Introduction to Data Science

Course Project

An interesting deep dive into a predictive analysis in reducing Public School Drop Out Rates and how they are impacted by State Tax Revenue, Household Income & Teacher Salary

SWAPNIL KURALE

OSAMA SYED

Group 92

I. DATA COLLECTION

A. We collected data from various online resources. Our main avenue of data was through the Federation of Tax Administrators (FTA) and the National Center for Educational Statistics (NCES) datasets. We primarily focused on 2017 data.

   https://www.taxadmin.org/2017-state-and-local-revenue

   https://nces.ed.gov/programs/digest/d18/tables/dt18_211.60.asp?current=yes

B. For Data Specific to New Jersey Household Wages there was a great online resource that compiled data from the American Community Survey Census. They made available the data in an easy to use CSV format which we downloaded.
   https://datausa.io/profile/geo/new-jersey#economy

C. For teachers salaries in NJ we found an article that sourced the NJ Department of Education and compiled the data for us which showed the average teacher salary in 2017 for each school district and each county.

   https://www.nj.com/education/2018/09/do_teachers_make_enough_heres_the_median_salary_in.html

D. For Graduation Rates in New Jersey we sourced data from the State of New Jersey Department of Education website for 2017 graduation rates.

   https://nj.gov/education/data/grate/2018/

E. For Crime Rates in New Jersey we sourced data from the New Jersey Department of State Police
   https://www.njsp.org/ucr/pdf/current/20180123_crimetrend.pdf

II. Data Format Description

A.

B. The 2017 tax csv file has 7 columns and 51 rows. In the dataset, the data was split into Total Source Collection and Total Tax Collection for every state and the District of Columbia.. We only focused on the Total Tax Collection because Source Collection is revenue collected from a states own resources.

C. The NCES csv file has 15 columns categorized by year range from 1969 - 2018 and divided by current inflation rate. The data includes average teacher salary for **public** elementary and secondary school for all 50 states and the District of Columbia. We focused on 2017-2018 NCES data. * we refer DC as one of “states’ below

D. The Data USA csv file 9 Columns of Data which shows the breakdown between 2013-2018 of average household income for each county in New Jersey by different Races and also a row which represents average household income for each county for all the races. The relevant data for us was 2017 data on average household income for total race since we wanted to focus on whether household income affects teachers salary and not bring race into play.

E. Data on teachers' salaries for each school district is a CSV with 6 columns. Columns useful for us are the County Name, District Code (Which we will use to do joins with other data sets) and 2017 salary.

F. Data on Graduation Rates for each school district is a CSV with 8 columns. Columns useful for us are are the County Name, District Code (Used to do a join with the teachers salaries data set), and graduation rates for students who graduated in 4 years and also 5 years in another column.

G. Data on the crime rates is a CSV with 3 columns. It includes the county name , the Crime Index- which represents the number of 8 specific crimes as specified by the FBI (Murder,Rape,Robbery, Aggravated Assualt,Arson,Burglary,Larceny Theft,Motor Vehicle Theft), the last column was the population of the county which would be used to calculate the crime rate using the crime index for the corresponding county.

III. Descriptive Analysis

A. On average, the Total Tax Revenue Collected by each state is $32.3 million and $4.95 thousand per Capita. States like California and New York were on the higher end while Alaska and Wyoming were on the lower end of tax revenue.

	Tax Revenue ($ Million)	Per Capita($ Thousand)
count	51.000000	51.000000
mean	32.340176	4.951412
std	43.521986	1.463632
min	2.720000	3.370000
25%	8.037500	3.924000
50%	17.685000	4.682000
75%	39.301500	5.452000
max	243.082000	10.717000

B. The top 5 average teacher salaries by states were New York, California, Massachusetts, District of Columbia, Connecticut. The lowest 5 average teacher salary states were Florida, Utah, Oklahoma, West Virginia, Mississippi. The overall average for teacher salary within the United states in 2017 was $57,282. Since there were outliers such as NY and MS, the median is a more accurate value of the teacher salary average, - $54,846. We can see the distribution between these two states is about $40,4778.

```
Highest 5 AVG SAL States
                   State  Avg_Salary ($ Thousands)
0               New York                     83.585
1             California                     81.126
2          Massachusetts                     79.710
3   District of Columbia                     76.486
4            Connecticut                     73.113

Lowest 5 AVG SAL States
              State  Avg_Salary ($ Thousands)
46          Florida                     47.721
47             Utah                     47.604
48         Oklahoma                     45.678
49   West Virginia                      45.642
50     Mississippi                      43.107

count    51.000000
mean     57.282137
std       9.760218
min      43.107000
25%      50.329500
50%      54.846000
75%      61.593000
max      83.585000
Name: Avg_Salary ($ Thousands), dtype: float64

57.282137254901954
54.846
40.477999999999994
```

C. We found that the average household income across all 21 NJ counties was about 80K with a range between 53k-115k . With a few additional metrics in the table below.

County Data Statistics

	Average Income
count	21.000000
mean	80461.380952
std	18307.213331
min	52627.000000
25%	66196.000000
50%	79173.000000
75%	90026.000000
max	114732.000000

D. We found the average teacher salaries for all 21 counties in New Jersey and the max average of a specific county. This shows the distribution which is represented in the following table

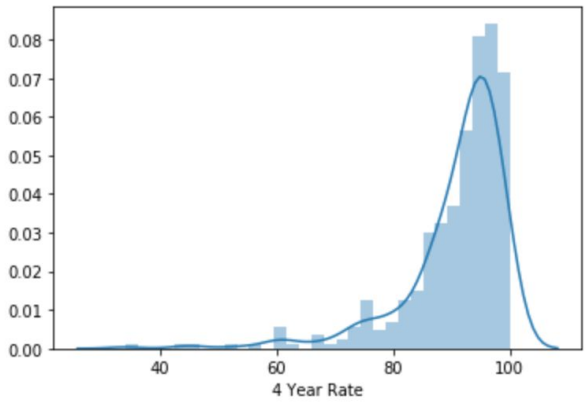
	Average Income	Teacher Salary
count	21.000000	21.000000
mean	80461.380952	63374.952381
std	18307.213331	2822.699036
min	52627.000000	58959.000000
25%	66196.000000	62010.000000
50%	79173.000000	63120.000000
75%	90026.000000	64199.000000
max	114732.000000	69145.000000

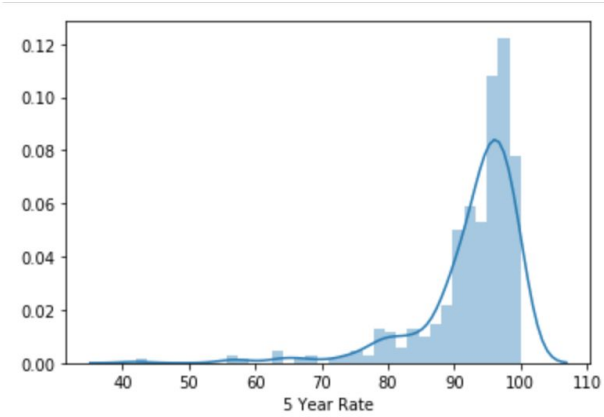
E. We had plenty of data on the 4 and 5 year graduation rates of each school district in New Jersey, 407 rows of Data to be precise, which

matched up with the school districts in the teacher salary data that we had. Performing a quick statistical analysis on the data we found the following results.

	4 Year Rate	5 Year Rate
count	407.000000	407.000000
mean	90.783784	92.648649
std	9.117862	7.655572
min	34.000000	42.000000
25%	88.000000	90.500000
50%	94.000000	95.000000
75%	96.000000	97.000000
max	100.000000	100.000000

Since we had 407 data points we also thought it would be a good idea to plot the distribution of values for graduation rates as shown below.





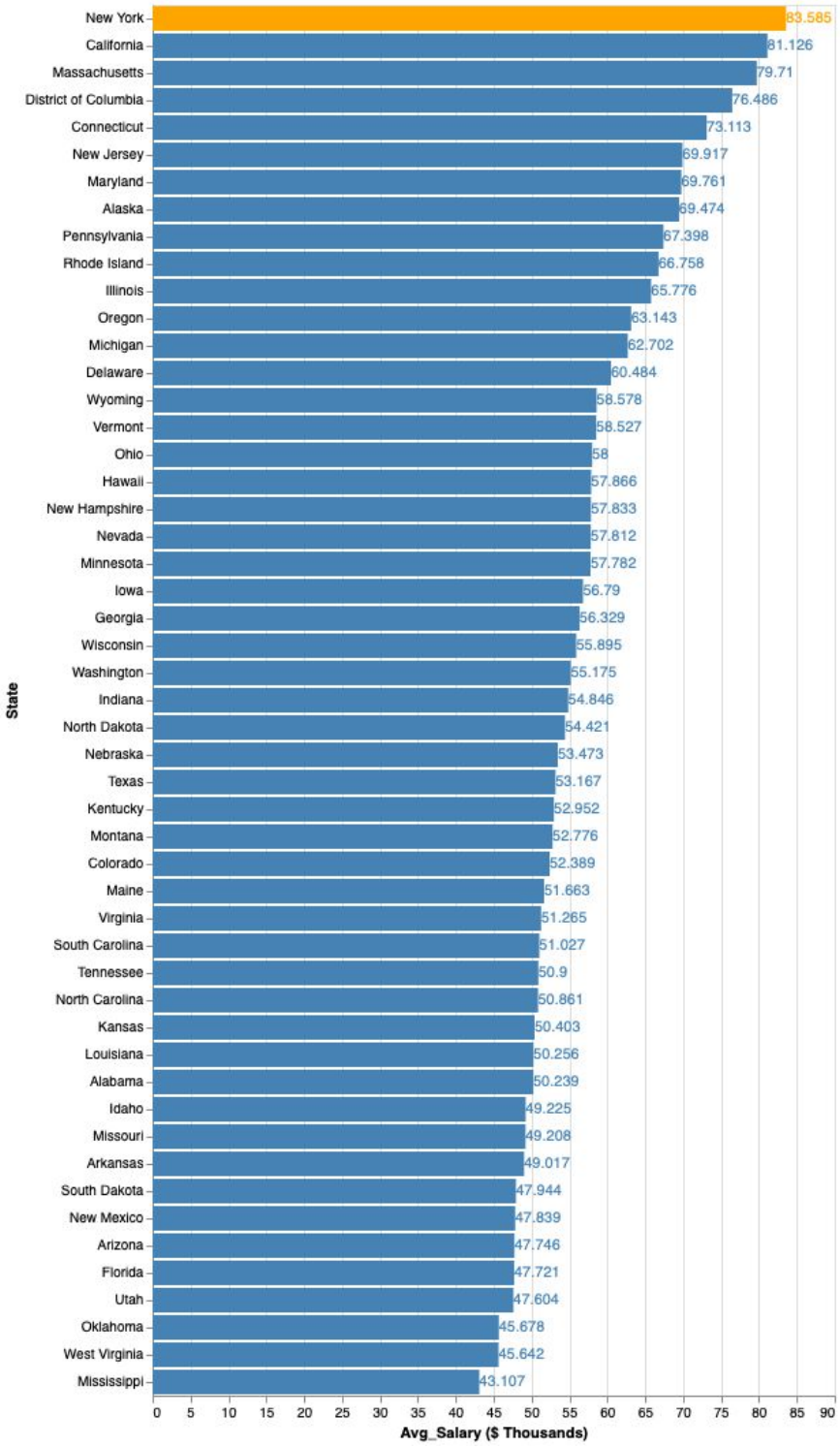
F. For the crime data we were able to source the crime index count in each county and using the population for each county we calculated the crime rate per 100,000 in each county.

	County	Crime Index	Population	Crime Rate per 100000
0	Atlantic	566	265498	213.18
1	Bergen	848	932420	90.95
2	Burlington	628	445610	140.93
3	Camden	1145	506224	226.18
4	Cape May	147	93129	157.85
5	Cumberland	425	151423	280.67
6	Essex	1702	796349	213.73
7	Gloucester	471	290961	161.88
8	Hudson	931	672827	138.37
9	Hunterdon	59	124712	47.31
10	Mercer	671	368168	182.25
11	Middlesex	956	827363	115.55
12	Monmouth	735	621990	118.17
13	Morris	312	493920	63.17
14	Ocean	608	595424	102.11
15	Passaic	803	504402	159.20
16	Salem	97	62883	154.25
17	Somerset	304	330573	91.96
18	Sussex	58	141197	41.08
19	Union	996	554738	179.54
20	Warren	124	105715	117.30

IV. Data Analysis, Visualization, and Insights

A. Since we had a wide range of average teacher salaries for the states., we decided to group the salaries with a range of {40000 - 55,000}, {55,001-70,000}, and {70,001- 85,000}. We noticed a unique trend that most states have an average teacher salary in the lower quartile range {40,000- 55,000} - 26 to be exact. {55,001-70,000} has 20 states, and {70,001- 85,000} has 5 states.

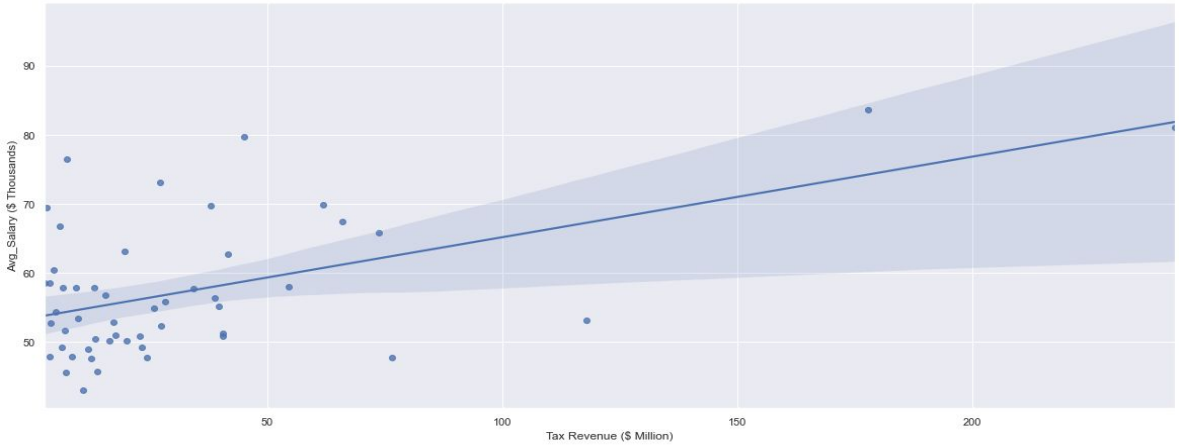
In fact, the total distribution for all 50 States and DC is as follows:

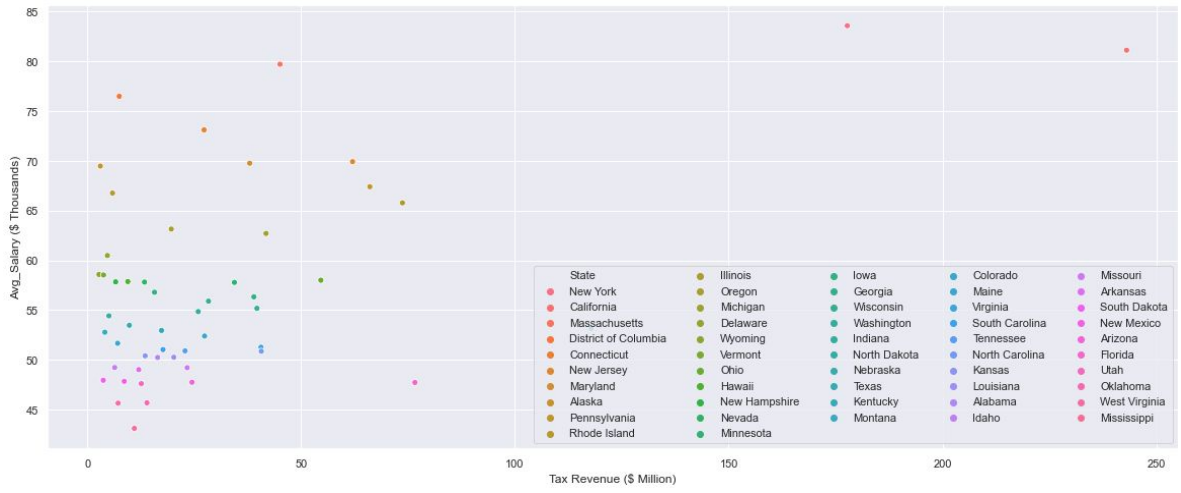


*Altair graph depicting average teacher salary by state

Based on our findings from State Tax Revenue, we were able to hypothesize that there is some relationship between the Average Teacher Salary and the Average State Tax Revenue. NY and CA have the highest Tax Revenue as well as the highest average teacher salary.

To compare these two quantitative values on a deeper level, we used a categorical scatter plot as well as a linear regression plot. If we compare these two graphs, we were able to deduct that there is some positive direct relationship between average teacher salary and average state tax revenue.

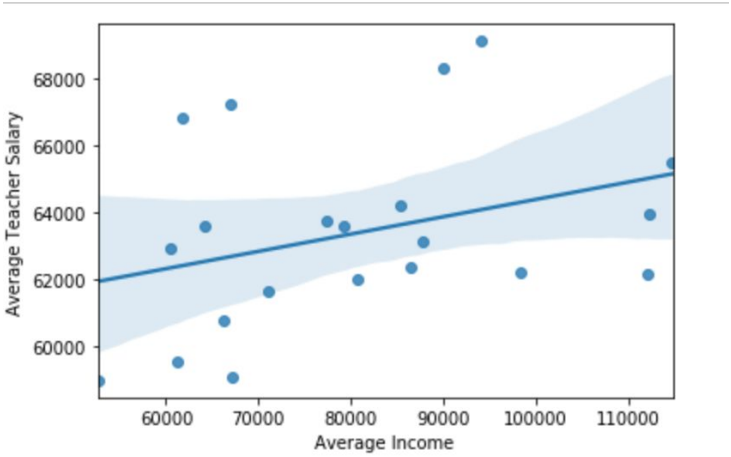




State	Avg_Salary ($ Thousands)	Tax Revenue ($ Million)
California	81.126	243.082
New York	83.585	177.751
Pennsylvania	67.398	66.070
New Jersey	69.917	62.027
Massachusetts	79.710	45.053

We further confirmed the claim by comparing the top 10 Average Teacher Salary and Top 10 State Tax Revenues. 5 states appeared on both lists
(California, New York, Pennsylvania, New Jersey, Massachucetts).

B. We took a deep dive into New Jersey statistics to see whether or not average household income in a county affected the teachers salary in that county. After plotting a graph of **Average Teachers Salary vs Household Income** we were able to make some interesting observations.



Including the linear regression line we are able to see that as household income increases in a specific NJ county the teacher's salary also seems to increase albeit not by a lot. Since public school teacher are paid with tax dollars, we deduced that there is a correlation between the Average Income, Average State Tax Revenue, and Average Teacher Salary.

C. Top 5 Average Income Counties and Teacher Salary

County	Average Income
Morris	114732
Somerset	112153
Hunterdon	112038
Monmouth	98270
Bergen	94107

County	Average Teacher Salary
Bergen	69145.0
Sussex	68318.0
Cape May	67241.0
Atlantic	66846.0
Morris	65524.0

D. Lowest 5 Average Income Counties and Teacher Salary

County	Average Income
Passaic	64233
Atlantic	61777
Salem	61250
Essex	60430
Cumberland	52627

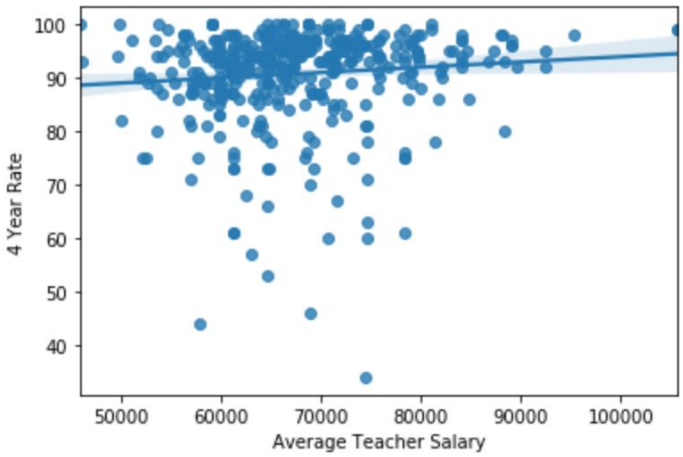
County	Average Teacher Salary
Ocean	61651.0
Camden	60758.0
Salem	59530.0
Hudson	59076.0
Cumberland	58959.0

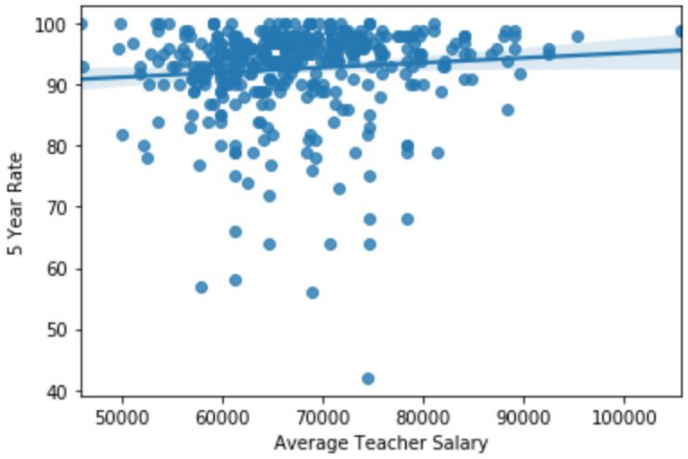
E. Furthermore, another hypothesis we wanted to explore was that teacher salaries impact the quality of education. Using the data that we accumulated, we measured 4 & 5 **Graduation Rates vs Teachers Salaries.** This metric had 407 data points that we were able to plot and we did this for both 4 and 5 year graduation rates.

District Code	Average Teacher Salary	4 Year Rate	5 Year Rate
3710	105650	99	99
3710	105650	99	99
3995	95303	98	98
1290	92432	95	96
1290	92432	92	95
...	...	...	...
6033	49970	82	82
6018	49800	100	100
6032	49690	94	96
6013	46000	93	93
6068	45900	100	100

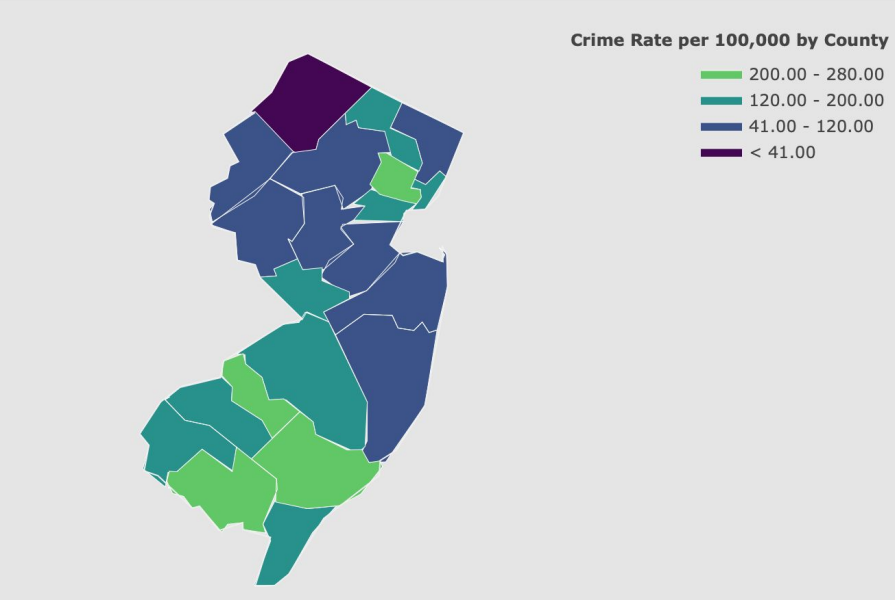
*5 years are those students who may have had to stay in high school for an extra year rather than the traditional 4 years.

Consequently, we noticed that a higher teachers salary resulted in a slightly higher graduation rate for both 4 and 5 year graduation rates. The linear regression line has a positive slope which indicates that.





F. Using our crime data we were able to make a visual representation of the crime rates in New Jersey



Using this representation we can see some of the most dangerous places to live (over 200 per 100,000 crime rate) in New Jersey are in South Jersey such as Atlantic, Camden, and Cumberland county and in North Jersey it is Essex County.

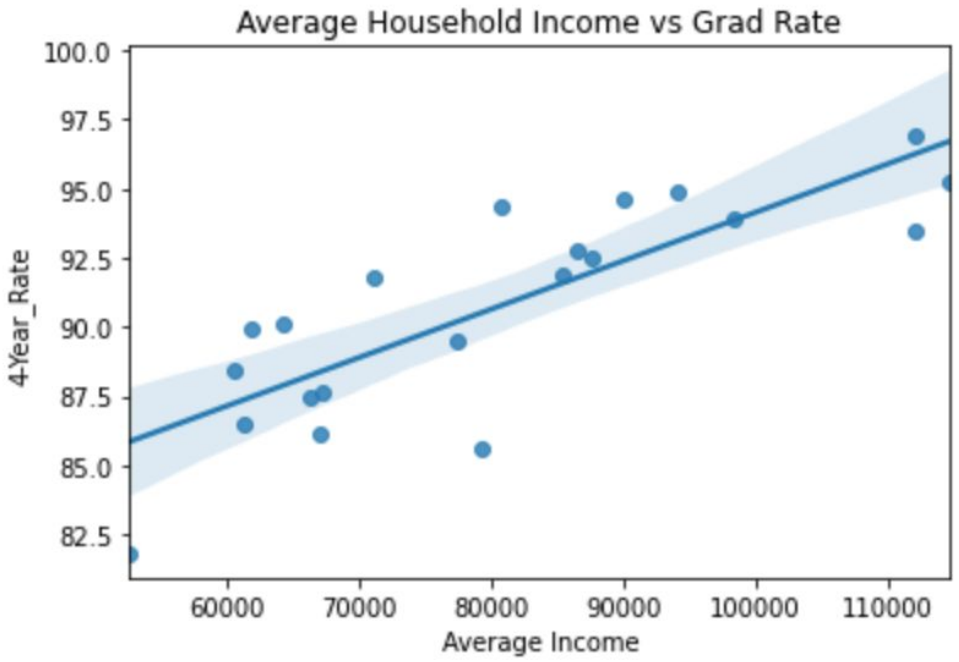
This data alone doesn't tell us much about the quality of education being impacted by the crime rate but in our machine learning analysis below we were able to draw some interesting insights.

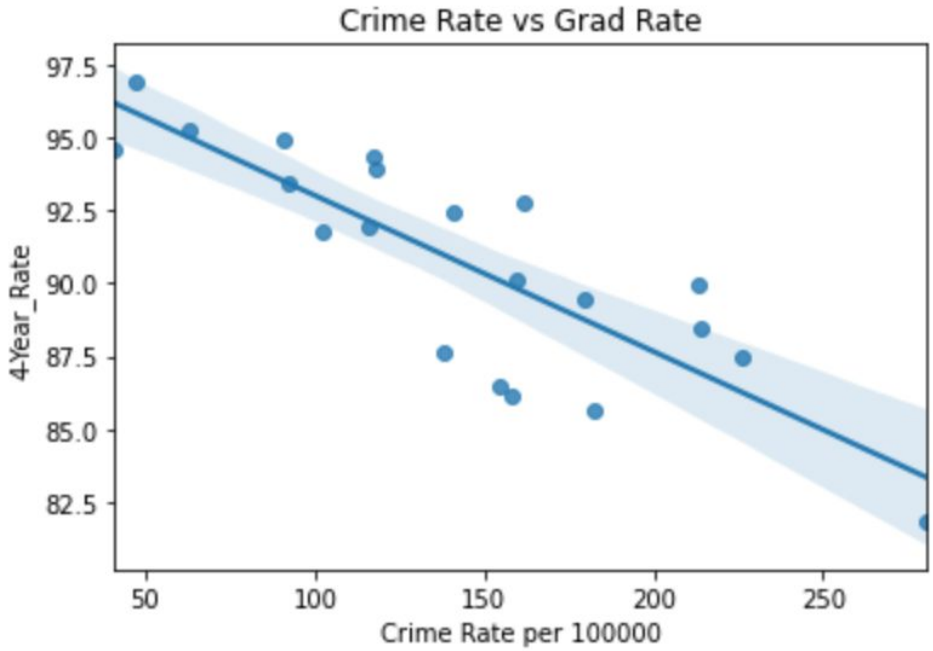
V. Machine Learning

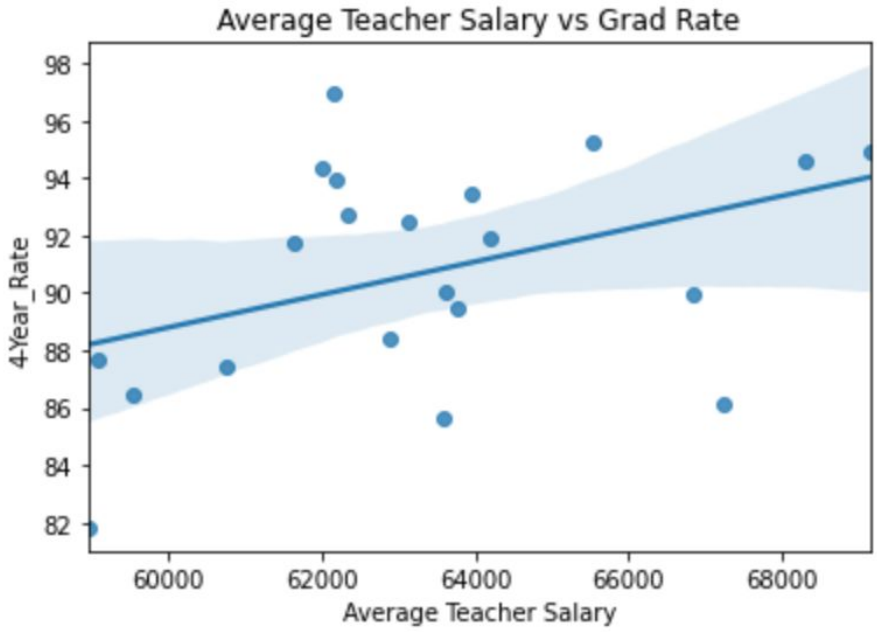
The main objective of our analysis was to see which factors affect graduation rate and whether or not we can apply a machine learning model to predict the graduation rate given a set of variables. We took a deep dive into the following variables:

1. Average Income
2. Average Teacher Salary
3. Crime Rate per 100,000

At first we individually examined the independent variables and compared each to see if there is a relationship between the variable and the graduation rate.







From our analysis of the three independent variables we came to the following conclusions:

1. Graduation rate is higher when household income is higher
2. Graduation rate is lower when crime rate is higher
3. Graduation rate is higher when teachers are paid more

Since each of these variables independently impacted graduation rate, we were interested in determining if these variables have a correlation together to impact graduation rates.

We used Regression Analysis that would predict the graduation rate of a county given all these 3 metrics together.

We trained our ML model with 80% of the data and tested it on the rest of the 20% to see how accurate our model was.

We compared three different types of Regression - Linear, SVR, and Polynomial.

```
Linear Accuracy Score: 0.8557064802959055
SVR Accuracy Score: 0.5197760534072351
Polynomial Accuracy Score: 0.9415487573817697
```

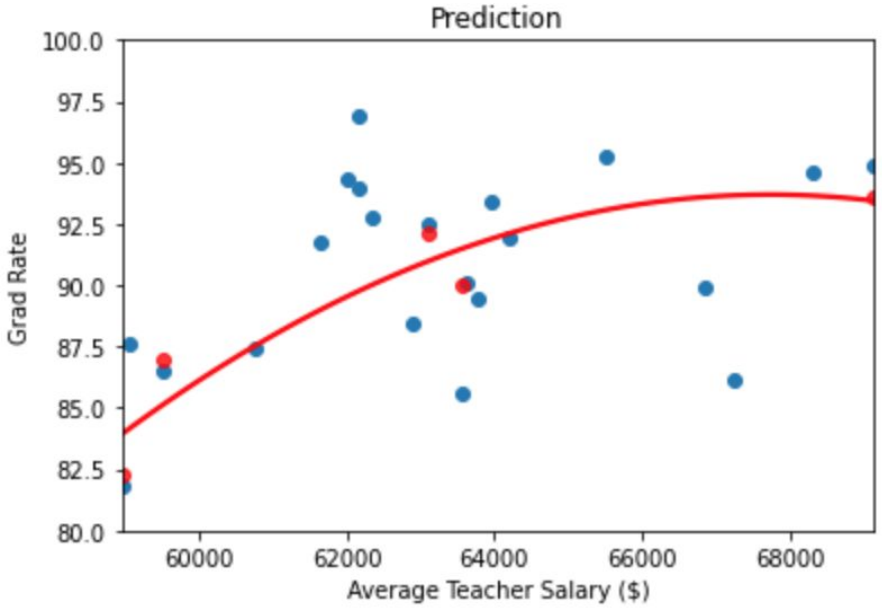
Since Polynomial Regression received the highest accuracy score, we implemented our predictive analysis with this regression.

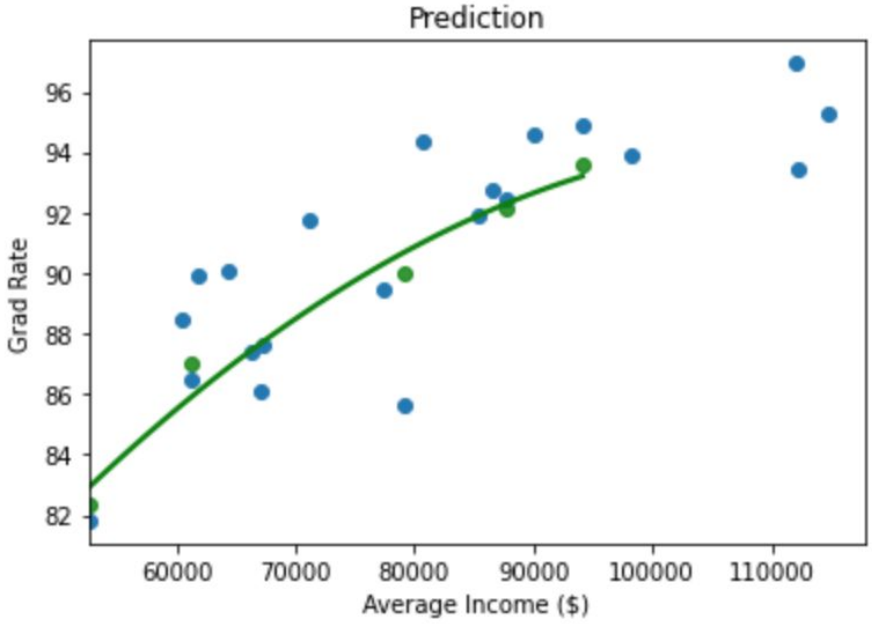
*The Coefficients are [Crime Rate, Household Income, Teacher Salary]

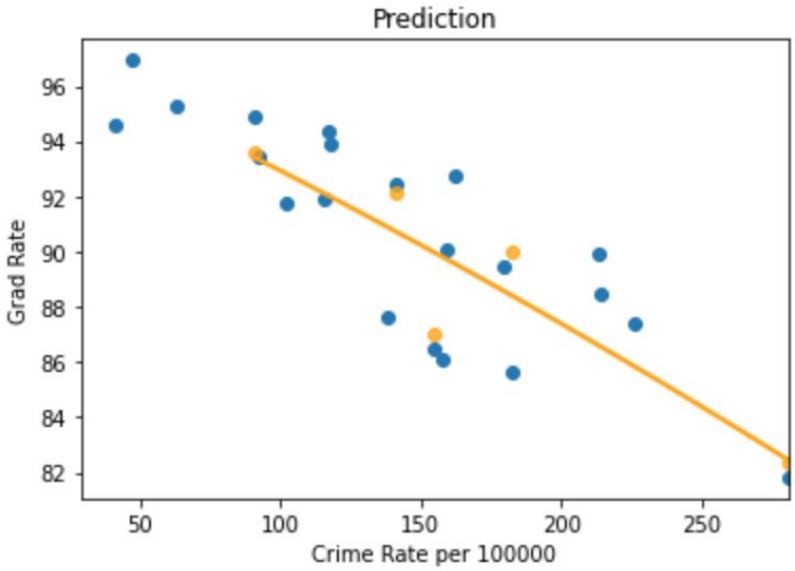
```
Coef: [    63.17 114732.    65524.  ]
 Predicted Grad Rate: 95.6146016280072
Actual Grad Rate: 95.25999999999998

Coef: [  138.37 67154.   59076.  ]
 Predicted Grad Rate: 88.66227704064974
Actual Grad Rate: 87.638

Coef: [   90.95 94107.   69145.  ]
 Predicted Grad Rate: 93.55484557723759
Actual Grad Rate: 94.91085106382978

Coef: [  226.18 66196.   60758.  ]
 Predicted Grad Rate: 86.51991560842737
Actual Grad Rate: 87.42833333333333

Coef: [   41.08 90026.   68318.  ]
 Predicted Grad Rate: 94.57134090318621
Actual Grad Rate: 94.58

R^2 Score: 0.9415487573817697
```

This regression gave us a high R^2 Score of .9415 which implies that our model has a strong correlation.

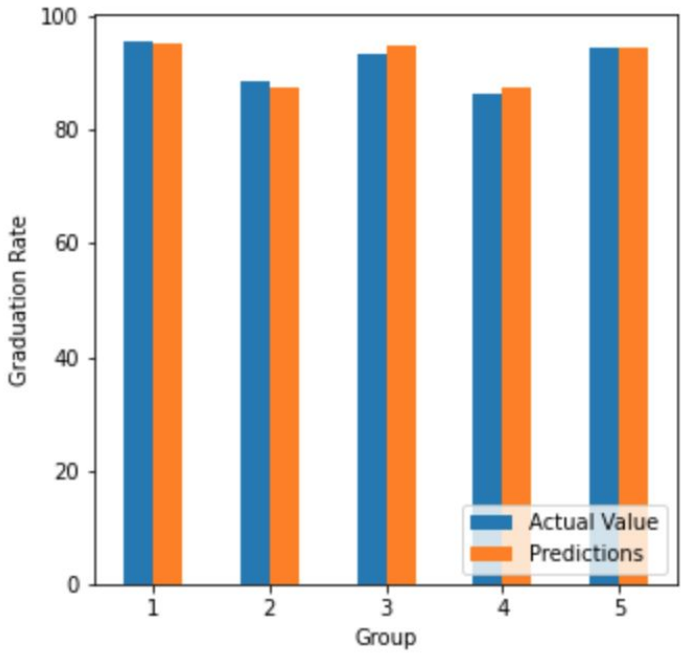
The below graphs show the predictive values from the ML in respect to the actual values taking the 3 independent variables into account.







In comparison to the actual graduation rates, the ML model accurately predicts the graduation rates given all 3 variables.



VI. Related Work

There are a few studies online that explore graduation rates in detail. Ours are a little different. We were able to find two studies.

One is from a blog post on the following website:

https://towardsdatascience.com/capstone-blog-62002ebacf6b

Our work differs because we look into different metrics used to predict graduation rates. The main difference is that the author used college stats vs. high school stats. We believe that education reform early on can affect quality of life better than later on. That is why we chose to look at high school graduation rates.

Mastercard Inc. also had a study on graduation rates but like the author above but like the author above there's also focused on college graduation rates. They also argued like we did that lower income students have lower graduation rates.

https://www.mastercardcenter.org/insights/data-science-can-help-boost-college-graduation-rates

One of the biggest contributors to the graduation rate that we found was the crime rate in the area. None of these studies mentioned above took that into account which we believe really set our study apart from the other ones.

VII. Conclusion

Using the model, we can closely predict the graduation rate of a county given the crime rate, household income, and teacher's salary throughout all 50 states.

We made the following conclusions:

- Higher Salary = Increased in Graduation Rates
- Higher Household Income = Increase in Graduation Rates
- Higher Crime Rate = Decrease in Graduation Rate

Higher tax revenues lead to higher teacher salaries which contribute to better graduation rates. We need some kind of tax reform, not necessarily higher taxes, but a change which can get teachers a better salary, help reduce crime in neighborhoods, and also a better system which can improve income for households.

In our research we limited our analysis based on counties. This may result in some inaccuracies in our finding based off of counties because of specific cities that have higher crime areas such as Newark and Camden than the rest of the county. In the future, we hope that various attributes such as household income can be found for each school district which would provide a better representation of school districts that need more reform.

We can further our research to delve into inner cities of NJ and all over America by focusing on how society can help reduce crime rates and poverty, while increasing graduation rates.

Overall, there definitely are more factors that will impact graduation rates that are deeply rooted in systematic issues within our nation. One particular issue is the practice of redlining throughout the 1960s which has had a lasting negative impact on our society till today. If state government, families, and educators work together to reform crime rates, teacher salary, and household income, we can progress to drastic changes in the future that are deeply seeded in our nation.

VIII. Acknowledgments

Libraries used

- Pandas
- Matplotlib
- Altair
- Scikit-learn
- Seaborn
- Numpy
- Plotly(choropleth)
- Shapely
- pysh
- geoPandas