The Qualitative Analysis on the Relationship Between Household Income and Tucson's Reported Crime

Joyce Dam

Louis Romeo

Claire O'Brien

University of Arizona

CSC 380 Principles of Data Science

Cesim Erten

12/10/2024

Introduction:

The correlation and causes of crime are complex in that there are many forms and types of crime with increasingly nuanced motives. There are also many compounding factors that contribute to the causes of crimes. The occurrences of crime and their type is one of the many concerns people have in regards to safety when traveling to unfamiliar locations. There is also great concern about identifying the factors that are the greatest causes of crime, so that by addressing those problems crime can be stopped at its roots. In this project, we aim to analyze a possible correlation between the household incomes and the multiple facets of crime in Tucson, Arizona. We will investigate crime and household income's relationship, starting at smaller locations in Tucson such as its wards and neighbors. Additionally, we will examine different types of crime and their occurrences at each neighborhood/ward, offering our insights on the relative safety at certain locations.

Objectives:

Our purpose of this project is to analyze, create visualizations, and convey the causal relationship between crime in Tucson and household income using data available at City of Tucson Data Hub. Using our analysis, we will discuss the occurrences of crime, the correlations of different types of crime with respect to household income and the possibility of household income used as a prediction for the occurrence of specific crime. We will aim to answer the following questions:

1. What kind of relationship does income level and frequency of crime pertain to?
2. How does income level affect what kind of crime is committed?
3. Can we use statistical models to predict crime rate using the median household income?

Related Works:

There is a long history of the study of the relationship between income inequality and crime levels. The exact factors considered when researching these (such as the population density of the area and the classification of the crimes in the dataset) can affect the results, but the general consensus is that income level does have an impact on crime rates. For example, in Patterson's 1991 article, "POVERTY, INCOME INEQUALITY, AND COMMUNITY CRIME RATES", Patterson studied a collection of residential areas and concluded that severe poverty was associated with higher rates of violent crime but not burglary, and that "relative poverty" does not have a significant correlation with violent crime or burglary. However, Patterson factored in numerous factors about the characteristics of the neighborhoods being surveyed and heavily emphasized social factors such as the degree of an area's social integration. Our research focused more on the direct relationship between income and crimes, and features greater division of types of crimes, beyond Patterson's main distinction of "violent crime" or "burglary".

In tangent, crime and economic equality may be more relevant in locality as opposed to national or country-level. According to Kang, many crimes taken are usually within an offender's residential area which likely follows that an offender's decision to commit crime most often correlates to the economic inequality in their own neighborhood. Not only that, but despite the logical argument that offender's would commit crimes in wealthier areas for higher gains, it is often not the case. This is theoretically due to wealthier neighborhoods having higher risk of detection and often higher levels of punishment. Aspects we found most relevant in this study had been the type of model used for the main regression, and the usage of log crime rate per capita to normalize data and to take into account the various population sizes in each neighborhood.

Methodology: To answer our questions in the Objectives, we will utilize the following datasets:

Tucson Neighborhood Population:

https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-population-statistics/about

Tucson Police Reported Crimes:

https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-reported-crimes/about

Tucson Police Arrests:

https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-arrests-2021-open-data/about

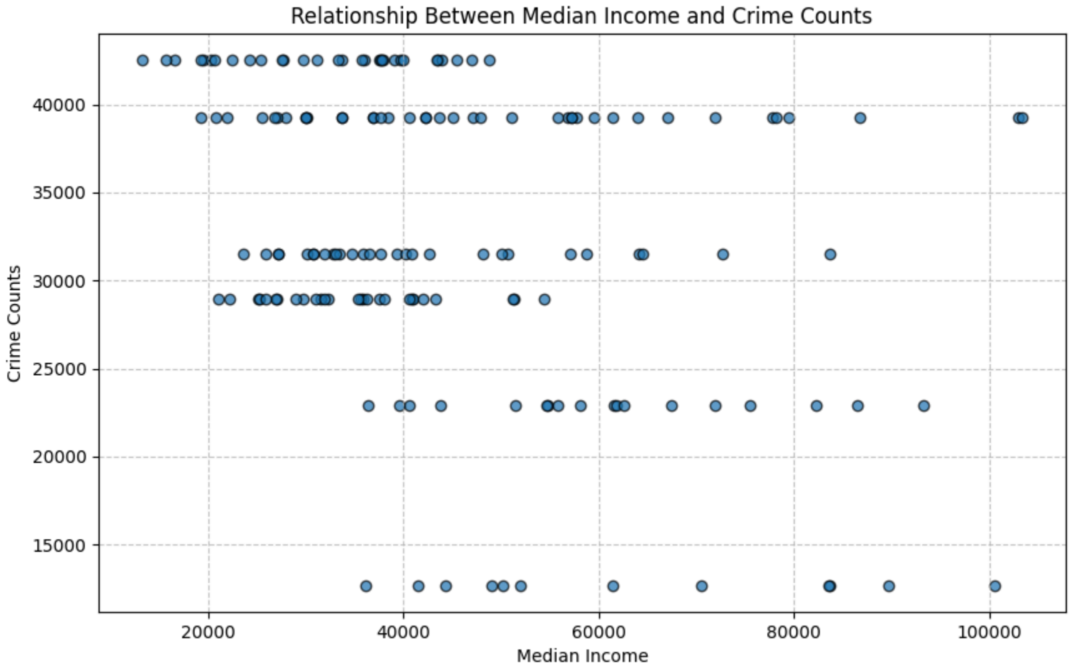
Neighborhood Income:

https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-income/about

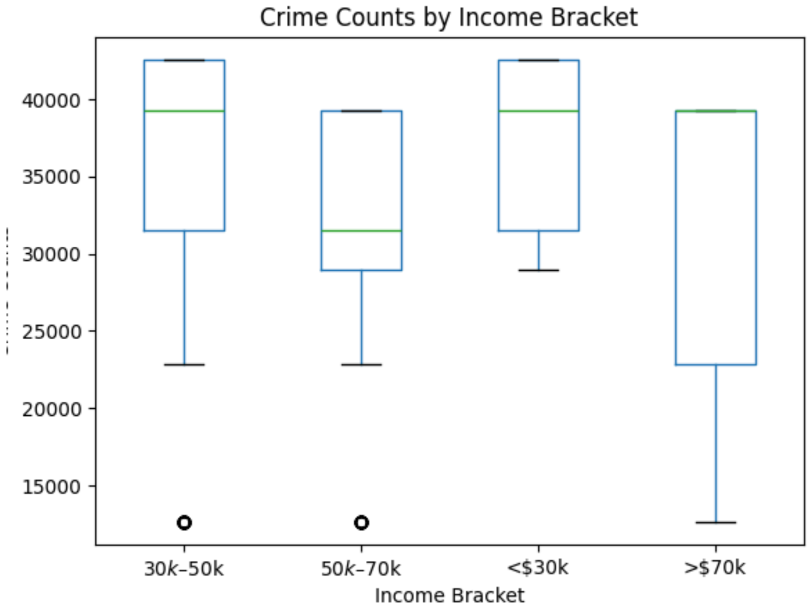
We will utilize common features of our datasets to compare, merge, and analyze our new datasets. We will accompany our analysis using visualizations such as scatter plots, 3D scatter plots, pie charts, and box plots to better our understanding of crime and income. To describe the relationship between income and crime, we will use Pearson's correlation coefficient and demonstrate it in our visuals. Furthermore, we will use techniques such as linear and logistic regression models to complete classification and prediction tasks, demonstrating the accuracy of our statistical models.

Results:

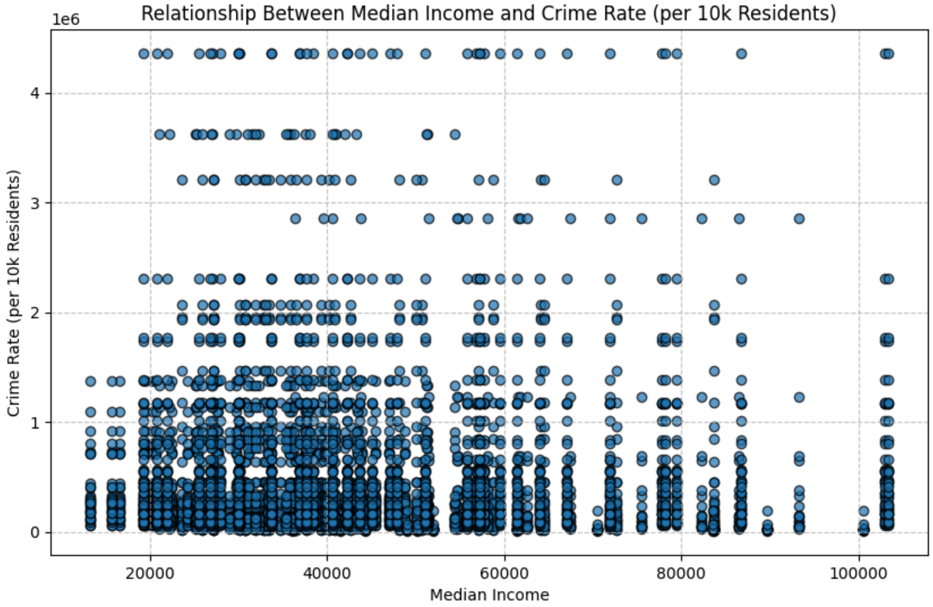
To address our first question, we examined the relationship between income levels and the frequency of crime through three visualizations: a scatter plot of median income vs. crime counts, a box plot of crime counts by income bracket, and a scatter plot of median income vs. crime rate per 10,000 residents. The visualizations highlight different aspects of the relationship, offering valuable insights into the dynamics of income and crime relation.



The scatter plot of median income vs. crime counts reveals no clear linear trend between income levels and total crime counts. The data points are widely scattered, with many wards showing similar crime counts regardless of their income levels. However, there are outliers where wards with very low or high income exhibit lower crime counts. This lack of a strong correlation suggests that income alone is not a significant determinant of total crime counts. Instead, other factors, such as population density or the types of crimes reported may be more substantial in influencing crime levels.

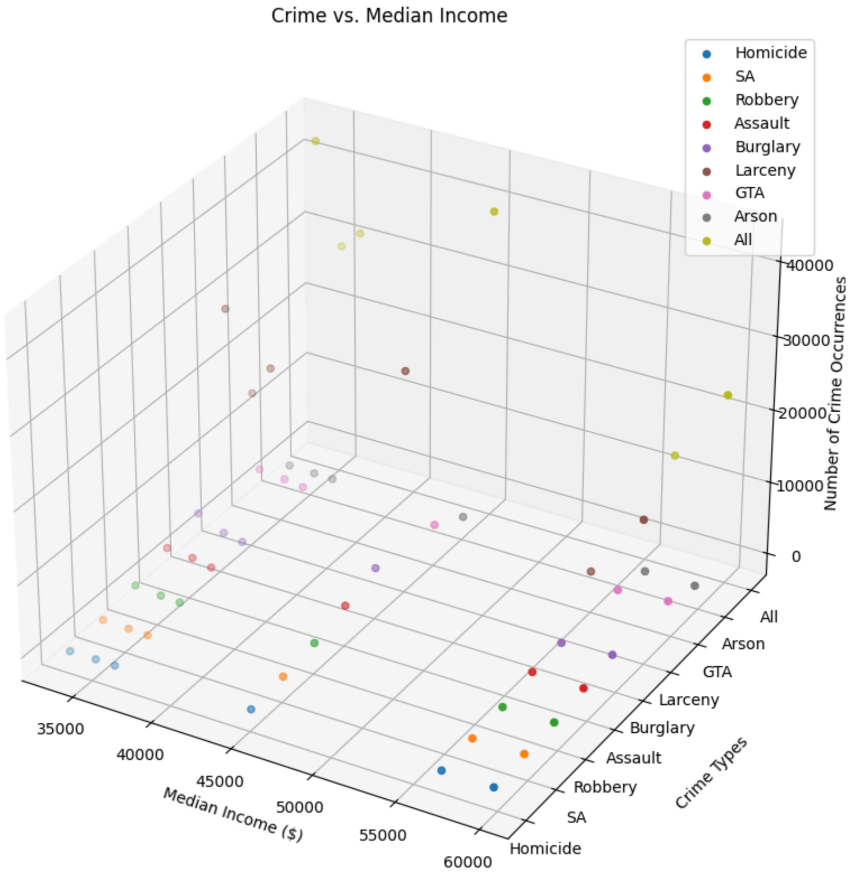


Secondly, the box plot of crime counts by income bracket provides a closer look at the distribution of crime counts within different income categories (Y-axis is crime count). The medians of crime counts are fairly consistent across all brackets, suggesting that income brackets do not drastically differentiate crime levels. However, lower-income brackets (<$30k) have a wider variability with some wards having both very high and very low crime counts. Higher-income brackets (>70k) still show notable variability. The presence of outliers indicates that some wards deviate significantly from the general trend. This variability within brackets highlights that local factors may have a significant impact on crime frequencies.

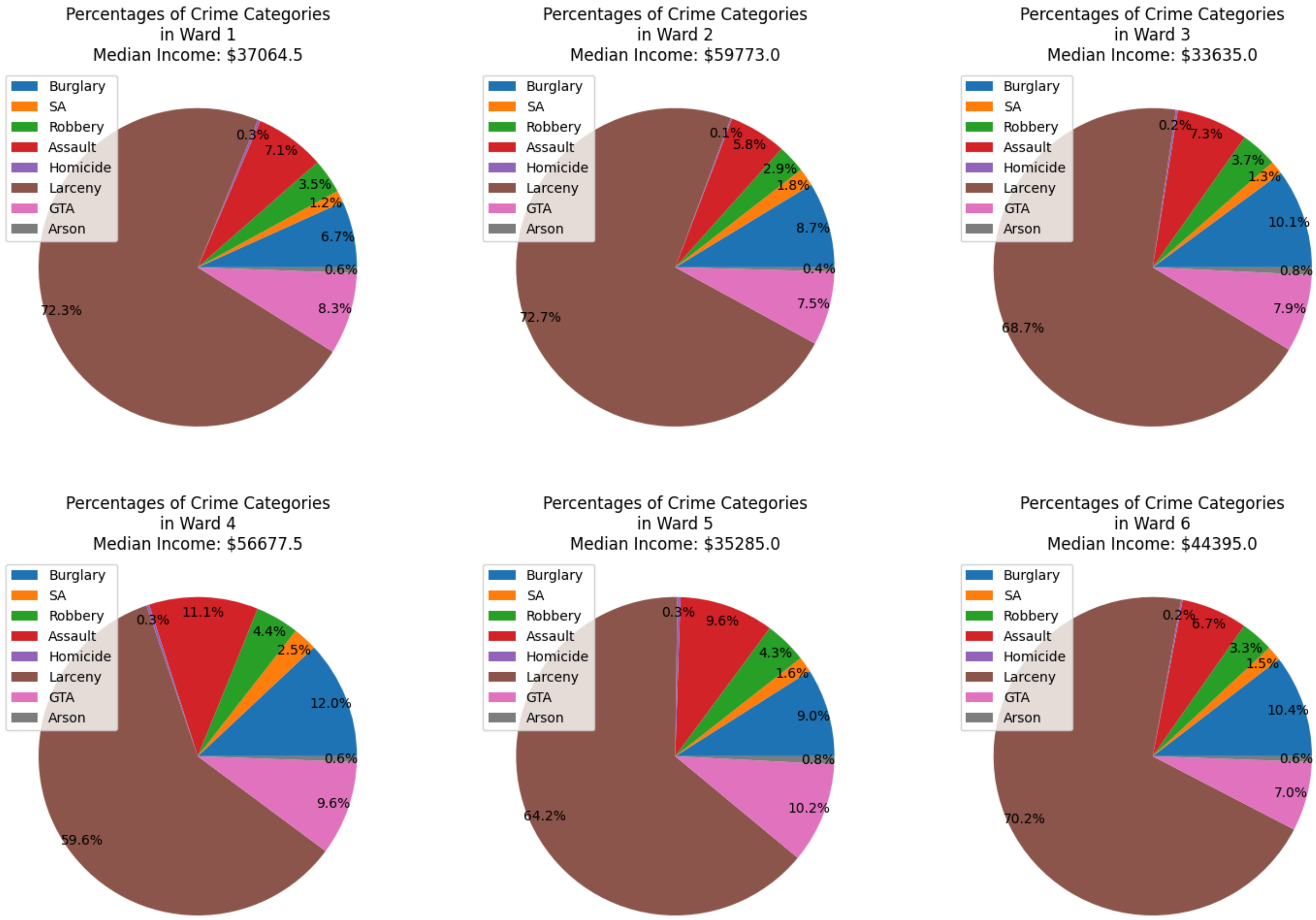


This scatter plot of median income vs. crime rate per 10,000 residents normalizes crime counts by population size showing a clearer measure of crime frequency relative to ward size. This visualization shows a wide distribution of crime rates at all income levels but wards with lower incomes (<$40k) tend to exhibit higher crime rates more frequently. Additionally, some higher-income wards (>$70k) show lower crime rates indicating potential socioeconomic factors that decrease crime in these areas. By normalizing the data it becomes apparent that lower-income wards face disproportionately higher crime rates, showing a potential disparity in crime prevalence linked to income levels.

To address the second question, we first created a 3D scatterplot to display median incomes of each ward from the dataset in relation to the number of reported incidents of each category of crime. While it is immediately apparent that the wards with lower median incomes (>$40,000) have higher overall crime counts, there does not appear to be a significant difference in the most popular types of crimes committed between the wards. Larceny is the most common crime regardless of income level, followed by a combination of Burglary, Assault, and GTA. This supports the hypothesis that income level has an impact on crime rate, but provides no evidence towards income level's role in the types of crime committed.



Next, to ignore the overall magnitude of the crime counts and instead view them in terms of their prevalence compared to other categories within each ward, we created pie charts for each ward that display the percentages of each category of crime. Generally, there are not major differences in the percentages of each crime category between the wards. The greatest difference appears in Ward 4, which has a slightly lower Larceny rate (59.6% compared to most of the other wards averaging about 70%) and a slightly higher Assault rate (11.1% compared the others' approximate 6-7%). Its Burglary rate is also the highest at 12.0%, though this may be insignificant because the Burglary rate varies slightly more between the wards, often by a few percent and without a discernible pattern in relation to the median incomes. While these differences might have been of interest since Ward 4 has the second highest median income of $56,677.50, Ward 5, the ward that exhibits the second greatest differences in crime makeup and that also differs from the other wards' percentages in a similar way as Ward 4 (Ward 5's notable percentages are Assault at 9.6% and Larceny at 64.2%), has a median income of $35,285.00 which is the second lowest of the wards. This seems to discredit the notion that Ward 4's different crime makeup is due to its higher median income. Overall, we were not able to find evidence that income level has any meaningful bearing on the types of crime being committed. Other factors such as population density and the number of perpetrators committing crimes outside of the ward they reside in may also be useful to consider when studying this relationship.



Percentages of Crime Categories in Ward 1 Median Income: $37064.5

Percentages of Crime Categories in Ward 2 Median Income: $59773.0

Percentages of Crime Categories in Ward 3 Median Income: $33635.0

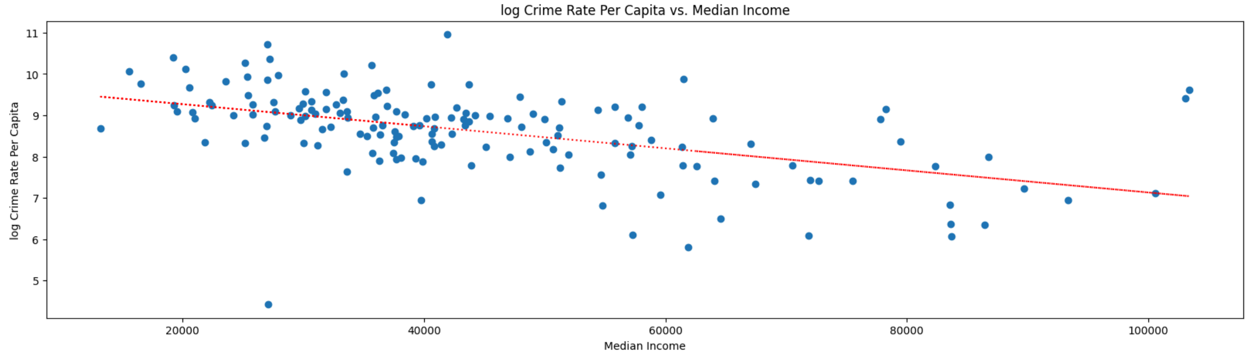
Percentages of Crime Categories in Ward 4 Median Income: $56677.5

Percentages of Crime Categories in Ward 5 Median Income: $35285.0

Percentages of Crime Categories in Ward 6 Median Income: $44395.0

To determine if we could use statistical modeling to predict the rate of crime based on median income, we decided to utilize the data frame of Tucson Police Arrests, Tucson Neighborhood Population, and Neighborhood Income. We merged our datasets into a singular data frame and plotted the relevant data points into a scatter plot for a visualization of our model, linear regression and logistic regression. We omitted outliers such as neighborhoods with no data on its population (Sombras Del Cerro), data that had no crime recorded (Catalina Vista/Blenman-Elm) and data that had lower population but higher crime frequency

(Downtown). We then normalized the crime rate per capita and recorded each plot point to a scatter plot, hence, our resulting visual.

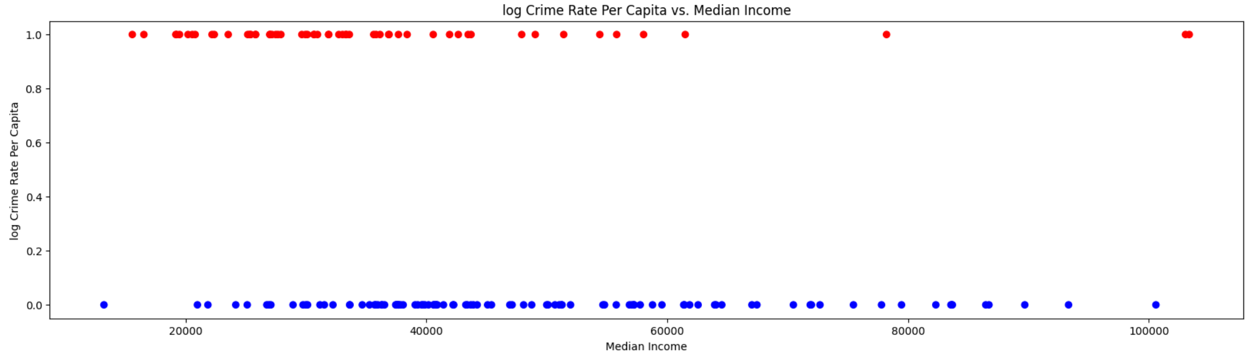


Using Pearson’s correlation, we found that the relationship between crime rate and median income has a negative correlation of .51, indicating that a low median income would generally have a high crime rate and vice versa. Using the linear regression model to determine predictability, we find that our mean-squared-error (mse) is 0.43 and our r-squared value is 0.35. In other words, our model is relatively fit as our mse is considerably low, and our accuracy of our predictions is 35%. The low r-squared value implies that the model does not account for other factors such as surveillance, social safety nets, or other economic programs.

Linear Regression model and correlation:

Correlation:	-0.506569321617064
Mean Squared Error:	0.43723435154843954
R-squared:	0.35243619597556397

For classification models and utilizing it for predictability, we created two classes known as high crime rate and low crime rate. We would categorize a crime rate if the crime rate itself was above or below the crime rate mean, which we found was 9.02.



In terms of accuracy, we found that our accuracy score was .78, which means the logistic regression model correctly predicts our outcomes 78% of the time. Meanwhile, our precision is 0.7, indicating that our model predicts our positive cases 70% of the time. On the other hand, our recall is relatively lower compared to our other scores, which lands around .64 (64%). While our model adequately identifies high crime rates, there does seem to be some lacking data/information. We can reasonably assume, like the linear regression model, that this model still does not consider other variables that either prevent crime or deter it. Regardless, the logistic regression model performs significantly better than the linear regression model.

Logistic regression and Crime Rate Mean:

log(crime_rate): 9.021566809191127

Accuracy: 0.78125

Precision: 0.7

Recall: 0.6363636363636364

Conclusion:

Our analysis consisted of trying to answer three questions, what kind of relationship does income level and frequency of crime pertain to, how does income level affect what kind of crime is committed, and can we use statistical models to predict crime rate using the median household income? Firstly, when looking at the relationship between income level and crime frequency, we identified that there was no clear linear trend between income levels and total crime counts but income brackets do not drastically differentiate crime levels with a few notable outliers. Most notably however, the normalization of crime rates reveals that lower-income wards tend to experience higher crime rates relative to their populations suggesting that socioeconomic disparities may contribute to variations in crime prevalence.

We were not able to find a meaningful connection between income level and the types of crime being committed. While lower-income wards had higher overall crime levels, the most common crimes stayed the same between different median income levels. The biggest variances in crime makeup took the form of the same crimes but appeared in a high-income and low-income ward respectively.

In terms of utilizing statistical models to determine if we could predict a crime rate based on the household median income, we found that while both linear and logistic models were adequate, the logistic regression model had higher accuracy and precision than the linear model. On the other hand, both models were still missing some variance and data, which we can assume was most likely due to additional factors such as economic programs, level of surveillance, or even social factors that deter crime. Despite lacking in some areas however, the logistic regression model can still be used to determine areas with relatively low crime rate and high crime rate with a 78% accuracy. Future studies could incorporate additional variables, such as education, unemployment, and housing conditions, to better understand the underlying causes of crime. Initiatives to improve economic opportunities, community engagement, and access to education in these areas may help mitigate crime rates and address socioeconomic disparities.

References:

Kang, S. (2016). Inequality and crime revisited: effects of local inequality and economic segregation on crime. *Journal of Population Economics*, *29*(2), 593–626. http://www.jstor.org/stable/44280406

Neighborhood Income. *City of Tucson Data Hub*, City of Tucson, https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-income/about. Accessed 9 Dec. 2024.

Patterson, E.B. (1991), "POVERTY, INCOME INEQUALITY, AND COMMUNITY CRIME RATES". Criminology, 29: 755-776. https://doi-org.ezproxy3.library.arizona.edu/10.1111/j.1745-9125.1991.tb01087.x

Police Arrests 2021. *City of Tucson Data Hub*, City of Tucson, https://gisdata.tucsonaz.gov/datasets/tucson-police-arrests-2021-open-data/about. Accessed 9 Dec. 2024.

Tucson Neighborhood Population. *City of Tucson Data Hub*, City of Tucson, https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-population-statistics/about. Accessed 9 Dec. 2024.

Tucson Police Reported Crimes. *City of Tucson Data Hub*, City of Tucson, https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-reported-crimes/about. Accessed 9 Dec. 2024.