Analysis of Crime in Tucson

Claire Lodermeier and Chitrangada Juneja

CSC 380: University of Arizona

8th of December 2024

Introduction

Background

In this project, we will be discussing and analyzing the different factors and characteristics that affect crime in Tucson, using open-source data available to the general public to model and predict crime trends. Crime is a multifaceted social issue that deeply impacts communities, shaping the economic health and well-being of cities. Our goal with this project was to research through data driven methods and reassess our understanding of crime in Tucson using relevant machine learning models. We gathered the datasets from several sources, the primary one being the City of Tucson's Open Data website. We also referred to a past report, research papers, and crime statistics whilst forming the basis and deciding the questions we wanted to address in this project. Through our research, we decided to explore the relationship between 4 independent factors-Income, Education, Unemployment, Downtown Distance- and Crime, through the years in Tucson. Our goal was to discover a relationship between any of the above factors that affected Crime in Tucson, and once found- discuss what its ethical implications and possible predictive policing would mean for Tucson citizens.

As we engage in this open-ended analysis, it is important to address the possibility of possible bias, data selection pitfalls and unintended consequences of our models' designs and predictions. Our ultimate aim with this project, was not to simply build accurate models, but to contribute to an informed understanding of crime in this city.

Objectives

The primary objective of this project is to identify and analyze several factors that play a role in influencing crime in Tucson. We examined data through the years, and through data analysis tools such as Exploratory Data Analysis (EDA), and Ridge and LASSO models, attempted to address the following questions:

1. *How does crime vary across neighborhood*s*? More specifically, with differing*
    a. *Income per Capita?*
    b. *Education Attainment?*
    c. *Unemployment Rates?*
    d. *Distance from Downtown?*
2. *How do these factors correlate, if at all? Has the crime rate been going down or up over the years?*
3. *Can changes in socio-economic conditions and policies at certain neighborhoods have a direct influence on the crime rate?*

By answering these questions through the use of analytical tools and visuals, we aim to gain a deeper understanding of how factors influence crime rates in Tucson. The tools we choose aim to identify key drivers of crime, highlight spatial and temporal trends, and show that socio-economic conditions and targeted policies can have a significant impact on reducing crime. We hope this analysis provides valuable insights for policymakers aiming to implement effective crime reduction strategies based on socio-economic conditions.

Related Work

There exist several studies and reports that explore the relationship between crime and social, political and geographical conditions; however, the ones we primarily referred to for this study were:

1. **City of Tucson: Poverty and Urban Stress Report, 2020**
    - This report offers detailed examinations of some of the factors we employed in our work- median income, poverty, and education in Tucson. It uses American Community Survey data to analyze trends in poverty, income, education and related stress indicators. It provided contextual socio-economic data that helped us identify high-stress geographic areas, which in turn aided us as we cleaned and reviewed different datasets in our analysis.
2. **"The Determinants of Crime in Tucson, Arizona" (Cahill, Mulligan**)
    - This research paper written in 2003 by faculty members at the Department of Geography and Regional Development, the University of Arizona, really reinforced that the questions we were asking were practically relevant, and that we were on the right track. It investigates the social, economic, and spatial factors that affect crime patterns in Tucson, and examines neighborhood-level variations in crime. Areas with high level of stressors were identified with higher crime rates, and integrating these insights into our work was essential in helping us ensure that our paper was theoretically sound.

Methods

Data Pipelining

1. ***Data Acquisition**:* The following datasets were collected and used in our overall evaluation:
    a. **Tucson Police Reported Crimes**: Dataset containing Tucson Police's reported crimes from 2017 to the present.
    b. **City of Tucson Ward Boundaries**: Shows the six wards of the City of Tucson, and their addresses.
    c. **Police Incident Reports**: We referred to data contained in Police Incident Reports ranging from 2014 to 2023. These contained references to crimes for which reports were filed, not just called in, and have geographical locations corresponding to where the Officer responded, which was within a 10-block radius to where the crime occurred. The datasets contained a lot of information, the most relevant to our work being the addresses, latitude, longitude, and ward of the crime occurrence.
    d. **Neighborhood Educational Attainment**: This dataset contains the various levels of education attainment according to neighborhoods in Tucson, aggregated from data last collected in 2019; we focused mainly on these 3 columns: Associate degree, bachelor's degree and Grad/Professional Degree.
    e. **Neighborhood Population Statistics**: Dataset that contains the population counts in Tucson by neighborhood and ward, aggregated between 2010-2019.

f. **Neighborhood Income per Capita**: This dataset was particularly important in providing information about the income per capita distributed throughout Tucson by neighbourhood and ward, aggregated between 2010-2019.

g. **Neighborhood Employment Demographics**: This dataset contains information about the different occupations people have held in Tucson, aggregated up to 2019. This enabled us to calculate the unemployment rate per neighbourhood.

2. ***Data Preprocessing***: This step included cleaning missing values, getting rid of redundant columns, dropping rows with no relevant data, and making sure all our datasets were on the same year-range as we conducted our investigation into the various factors. In several of the datasets, such as the Educational Attainment, we had to merge and sum columns for them to be relevant and scaled in accordance with the dataset they were being compared against.
3. ***Feature Selection***: Features relevant to the specific questions we asked were examined, once the data was cleaned. The goal for this was to identify factors most relevant to predicting crime rates as we defined them, while also noting their limitations in generalizing to test data.

Models and Analysis Techniques

1. **Exploration**

- An initial inspection of the Tucson reported crimes database didn't yield meaningful results, since there was little information about each datapoint. Looking at the available

data about neighborhoods and wards led us to questions about how income, education, employment and location are all related to crime rates in different areas.

- Once we had established that an exploration of geospatial and temporal factors would yield meaningful results for our proposed questions, we decided to form our tailored datasets for our models.
- After our data cleaning, elementary visualizations aided us in gaining a sense of what possible hypotheses we could be looking at of interest.

2. Datasets Used

- In our initial dataset, details about neighborhoods/ locations of the crime were sparse, and since we required a generalization that gave use enough data points to train a model, we turned to the actual police incident reports, from 2009, up to 2023.
- We cleaned this data and decided to use neighborhoods as the basis for relating crime rate to our features of interest.
- We first decided to investigate the relationship between crime rate and distance from downtown Tucson, as it seemed in our initial exploration that crime was very concentrated in that region.
- The datasets by themselves were not useful; however, data cleaning yielded results we could integrate into our work. In particular, getting rid of missing or invalid values, filtration of the data, and calculations made for important features such as crime rate and Education Ratio, helped us narrow our focus.
- We established our four features, Income **(INCOME**), Unemployment (**UNEMPL**), Average distance from Downtown (**AVG_DIST**), and Ratio of Educated Citizens (on

the basis of neighborhood) (**EDU_RATIO**), and created our datasets tailored to our needs. We also generated all possible subsets of the various combinations of these features to explore those relationships as well.

3. Analytical Tools Employed

The primary programming language for our work was Python, alongside relevant libraries such as Pandas, Seaborn, GeoPy and NumPy to aid us in our data processing, visualization, and model design. Colab Notebook by Google served as the primary development environment, facilitating collaboration and easy documentation.

Visualization: We used a variety of visualization tools that helped us notice correlations and understand the data, such as:

a. Bar Plots: to visualize the number of crimes in different wards, to compare crime rates among top 10 neighborhoods with the highest crime rates.

b. Scatterplots: To visualize crime rates versus the selected features, i.e. INCOME, UNEMPL, AVG_DIST, and EDU_RATIO.

c. Histograms: To display and understand the distribution of crime incidents based on their distance from downtown. This was particularly useful in understanding the change in density of crime across different ranges of distance.

d. Line Plots: To show temporal trends in crime rates for different neighborhoods from 2014 to 2019.

e. Kernel Density Estimation (KDE) Plots: We created a smoothed density curve of crime incidents' distance from downtown, accompanying our histogram for the same, as an alternate visual.

Models Designed: The models chosen for this project reflect our gradual understanding of the complex dynamics between factors that affect crime in Tucson. After our visualization of the respective data was complete, we decided to use the following techniques in our models to test and train the data:

a. Simple Linear Regression: A straightforward starting point to explore relationships between independent variables (such as income and unemployment) and the dependent variable -crime rate. This helped us understand the baseline for understanding the strength of relationships.

b. Polynomial Regression: To capture non-linear relationships between the features; this improved the model's flexibility and was a clear indicator that there is no simple relationship to explore here.

c. Ridge Regression: Learning from the results of our models, we decided to use Ridge Regression to further the accuracy by penalizing large coefficients and reducing overfitting of the data.

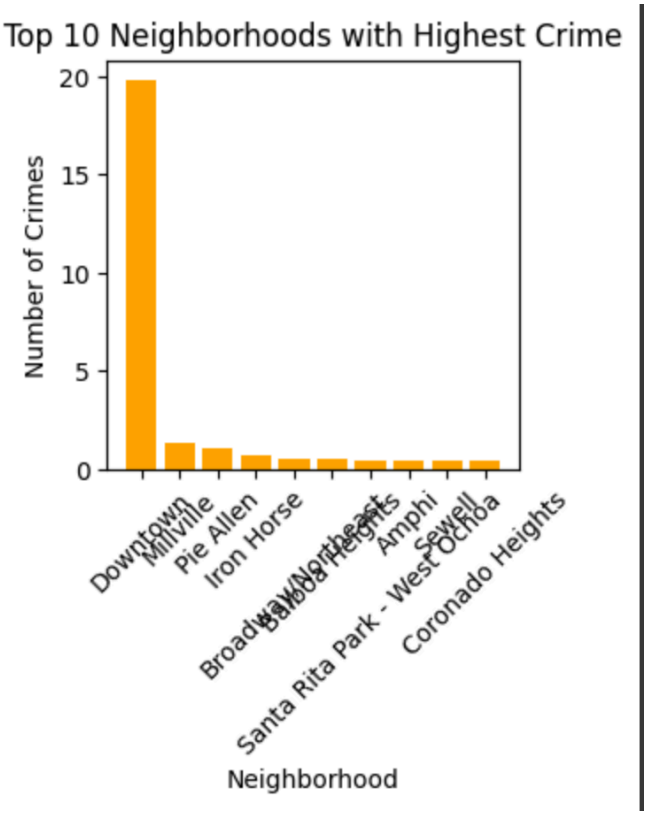
d. Lasso Regression: This enabled us to further refine our work by performing feature selection by shrinking the coefficients of less relevant features to zero. While this did simplify our model, there did not seem to be any standout key predictors.

Results

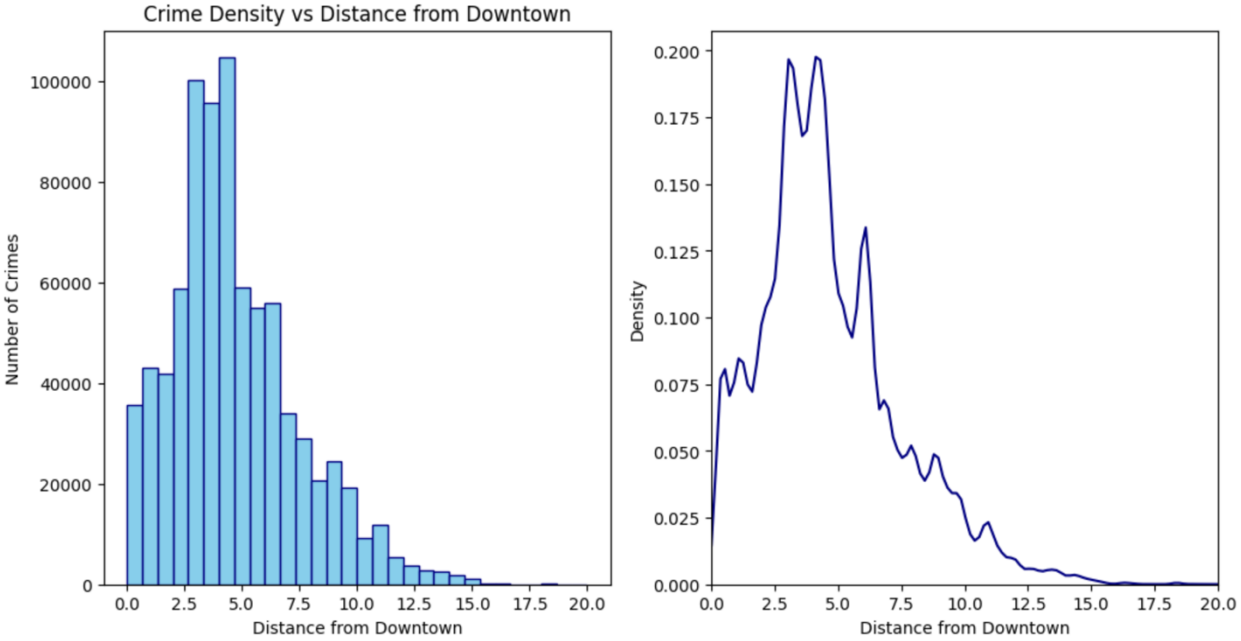
This section highlights the outcomes of each modeling approach, evaluates the models' ability to generalize to new data, and discusses the implications of the findings in addressing the problem at hand. The result of analysis provides a comprehensive evaluation of the relationship between crime rates and socioeconomic factors, using predictive modeling and statistical techniques as outlined above. The workflow includes data preprocessing, exploratory data analysis (EDA), feature selection, and the implementation of multiple regression models to identify the most influential predictors. Key insights were derived through both quantitative metrics and visualizations, allowing for an in-depth assessment of model performance and its alignment with the data.

Analysis on Crime Rate in Downtown Neighborhoods

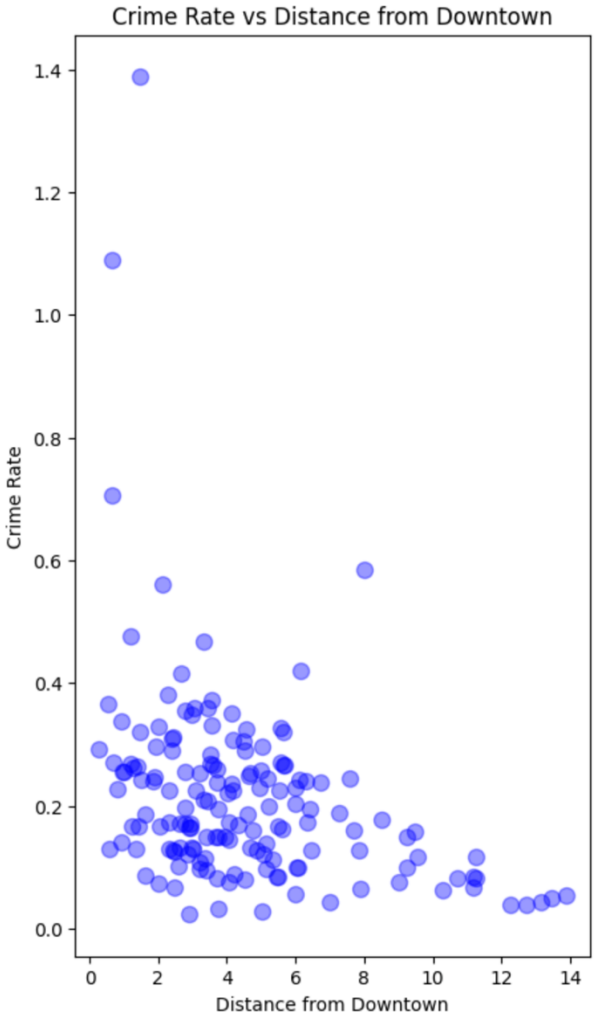
From the initial data clean, we could infer those neighborhoods in downtown Tucson seemed to have a higher concentration of crime. We decided it was reasonable to hypothesize that proximity to downtown could influence crime rates, whether due to increase opportunities for crime, heightened law enforcement presence, or socioeconomic disparities. Crime rates often show spatial variability within cities, with downtown areas frequently having unique dynamics due to population density, economic activity, nightlife, and access to resources. By including this variable, the analysis captures a critical dimension of the data that enhances its applicability and interpretability.

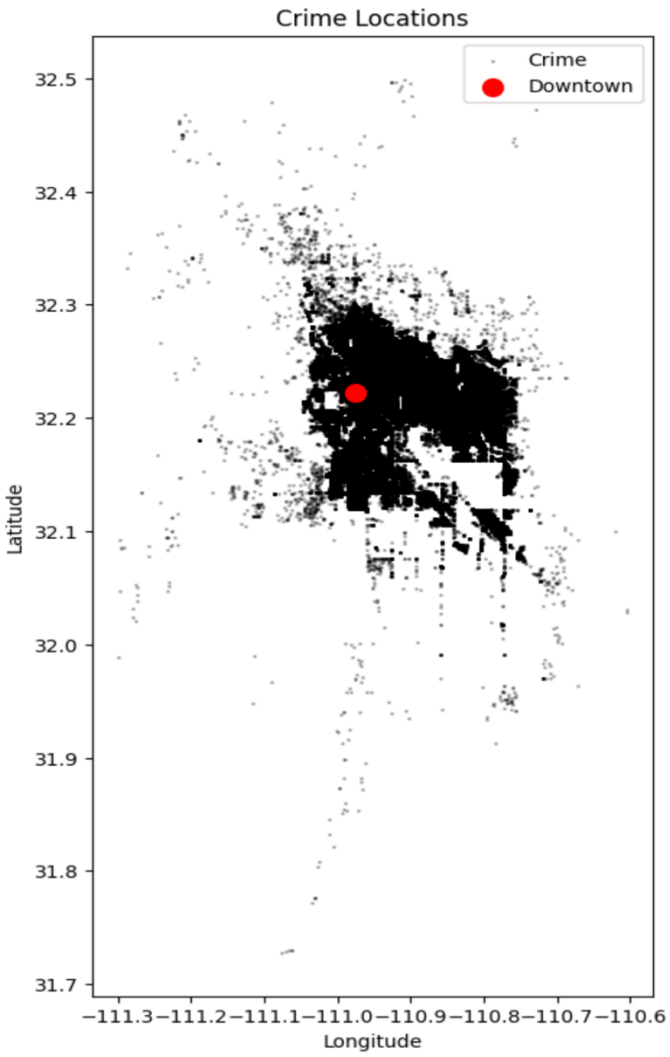


The above bar graph provides an initial visualization about the distribution of crime across neighborhoods. Clearly, downtown appears to have a significantly high value- thus, we then proceeded to add data about each incident's distance from the city center and create visualizations for the same. By computing latitudes and longitudes, alongside ward information, we inferred distances from downtown for each individual crime and created a brand-new column in our dataset highlighting each crime's distance from downtown. The histogram and KDE plot below highlight this information. Clearly, there was a certain radius from downtown that was related to distribution of crimes.



Next, we decided to compute crime rates, defined essentially as a ratio of reported crimes to the population of a given area, normalized to allow for meaningful comparison across regions. A scatterplot for the same allowed for better visualization of our data.





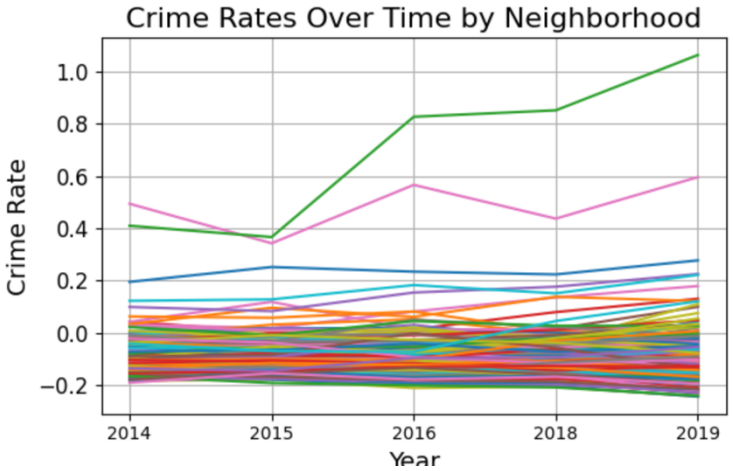
The graph on the right reflects the information given by the scatterplot by directly plotting the data onto the map of Tucson, clearly validating our claim that using distance as a feature in the analysis is reasonable.

The results from our initial foray into investigating crime in Tucson highlighted how distance from downtown, or rather, any analogous geographical variables, are significant features in predictive models for the same. From our visualization, we concurred that distance from downtown has a negative correlation with crime rates, likely due to lower population density and economic activity in peripheral areas. Next, we explored the temporal relationship between crime rates and neighborhoods.

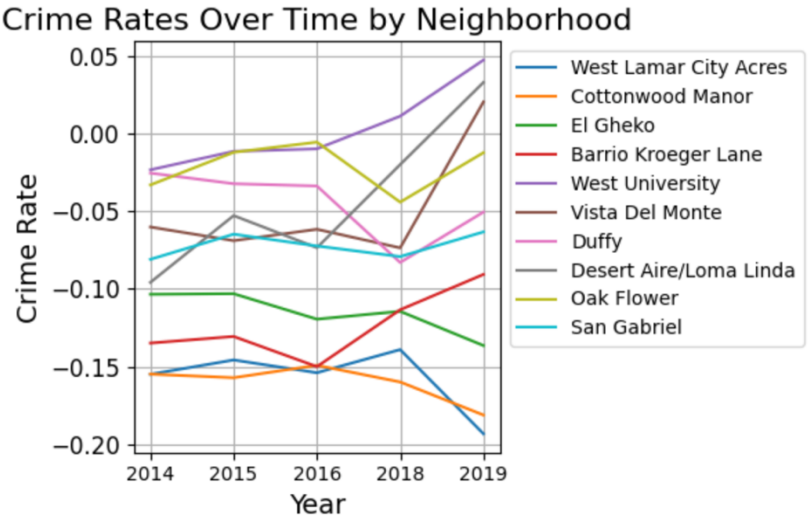
Analysis of Crime Rates in Neighborhoods through Time

Crime is a dynamic and changes over time, and throughout our investigation, we slowly realized that the real-world drivers of crime rates are more complex than the variables included in these datasets. Analyzing crime rates in neighborhoods through time was the most logical next step in establishing a geospatial awareness for our investigation. This step was also important because the results would also clearly reflect the results of past interventions to reduce crime, such as policy changes, increased patrolling or community programs in certain neighborhoods.

Initially, we tried to push all our data into a single graph. This resulted in a cluttered data set that was difficult to use for inferences.



We decided to partition our data in such a way that every time we generated a graph for Crime rate over time, we would get 10 random neighborhoods. This ensured our visualizations were not biased in any way.



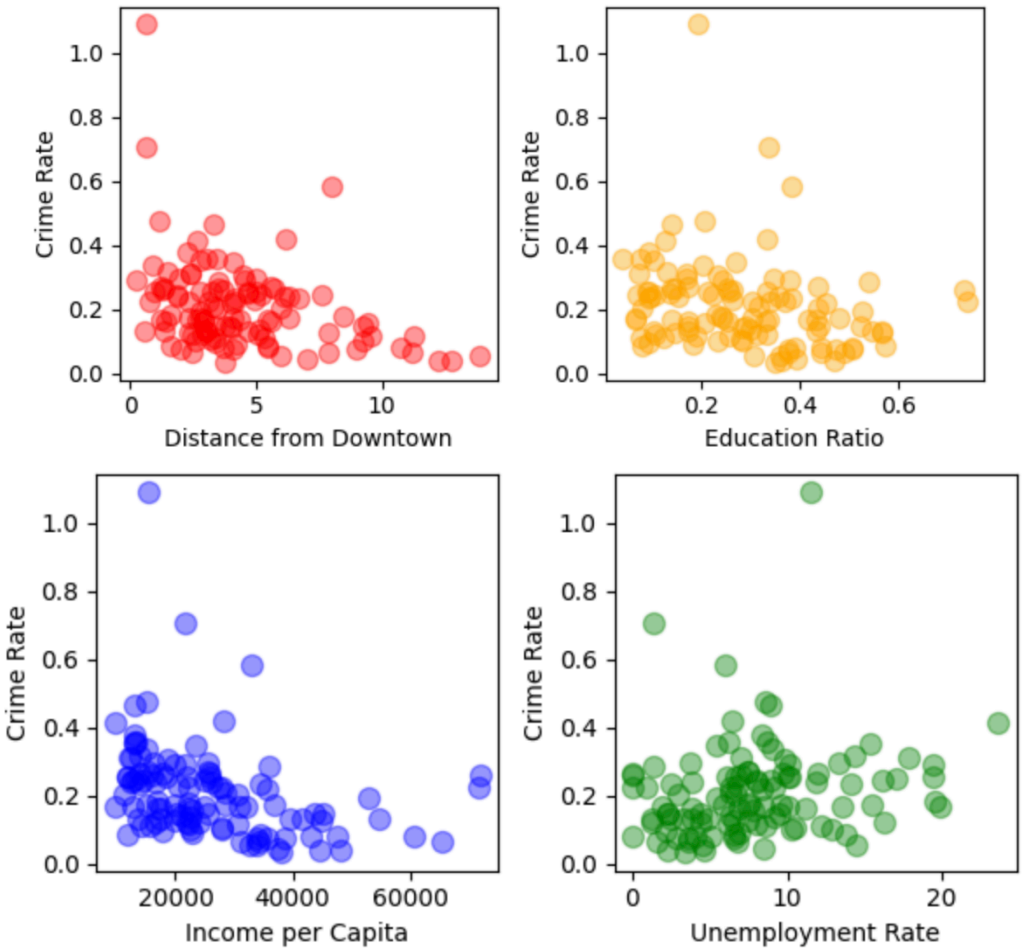
The graph above helped us infer two things- one, that crime rates hadn't been positively affected by any policy changes; if anything, they all averaged out to be almost constant, if not increasing. Two, different graphs generated all supported our first initial analysis that crime rates were negatively correlated with distance from downtown (AVG_DIST). Taking the graph above as an example, West University, Barrio Kroeger Lane and Vista Del Monte are all neighborhoods that are in close proximity to Downtown Tucson and show a clear rising crime rate over the period of 2014 to 2019. Neighborhoods like San Gabriel, while less in number, show relatively stable crime rates, despite their proximity to Downtown Tucson, implying that local neighborhood policies may influence crime.

It is important to note that overall, crime rates have not decreased through the years; this is an important takeaway from our research, as it can positively influence decisionmakers and officials. Preventive strategies, rather than solely punitive responses, may hold the key to addressing crime over time.

Analysis of Regression Models

After much data cleaning and preprocessing and splitting our training and testing datasets into a 70-30 ratio, we began to lay down the structure for our regression models.

First, we wanted visualizations for crime rate versus each of our features, to provide a backbone for the questions we were asking.

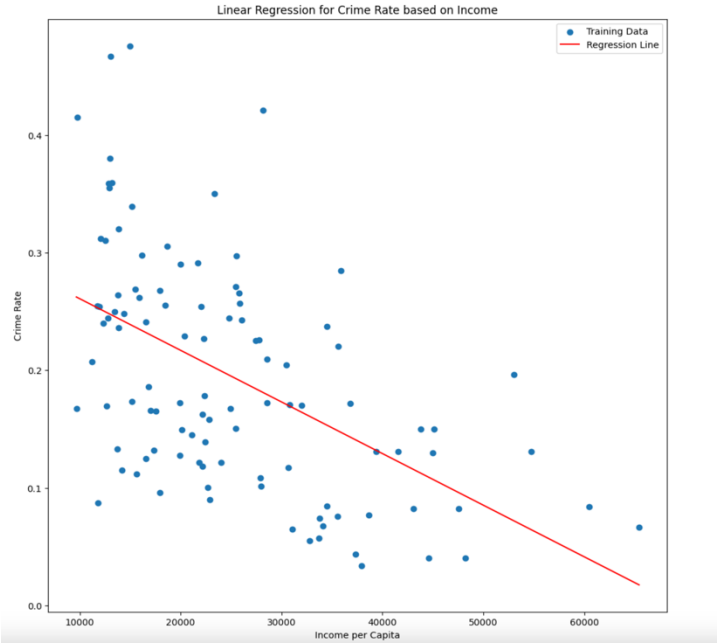


We create a visualization that helps explore the relationship between the crime rate in Tucson neighborhoods and factors like income, unemployment, distance from downtown, and education levels Based on these scatterplots, neighborhoods with crime rates greater than 0.5 are outliers.

Throughout the following model discussions, the features will be presented as they are, followed by the variables that represent them in the code; for example: UNEMPL for Unemployment. We will talk extensively about R^2 scores-which basically indicate the proportion of variance in our dependent variable (crime rate), that can be explained by our independent variables (our four features). We will also talk about Mean Squared Errors (MSE), which basically evaluates the performance of a regression model, and Residuals, which are the differences between the actual values and the predicted values from the model.

Linear Regression

Next, we began with a simple linear regression model to explore any correlations between income per capita (INCOME), and crime rates (CRIME_RATE), and produced a scatterplot for the same. The dataset is shuffled to ensure zero bias.



Findings: Points were spread weakly around the regression line- implying that while there was a relationship between income per capita and crime rate, it was weakly modeled by simple linear regression. The negative slope indicates that crime rates tend to decrease as income increases, which makes sense, since richer neighborhoods tend to have lower crime levels due to higher security levels and surveillance capabilities. We expected to see this relationship. The R^2 score was 0.23, which indicated a weak albeit present relationship between crime rate and income.

Polynomial Regression

Our results from linear regression made us ask ourselves if the relationship between these features and crime rate was perhaps more complex than we assumed; to test this, we designed a polynomial regression model, by squaring our independent variable, INCOME. We noticed an improved R^2 score of 0.3, which then prompted us to explore more complex models, with more features interacting.

Feature Selection

We decided to calculate R^2 scores for different subsets of the features, in both our linear and polynomial regression models. The feature selection process in the PDF for both the linear and polynomial regression models involved testing different subsets of features to determine which combination best explains the variance in the target variable (crime rate). The results for both were as expected- crime rate by neighborhood was influenced by all our selected features, and our models yielded the most accurate results when we used multi-variables. Our R^2 scores increased as well, with a 0.35 and 0.4 value respectively. This suggests that the features

selected for the model are all impacting crime rates in some nonlinear way but are not the only factors contributing to the complexity of crime rate. The R^2 scores for our best linear and polynomial regression models did not have a strong result – we observed low values in the range of 0.25-0.30, with Mean Squared Error values of 0.0075 in both. This informed us that while our model was predicting crime rates well, the relationship between our factors and crime rate itself was far more complex than anticipated.

To determine if our model was overfitting, we decided to implement Ridge and Lasso Regression models, and check those with our training and testing data.

Lasso Regression

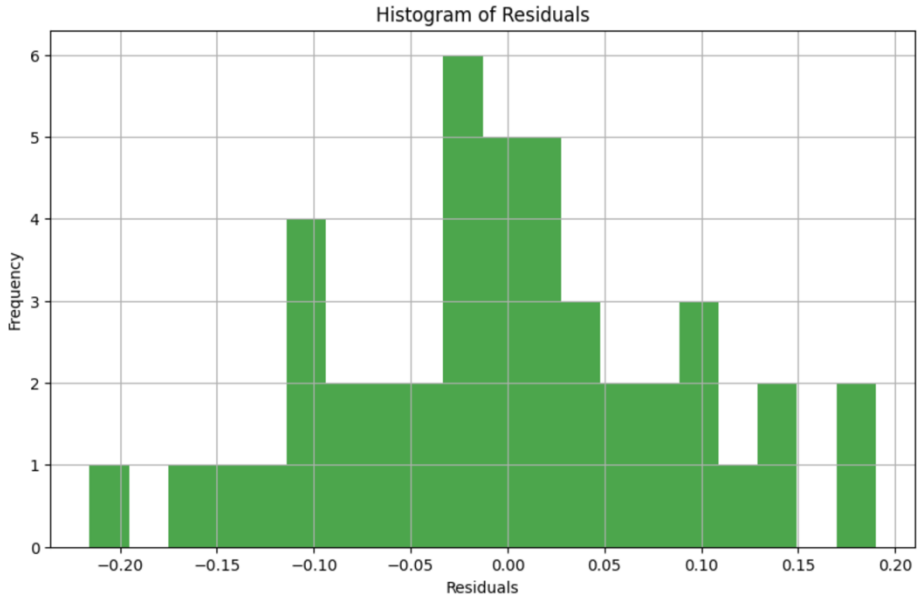
Our Lasso Regression Model helped inform our decision to proceed with the ridge regression model, with alpha values of 0.015 and a Mean Squared Error of 0.006- indicating that our model requires lower levels of regularization (indicating it is a good fit), and that on average, our model was quite accurate at predictions of crime rate based on our 4 parameters.

Ridge Regression

Features like income, unemployment, education ratio, and distance from downtown may have interdependencies, leading to multicollinearity. This makes the coefficients in ordinary least squares (OLS) regression unstable. Ridge regression spreads the importance across multiple features by shrinking their coefficients, thus balancing the trade-off between bias and variance. Our Ridge Regression model was by far the best in terms of working with the training data- A good R^2 score of 0.712 with an alpha value of 0.625. The α value represents the regularization parameter in Ridge regression that controls the strength of the penalty applied to the size of the model's

coefficients. At α=6.25 the model applies a moderate penalty to the coefficients, striking a balance between underfitting (not capturing enough detail) and overfitting (capturing noise). The Ridge regression model successfully reduces overfitting compared to unregularized linear regression, as evidenced by the higher R^2 score in cross-validation. This result suggests that Ridge regression is an appropriate choice for the data.

When evaluated with the testing data, however, we noticed a low R^2 score- this indicated to us that our model required a larger dataset and was missing critical predictors. Overfitting could also be a contributor, but that would be unlikely as our Ridge Regression model's purpose was exactly to prevent that.



To better understand the results of our ridge regression model, we created a histogram to represent the distribution of residuals from the model's predictions on the test dataset. A good model is characterized by residuals that are centered around zero (indicating unbiased predictions) and spread symmetrically (no systemic over- or under-prediction). Clearly, the

residuals range from -0.2 and +0.2, indicating the model predictions are close to the actual values, and that since the residuals are mostly centered near zero, the model's predictions are not significantly biased. The histogram appears symmetrical, which indicates that the model's errors are evenly distributed—neither predominantly positive (underprediction) nor negative (overprediction). The highest frequency of residuals is around −0.05 to +0.05, meaning the majority of predictions are fairly close to the true values.

Therefore, from all these details, we infer that the issue with our testing lies in our lack of enough datasets comprising the wide variety of data we trained our model on, along with a need for additional features and more complex models to predict crime rates.

Conclusion

On the basis of the analytical tools described above, we found the following answers to the questions we asked:

1. How Does Crime Vary Across Neighborhoods?

- **Income per Capita:** Crime rates tend to decrease with increasing income, but the relationship is non-linear, as suggested by the inclusion of quadratic terms in the model. Wealthier neighborhoods generally experience lower crime rates, but other factors can moderate this effect.
- **Education Attainment:** Higher education levels are associated with lower crime rates, although the relationship is not entirely linear and varies across neighborhoods.
- **Unemployment Rates:** Higher unemployment is linked to higher crime rates, supporting the notion that economic stability reduces criminal activity. However, the effect weakens in some areas, indicating other factors at play.
- **Distance from Downtown:** Neighborhoods closer to downtown experience higher crime rates, likely due to higher population density, economic activity, and transient populations.

2. Correlation Between Factors and Crime Trends

- Socioeconomic factors like income, education, and unemployment are moderately correlated with crime rates, but no single factor completely explains the variability.
- Over the years, crime rates have fluctuated across neighborhoods, with some areas experiencing rising trends (e.g., Duffy, West University) while others, such as Cottonwood Manor, show consistent declines. This suggests varying impacts of localized conditions and interventions.

3. Influence of Socioeconomic Conditions and Policies

- Changes in socioeconomic conditions and targeted policies can directly influence crime rates in certain neighborhoods. For example, neighborhoods with rising income and education levels or effective crime-reduction programs have shown declining crime rates. Conversely, areas with persistent unemployment or inadequate policy support remain hotspots for crime.

The results reveal that while factors like **income, unemployment, distance from downtown, and education ratio** partially explain crime rate variability, the best-performing model—a Ridge regression with an optimal regularization parameter (α=6.25) achieved a maximum R^2 score of 0.712 during cross-validation. However, the low R^2 on the test set (0.1857) highlights challenges in generalization, suggesting the presence of unmodeled complexities or missing features.

These findings underscore the need for a more holistic approach to understanding and addressing crime, combining robust data collection with advanced modeling techniques. Future analyses could incorporate additional variables, such as policing practices, housing stability, or cultural dynamics, and explore non-linear models or machine learning approaches for improved predictive power.

In conclusion, while this study successfully identifies meaningful trends and relationships, it also highlights the complexity of crime as a social phenomenon and the importance of tailoring solutions to the unique characteristics of neighborhoods. By combining these insights with targeted interventions and community engagement, policymakers and stakeholders can work toward creating safer, more equitable communities in Tucson.

References

1. City of Tucson GIS Data. "Tucson Police Reported Crimes." Available at: https://gisdata.tucsonaz.gov/datasets/tucson-police-reported-crimes/explore
2. City of Tucson GIS Data. "Tucson Neighborhood Boundaries." Available at: https://gisdata.tucsonaz.gov/datasets/0657741d9e4a4289912330077707ad39_15/explore?location=32.154769%2C-110.881650%2C9.81
3. City of Tucson Open Data. "Tucson Police Incidents 2023 Open Data." Available at: https://data-cotgis.opendata.arcgis.com/datasets/cotgis::tucson-police-incidents-2023-open-data/explore?location=32.221543%2C-110.989100%2C8.21&showTable=true
4. City of Tucson GIS Data. "Reported Crimes by Location." Available at: https://gisdata.tucsonaz.gov/datasets/a4ed8b6bf0ad4515bfc43df83175e40f_0/explore?location=32.197880%2C-110.889177%2C10.33&showTable=true
5. City of Tucson GIS Data. "Reported Crimes by Ward." Available at: https://gisdata.tucsonaz.gov/datasets/59f033d07eae41b0bdc21db87375d721_0/explore?filters=eyJXQVJEIjpbMSw1LjU0XX0%3D&location=32.197917%2C-110.889177%2C10.92&style=WARD
6. City of Tucson GIS Data. "Unemployment Rates and Census Data." Available at: https://gisdata.tucsonaz.gov/datasets/ac0ce3fbf3f24312b170f3bff94f8956_0/explore?filters=eyJVTkVNUFJUX0NZIjpbMCwyMy42XX0%3D&location=32.197552%2C-110.889177%2C10.33&showTable=true

7. City of Tucson GIS Data. "Demographic Data for Tucson Neighborhoods." Available at: https://gisdata.tucsonaz.gov/datasets/9978db16633f4d45ac8a6b31d82f95f7_0/explore?location=32.197879%2C-110.889177%2C10.45
8. City of Tucson. "Poverty and Urban Development Report 2020." Available at: https://www.tucsonaz.gov/files/sharedassets/public/v/1/living-and-working/housing-community-development/documents/povreport2020.pdf
9. Mulligan, G., & Vias, A. (2001). *The Determinants of Crime in Tucson, Arizona.* Available at: https://www.researchgate.net/profile/Gordon-Mulligan-2/publication/250171621_The_Determinants_of_Crime_in_Tucson_Arizona_1/links/53d87b140cf2631430c32209/The-Determinants-of-Crime-in-Tucson-Arizona-1.pdf
10.