Prediction of Tucson Crime Based on Neighborhood Income and Demographics

Austin Raben, Cameron Rodgers, Elias Zheng
Department of Computer Science,
University of Arizona,
1200 E University Blvd, Tucson, AZ 85721
austinraben@arizona.edu

***Abstract* - Crime prediction is a valuable tool to assist in the allocation of policing and crime prevention resources to the places that need it the most. Previous research suggests that crime rates in lower-income areas are much higher than higher-income areas, however crime prevention resources are usually not dedicated to said lower-income areas. [1] As a large urbanized area with diverse neighborhoods of varying wealth indexes with publicly available, curated data, Tucson appears to be an ideal model to apply the model of income-based crime rate prediction in the United States. By drawing from these predictions, limited crime prevention resources can potentially be directed to the proper areas where they will make the most impact.**

I. Introduction

Tucson is a medium-sized city located in southeastern Arizona, and is classified as a city of the Tucson Metropolitan Statistical Area (MSA) under the United States Census Bureau. It has a population of ~543,000, with income inequality levels below the national average. [2] When compared to other metropolitan areas in the western United States, Tucson has a high rate of homicides and firearm fatalities and a higher-than-average rate of property crime. [3] Importantly, the city of Tucson has an extremely well maintained and publicly available database of city data available for free at https://gisdata.tucsonaz.gov/.

The contents of this analysis will utilize only crime and income data from neighborhoods within the city of Tucson itself due to the dataset. Dynamics from the greater Tucson MSA will be excluded from this analysis and prediction.

Based on the unique demographic and crime dynamics in the city of Tucson, we are interested in whether or not the wealth of a neighborhood affects the likelihood of violent crimes committed. (Question 1, Q1)

We believe that due to the difference in wealth resulting in a lack of access to necessary resources, low-income neighborhoods will experience a greater likelihood of occurrence of violent crimes compared to neighborhoods with higher incomes.

Tucson additionally has a high level of incidents involving firearms compared to other MSAs in the western United States. [3]

To this end, we are interested in whether or not the wealth of a neighborhood affects the likelihood of crime involving weapons. (Question 2, Q2)

We believe that due to a lack of resources and exposure to crime in low-income neighborhoods, the likelihood that a weapon is used to carry out a crime will be greater compared to that of neighborhoods with higher incomes.

In order to answer these questions, we trained two separate machine-learning prediction models.

II. Related Works

P. Fajnzylber et. al previously investigated the correlation between violent crime and income inequality by studying the national homicide rates of various countries compared to the national income inequality. When controlled for additional variables such as education level, unemployment, youth population, and gross domestic product, Fajnzylber et. al found that income inequality served as an accurate predictor of crime at the national level. [4]

However, in a meta-analysis of 43 similar studies concerning the correlation between income inequality and crime, M. Pazzona concludes that the true impact of income inequality on crime is minimal, suggesting that income inequality may affect crime in different ways than previously thought. [5]

In regards to the findings of Fajnzybler et. al, J. Brush performed a study with similar goals to Fajnzylber et. al, performed both an analysis at the national level and the county level of the United States. Brush's model found that although income inequality could accurately predict crime rate disparities across different counties nationally, it does not play the same relationship at the county level. This suggests that perhaps income inequality does not serve as an accurate predictor of crime-rate at the microeconomic level. [6]

In regards to our models, what has not been addressed by previous models is the relationship between income inequality and rates of violent crime to non-violent crime. Our model in Q1 attempts to investigate whether or not there is a link between income inequality and likelihood of occurrences of violent and non-violent crime as opposed to crime in general. In Q2, our model attempts to extend this idea of using income to predict the severity of crimes, which is also novel.

Additionally, most other models do not focus on analysis of local neighborhoods and instead start their analysis from the county-level, or most commonly work at the state/provincial to national level. Our study attempts to apply the same concepts and insights at a much more micro level.

Finally, our models do not utilize the Gini coefficient to analyze income inequality. Instead, we utilize the Esri wealth index, as reported by the Tucson dataset, which measures the wealth of an area against the national level, where a value of 100 represents the national average. [7] This allows our models to utilize local insights as the Esri wealth index makes the income level disparity visible, whereas the Gini coefficient would mark two regions with the same income inequality but different income levels as the same. Therefore, the Ersi wealth index makes visible both income inequality and income level disparities across neighborhoods.

III. Methods

A. Data Pipeline

Data was sourced from the publicly available Tucson GIS datasets. The datasets used in this analysis were:

1. Tucson 2024 Police Incidents (https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-incidents-2024-open-data/about)
2. Tucson 2010-2019 Aggregated Neighborhood Income (https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-income/about)
3. Tucson 2010-2019 Aggregated Neighborhood Population (https://data-cotgis.opendata.arcgis.com/datasets/cotgis::neighborhood-population-statistics/about)
4. Tucson 2010-2019 Aggregated Neighborhood Age Demographics (https://data-cotgis.opendata.arcgis.com/datasets/cotgis::neighborhood-age-demographics/explore?showTable=true)
5. Tucson 2010-2019 Aggregated Neighborhood Race Demographics (https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-race-demographics/explore?location=32.197622%2C-110.889177%2C10.76&showTable=true)
6. Tucson 2010-2019 Aggregated Neighborhood Education Attainment (https://data-cotgis.opendata.arcgis.com/datasets/cotgis::neighborhood-educational-attainment/about)

These datasets were cleaned and relevant characteristics from utilized datasets were combined into two tables: one for the neighborhood dynamics and the other for police incident data.

The cleaning process involved removing tuples with null values or human formatting errors / inconsistencies due to human-entered data. Notably, some police incident tuples contained 'NONE' string values for certain relevant categories; these were omitted from the final dataset. Additionally, some neighborhood names in the police incident data did not line up with the neighborhood names in the neighborhood demographic datasets. Tuples which pertained to these neighborhoods were omitted.

There were several constructed columns as well, created from aggregating separate fields in the original data. Each field concerning education was obtained by aggregating relevant columns together (ie. Population with an undergraduate and population with a graduate degree were grouped together as simply college degrees). Additionally, populations with 'some high school' were grouped with the 'no high school' group.

In the police incident table, whether or not there was a weapon involved was determined by looking at if the weapon description fields were null or not. A non-null indicated weapon involvement.

The schema of the neighborhood table is as follows:

Neighborhood Name
Esri Wealth Index
Total Population
Population Density
Population of Families
Average Household Size
Esri Diversity Index
Median Male Age
Median Female Age
Population without High School Education
Population with High School Diploma / GED
Population with College Degree

Table 1, Neighborhood Data Schema

The schema of the police incident table is as follows:

Incident ID
Neighborhood Name
Crime Type
Crime Category
Weapon Involved

Table 2, Police Incident Data Schema

After data cleaning and processing, each model could filter out unwanted fields in their own pre-processing stage. For example, in Q1, weapon involvement is not used. Additionally, crimes which were neither categorically violent or nonviolent were omitted from the training and test sets.

For both models, the training and test data follow an approximately 80-20 split of the total dataset.

B. Model Training

For Q1, we created a linear model utilizing elastic net linear regression with its alpha and L1 to L2 ratio tuned on the training data using 10-fold K-fold cross validation to predict the likelihood of violent crime [P(V)] in a neighborhood based on the wealth index of the neighborhood. We controlled for the potential influence of other neighborhood dynamics such as education, population, and age by including them as predictors in our model (see Tab. 1).

For Q2, we utilized a similar elastic net linear regression model with alpha and L1 to L2 ratio tuned on the Q2 training data utilizing 10-fold K-fold cross validation. In Q2, this model predicts the likelihood of weapon involvement [P(W)] in a neighborhood, also based on the wealth index of the neighborhood. Similarly to Q1, we controlled for the potential influence of other neighborhood dynamics by including them as predictors in our model (see Tab. 1).

We chose the elastic net model because it can easily handle potential multicollinearity and overfitting to the large number of neighborhood dynamics that were included as features, thanks to its usage of both ridge (L1) regularization and lasso (L2) regularization. If the wealth of a neighborhood is an important predictor for the percentage of crimes that are violent, then the elastic net model should highlight this coefficient.

Due to the sensitivity of regularization, we employed Z-score normalization to perform feature scaling which ensures that features containing larger numbers are not overrepresented in the model. Since the dependent variable is a percentage clamped to the interval [0,1], all independent variables were scaled to reflect this range.

C. Model Testing and Validation

Model testing was performed by fitting the model on the normalized X and making a prediction based on the X-test set.

For model validation, we looked at the R^2 score, mean squared error (MSE), and root mean squared error (RMSE) as an indicator of model performance. Analyzing the magnitude of the model coefficients allowed us to see the role that neighborhood wealth was playing in the prediction the model was making and to identify covariates.

IV. Results

A. Q! Model Prediction Results

The test performance indicator values are as follows:

R^2	0.0992
MSE	0.0050
RMSE	0.0710

Table 3, Q1 Model Performance Indicators

The coefficients for the models are as follows:

Total Crimes	-0.0341
Wealth Index	-0.0225
Total Population	0.0000
Population Density	0.0038
Family Population	0.0033
Avg. Household Size	0.0000
Diversity Index	0.0044
Median Male Age	0.0000
Median Female Age	-0.0024
Not H.S. Educated Pop.	0.0000
H.S. Educated Pop.	0.0167
College Education Pop.	0.0000

Table 4, Q1 Model Coefficient Magnitudes

Test indicators and coefficients are normalized, thus they are scaled in respect to P(V).

We can see based on the coefficient values that population density, family population, diversity index, median female age, and high school educated population were covariates that were controlled for by including them as predictors of the model.

The total number of recorded crimes, the wealth index, and the high school educated population were the strongest potential predictor of P(V). Wealth index and total crimes share an inverse relationship with P(V): as their values go up, the value of P(V) tends to decrease. In the case of the high school educated population, there is a direct relationship between the high school educated population and P(V): as the high school educated population increases, P(V) increases.

Our low R^2 value indicates that the Q1 model does not necessarily correlate well with the dependent variable, and our indicators do not fully explain its variance well. However, the low MSE value suggests that the Q1 model's predictions are quite accurate compared to ground truths. Our RMSE value indicates that the predictions made by the Q1 model were off by an average of 7 percentage points.

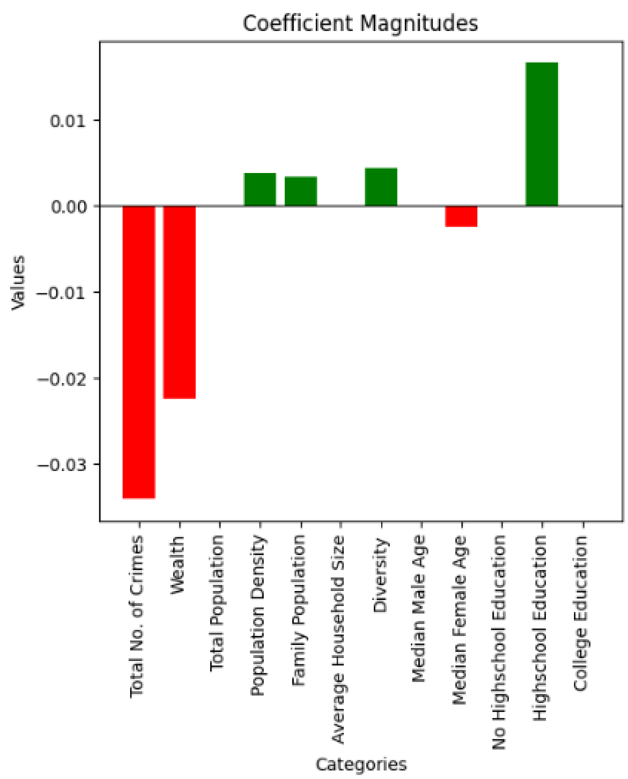

Figure 1, Q1 Model Coefficients Plotted

B. Q2 Model Prediction Results

The test performance indicator values are as follows:

R^2	0.1548
MSE	0.0027
RMSE	0.0520

Table 5, Q1 Model Performance Indicators

The coefficients for the models are as follows:

Wealth Index	-0.0092
Total Population	0.0000
Population Density	0.0136
Family Population	0.0003
Avg. Household Size	0.0058
Diversity Index	0.0000
Median Male Age	0.0000
Median Female Age	0.0000
Not H.S. Educated Pop.	0.0000
H.S. Educated Pop.	0.0010
College Education Pop.	0.0000

Table 6, Q1 Model Coefficient Magnitudes

Once again, please note that test indicators and coefficients are normalized and thus scaled in respect to P(W).

Based on the magnitudes of associated coefficients, it seems that population density, family population, average household size, and high school educated population were covariates in this problem, which our model controls for.

Wealth seems to serve as a strong potential indicator of P(W) in our model, but population density is the strongest predictor of P(W).

Wealth displays an inverse relationship with P(W), similar to wealth in P(V). As the wealth index increases, P(W) goes down. In Q2, every covariate had a direct relationship with P(W): an increase in value meant an increase in P(W).

Our Q2 model observes better fit to the problem than our Q1 model with an R^2 score of 0.1548. This is still a poor fit to the problem, and suggests much of the variance in the dependent variable is unexplained by our Q2 model. MSE is very good at 0.0027, meaning our Q2 model's predictions were also quite accurate when compared to ground truths. Finally, the RMSE value indicates that predictions made by the Q2 model were off by an average of 5 percentage points.

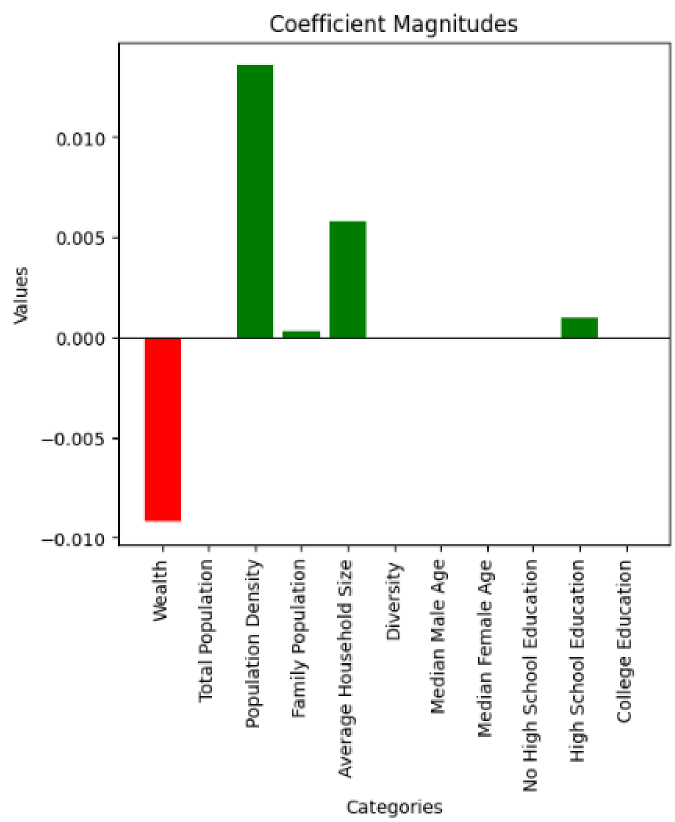

Figure 2, Q2 Model Coefficients Plotted

V. Conclusion

A. Discussion of Results

Both Q1 and Q2 models were able to predict P(V) and P(W) with good accuracy respectively, although our low R^2 scores suggest that they were not sufficiently correlating the variance of the dependent variable with that of the independent variables.

This could be for a variety of reasons, however we believe that it can be chalked up to two primary factors.

Firstly, the dataset is relatively small and full of outliers. As illustrated by Figure 3, the majority of crime occurs in neighborhoods with a wealth index between ~10 and ~75. Neighborhoods outside of this range become outliers for crime rates and thus P(V) and P(W) because of a lack of crime, which may play a role in throwing off the R^2 value.

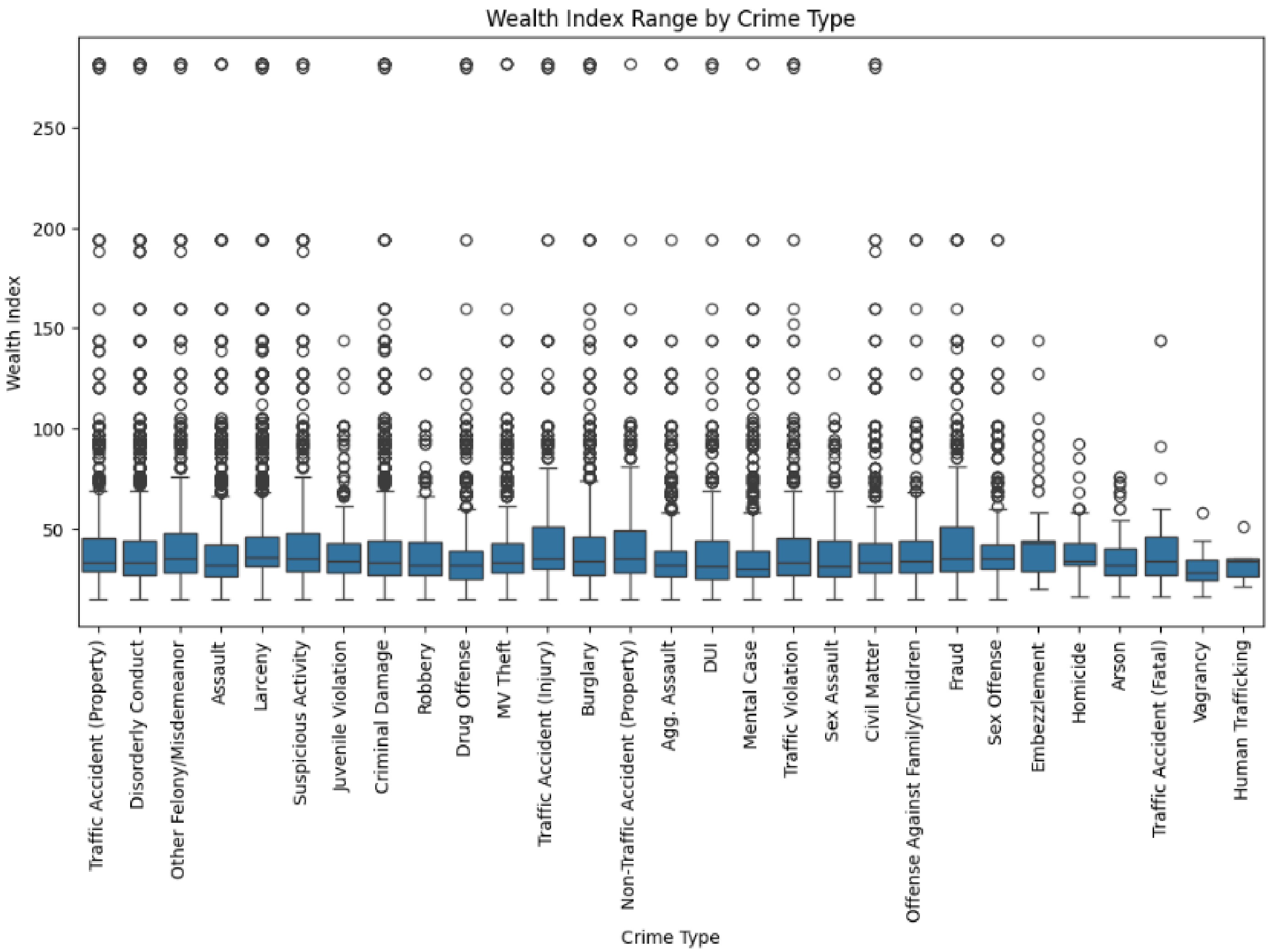

Figure 3, Distribution of Crimes by Wealth Index

Pivoting to the topic of the Q1 model, Figure 3 also illustrates the fact that the wealth range at which nonviolent crimes and violent crimes occur shares significant overlap. This means that for neighborhoods outside of the interquartile ranges of crimes wealth serves as a decent predictor for P(V), within the interquartile range it is an extremely poor predictor because we glean no insights from neighborhood wealth due to the significant overlap between violent and nonviolent crimes.

This suggests that due to the nature of higher income neighborhoods as outliers in our data, they skew the fit of our model such that it does not correlate well with the problem.

We think, though, that given the presence of wealth as a primary predictor in the Q1 model, that we generally accept the alternative hypothesis: given the relationship that wealth has with P(V), there is observably an effect that neighborhood wealth has on violent and nonviolent crime, although wealth is not a good predictor of P(V) alone.

In regards to the Q2 model, it appears to have been a much better fit to its respective problem than the Q1 model. This suggests that there may be a more evident link between neighborhood wealth and P(W), however population density appears to play a major role as well.

Part of this may be because the elastic net model tends to not drop predictors which are correlated with each other: when we think of low-income neighborhoods we typically think of multi-family homes and apartments, whereas high-income neighborhoods may be suburban in nature with a much lower population density, which correlates these two features. So in reality, we are still using wealth as a primary predictor.

We find that similar to Q1, the same issue of using wealth as a predictor arises for the Q2 model. As illustrated by Figure 4, for neighborhoods which fall in the interquartile range it is impossible to reasonably determine P(W) with wealth alone, as in the interquartile range of neighborhood wealth for weapon usage and no weapon usage overlap almost completely.

Outside the interquartile range, we can see that higher-income neighborhoods as indicated by outliers are much sparser in weapon usage than no weapon usage. This suggests that for these neighborhoods, wealth serves as a better predictor of P(W) than it does in the interquartile range.

Although Q2 suffers from the same issues, there is an observable inverse relationship with wealth and P(W) as well, meaning that at the very least we can reject the null hypothesis that there is no correlation whatsoever.

It appears that in both Q1 and Q2, wealth plays a role, but does not paint the entire picture as it only explains 10-15% of the variance of the dependent variables, P(V) and P(W).

An interesting take-away would be that at the micro-level, income inequality and wealth at first glance does not appear to be a good predictor of crime dynamics.

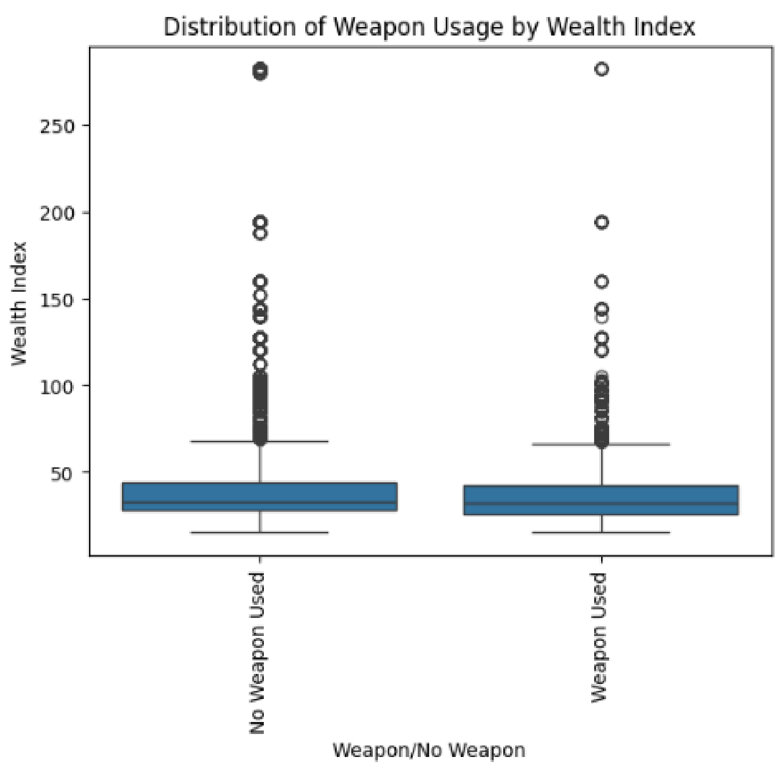

Figure 4, Distribution of Weapon Usage by Wealth

B. Limitations

A limitation that may have contributed to low test-accuracy in the Q1 and Q2 models is the simplicity of the linear model in being unable to capture the underlying complexities of the dataset. A more complex, nonlinear model such as a random forest may exhibit better accuracy as it is better able to capture the complex interactions between demographics and crime. If we had learned about these models and employed them, we may have seen different test results.

Certainly, our low R^2 scores may indicate that there is a nonlinear relationship which is not able to be captured by linear regression alone.

However, a worthy consideration is that when looking at the residuals of our models, it appears that the elastic net model is actually doing quite a good job at modeling the problem at hand.

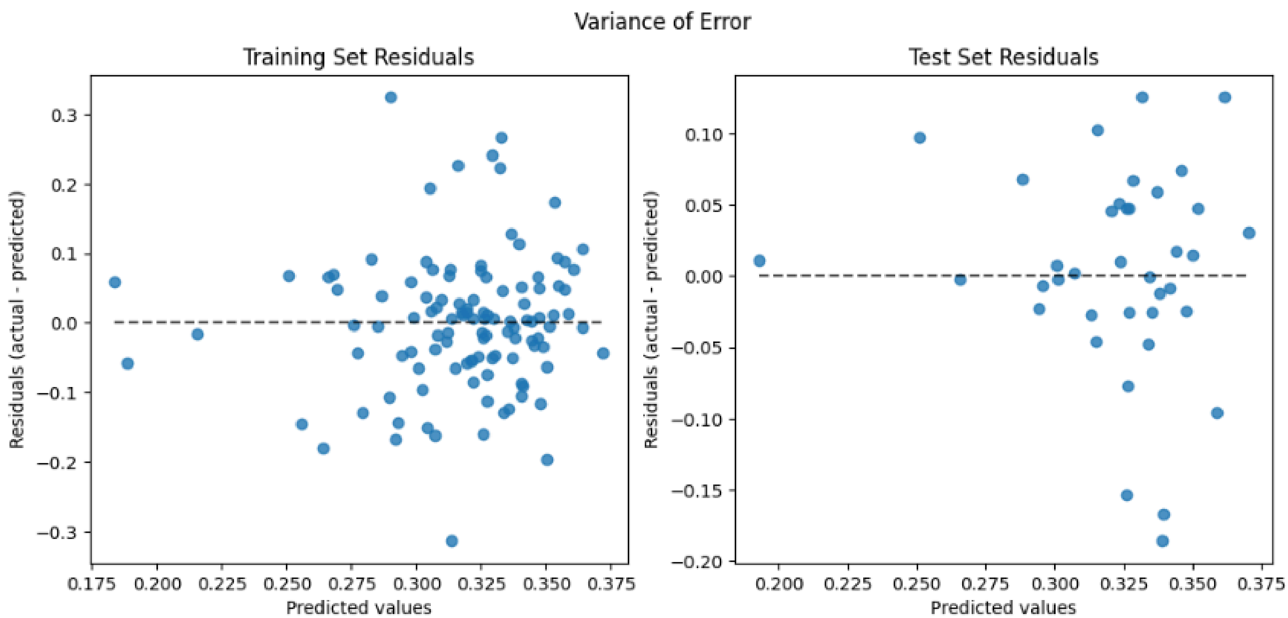

Figure 4, Residuals of Q1 [ABOVE]|| Figure 5, Residuals of Q2 [BELOW]

As shown in Figure 4 and Figure 5, the residual plots of both models showcase a typically desirable variance of error distribution: data points are evenly distributed, tending to cluster towards the middle of the graph. In our training data, the clusters tend to be around the low values on the y-axis.

Thus it may not be that linear regression models are too simplistic to capture the intricacies of the relationships between predictors, but rather that there simply was a small and sparse dataset which was further compounded by outliers in the form of neighborhoods with extremely high or extremely low wealth.

C. Future Work

Further expanding the dataset to encompass the greater Tucson MAS would likely alleviate some of the fitting issues that we observed during the course of this study. Unfortunately, the plausibility of that is rather low because of idiosyncratic data reporting standards, a significant amount of effort in data normalization would have to be invested in order to repeat this same study with that dataset.

An alternative to seeking out more data in the Tucson MAS is to find a MAS with a larger dataset of neighborhoods and greater availability of data. It will certainly be interesting to see whether or not a larger dataset will boost our R^2 score.

Finally, an expansion on this project would be to attempt to train nonlinear models such as a random forest model to undertake the same regression task. It is entirely possible that there are unforeseen dynamics between the data that cannot be adequately captured with simple linear regression.

References

[1] “Adverse Childhood Experiences as a Risk Factor for Anti-Social Behaviour Among Young Adults in the Western Cape, South Africa | Acta Criminologica : African Journal of Criminology & Victimology,” Acta Criminologica : African Journal of Criminology & Victimology, 2021, https://hdl.handle.net/10520/ejc-crim_v34_n1_a1.

[2] “Tucson, AZ | Data USA,” Datausa.io, 2017. https://datausa.io/profile/geo/tucson-az/

[3] “Public Safety | MAP AZ Dashboard,” mapazdashboard.arizona.edu. https://mapazdashboard.arizona.edu/quality-place/public-safety

[4] P. Fajnzylber, D. Lederman, and N. Loayza, “Inequality and Violent Crime,” The Journal of Law and Economics, vol. 45, no. 1, pp. 1–39, Apr. 2002, doi: https://doi.org/10.1086/338347.

[5] M. Pazzona, “Revisiting the Income Inequality-Crime Puzzle,” World Development, vol. 176, p. 106520, Apr. 2024, doi: https://doi.org/10.1016/j.worlddev.2023.106520.

[6] J. Brush, “Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties,” Economics Letters, vol. 96, no. 2, pp. 264–268, Aug. 2007, doi: https://doi.org/10.1016/j.econlet.2007.01.012.

[7] “Essential vocabulary—Esri Demographics Reference | Documentation,” Arcgis.com, 2020. https://doc.arcgis.com/en/esri-demographics/latest/reference/essential-vocabulary.htm#ESRI_SECTION1_61CB470990084126AF39D2AED9F4B94C (accessed May 06, 2025).

[8] “Tucson Crime Rates and Statistics - NeighborhoodScout,” www.neighborhoodscout.com. https://www.neighborhoodscout.com/az/tucson/crime