Data Analytics Project

27 October 2018

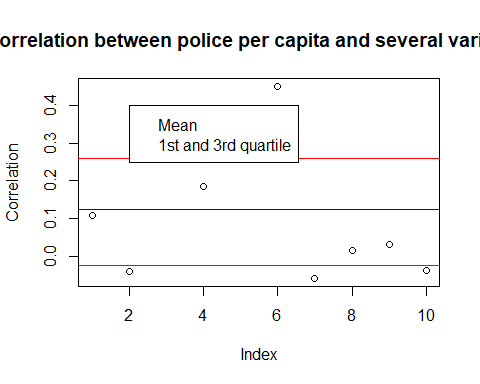
Introduction

In this project we(Nathan Brunel, Ronan Doran, John De Guzman) intend to analyse and extract appropriate information to either support or disprove our listed hypotheses. The statistics we selected are variables surrounding crime in North Carolina. We chose this data as we can extrapolate interesting points and correlations from it. It is our goal to clearly convey which variables relate to crime, we anticipate accomplishing this by grouping data sets, visualizing results, and analysing them collectively.

# Nathan Brunel (X00139550)

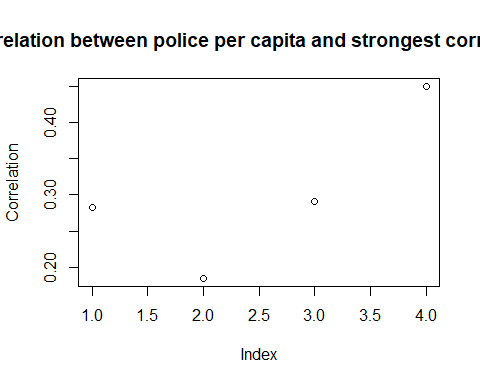
Hypothesis one. Which of the non economic factors most closely relates to the amount of police per person in each state. To accomplish this I created a vector made up of the correlation values up to 10.

(Some graphs have been compiled incorrectly, they work perfectly when running in R)



The above plot illustrates the correlation between the police population and the index. The index position stands for a selection of seperate values. 1 Is the average Tax paid. 2 Is the average population density. 3 Is the amount of face to face crime. 4 Is the overall crime rate. 5 Is the probability of arrest. 6 Is the probability of convictio after arrest. 7 Is the probability of prison sentence. 8 Is the average prison sentence. 9 Is the percentage minority. 10 Is the percentage of young males.

From this plot we can gather that index number 1,2,7,8,9, and 10 all have below average correlation with police. Doing the plot again without this values will allow us to more closer study the variables with strong positive correlations.



The above plot is the same but adjusted to only inlcude the strongest correlations. This plot clearly reinforces that there is a strong link between Face to face crime, crime rate, arrests and the strongest correlation is convictions.

In summary of hypothesis one I have shown which of the selected values most closely correlates with police population in the state. Given that police are almost entirely paid for by taxes I had expected there to be a stronger link between the areas that paid the most and the police population. Instead the correlation between tax and police was one of the weakest measured being only 0.1. The strongest four where listed above (not in order) which were all causes or results of police. Unfortunately this test uncovered nothing unexpected. Hypothesis two. The states with the most crime correlates with those that have the most police and determining if getting the averages for each state has a significant effect on accuracy. To do this I a vector storing the average crime for each county through the period of data (1981-1987) available and then doing the same for police population. After these were done I checked the correlation between between the correlation of averages though the period and the original values

## [1] 0.1194625

## [1] "This value is the average of each county over the period compared to the average police presence"

## [1] 0.1848264

## [1] "This value is the unedited comparison measuring the link"

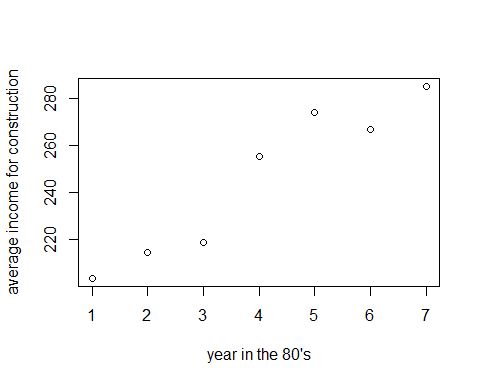
## [1] "The difference is "

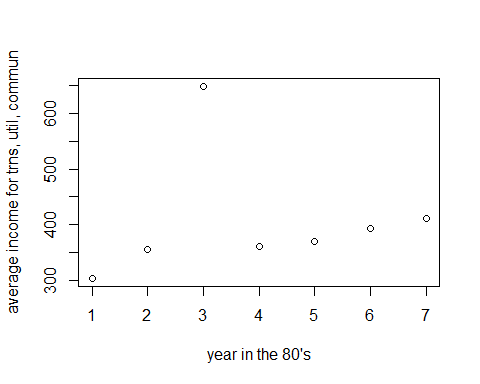
## [1] 0.06536396

In relation to unedited comparison the error 35.37% off.

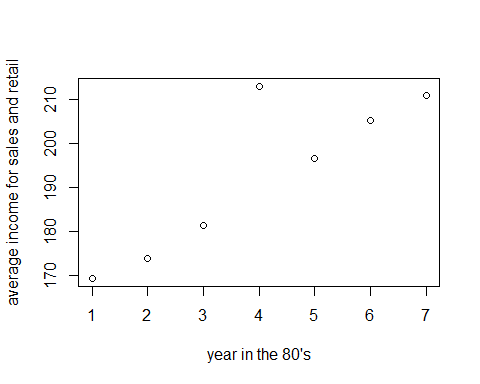
In summary the difference is substantial from adjusting the values to averages over the same period and data. This occured on a very small set if the data encompassed more years or more locations the accuracy would have exponentially decreased.

Hypothesis three. Which of the industries average weekly income was most affected throughout the period regardless of positve or negative effect.

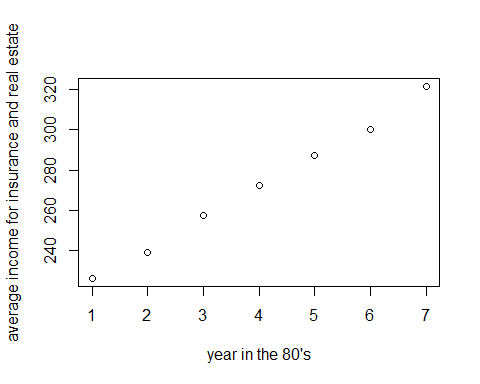
 This plot illustrates the increase in weekly income for people working in construction.

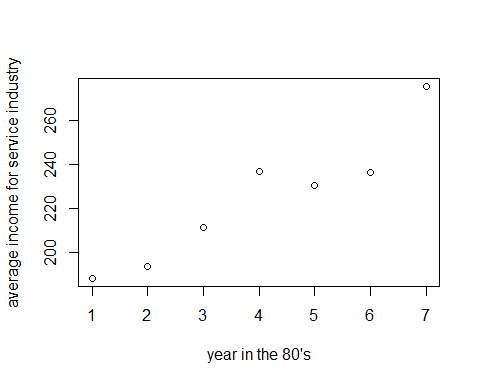
 This plot illustrates the income for people working in transportation, utilities and communication. Notice the unusual spike in 83. The max recorded weekly income is $ 3041.96 and the minimum was $139.49. The summary below displays information on the weekly income for this year I believe there may have been an error in recording this data as these values are significant outliers.

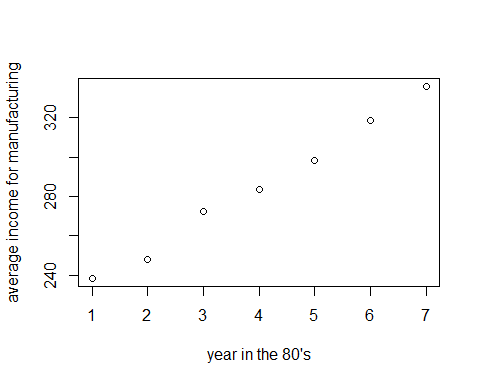
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 139.5 331.9 368.7 648.7 607.2 3042.0

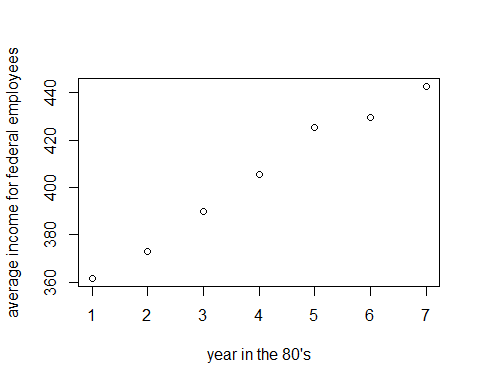
 This plot illustrates the income for people in retail and sales. Similarily to the previous plot there is a spike occuring in 1984. THe max is once again an outlier as demonstrated below so it is safe to consider this a result of human error.

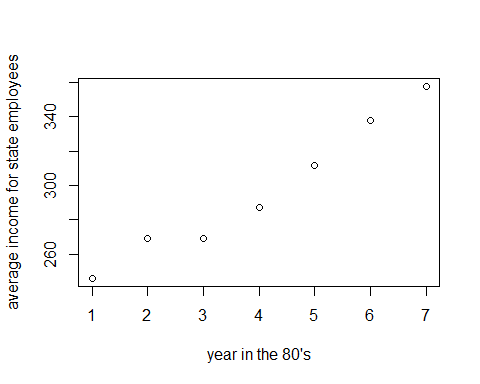
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 138.2 171.1 186.6 212.9 204.4 2242.7

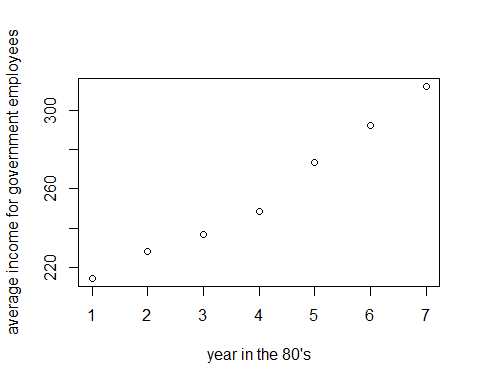
 The above plot illustrates the increase in income for people employeed in insurance and real estate.

 The above plot illustrates the increase in income for people employeed in the service industry.

 The above plot illustrates the increase in income for people employeed in insurance and real estate.

 The above plot illustrates the increase in income for people employeed in insurance and real estate.

 The above plot illustrates the increase in income for people employeed in insurance and real estate.

 The above plot illustrates the increase in income for people employeed in insurance and real estate.

These plots successfully confirm the rapid growth in income for the people of North Carolina between 1981 and 1987 but which of them was most effected can be found by finding the range and taking the smallest average from the largest average and the result will be the overall increase. Since I have outlined my suspicion for human error in two of the values and will do them two ways to try and eliminate the errors.

## [1] 2258.976

## This is the increase in construction

## [1] 3013.1

## [1] "This is the increase in trns, util, commun"

## [1] 2225.873

## [1] "This is the increase in whole sales and retail trade"

## [1] 505.9498

## [1] "This is the increase in finance, insurance and real estate"

## [1] 2175.224

## [1] "This is the increase in service industry"

## [1] 545.02

## [1] "This is the increase in manufacturing"

## [1] 342.55

## [1] "This is the increase in federal employees"

## [1] 374.98

## [1] "This is the increase in state employees"

## [1] 224.5

## [1] "This is the increase in local governments employees"

By this method $3013.1 is the largest incerease in wages. However if we instead take the first value from the last and assume that it was human error we get the same result for all of them except the construction and “trns, util, commun”. Instead we get

## [1] 81.69573

## [1] " for constrcution and"

## [1] 107.3279

## [1] " for transport untilities and communication"

In summary the industry most affected is manufactoring with over $540 increase over the period.

Hypothesis four. Counties with below average income will have higher crime rates. To test this I will sort the counties by whether they have a crime rate greater or less than the overall average. After splitting the counties into appropriate data frames I then compare the average crime rates.

## The average income of dataset: $ 287.2524

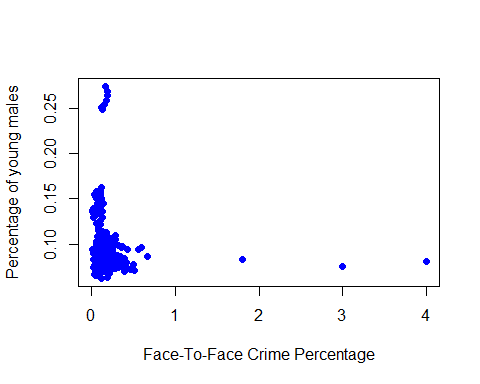
## The average crime rate for high income counties is 0.02596842

## The average crime rate for low income counties is 0.02110641

In summary the results clearly show the hypothesis is incorrect. The lower income counties have slightly less crime on average. This could be due to the rural areas having a lower crimerate and a lower average income. This seperate

# Ronan Doran (X00136708)

# Hypothesis: Young males are responsible for more violent crimes.



In this comparison, I compared the percentage of young males (Age 0-18 approximately). This graph shows that; the lower the amount of young males, the more direct the crimes are towards others. This could be down to the males forming gangs to pursue crime.

As of the making of this data set, Jacksonville in 1981 existed the highest ‘Young Males to Population’ percentage at an extraordinary 27.435%.

## [1] 0.2743584

When I tried to find the max percentage of the crimes to find which ones were face-to-face, I found a bug in the dataset. For an unknown reason, the value was 400% which is Impossible. I had to assume this was a mistake and I made a script to find the max of a new dataset which had all the values of the previous one except there was a constraint for < 100%. I then found out that at one point in the time frame of this dataset, the violence of crime reached 66.6% [1], meaning that of all crimes committed in that county in that year, 2/3rds were confrontational/direct crimes. This happened in the county identifier ‘185’ (John in my group also found evidence in 1985 of county 185 having the highest conviction rate, this leads us to think that this is the most violent county) which is known as Warrenton in 1981 [2]. (Using the link provided in references for County Name: County ID).

[1]

## [1] 0.6666667

[2]

## [1] 185

At the time that this massive crime statistic abrupted, the percentage of young males was at 8.59%.

## [1] 81  
## [1] 185  
## [1] 0.085924

This is below the overall average 8.89%.

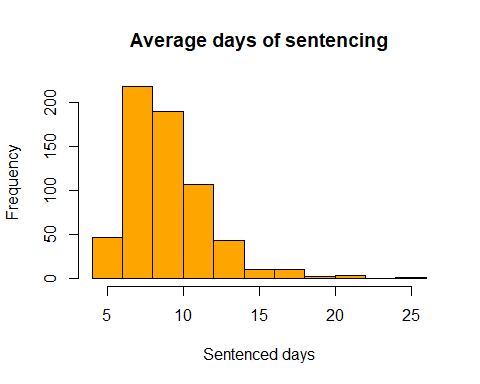
## [1] 0.08897389

Another piece of evidence I found against the hypothesis was that the correlation of both data columns is -0.04259. This is a weak negative correlation, meaning that when one of the variables increases, the other variable is affected negatively. (If X rises, Y shrinks). This doesn’t happen much but it does happen ever so slightly.

# Hypothesis: Young males receive longer prison sentences than the rest of the population.

The total average sentencing is 8.9 days across all counties from 1981 to 1987.

## [1] 8.95454



As this histogram and the mean both identify, the majority of the sentencing time last just over a week with the odd exception lasting up to 20 days (4 instances).

## [1] 4

This means that 4 counties have an average sentencing time greater than 20 days. These counties and time frames are:

## [1] "Average sentencing time in county: 39, 1983 is: 20.98 days.Percentage of young males in this year: 0.0720171"  
## [1] "Average sentencing time in county: 81, 1981 is: 20.52 days.Percentage of young males in this year: 0.0920546"  
## [1] "Average sentencing time in county: 115, 1981 is: 25.83 days.Percentage of young males in this year: 0.0856251"  
## [1] "Average sentencing time in county: 115, 1987 is: 20.7 days.Percentage of young males in this year: 0.0725349"

From this query, I noticed that county ‘115’ also known as Marshall (<https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina> ), has an average sentencing higher than 20 days in 2 of the 8 years measured. I compared these results to the percentage of young males in these eras and found little correlation. The percentages are: In 1981 - 8.5% In 1987 - 7.2%

I say that there is little correlation because young males can still cause some of the crime, but crime is not directly composed of young people. They are both independent.

## [1] 0.0580959

The above command results in an answer of 0.058. This answer is much closer to 0 than it is to 1. Meaning that the correlation is less relevant. This is however proof that young people do commit some crime. This clarifies that my hypothesis is not true.

# Hypothesis: Population will increase at a higher rate over the years

To start this, I decided to pick a number above the mean of the dataset. I picked the number ‘7’ and compared the density values to this. I wanted to find a small portion of values in which I could calculate the population for. I found an external link, provided in the references, to give me the population of the county Charlotte from the years 1981 - 1987.

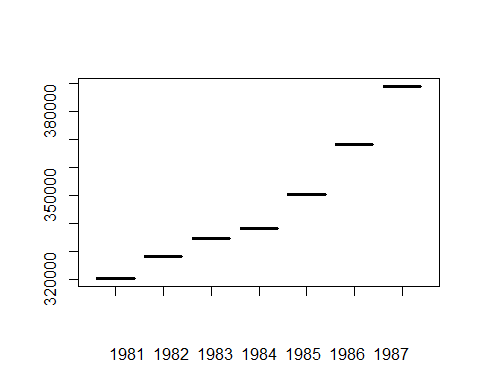
## [1] "119 1981"  
## [1] "119 1982"  
## [1] "119 1983"  
## [1] "119 1984"  
## [1] "119 1985"  
## [1] "119 1986"  
## [1] "119 1987"

## [1] 7

From here, I can calculate the population of the county and see if it rises over the time period.

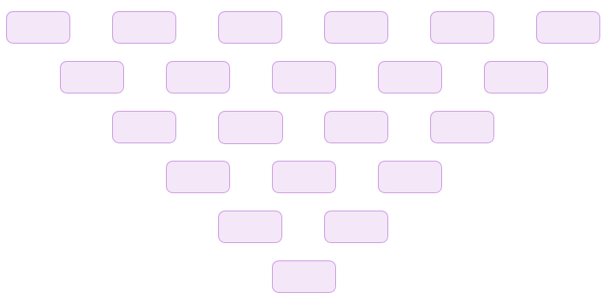
## [1] "7759 between period 1"  
## [1] 7759  
## [1] "6632 between period 2"  
## [1] 14391  
## [1] "3325 between period 3"  
## [1] 17716  
## [1] "12235 between period 4"  
## [1] 29951  
## [1] "17870 between period 5"  
## [1] 47821  
## [1] "20783 between period 6"  
## [1] 68604  
## [1] "NA between period 7"  
## [1] NA

This Box and Whisker diagram shows the exponential increase of the population over the data set duration.



The “num” variable above was used to total the differences. This information is showing the increase of population between each of the years. At 1987, the difference over the course of the previous 7 years is an increase of 68,604 people.

I got the difference between the population for each year and reduced each data set until I eventually got 1 number (see the below picture). The final piece of data was the closest matched variable for all of the years. I used this final variable to scale my way back up through the chart which gives a new value on the very right for the year 1988. This value matched the current growth over the years as my result ended up with an increase of 27 thousand people in one year. From this link,(<http://ww.charmeck.org/planning/ResearchGIS/HistoricalData/Charlotte_Population_and_Area_since_1980.pdf> ), you can see that from 1987 to 1999, the population barely changed (couple of hundred people over the space of 2 years), whereas in 2000 - 2001, the increase was 36k in one year. This shows how unpredictable the population can be. This proves that statistics is pure estimation based on maths and not real life scenarios.)



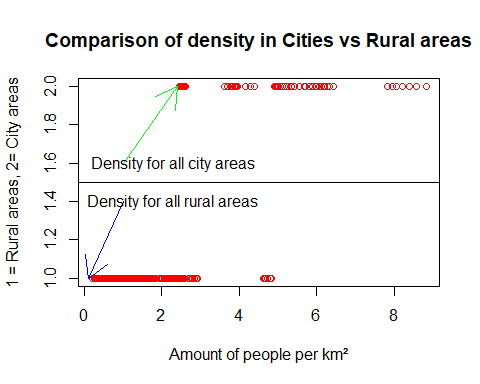
This is a picture i made as a demonstration for what i’m trying to say above

The ending value I got was an increase of 1.0712 of the current population for that year. Seeing that the population in 1987 was 388,955 people, to get the estimated population for 1988, I multiplied 388,995 \* 1.0712 and this got me the result. I made a R script [1] to test the population in 30 years from then (2017’s population) and got an ending result of 805,686 people. The actual population of Charlotte in 2017 was 859,035. Although it isn’t perfect, I do believe that this is an accurate rate of increase based on the fact we only have 7 years of prior information to work from.

[1]

## [1] 805686.4

# Hypothesis: Is there a correlation between Density and living in Cities?



## Average people per 100m² in CITY: 5.150256

## Highest amount of people per 100m² in CITY: 8.827652

## Average people per 100m² in RURAL: 1.018823

## Highest amount of people per 100m² in RURAL: 4.834734

The maximum density in the Rural areas is still less than the average of the density in City areas. This is a definite assist towards proving the hypothesis that there are more people per km² in cities rather than in rural areas.

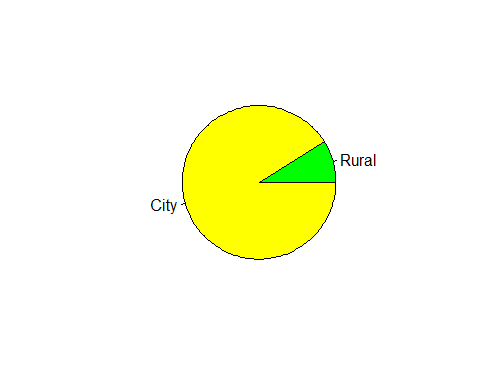
I then went on to test the correlation between the density and the smsa value. The excel speadsheet said it was a categorical column as the options were “no” or “yes” but when i tried to test against this, it said that the values were “1” and “2” for the values respectively. I made the code below to make a new collection, assigning the same values to a new collection which i could then get the correlation for (because both values are now numerical).

## Correlation for Density vs Location: 0.8173011

This is a strong positive correlation, meaning that the higher the density of the location, the higher the chance you are living in a city.

Pie chart showing the ratio between Rural:City

## ( 574 , 56 )



# John De Guzman (X00137820)

Firstly I extracted specific datas from the dataset to a data frame in order to test my hypotheses.

### Hypothesis 1: Crime conviction rates increases throughout the 1980s.

Here we will look into the probability of crime convictions in the 80s. We will try to find which out of the year sin the 80s.

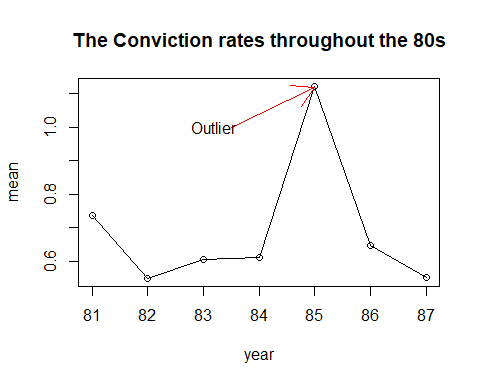
First things first lets look into the data set we will be comparing. Here we have the years and mean of crime conviction rates for the whole 7 years. If we display the two datas without modifying the rows then we would be displaying over 600 records of data. That is inefficient. So now that we displayed the data. We can now visualize it plotting it in a graph.

## year mean  
## 1 81 0.7362951  
## 2 82 0.5486328  
## 3 83 0.6048943  
## 4 84 0.6123345  
## 5 85 1.1218061  
## 6 86 0.6454996  
## 7 87 0.5508611

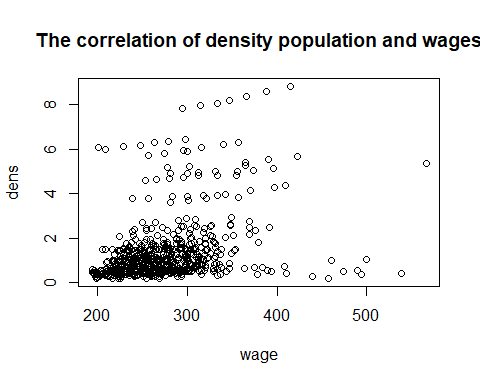
Here we can see the data in graph form below. It appears to spike on the year 85. This could be an outlier as shown in the graph and when displaying the summary of the crime conviction rates. The third quartile of the crime conviction rates is significantly less than the max, the percentage error of this is 5821%. This indicates that the max is an outlier.

summary(crime.data$crimeConvict\_rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.06838 0.34769 0.47437 0.68862 0.63560 37.00000

 Although there was a significant spike in 1985. The graph shows the crime rate decreases rather than increasing. This proves my hypothesis wrong.

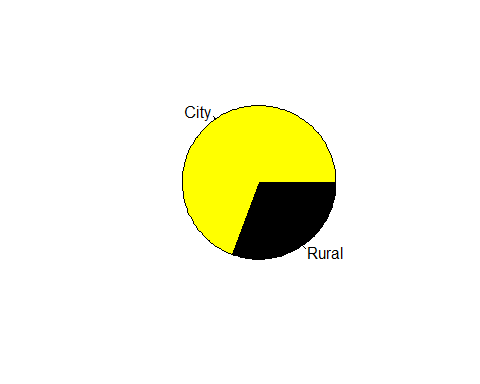
### Hypothesis 2: The lower the density population the higher the wages.

In this second hypothesis I intend to demonstrate that the relationship between the density and the wages is a strong negative association. I created a data frame that stores wage and population density. I then plotted it as a scatterplot as shown below. 

The diagram above shows that the lower the population desnity the lower the wages are. This indicates a strong positves correlation between these variables. In summary this supports my hypotheis.

### Hypothesis 3: There is a higher chance of crime in the city than the countryside.

In this final hypothesis I will illustrate that the there is a higher chance of crime in the the city. In the data I extracted the crime rate and the SMSA, Standard Metropolitan Statistical Area. I then calculated the mean of the crime rate that equals to false if it is rural. Then the same for city letting it equal to true.



## 0.06402297 city

## 0.02842317 rural

The pie chart shows that there is significantly higher crime rate in the city rather than in rural areas. Also showing the mean above, the mean of the city is higher than the rural area. This proves the hypothesis that there is higher chance of crime happening in the city.

We did a summary of all our investigations to clarify the results.

# Hypotheses;

**1: Which of the non-economic factors most closely relates to the amount of police per person in each state.**

Convictions most closely reflects the size of the police population.

**2: The states with the most crime correlates with those that have the most police and determining if getting the averages for each state has a significant effect on accuracy.**

The percentage error is over 35%. This indicates that using averages greatly reduces the accuracy even over a relatively small selection of data.

**3: Which of the industries average weekly income was most affected throughout the period regardless of positive or negative effect.**

Manufactoring, Was most affected throughout the selected timeframe.

**4: Counties with below average income will have higher crime rates.**

This is incorrect the regions with below average income had a lower average crime rate than regions with higher income.

**5: Young males are responsible for more violent crimes.**

I found a negative correlation where my hypothesis indicates it should be positive. I found the county that had the highest amount of crime, the % of young males was below the average % so this proves that there is   
 not much linking these 2 variables together.

**6: Young males receive longer prison sentences than the rest of the population.**

Similar to the last hypothesis, I found the correlation between both the percentage of young males and the average prison sentencing for that era. This resulted in little correlation which meant that young males did   
 commit some of the crimes, but crimes are not mainly composed of young males.

**7: Population will increase at a higher rate over the years.**

For this query, i did a case study on the highest populated county in North Carolina (Marshall). I made a script to calculate the expected growth based on the previous 7 years and i used this information to formulate what the population in 2017 would be. I believe that my result was quite accurate based on the fact i didn't have much previous data to work with.

**8: Cities are more densely populated than rural areas.**

There was a strong positive correlation of .807 between these variables. This means that the higher the density, the higher the chance the location is a city.

**9: Crime conviction rates increase throughout the 1980s.**

Crime conviction rate decreases throughout the 1980s due to the low rates in the last years.

**10: There is negative correlation between wages and population density.**

The lower the population density, the higher the wages get for each individual.

**11: There is a higher chance of crime in the city than in rural areas.**

There is a higher chance crime rate in the city than in rural areas as shown in the piechart.