Steven Casey
Stacy Kirchner
Keldin Maldonado
Christopher Varela

Github: https://github.com/sevenwhiteclouds/383-population-growth-decline
Data source: Violent Crime Rate - Dataset - California Open Data

Introduction

Why was the project undertaken?
What was the research question, the tested hypothesis or the purpose of the research?

The project was undertaken to determine if violent crime rate would have influenced the change of population of an area. It is important to understand the impact a high or low violent crime rate would have on an area. It is also equally important to understand what influences an increase or decrease in the population of an area.

Selection of Data

What is the source of the dataset? Characteristics of data?
Any munging or feature engineering?
The dataset is retrieved from CA.gov open data portal. There are originally 27 columns with 49,227 rows consisting of object and float64 data. Several columns were dropped because they have irrelevant data that is not needed for our hypothesis. Rows that had Nan values for ‘rate’ or ‘dof_population’ were dropped because they are useful without data. There were two additional columns added, ‘Previous_Denominator’ and ‘Pop_Change_Pct’. ‘Previous_Denominator’ groups the data set by ‘geoname’, which is the area, and gets the ‘dof_population’ from the previous row. This allows the current row to have the population of the previous year. This new column is then used to create ‘Pop_Change_Pct’, which is the percentage of change in population from the previous year to the current year. The first years that do not have a previous year to look at are given 0 for ‘Pop_Change_Pct’.

Methods

What materials/APIs/tools were used or who was included in answering the research question?

The tools used were Pandas, Seaborn, Matplotlib, Jupyter Notebook, and Scikit-Learn. Three different algorithms that were explored were Linear Regression, Random Forest, and Polynomial Linear Regression.

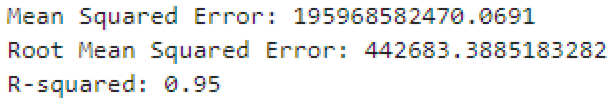
Results

What answer was found to the research question; what did the study find?
Was the testested hypothesis true? Any visualizations?

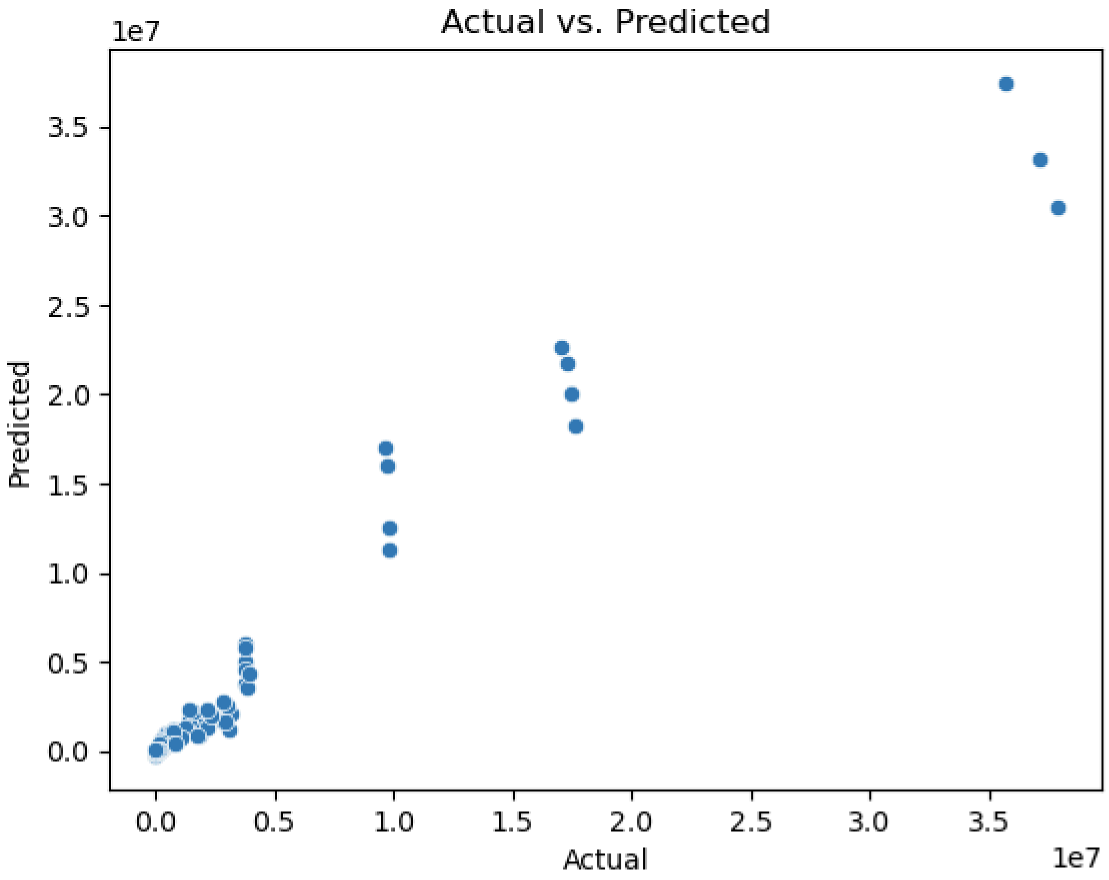
It was difficult with this dataset to be able to get a full understanding of the role that the crime rate was having on the population. When trying to predict the population growth which is represented by ‘Pop_Change_Pct’ using the crime rate, ‘rate’, there would be a good RMSE but the R-square would be 0. To address this problem we tried adding more features to see if the population could be predicted instead based on the crime rate, number of crimes, and population growth. Switching to these factors would then give us a good R-square but then a high RMSE. There was a tradeoff between goodness of fit and predictive accuracy, but we were unable to confidently prove our hypothesis. It is possible that there are other factors that have an impact on population growth and that by only focusing on crime rate, we were underfitting.

<u>Linear Regression:</u>

<u>Target - ‘dof_population’</u>



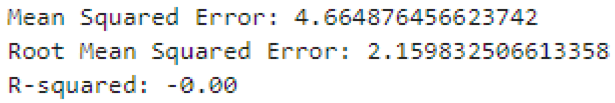
```
Mean Squared Error: 195968582470.0691
Root Mean Squared Error: 442683.3885183282
R-squared: 0.95
```



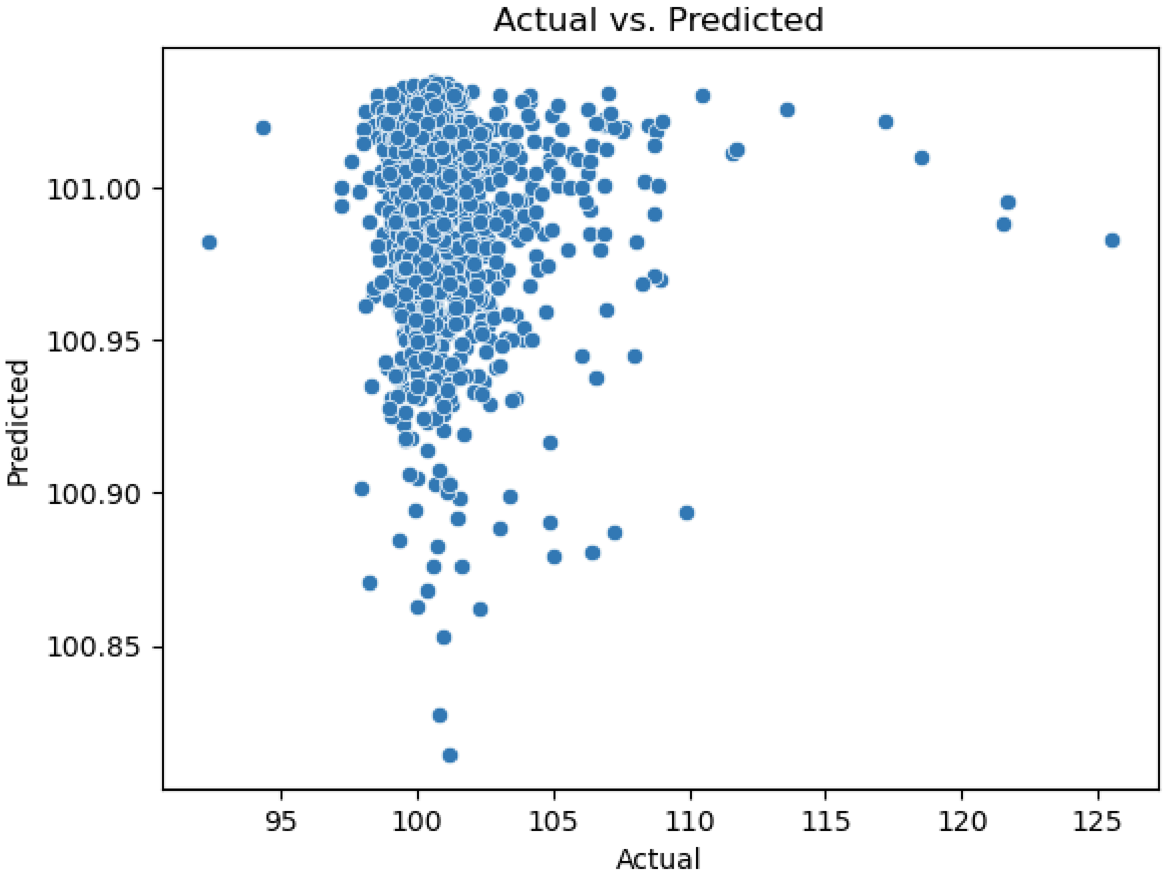
Actual vs. Predicted

When using the crime rate, numbers of crimes, and population growth, would be able to predict the population. The model was able to get a R-squared value of 0.95 but a RSME of 442683. The RSME is very high, which means the model’s predictions are very off from the actual values.

Target - ‘Pop_Change_Pct’

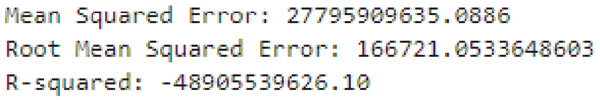


```
Mean Squared Error: 4.664876456623742
Root Mean Squared Error: 2.159832506613358
R-squared: -0.00
```

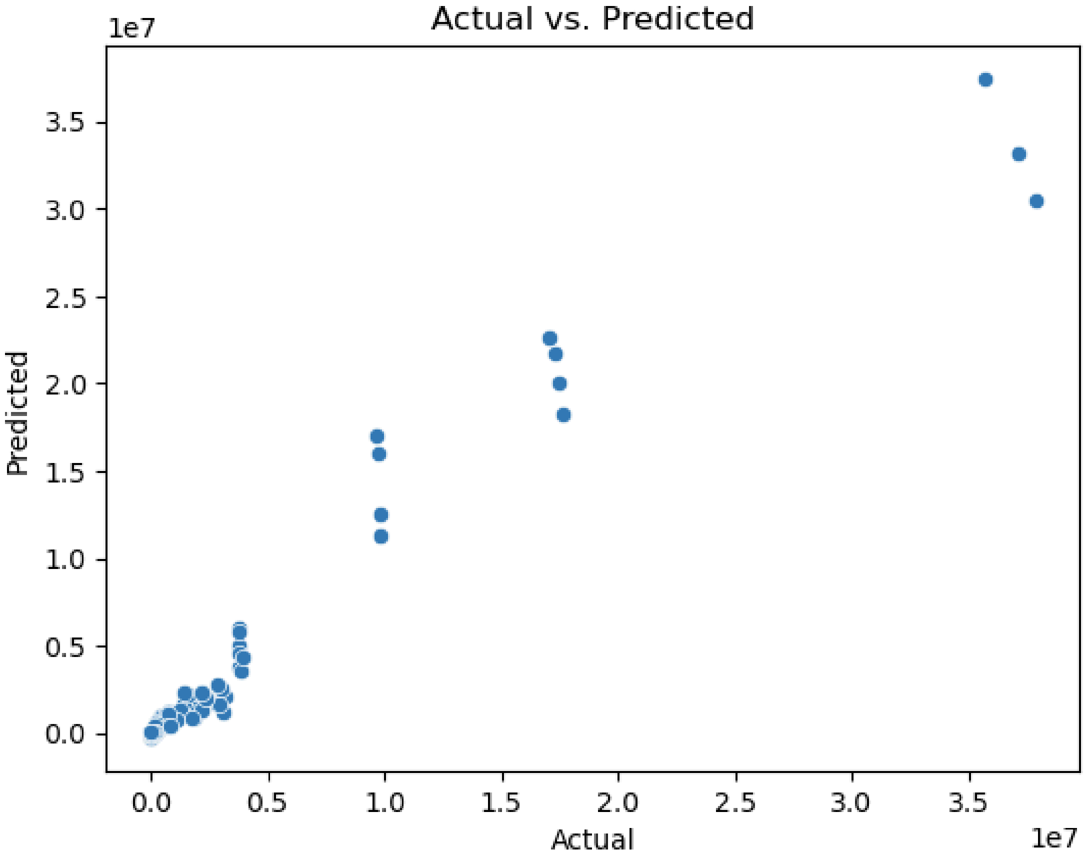


When using crime rate as the predictor and population growth as the target the model was able to predict the value with a 2.16 RSME, despite having a R-squared value of 0. This shows that the model is able to predict the values, but there is no correlation between the predictor and target.

Random Forest:
Target - ‘dof_population’

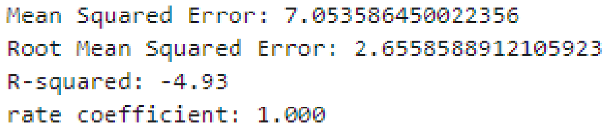


```
Mean Squared Error: 27795909635.0886
Root Mean Squared Error: 166721.0533648603
R-squared: -48905539626.10
```

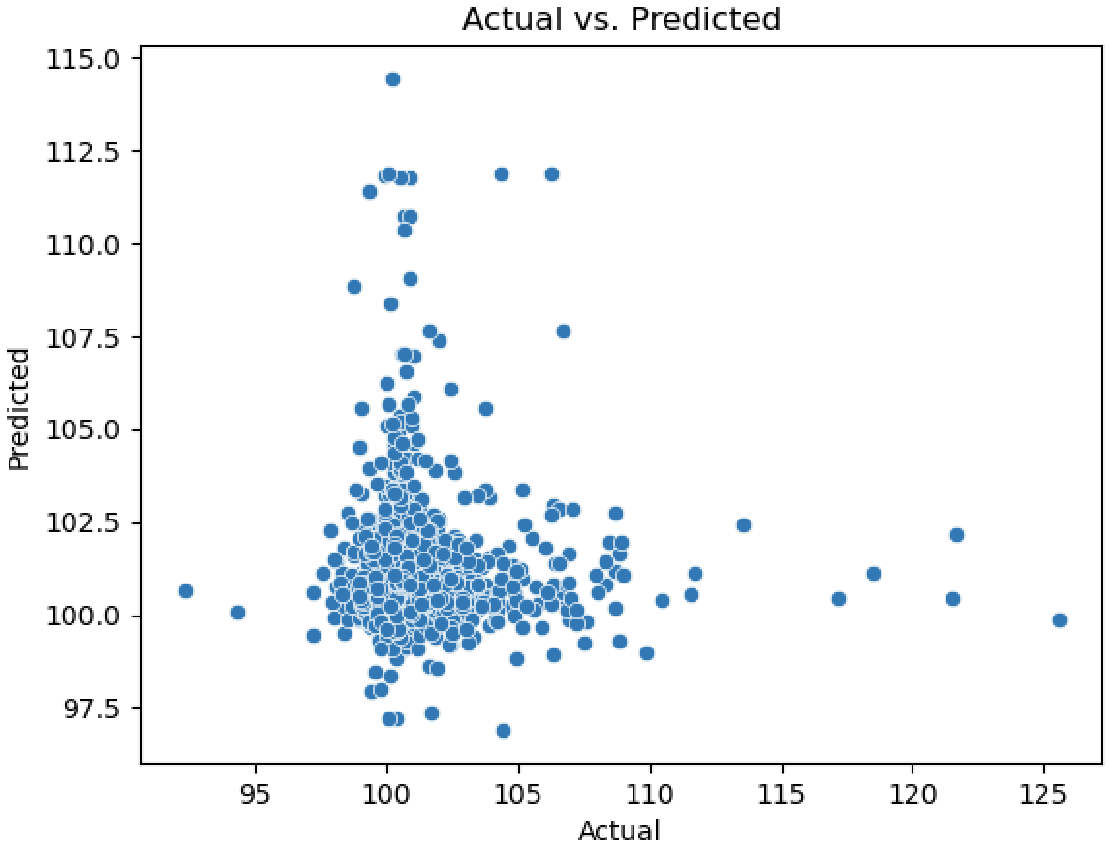


There was a extremely negative R-squared value when predicting population (‘dof_population) using the features rate', 'numerator', and 'Pop_Change_Pct', but this model was able to have a lower RMSE when compared to Linear Regression for the same predictors and target.

Target - ‘Pop_Change_Pct’

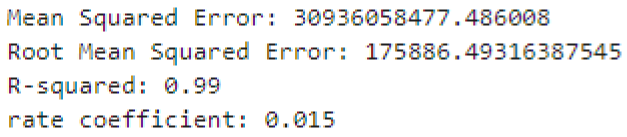


```
Mean Squared Error: 7.053586450022356
Root Mean Squared Error: 2.6558588912105923
R-squared: -4.93
rate coefficient: 1.000
```

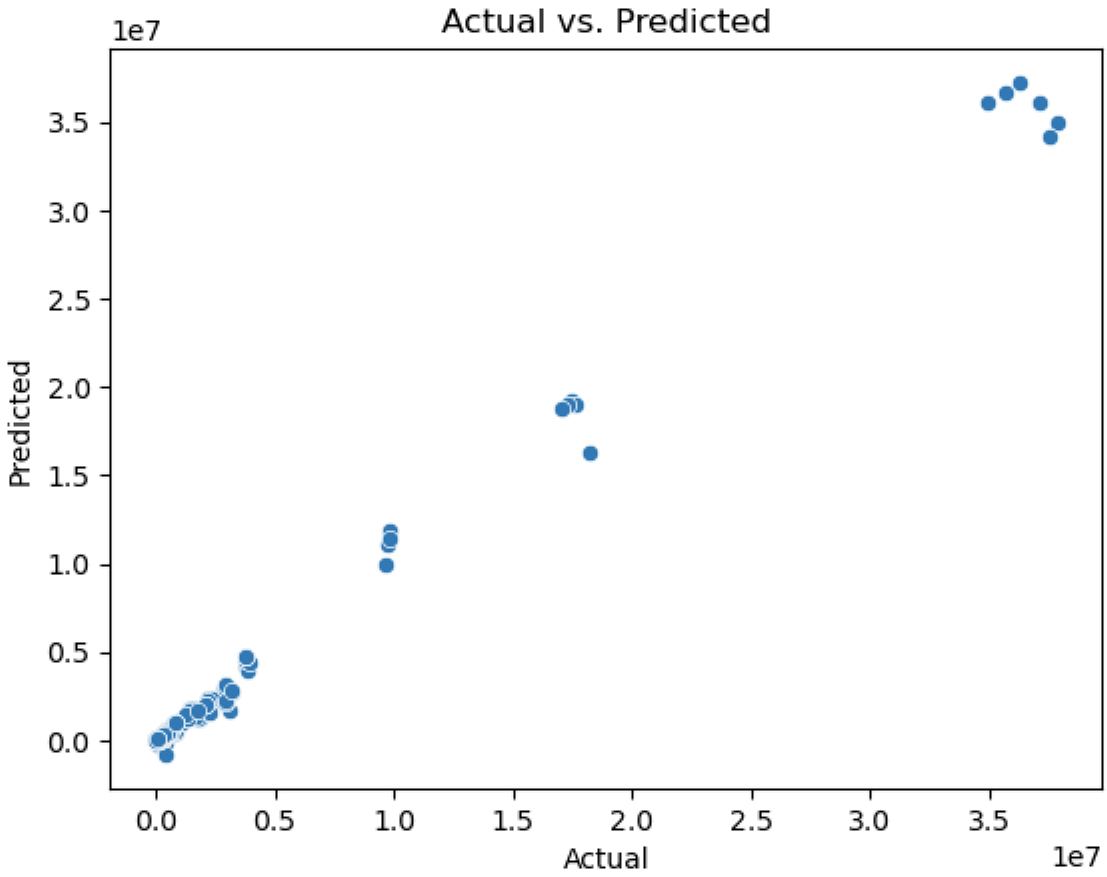


When using rate as the predictor and ‘Pop_Change_Pct’ as the target there was an R-squared value at -4.93 with a RSME of 2.66, which means there is no relationship between the two. However, the low RSME shows that the prediction is only off by about 2.66 when compared with the test values.

Polynomial Linear Regression:
Target - ‘dof_population’

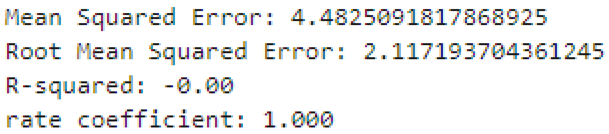


```
Mean Squared Error: 30936058477.486008
Root Mean Squared Error: 175886.49316387545
R-squared: 0.99
rate coefficient: 0.015
```

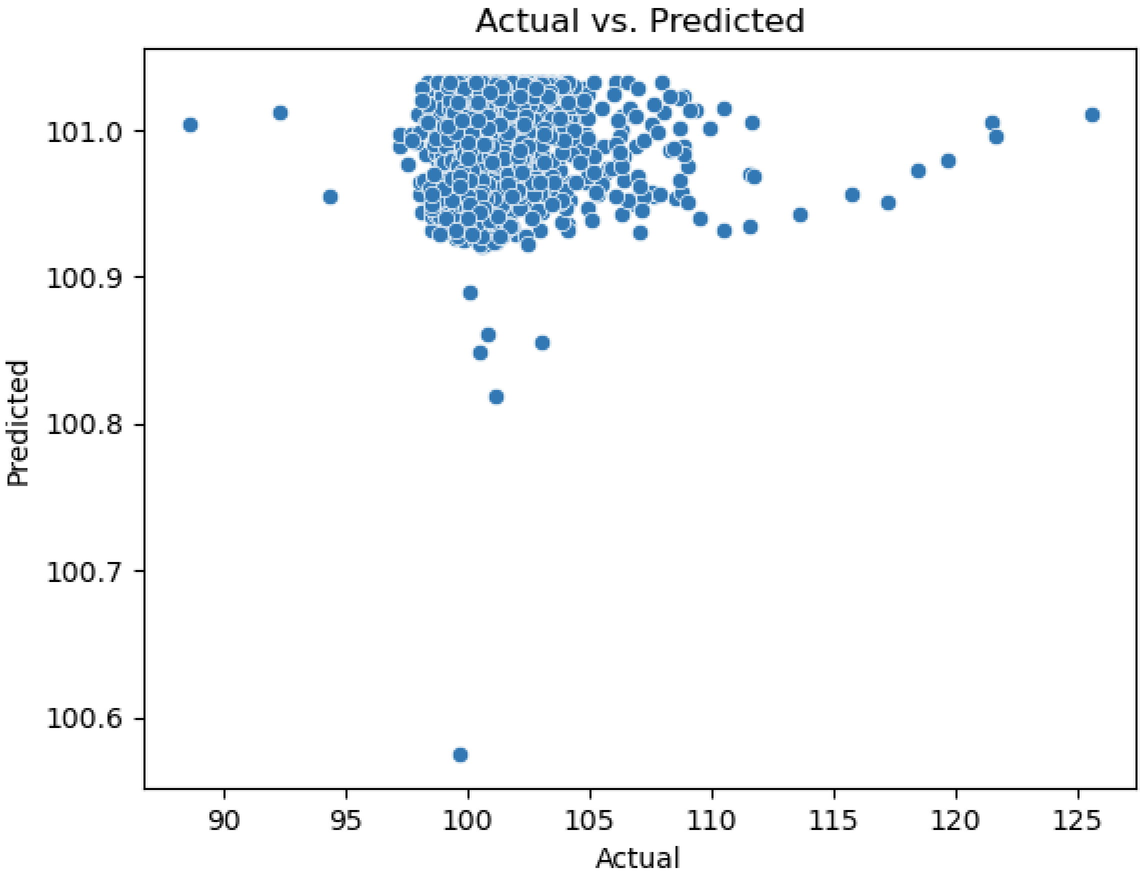


This appears to be the best model for predicting population using features: 'rate', 'numerator', and 'Pop_Change_Pct'; since it has the lowest RMSE and the highest R-squared.

Target - ‘Pop_Change_Pct’

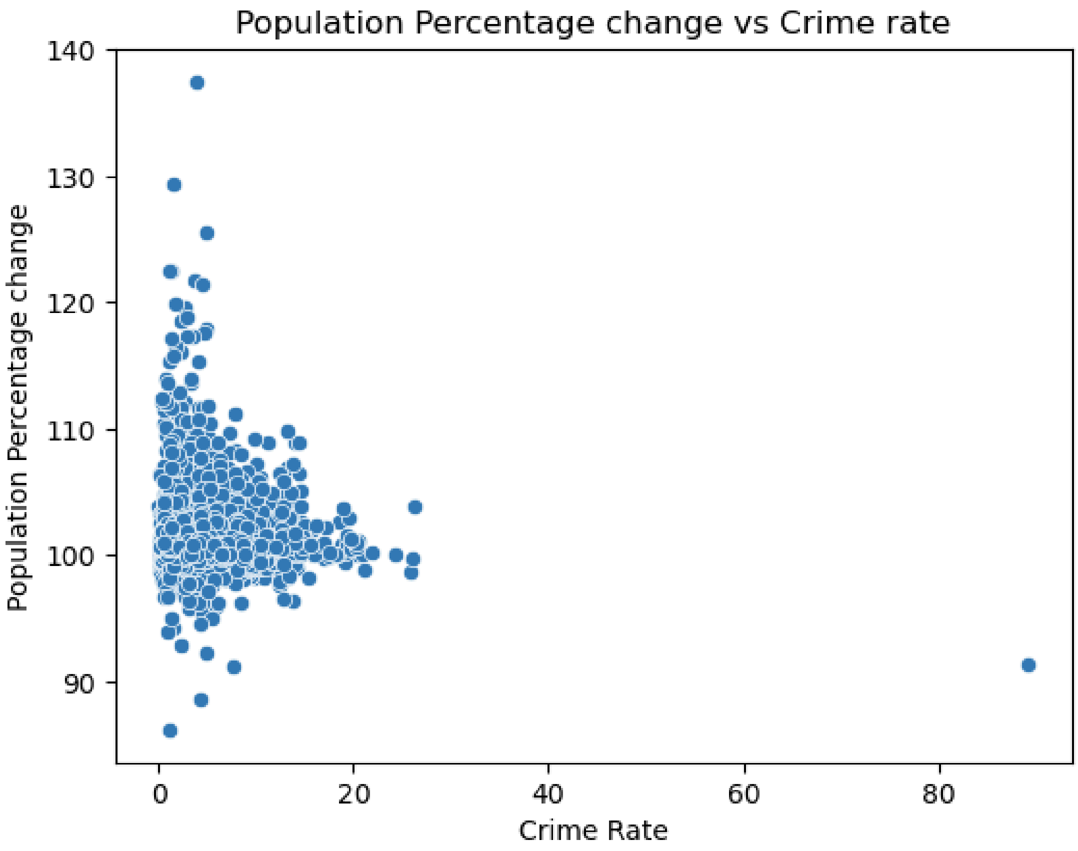


```
Mean Squared Error: 4.4825091817868925
Root Mean Squared Error: 2.117193704361245
R-squared: -0.00
rate coefficient: 1.000
```



When using ‘rate’ to predict ‘Pop_Change_Pct’, there were similar results to regular linear regression. There was a 0 r-square and a low RSME of 2.12. The graph for polynomial regression shows that there is a more extreme outlier that is skewing the graph.

Negative correlation:



However, this graph appears to show some negative correlation between population growth and crime rate since as the crime rate decreases the percentage of population change increases. There is also an outlier with a high crime rate.

Discussion

The data set used was not sufficient enough to comfortably predict future population changes. Ultimately, there are many more variables that would be needed to predict this. For example, world population changes such as births and deaths would need to be accounted for along with economic changes. Other research has been conducted on the topic of California's population. A study done in October 2023 by the Public Policy Institute of California focused on why California's population was decreasing. The predictors used in the study were the number of births, number of deaths, and immigration (Johnson et al., 2024). California's population heavily followed the birth rate and deaths of California residents. In the future, researchers could investigate trends between population growth and available housing, especially in specific locations. Using population data, it is also possible for researchers to figure out what is contributing to the world population increasing or decreasing. The tools we used for this assignment were very helpful in creating a final product. Pandas was used for creating DataFrames which were essential for holding and moving though data. Seaborn and Matplotlib were used for plotting data and making it look more aesthetically pleasing. Scikit-Learn was heavily used to create three different machine learning models that learned from the data held within the pandas DataFrame. Together, the tools create a powerful package that made this experiment possible.

Summary

The biggest finding was that crime rate alone can not determine population growth. Using just the crime rate as the predictor, a good RMSE was able to be achieved, which indicates that the predictions were close to the actual values. However, since the R-squared value was 0, it meant that the model does not explain the variance in the population growth. Despite getting accurate predictions, the data was not able to fully understand the pattern of population growth. This shows that crime rate is not a significant factor when it comes to population.

References

Johnson, H., McGhee, E., Subramaniam, C., & Hsieh, V.(2024, February 7). *What's behind California's recent population decline-and why it matters.* Public Policy Institute of California. https://www.ppic.org/publication/whats-behind-californias-recent-population-decline-and-why-it-matters/