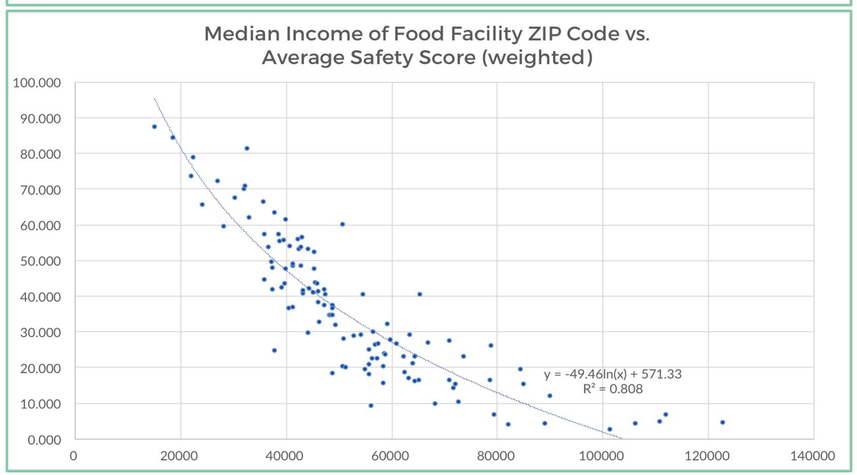
Sources: American Fact Finder, Western PA Regional Data Center

Inspection data were collected from **2012-17**, and included more than **200,000** rows in total. We summed the low, medium, and high-risk violations per food facility, created a **weighted safety score** based on the number of violations of each type, and incorporated the **median income** of the ZIP code of the facility into the dataset used for analysis.

Results







Analysis

- The average number of high-risk violations shows a **strong negative correlation** with median neighborhood/ZIP code income; $r^2 = 0.708$ (70.8% of the variability in the data can be explained by the linear association with income).
- Similarly, the overall unweighted safety score shows a strong negative correlation with median income; $r^2 = 0.666$. Since the safety score is the number of violations per food facility, a **lower score is better** as far as safety is concerned.
- The negative correlation with median income becomes stronger ($r^2 = 0.808$) when the scores are weighted — 1 point for low-risk violations, 2 for medium-risk, and 3 for high-risk. However, the trend is more **logarithmic** than linear (a linear association has $r^2 = 0.709$). Interestingly, the weighted score is usually close to the unweighted score multiplied by the number of high-risk safety violations for a food facility.

Challenges

- The data span 5 years, so there may be **inconsistencies over time** in the quality and procedure of inspections.
- Allegheny County has not established a clear method of categorizing safety violations, so the distinction between risk levels could affect the interpretation of the data.
- Safety inspections can be **subjective**, and assessment of a food facility may differ between department officers.
- **Lurking variable:** number of food facilities per ZIP code.

Conclusion

- **Food facilities in lower-income areas are likely less safe.**
- The Health Department should provide funding to improve food safety in areas that receive worse safety scores.
- A more objective inspection system should be developed to avoid bias from perception of a food facility's surroundings.

Does increased walkability of Pittsburgh townships correlate with the income of people living in them?

Background and Purpose

Purpose: To determine the relationship between median household income and walkability of townships in or surrounding Pittsburgh

We picked this topic because of its ambiguity. In areas that are walkable, it is probable that wealthier citizens will be attracted and thus will fund gentrification, leading to decreasing distances between amenities. On the contrary, car-dependent, suburban areas are a traditional sign of wealthier households, and these oppose the notion that walkability corresponds to a greater income.

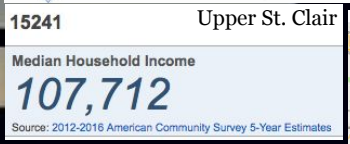
Gathering Data

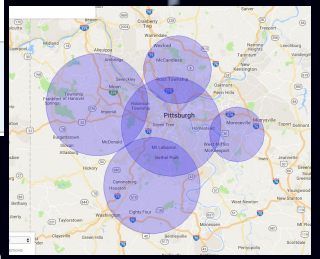
We collected data from different zip codes in and surrounding Pittsburgh. We later split this into larger geographical regions.

Walk Score® data: We used www.walkscore.com to determine walkability for each zip code

- A Walk Score is a value between 0 and 100 (100 being the best) which represents analysis of walking distance to "nearby amenities," population density, block length, intersection density, and more.

Income data: We gathered median household income data from the US Census Bureau by zip code

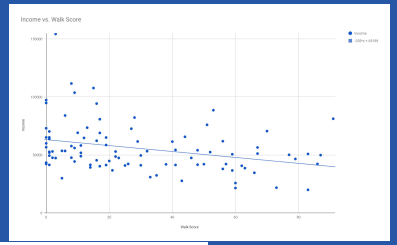


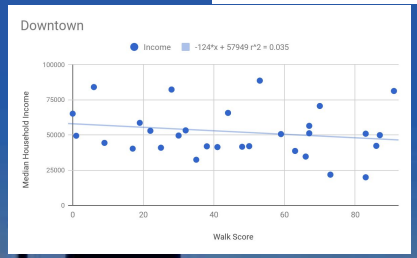


Map (right) shows shows the population density of the 5 regions of Pittsburgh that we studied. This is important to understanding regional dynamics corresponding to income and walkability.

Data and Results

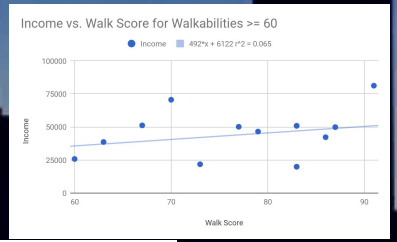
Graph (right) displays the relationship between the income and walk score of all Pittsburgh zip codes.

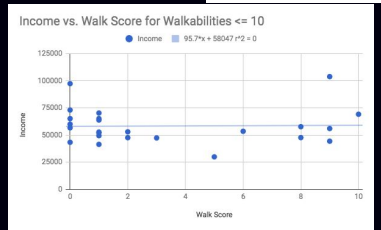




Graph (left) shows the relationship in downtown Pittsburgh zip codes. This is 1 of 5 regions that we analyzed.

Graph (right) portrays the relationship between income and areas with very high walkability (≥ 60).





Graph (left) shows the relationship between income and areas with low walkability (≤ 10).

Analysis and Conclusion

Correlation: There is a negative relationship between income and walkability.

Challenges in gathering data:

- The data was accessible but not consolidated. We had to type in zip codes one by one to obtain each data point.
- Grouping zip codes into smaller regions proved difficult because of unclear division lines. We ended up using Interstate roads to make divisions.
- For data regarding income, participants may have been reluctant to admit their true income.

Conclusion and Discussion:

- We observed a negative correlation for all of the zip codes in and surrounding Pittsburgh overall.
- After analyzing extreme walk scores and each of the 5 regions, there was no clear pattern that seemed to drive the overall weak, negative trend we were seeing.
- Pittsburgh was founded in 1758. A mixture of existing infrastructure and social/economic status with new opportunities in downtown could be the source of low correlation.

Recommendations:

- College students tend to thrive in high walkability areas as a result of the low cost of living. Being a broke college student, it is easier to live in an area in which a car is not a necessity.
- It would be unwise for businesses that attract exclusively wealthy crowds to buy property in areas with high walkability.
- Alternatively, businesses that offer affordable products should buy property in areas with high walkability.

Upper St Clair - Kriti Shah, Brooke Christiansen, Jack Clark, Jack Gordley, Dylan Jenny, Meghan Joon, Benny Pribanic, Hannah Trivedi, Mallika Matharu, and Dina Leyzarovich

The Price Efficiency of Bike Riding

Purpose

- The purpose of this project is to **determine if Healthy Ride's free bike sharing solution is a feasible alternative to other modes of transportation** in Pittsburgh.

Criteria

- For Data (We Assumed):
  - Acceptable Weather
  - The Individual Commuting is Healthy
  - Average Traffic Conditions
- For Locations:
  - Pittsburgh Places of Interest

Healthy Ride

Costs of Transportation

- Biking (Healthy Ride)
  - Under 15 min: **Free**
  - After 15 min: $2 per 30 min
- Walking
  - Free
- Bussing (Port Authority)
  - $2.75 per ride

Mining Data with Python

- Performed using:
  - Google Maps Distance Matrix API
  - Google Maps Geocoding API
- Major Python Classes
  - Data_Accessor.py
    - Given list of Locations -> Output Travel Times to .csv
  - Location.py
    - Given Street Address -> Return the Latitude & Longitude
  - Trip.py
    - Calculates Trip times for different types of Transportation, finds the closest HR Station, and the Distance between 2 Locations

A Snippet of Trip.py

```
def distance_between_locations(origin , destination , mode_of_transport ):
[ . . . ]
def time_to_walk(origin,destination,):
  darn = distance_between_locations(
    origin=origin,
    destination = destination,
    mode_of_transport=1
  )
  return darn
def time_needed_transit(origin,destination,):
[ . . . ]
def time_to_bike(origin,destination,):
  #Calculate the amount of time it will take to walk to the nearest bike station
  darn = distance_between_locations(
    origin=origin,
    destination=destination,
    mode_of_transport=3
  )
  return darn
#def check_time(str,):
  #convert hours, min, into an int
def find_nearest_station(origin,list_of_stations):
  minimum=0
  station_number=0
  for station in range(len(list_of_stations)):
    distance_between_points =
haversine(origin.get_lng(),origin.get_lat(),list_of_stations[station].get_lng(),list_of_stations[
station].get_lat(),)
    if(station == 0):
      station_number = station
      minimum = distance_between_points
    else:
      if(minimum>distance_between_points):
        minimum = distance_between_points
        station_number = station
  #call the list value of station_number
  return station_number
class Trip:
  origin_station = Station
  destination_station= Station
  bike_trip=''
  transit_trip=''
  walking_trip = ''
  #Given a starting location and an end destination find the nearest station
  def __init__(self, origin,destination,list_of_stations,):  # special method __init__
    self.origin_station = list_of_stations[ find_nearest_station(origin,list_of_stations) ]
    self.destination_station =
list_of_stations[find_nearest_station(destination,list_of_stations)]
    #This returns a concatinated string
    #Above this comment is not the problem
    print(self.origin_station)
    print(self.destination_station)
    self.bike_trip = time_to_walk(origin =
origin,destination=self.origin_station)+time_to_bike(self.origin_station,self.destination_stati
on)+time_to_walk(origin=self.destination_station,destination=destination,)
    self.transit_trip = time_needed_transit(origin=origin,destination=destination)
    self.walking_trip = time_to_walk(origin=origin,destination=destination)
```

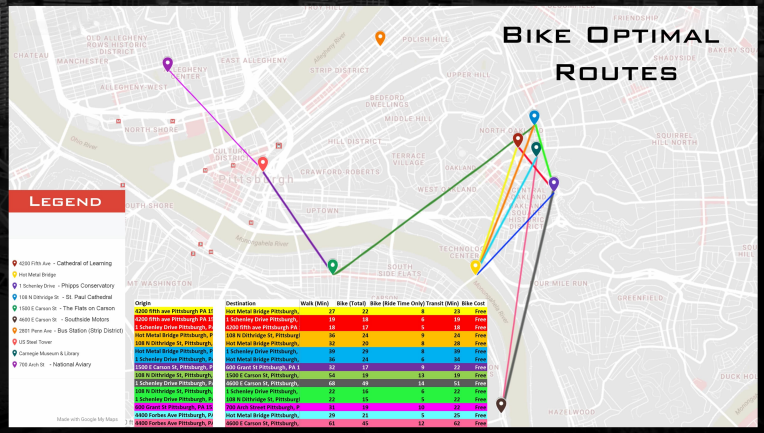
Data

- Sources:
  - Bus Stop & HR Bike stations -> Western PA Regional Data Center
  - Travel Times -> Google Maps API

Origin	Destination	Walk (Min)	Bike (Total)	Bike (Ride Time Only)	Transit (Min)	Bike Cost	Best Examples
4200 fifth ave Pittsburgh PA 15213	Hot Metal Bridge Pittsburgh, PA	27	22	8	23	Free	T
4200 fifth ave Pittsburgh PA 15213	15000 Carson St, Pittsburgh, PA 15203	50	21	13	16	Free	F
4200 fifth ave Pittsburgh PA 15213	44600 Carson St, Pittsburgh, PA 15210	60	46	15	36	Free	F
4200 fifth ave Pittsburgh PA 15213	108 N Dithridge St, Pittsburgh, PA 15213	6	7	1	6	Free	F
4200 fifth ave Pittsburgh PA 15213	2801 Penn Ave, Pittsburgh, PA 15222	45	31	15	22	Free	F
4200 fifth ave Pittsburgh PA 15213	600 Grant St Pittsburgh, PA 15219	19	18	6	19	Free	T
4200 fifth ave Pittsburgh PA 15213	600 Grant St Pittsburgh, PA 15219	52	23	15	17	Free	F
4200 fifth ave Pittsburgh PA 15213	4400 Forbes Ave Pittsburgh, PA 15213	6	14	4	6	Free	F
4200 fifth ave Pittsburgh PA 15213	700 Arch Street Pittsburgh, PA 15212	82	33	24	40	NF	F
Hot Metal Bridge Pittsburgh, PA	4200 fifth ave Pittsburgh PA 15213	31	26	9	23	Free	F
Hot Metal Bridge Pittsburgh, PA	15000 Carson St, Pittsburgh, PA 15203	32	30	13	21	Free	F
Hot Metal Bridge Pittsburgh, PA	44600 Carson St, Pittsburgh, PA 15210	32	48	8	32	Free	F
Hot Metal Bridge Pittsburgh, PA	108 N Dithridge St, Pittsburgh, PA 15213	36	24	9	24	Free	T
Hot Metal Bridge Pittsburgh, PA	2801 Penn Ave, Pittsburgh, PA 15222	64	48	24	36	NF	F
Hot Metal Bridge Pittsburgh, PA	1 Schenley Drive Pittsburgh, PA 15213	39	29	8	39	Free	T
Hot Metal Bridge Pittsburgh, PA	600 Grant St Pittsburgh, PA 15219	48	34	17	21	NF	F
Hot Metal Bridge Pittsburgh, PA	4400 Forbes Ave Pittsburgh, PA 15213	32	25	6	24	Free	F
Hot Metal Bridge Pittsburgh, PA	700 Arch Street Pittsburgh, PA 15212	77	44	26	42	NF	F

- If Bike (RTO) < 15, The Trip is Free
- If Bike (RTO) < 15 & Bike (T) < Walk & Bike (T) <= Transit, This is an Opitmal Bike Route

Bike Optimal Routes



Analysis and Conclusions

- In 15 of our tested routes, **Biking** (through HR) **is a fast, free, and feasible alternative** to Walking or Transit
- Biking is not a feasible alternative in areas of challenging terrain (Polish Hill – Outlier)
- Expansions: App that Determines fastest Transportation (Is HR feasible in the User's Location?)
- Recommendation: Expansion of Free Bike sharing services in Pittsburgh and Beyond

Central Catholic High School - 2018

Jackson Kaib, John Lynch, Sean Hazlett, Colin Pietropaolo, John Eakin, Roark Rembold

The Correlation Between Internet Speeds and Obesity Rates

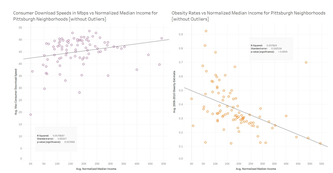
Introduction

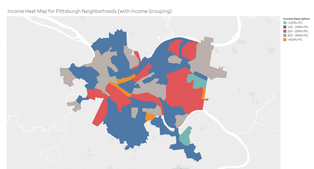
The past ten years have yielded some of the fastest obesity growth rates across the US. With Internet usage, prevalence, and speeds also rapidly increasing simultaneously, this project explores any possible correlation between Internet speeds and obesity rates.

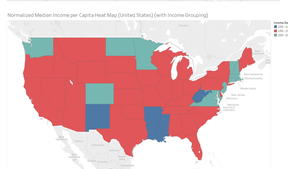
Therefore, our Null Hypothesis (H_0) is that there is no statistically significant correlation between internet speeds and obesity rates.

Datasets / Methodology

	Name	Source	Purpose
Pittsburgh	2017 TIGER / Line Shapefiles	US Census Bureau	Matches PA block codes to shapefile metadata. Used for heat map construction.
	Allegheny County Obesity Rates	Western PA Regional Data Center	Provides obesity rates as a percentage of obese in a specific Census block group.
	Pittsburgh ISPs by Block	FCC / Western PA Regional Data Center	Provides maximum advertised consumer and business upload and download speeds by Census block.
	PGH Snap Census Data—Education and Income	Western PA Regional Data Center	Provides median income for Pittsburgh Neighborhoods in 2010.
	PGH Snap Census—Housing	Data.gov	Provides number of occupied housing units per Pittsburgh neighborhood. Used for income normalization.
	PGH Snap Census—Neighborhood Census Data	Data.gov	Provides population per Pittsburgh neighborhood. Used for income normalization.
United States	Adult Obesity in the United States	State of Obesity	Provides adult obesity rates as a percentage of state population.
	United States - Average household size, 2009-2013 by State	Index Mundi	Provides average household size by state. Used for income normalization.
	USA Average Peak Internet Speeds	Fastmetrics	Provides average internet speeds for each state.
	US Median Income—2016	US Census Bureau	Provides median income by state.



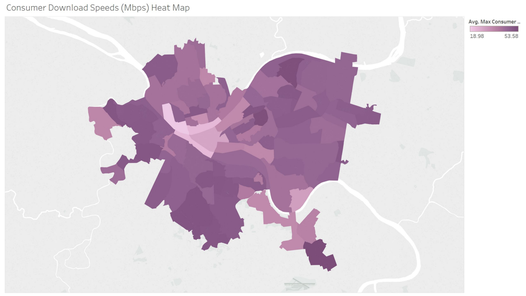


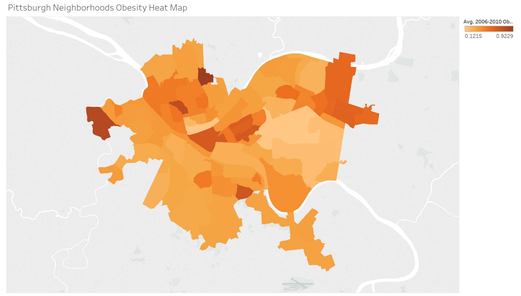


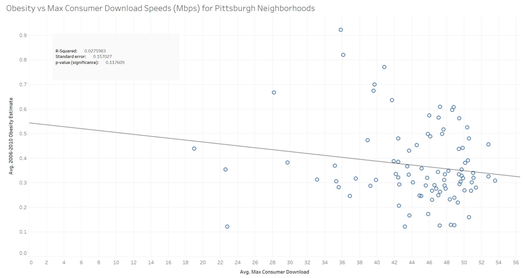
As displayed above, income confounds both obesity rates and internet speeds ($p < 0.05$ for all graphs). As such, we analyzed aggregate data as well as income-grouped data based upon the Federal Poverty Level (FPL), which is established at $12,060 for a single person household. Median household incomes—for both neighborhoods and states—were divided by their respective average household size to normalize the dataset (i.e. all incomes in our analysis are relative to the FPL of a single person household).

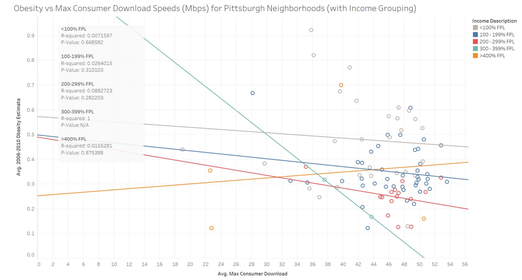
Visualizations and Results

Pittsburgh

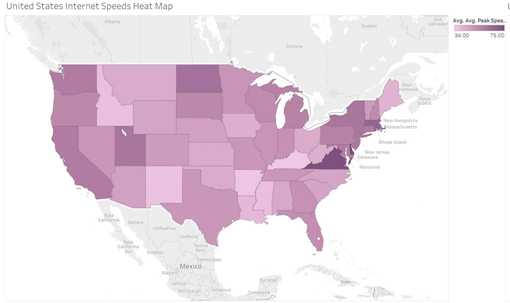


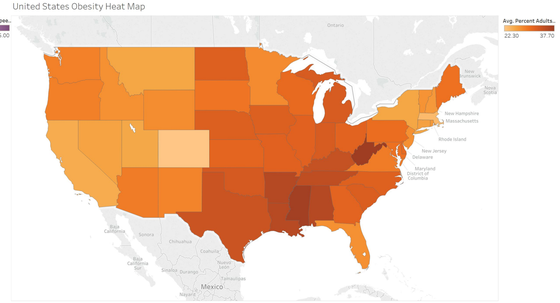


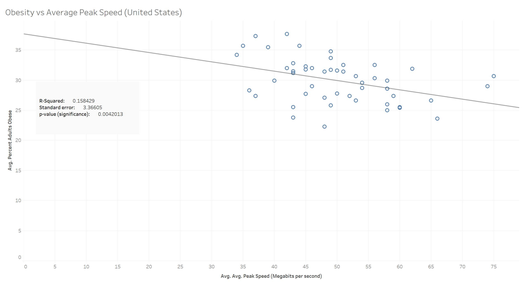


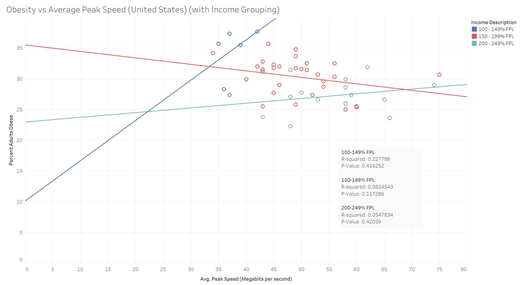


United States









North Allegheny Team #1
Akshath Jain, Steven Lu, Ethan James, Janet Wang, Diane Lee

Analysis/Conclusion

Analysis

At the local level, there appears to be a very weak negative correlation of aggregate data. However, given that $p = 0.12 > 0.05$, this result is not statistically significant. When this same data is divided into distinct income groups (<100% FPL, 100-199% FPL, 200-299% FPL, 300-399% FPL, >400% FPL), there appears to be very weak negative correlations for the <100% FPL, 100-199% FPL, and 200-299% FPL groups and a very weak positive correlation for the >400% FPL group. The results for the 300-399% FPL group can be disregarded as it only contains two data points.

When the aggregate data at the national level is analyzed, there appears to be a statistically significant ($p = 0.0042 < .05$) correlation between internet speeds and obesity rates. When the data is grouped by income (100-149% FPL, 150-199% FPL, 200-249% FPL), there is a weak negative correlation for the 150-199% FPL income group and weak positive correlations for the 100-149% FPL and 200-249% FPL groups.

Discussion

While our data points towards no correlation between between Internet speeds and obesity rates, other researchers have found significant results regarding the connection between the presence of technology and obesity. For example, A 2009 study by the Central Queensland University determined that participants with a high leisure-time Internet and computer usage were 1.46 times more likely to be overweight than participants with less leisure time Internet usage. Another report from the Milken Institute in 2012 showed that a 10% increase in national spending on technology was correlated with a 1% increase in that nation's obesity rate. So while our results indicate that internet speeds and obesity are not related, internet usage and obesity definitely are.

Possible sources of error are most likely caused by our data sets. For example, the Internet speeds data set for Pittsburgh provided maximum advertised speeds for a particular census block, which may not be entirely representative (the US data set was much more accurate—data was accumulated from several billion Internet Speed tests).

Conclusion

The final results of our data analysis show that there exists no significant relationship between internet speed and obesity. Thus, we cannot reject the null and the two variables are likely independent.

Challenges

- Finding credible and representative datasets for both the local and national levels
- Determining how to judge a household's income
- Deciding cutoffs for income grouping
- Data visualization using Tableau

Recommendations

1. ISPs should reduce prices to increase high speed Internet access in lower-income neighborhoods.
2. Further research in area to find more definitive conclusions.