IDS 570 (Fall 2017)

**ANALYSIS OF CRIME RATE IN NORTH CAROLINA**

By Dubish, Jeffrey

Dykstra, Jonathan

Mehta, Deep

Pasad, Hetvi

Sharma, Neha

Sharma, Priyanka

**Introduction**

This research project was commissioned by the consulting group, Consulting Group Room 208, for the state of North Carolina in order for its policy makers and the state’s business managers to make better informed decisions regarding crime rate and the factors that influence it.

Crime rate has a wide effect on social issues. It affects a community’s real estate values, education levels, wealth, poverty rates, upward mobility, economic opportunities, and overall quality of life. On a state level, lower crime rates lead to less money spent on the justice system, leading to smaller or better spent government budgets. We chose to analyze North Carolina’s crime rate because we believe the state is a strong cross section of the average United State. North Carolina has approximately 10.1 million people. There are two major cities, Raleigh and Charlotte, as well as suburban, rural, mountainous, and coastal communities each with their respective and different economies. The State’s main industries are its energy sector, textiles, livestock, and tobacco. The median income is $50,800, which places it at the top of the bottom third of the States and more than 13% of households are below the poverty line. However, its economy has grown 17.9% in the last two decades and Forbes labels North Carolina as “the 3rd best state for business.” Clearly, the state has forward momentum, which makes it a great case to study and develop useful hypotheses from.

We have been tasked to identify the factors that most influence crime rate and develop a model that can predict the crime rate. We hope to be able to identify the most influential aspects of crime as a way to keep the citizens safe without excessive and unnecessary resources. Based on the background population information and our own intuition, we have identified three main areas of focus:  Police per capita, population density, and income levels.

Specifically, we expect an inverse relationship between the variables of crime rate and police per capita (**Hypothesis 1**). Identifying a relationship between these two variables will allow us to inform the state and local governments where to staff more (or fewer) police. We also expect a positive relationship between crime rate and population density (**Hypothesis 2**). This hypothesis is largely based on the intuition that citizens tend to become unrulier when they are constantly in close quarters. Thus, being able to identify the amount of influence this has on our dependent variable will help us inform the governments with their residential zoning and planning. Finally, we expect an inverse relationship between crime rate and income levels (**Hypothesis 3**). Identifying a relationship between crime rate and average weekly wages would help inform the necessity (or lack thereof) of economic stimulus programs that can assist lower wage groups.

**Data**

Our team was able to secure a suitable dataset so that we could attempt to answer these hypotheses. We selected this dataset because it contained the crime rates, police per capita, and weekly wages. It also had a couple factors that we could then use to subset the data to better identify correlations. This gave us the advantage of being able to target other factors that may be influencers to the dependent variable. For example, the dataset included county, region, wage by industry, as well as arrest and conviction probabilities. The dataset chosen utilized crime rate data from the FBI and wage data from the Bureau of Labor & Statistics. Both of these sources promise high quality data and have data teams that test and groom the data for quality assurance.

While our dataset is rather robust, it does have a few limitations. Firstly, the data obtained is from the 1980’s. It may be a difficult assumption that some of the data have held constant since then. However, for the sake of this analysis, we believe that any observations discovered would still be valuable in predicting an outcome. Further, it will be useful for future back testing that can help data scientists develop more predictive models. Another limitation is the number of years of records. Attempting to aggregate data to spot any yearly trends will limit our output to seven data points. However, as the data is aggregated yearly, we will not have any noise from seasonal crime rate trends.

**Detailed description of data**

The dataset follows statistics for 90 North Carolina Counties over seven years in 22 variables; resulting in 630 total observations. The variables are a mix of numeric and character type sets. The character types will be used as factors for this analysis. There are also numeric ranges within the dataset that can be set to factor ranges.

The dataset was secured on October 22, 2017:

https://vincentarelbundock.github.io/Rdatasets/datasets.html

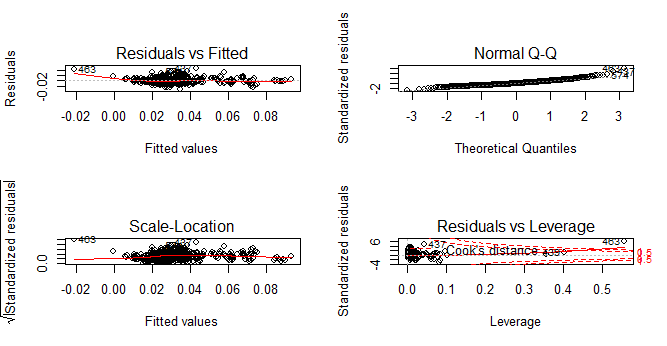
The specific dataset can be found at:

<https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/Ecdat/Crime.csv>

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| county | Factor | County Identifier |
| year | Factor | Year from 1981 to 1987 |
| Crmrte | Numeric | Crime rate |
| Prbarr | Numeric | Probability of an arrest |
| Prbconv | Numeric | Probability of conviction |
| Prbpris | Numeric | Probability of prison sentence |
| Avgsen | Numeric | Average Sentence, days |
| Polpc | Numeric | Police per Capita |
| Density | Numeric | Hundreds of people per kilometer |
| Taxpc | Numeric | Tax per Capita |
| Region | Factor | One of ‘other’,’west’,’central’ |
| Smsa | Factor | ‘Yes’, if it is a State Metropolitan |
| Pctmin | Numeric | Percentage of Minority |
| Wcon | Numeric | Weekly wage in Construction |
| pctymle | Numeric | Percentage of Young Males |
| Wtuc | Numeric | Weekly wage in Transportation, Utilities and Communication |
| Wtrd | Numeric | Weekly wage in whole sales and retail trade |
| wfir | Numeric | Weekly wage in finance, real estate and insurance |
| Wser | Numeric | Weekly wage in service industry |
| Wmfg | Numeric | Weekly wage in manufacturing |
| Wfed | Numeric | Weekly wage of federal employees |
| wsta | Numeric | Weekly wage of state employees |
| wloc | Numeric | Weekly wage of local government employees |
| mix | Numeric | Offence mix: face to face/other |

**Data Cleaning**

First, we removed irrelevant variable which redundantly indicated serial numbers. We then merged an additional column matching County Names corresponding to the respective counties by their codes. Next, per the graphs below, we removed 11 outliers, 6 of which belonged to Pender County which had an abnormality in crime rate where it increased three-fold between four years before dropping three-fold again. Moreover, the police per Capita dropped approximately 90% over subsequent years. We assumed either there was an error in data recording or the county experience such extremes in changes that it would warp our data.



**Research Question**

What factors impact crime in North Carolina over time?

**Purpose**

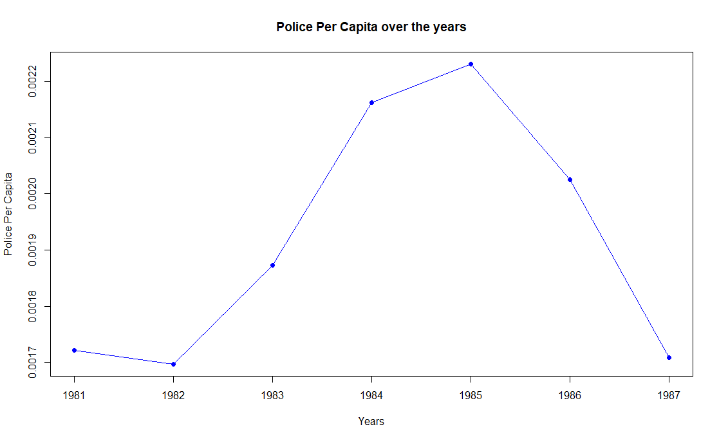
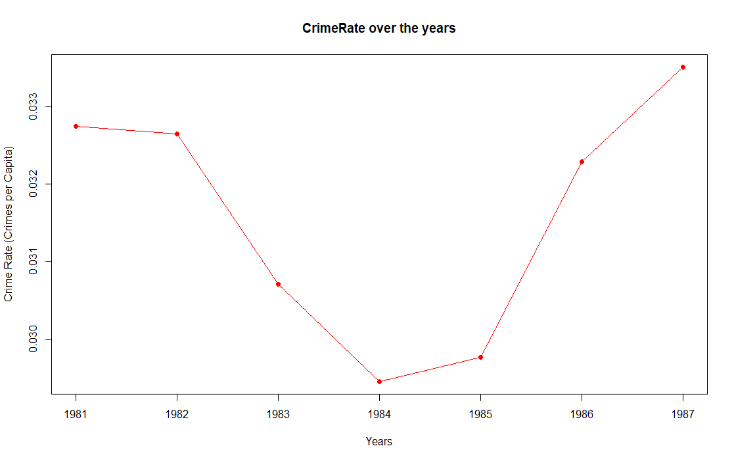
The purpose is to understand which factors most predict the crime rate. Understanding what impacts crime rate helps governments (federal, state, and local) determine where to allocate police assets, economic assistance programs, and city planning. It also helps businesses plan expansion and contributes to their financial modelling.

**Analysis and Results**

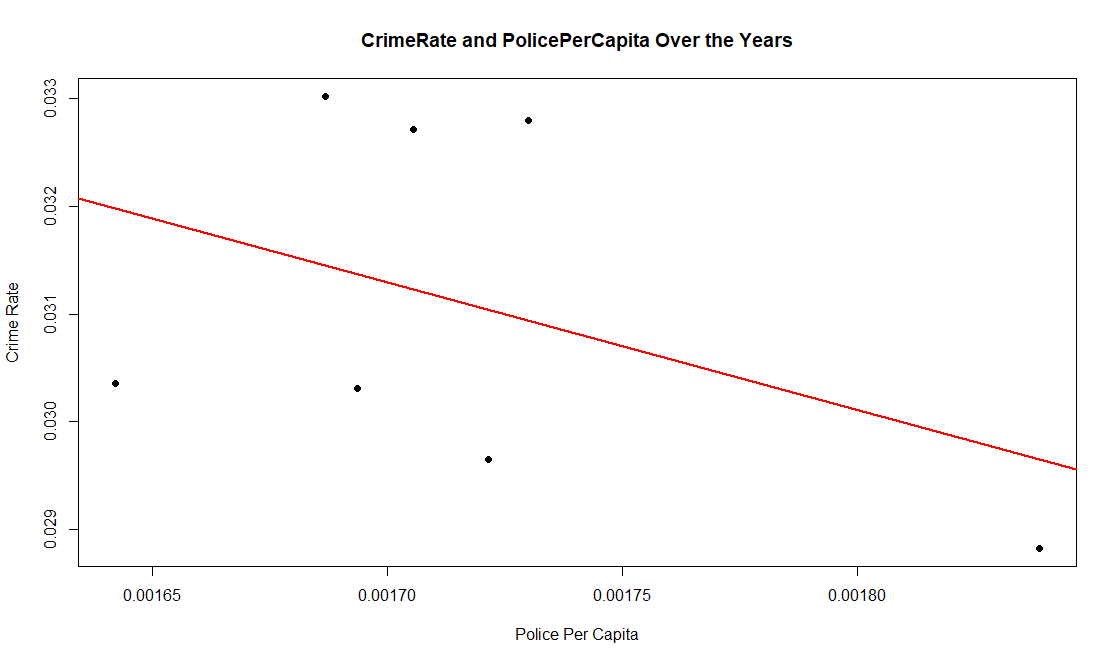
**HYPOTHESIS 1: Crime Rate has an inverse correlation with Police Per Capita**

**Methodology**:

We performed a preliminary bivariate analysis of police per capita and crime Rate that return somewhat ambiguous results. We clearly observed at least two trends between them but it was difficult to determine the linear trend while evaluating the data as a whole. We decided to individually trend police per capita and crime rate annually. An inverse relationship was visibly apparent. The graphs below clearly show the two variables move counter to each other.



During a bivariate analysis where the data points over the years were plotted together, these opposite trends nullified each other. Graphically, this produced a straight trendline showing no correlation. Therefore, we decided to aggregate police per capita by year and crime rate by year so that they could be properly plotted against each other, as seen in the graph below. To better understand the trends at a granular level, we broke Crime Rate and Police Per Capita down by region (west, central, other).



**T-test for Crime Rate and Police per Capita over the years:**

P-value at 99% Confidence Interval = 0.000000007381

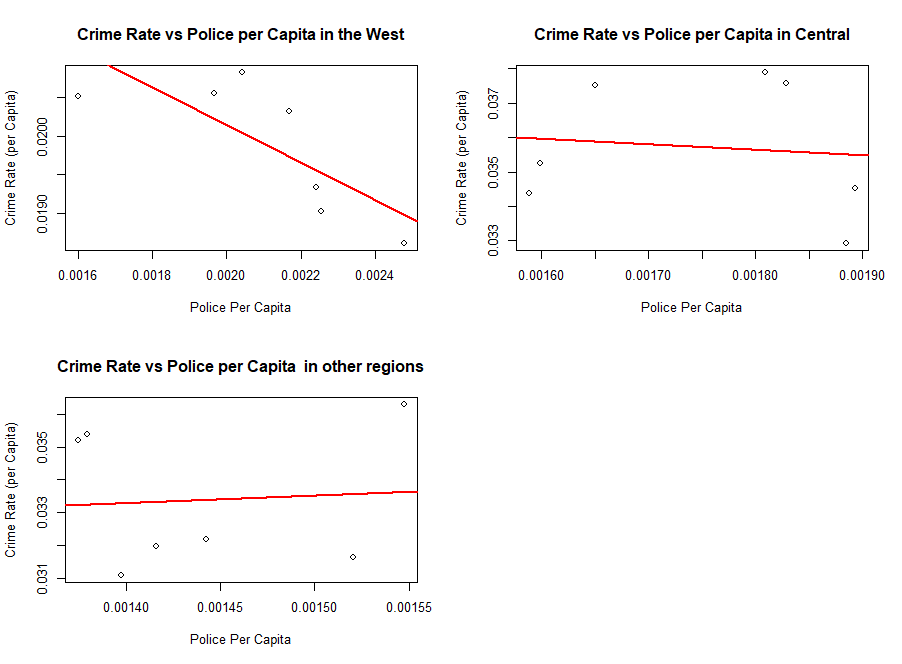
Since the p-value is less than 0.01, we reject the null hypothesis and accept the alternate hypothesis at a 99% confidence interval. We conclude that there is a relationship between Crime Rate and Police Per Capita.

**Correlation Test:**

From a linear correlation test, we found the correlation value is -0.42. Our suspicions were correct, there is a strong inverse correlation between Crime Rate and Police Per Capita.

**Correlation Tests broken down by region:**

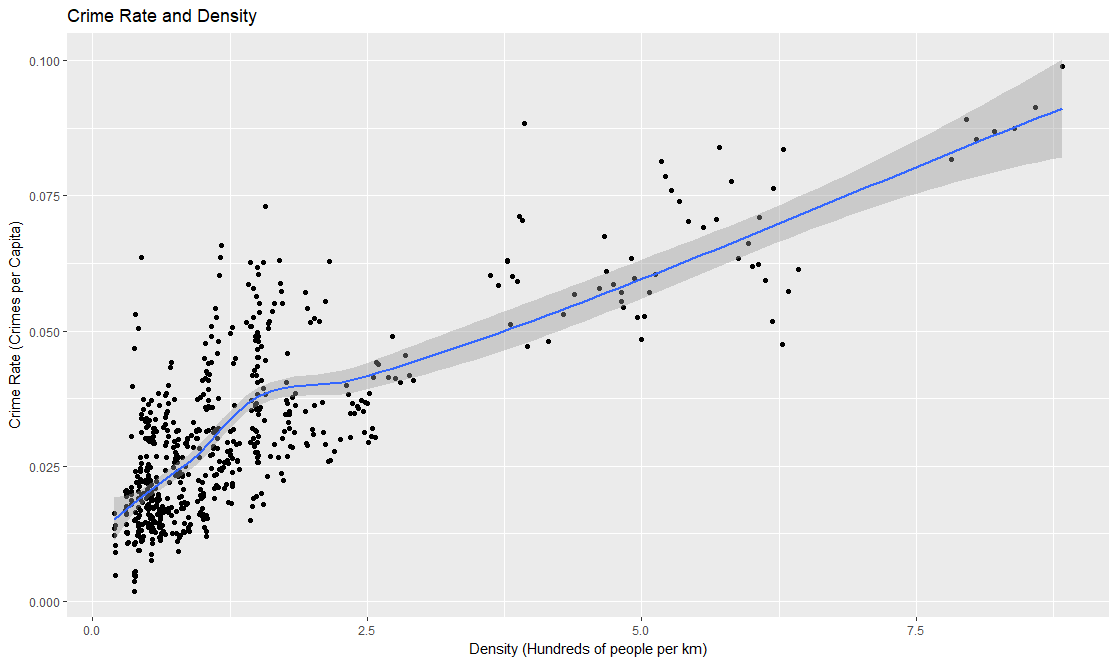
When breaking the crime rate vs. police per capita down by region we found a correlation value of -0.77 in the West; a correlation value of -0.10 in the Central region; and a correlation value of 0.07 in the counties of the Other region. From this, we concluded that increasing the number of police per capita as a method to reduce crime is most effective in the West region and that in the counties included in the Other region, there are likely more effective ways to combat crime than increasing police per capita.



**HYPOTHESIS 2:** **Crime Rate and Density:**

**Methodology:**

Univariate analysis of density shows that the data is normally distributed. Bivariate analysis of Crime Rate and density across a scatterplot (pictured below) graphically showed an expected strong positive linear correlation, as we expected. To confirm the relationship, we performed a t - test.



**T-test**:

P-value at 99% Confidence Interval = 0.00000000000000022

Since the p-value is less than 0.01, we reject the null and accept the alternate hypothesis at a 99% confidence interval. This confirms a relationship between crime Rate and density.

**Correlation Test**:

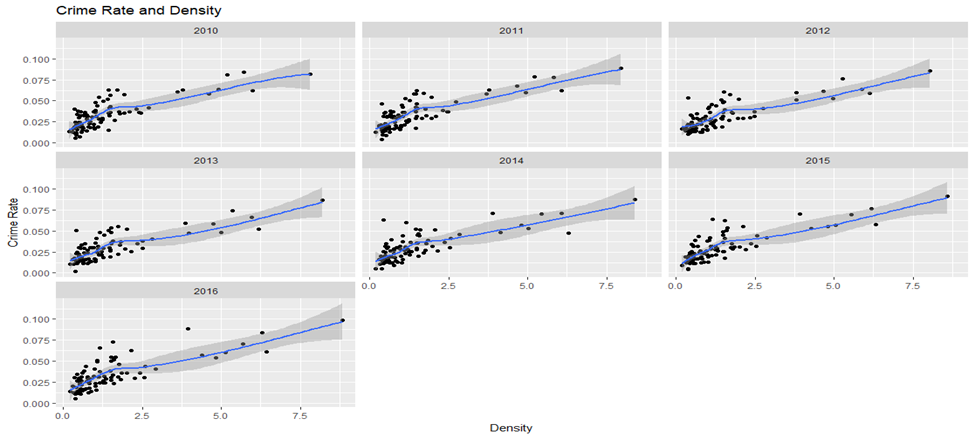
After a linear correlation test, we found the correlation value is 0.75; there is a strong positive correlation between Crime Rate and density.

**Linear Regression Modeling**:

Since the coefficient value of density is 0.008, we conclude that, when keeping all the other variables constant, an increase of 1 person per sq. mile results in the crime rate increasing by 0.008.

**Crime Rate and Density over the years:**

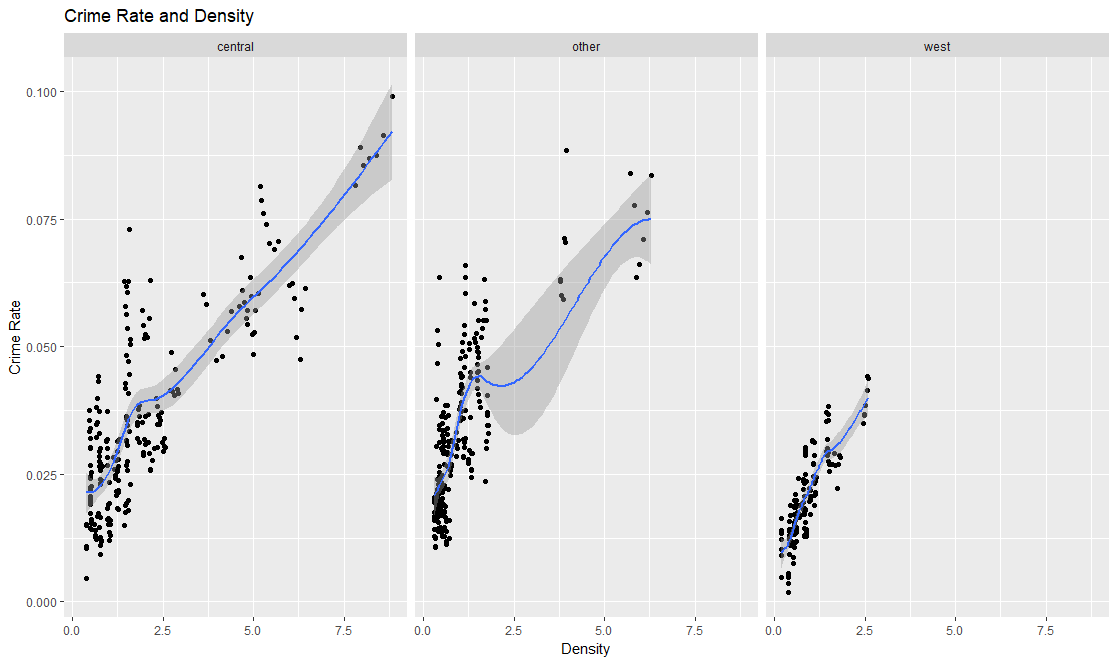
To ensure that density was not affected by year over year changes, we conducted further bivariate analysis for each year. We concluded that there was not enough significant change in density year over year as you can see from the constant trends below.



**Crime Rate and Region, Crime Rate and SMSA:**

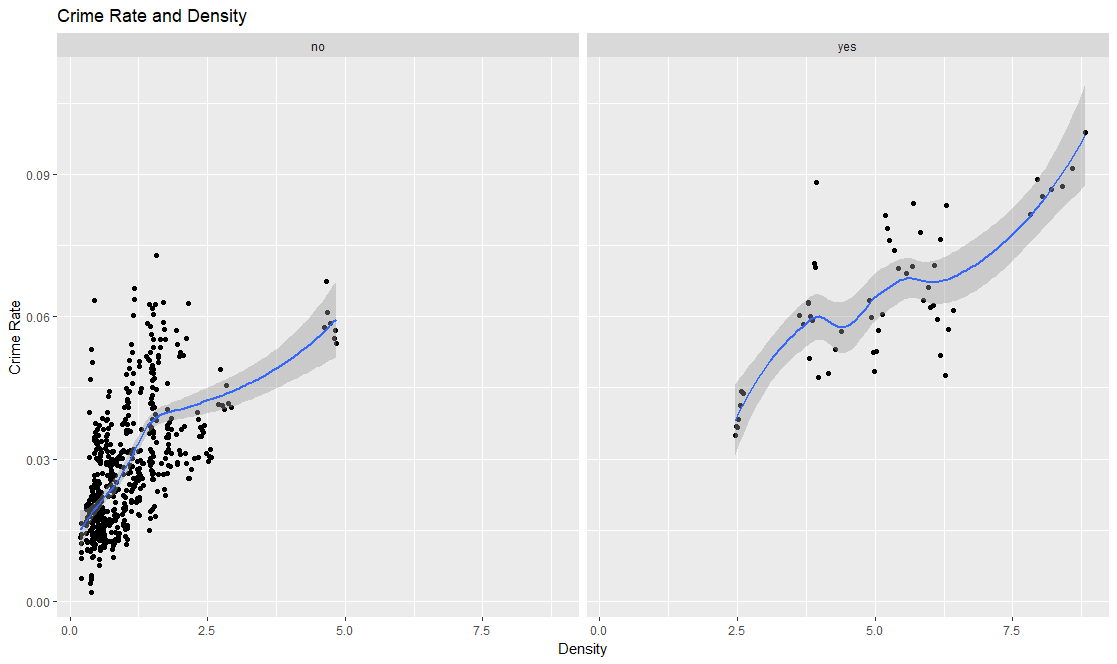
Further, because we saw that region played a factor in the correlation between crime rate and police per capita we decided to break down density by region and SMSA (whether a county is a metropolitan or not). Though the West region has the least dispersion of densities, we concluded that their respective correlations are close enough in relation to each other not to warrant any further investigation.

**Crime Rate vs. Density by Region**



We then further examined crime rate by metropolitan county or ‘not’ a metropolitan county and found substantiating evidence that crime rate is correlated with counties being a metropolitan, as we expected to find given they have higher population densities than non-metropolitan counties.

**Crime Rate vs. Density by SMSA**



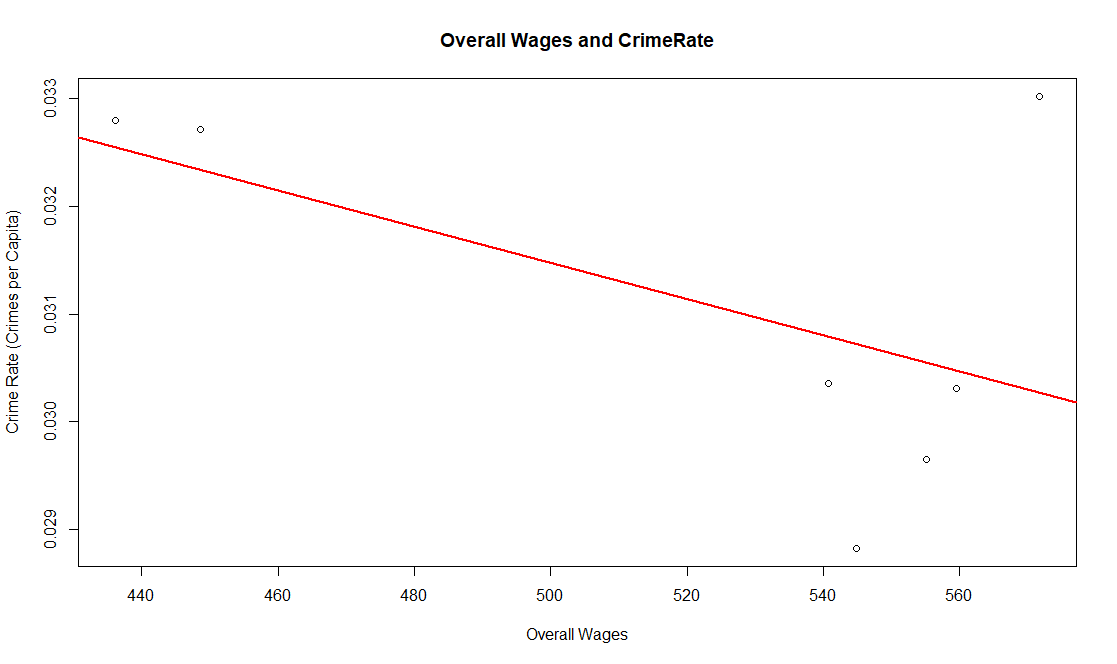
**HYPOTHESIS 3:** **Crime Rate has an inverse correlation with wages:**

**Methodology:**

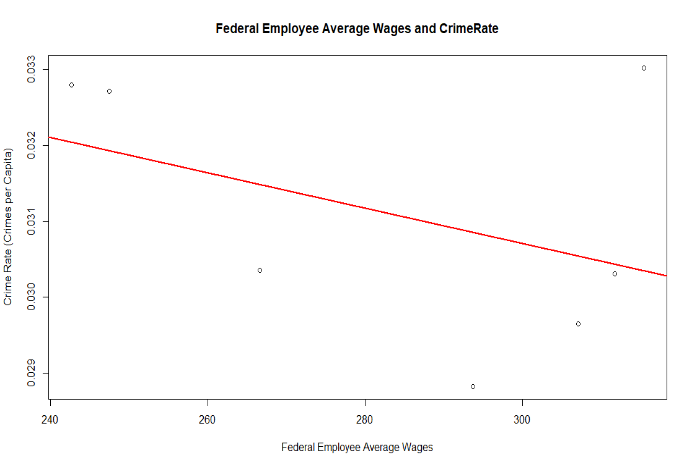
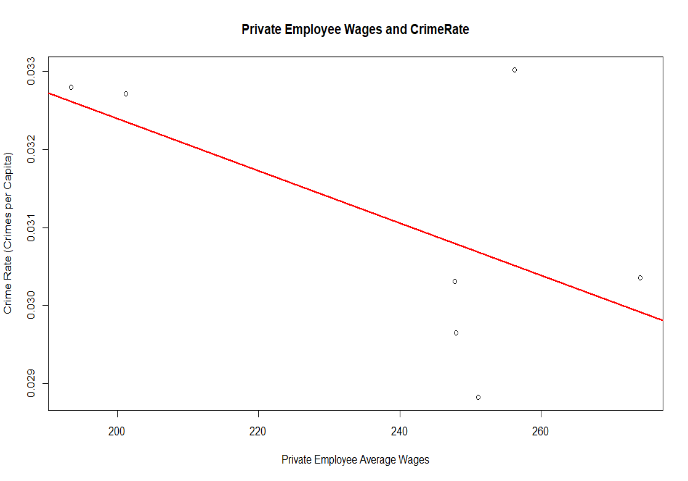
First, we ran correlation tests for each individual wage. Since there are 9 different types of wages in our dataset, we discovered that no specific wage group was a good predictor of crime rate overall. To better analyze wages, we concentrated them into two categorical groups government and private wages. The government wages include weekly wage of federal, state, and local government employees. All other wages were included in private wages. Furthermore, for every county the average wage across each category has been computed each year in order to investigate year over year changes.

For fair comparison of the relationship of wages with crime rate, the wages across years 1982 to 1987 have been standardized to real 1981 wages using the Consumer Pricing Index (1981 is the base year for which wages across remaining 6 years have been computed) to minimize the effect the inflation. We have named these columns as ‘rprivwageavg’ and ‘rgovwageavg’ for private and federal respectively.

Upon examining the aggregate data of all wages and crime rate for 7 years, we see that there is a negative relationship between them. There are 7 data points on the graph, one for each year.



Examining the relationship of crime rate with private wages and government wages separately, we find that there is a stronger negative correlation between private wages and crime rate than government wages and crime rate, as seem in the comparison below.



**T-test:**

P-value at 99% Confidence Interval = 000000005436

Since the p-value is less than 0.01, we rejected the null hypothesis and accept the alternate hypothesis at 99% Confidence Interval. We confirmed that there is a relationship between crime rate and wages.

***Correlation Test*:**

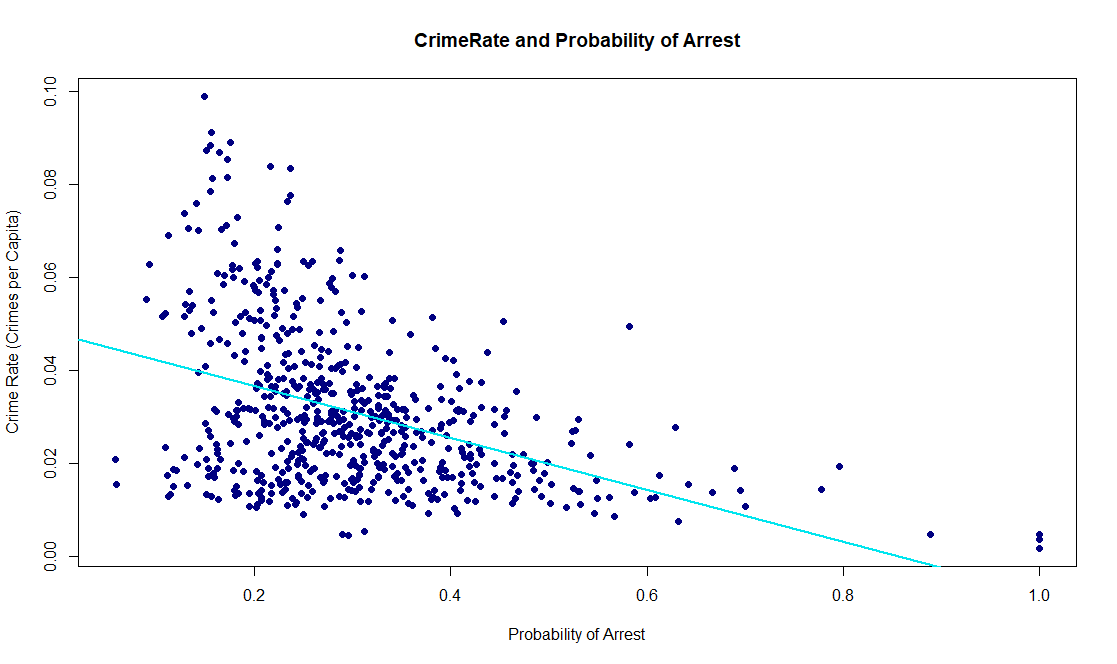
From a correlation test, we found that the correlation value of -0.54, a strong negative correlation between crime rate and wages. Holding constant for private wages, we found a correlation value of -0.58; and holding constant for government wages, we found a correlation value of -0.42.

Private and Federal wages are not significant in the model.

**OTHER OBSERVATIONS:**

***Crime Rate and Probability of Arrest:***

Crime Rate and Probability of arrest have an expected negative relationship.



**T-test:**

P-value at 99% Confidence Interval = 0.00000000000000022

Since the p-value is less than 0.01, we reject the null hypothesis and accept the alternate hypothesis at a 99% confidence interval. There is a relationship between crime rate and probability of arrest.

**Correlation Test**:

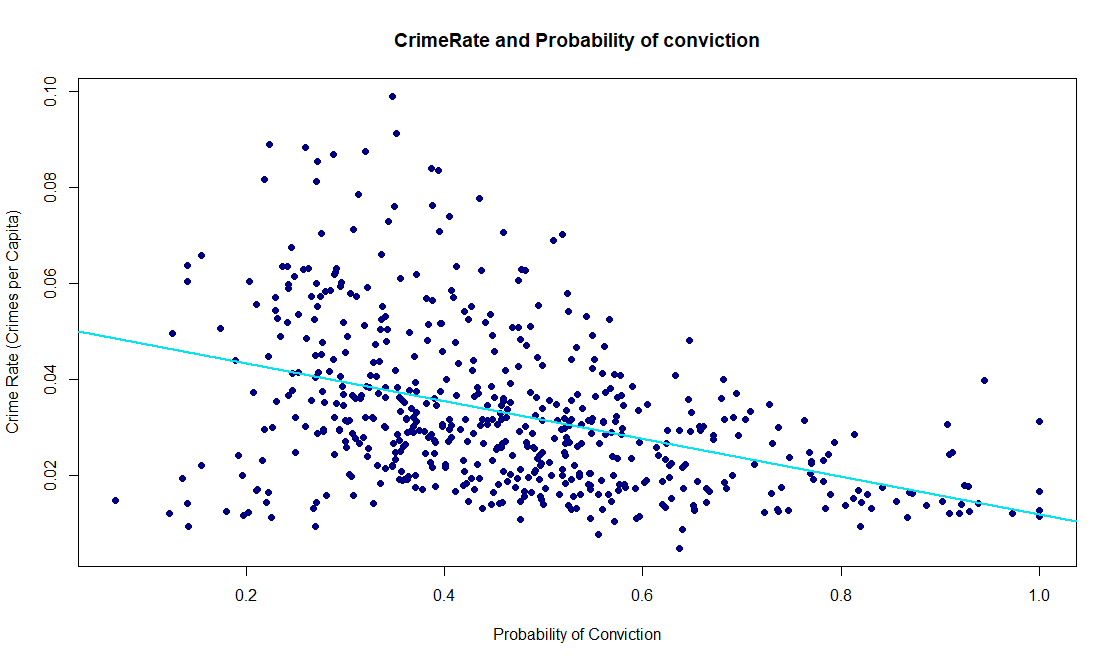
From of correlation test, we found a correlation value of -0.41. There is a negative correlation between crime rate and probability of arrest

**Linear Regression Modeling:**

Since the coefficient value of probability of arrest is -0.03, we can conclude that by keeping all the other variables constant, an increase of 1 unit of probability of arrest, there is a decrease of Crime Rate by 0.03.

***Crime Rate and Probability of Conviction:***

Similarly, we found that crime rate and probability of conviction have a negative relationship.



***T-test:***

P-value at 99% Confidence Interval = 0.00000000000000022

Since the p-value is less than 0.01, we reject the null hypothesis and accept the alternate hypothesis at a 99% confidence interval. We are confident there is a relationship between crime rate and probability of conviction.

**Correlation Test:**

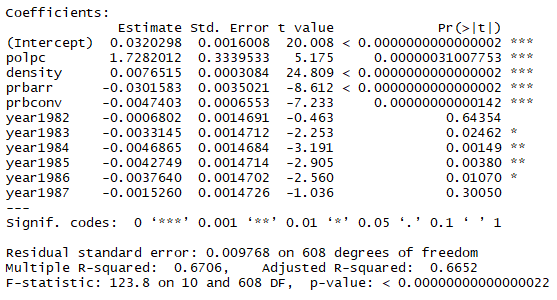
From a correlation test, we found that the correlation value of -0.41. There is a negative correlation between crime rate and probability of conviction.

**Linear Regression Modeling**:

Since the coefficient value of probability of arrest is -0.004, we concluded that when keeping all the other variables constant, an increase of 1 unit of probability of conviction, there is a decrease of crime rate by 0.004.

**Modelling**

Finally, we developed a model which to predict a county’s crime rate given the data types. The statistically significant variables are as follows: police per capita (polpc), density, the probability of arrest (prbarr), and the probability of a conviction (prconv). The table below includes individual years to show that 1982,1983,1986,1987 were not as statistically significant to our model at 99% confidence interval as 1984 and 1985. The adjusted R-squared value for our final model is 0.66. This means that the model can predict 66% of a county’s crime rate given the variables we chose.



**Conclusion**

Each of our hypotheses were observed to reject the null with 99% confidence. From this, we believe our model is providing meaningful insight to Consulting Group Room 208 and will translate well to helping local and state governments best address the common issue to reduce crime rates. It can also be confidently used to estimate future crime rates given newer data that is useful for business development planning.

To best reduce crime rates, government officials should be mindful to provide a larger police presence, remembering which regions that decision framework works strongest in. Further, we have shown that simply adding police is not enough. Police departments can save on man power by empowering their police to be more effective as there is a stronger relationship with the probability of arrest. Additionally, town planning should provide adequate levels of residential zoning space to prevent excessive population density.

Earnings/wages provide a smaller relation to crime rates. This suggests that economic incentive or stimulus programs may have a less measurable impact on crime rate, but because wages negatively correlate with crime rate, there is still a benefit to business and government paying higher wages within the counties. As tax rate was not a strong predictor in our model, programs to bring in business, such as lowering business taxes, may be more beneficial that economic stimulus.

Overall, there is still much work to be done. Measurement tools since the beginning days of sociology and crime have changed wildly since the early days of criminology and sociology 1920s at the University of Chicago. The advent of computers allowed data science to merge with sociology in a way that allow sociologists to quantitatively measure their theories. Economists and financial planners use sociological data to plan future development and to target audiences. We have provided a quantitative data analysis that is useful to decision makers at the nexus of these fields. We are confident that we are providing a tool that is useful for professional sociologists, economists, and business decision makes to use to estimate the impact of government policy and changing community metrics.