A

Group Project Final Report on

Analysis of crime and categories of educational levels involved in cities of United States.

By

Chandu Yerragopu

Saran Prasad Balasubramaniam

Wen Xie

Under the esteemed guidance of

Course: ISQS 5347 Advanced Statistical Methods

Prof: Alireza Sheikh-Zadeh, Ph.D.

TEXAS TECH UNIVERSITY

Rawls College *of* Business™

Table of contents

1 INTRODUCTION 3

2 BACKGROUND AND ASSUMPTIONS 4

3 DATA, ANALYSIS AND RESULTS 5

3.1 DATA SOURCE: 5
3.2 DESCRIPTION OF DATA 5
3.3 DATA CLEANING 6
3.3.1 Understanding the dataset 6
3.3.2 Cleaning the dataset 6
3.3.3 Identifying the reuiqred columns for analysis 6
3.4 DATA ANALYSIS - CONCEPTS USED: DESCRIPTIVE STATISTICS, MULTIPLE LINEAR REGRESSION AND CONDITIONAL PROBABILITY 7
3.4.1 Descriptive Statistics 7
3.4.2 Multiple Linear Regression 11
3.4.3 Conditional Probability 13
3.4.4 ANOVA 16

4 CONCLUSION AND FUTURE RESEARCH 17

4.1 CONCLUSION 17
4.2 FUTURE RESEARCH 17

5 REFERENCES 18

6 APPENDIX 19

6.1 MULTIPLE LINEAR REGRESSION 19
6.2 CONDITIONAL PROBABILITY 19

1 Introduction

One of the basic requirement to the well-being of any modern urban settlement is the low or no crime rate in and around that locality. Studies in the past show that participation in illegitimate activities often have been guided by the preconception that since crime is abnormal behavior, its causes must be sought in deviant factors and circumstances determining behavior. Criminal behavior has traditionally been linked to an offender's allegedly unique motivation, which in turn has been ascribed to a unique "inner structure" (e.g., deviations from physiological and mental health, spiritual degeneration), to the impact of exceptional social or family circumstances (e.g., political and social anomalies, war conditions, the disruption of family life), or to both.

This project aims at understanding the relation between education and crime. For the issues raised have frequently centered upon the role 'of education in determining or affecting the motivation and propensities of juvenile delinquents.' A reliance on a motivation unique to the offender as the major explanation of crime does not, in general, lead to the formulation of predictions regarding the outcome of objective circumstances.

We have the data with different levels of education and age. The detailed description of the each of the columns is mentioned in the section 2.2 of this document. We have analyzed this data for various education levels to find the relation between the education level of the individuals. We have used the statistics and probability concepts such as conditional probability – Bayes theorem, the concepts of descriptive statistics, multiple linear regression, ANOVA for the analysis of the data and answer various questions and finally obtain logical insights from the data.

2 Background and Assumptions

"There is a direct correlation between education, stable families and incarceration and crime" - J.C Watts.

The above quote is one of the main motivations for the starting of this project by us. The progress of any country or a locality will always have a direct connection with the educational level of people in that area which is again related to the number of options available for education. We started our project with this idea of understanding the relation between two broad areas namely the education and crime.

The scope of this project is limited as we are analyzing the data by taking only a subset of the actual data. It is because, various other factors such as gender ratios, health, economical levels also have impact on the crime rates in an area. We chose to analyze the education as and the crime rate, thus all other factors which may directly or indirectly impact the crime rate is not considered in our analysis. In case, at some point of time if we want to make a comment on the same, we assume that health, gender and economy are having no impact or very less impact.

3 Data, Analysis and Results

3.1 Data Source

For finding out the relation between the education levels and the crime, we have obtained data from the below website.

http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

3.2 Description of Data

The data (X1, X2, X3, X4, X5, X6, X7,X8,X9,X10,X11,X12) are for each city in US. The description of the data attributes is as follows:

Cities = list of cities in United States

X1 = total overall reported crime rate per 1 million residents

X2 = reported violent crime rate per 100,000 residents

X3 = annual police funding in $/resident

X4 = % of people 25 years+ with 4 yrs. of high school

X5 = % of 16 to 19-year-olds not in high school and not high school graduates.

X6 = % of 18 to 24-year-olds in college

X7 = % of people 25 years+ with at least 4 years of college

X8 = Total population in the city in thousands

X9 = % of Male population in the city

X10 = % of Female population in the city

X11 = % of male population involved in the crime

X12 = % of Female population involved in the crime

3.3 Data Cleaning

For analyzing the relation between the education and crime rates, we plan to use the concept of Multiple Linear Regression. Here is the list of steps followed for analysis:

a) Understanding the dataset

b) Cleaning the dataset

c) Identifying the required columns for analysis and suitable models

3.3.1 Understanding the dataset

The purpose of the project is to identify the relation between the education levels of people and the crime rates reported. The data and the attributes should be understood clearly in order to proceed with analysis.

3.3.2 Cleaning the dataset

The data obtained for analysis many not be clean and ready to use. In this project, we have found some missing values in the data, which we have suitably dealt with by excluding the rows with too many null values, etc. Proper care has been taken in this process to preserve the original characteristics in the initial data.

3.3.3 Identifying the reuiqred columns for analysis

As the primary aim of the project is to identify relation between education levels and crime rate, we have taken only the below columns for our analysis.

Cities = list of cities in United States

X1 = total overall reported crime rate per 1 million residents

X2 = reported violent crime rate per 100,000 residents

X3 = annual police funding in $/resident

X4 = % of people 25 years+ with 4 yrs. of high school

X5 = % of 16 to 19-year-olds not in high school and not high school graduates.

X6 = % of 18 to 24-year-olds in college

X7 = % of people 25 years+ with at least 4 years of college

3.4 Data Analysis - Concepts Used: Descriptive Statistics, Multiple Linear Regression, Conditional Probability, ANOVA

3.4.1 Descriptive Statistics

Descriptive statistics is generally the first step in understanding any data. They tell us about various details about the data and we have answered various questions and obtained answers for the same using R.

R-Code:

```
> project_data<- read.csv("E:/Data Science/ASM/Project/asm_project_data.csv")
> d=data.frame(project_data)
> #slicing the data frame to get only required columns
> d1=d[c("Cities","X1","X2","X3","X4","X5","X6","X7")]
> head(d1)
```

Output:

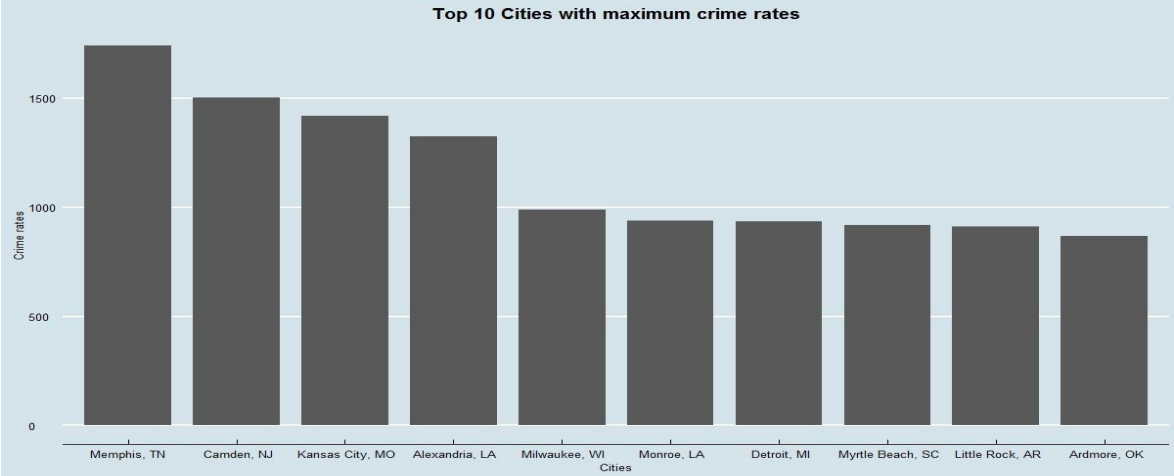
```
> head(d1)
              Cities  X1  X2 X3 X4 X5 X6 X7
1     Fall River, MA 478 184 40 74 11 31 20
2       Hartford, CT 494 213 32 72 11 43 18
3  Niagara Falls, NY 643 347 57 70 18 16 16
4      Kalamazoo, MI 341 565 31 71 11 25 19
5        Lansing, MI 773 327 67 72  9 29 24
6      Bridgeton, NJ 603 260 25 68  8 32 15
```

```
> summary(d1)
```

```
> summary(d1)
              Cities         X1              X2              X3              X4
  Alexandria, LA   : 1   Min.   : 341.0   Min.   :  29.0   Min.   :16.00   Min.   :42.0
  Ardmore, OK      : 1   1st Qu.: 497.0   1st Qu.: 230.8   1st Qu.:30.00   1st Qu.:49.0
  Atlanta, GA      : 1   Median : 654.5   Median : 454.0   Median :34.50   Median :59.0
  Atlantic City, NJ: 1   Mean   : 718.0   Mean   : 616.2   Mean   :37.76   Mean   :58.8
  Baltimore, MD    : 1   3rd Qu.: 820.5   3rd Qu.: 822.5   3rd Qu.:42.25   3rd Qu.:67.0
  Bessemer, AL     : 1   Max.   :1740.0   Max.   :3545.0   Max.   :86.00   Max.   :81.0
 (Other)           :44
       X5              X6              X7
 Min.   : 4.0   Min.   : 7.00   Min.   : 8.00
 1st Qu.:11.0   1st Qu.:21.25   1st Qu.:11.00
 Median :14.0   Median :25.00   Median :12.00
 Mean   :15.4   Mean   :29.90   Mean   :13.82
 3rd Qu.:19.0   3rd Qu.:34.25   3rd Qu.:15.75
 Max.   :34.0   Max.   :81.00   Max.   :36.00
```

Also, to find out some of the interesting information in the above data, we ask questions and find the answers for the same. Some of the questions are as below:

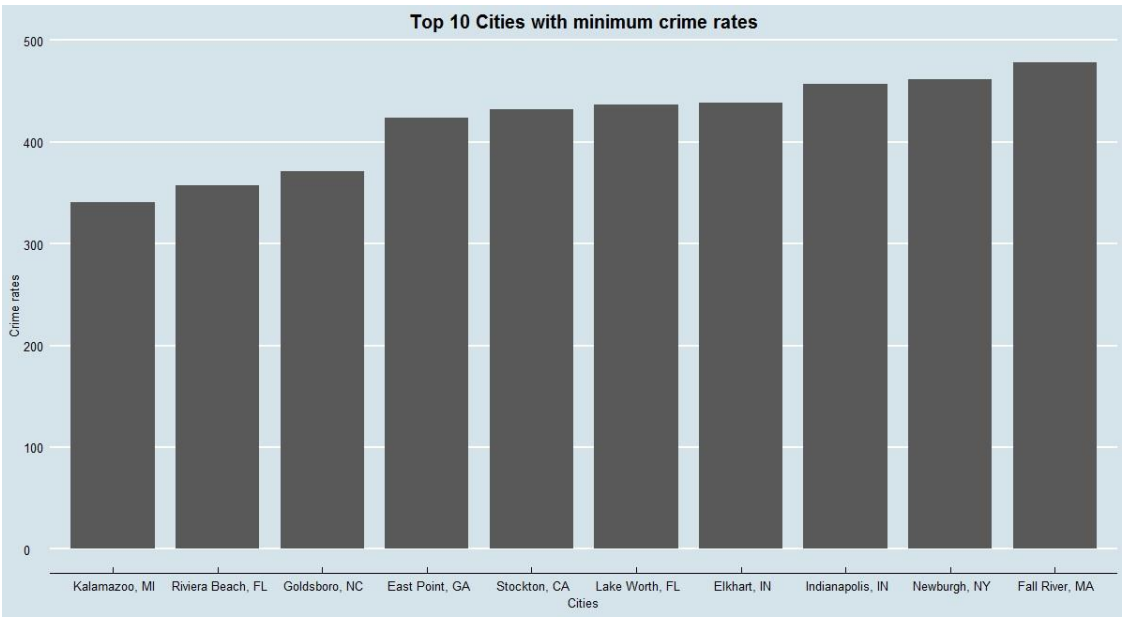
1. What are the top 10 cities with highest crime recorded?



Insight:

This helps us to identify the list of top 1o cities with highest crime record. This data can be used to start taking necessary actions in these cities first as the rate of crime is too high here.

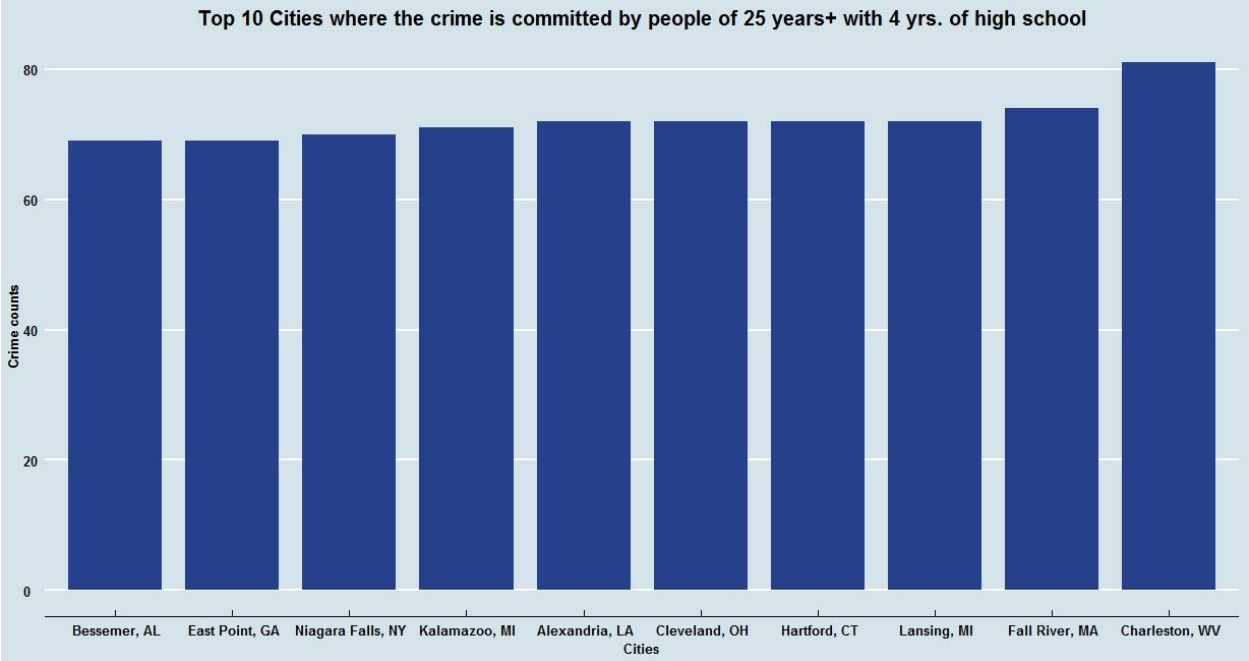
2. What are the top 10 cities with least crime recorded?



Insight:

Of the list of 50 cities that we are trying to analyze as part of this project, we identify these 10 as the cities with least crime, which means that these cities are better to live among the total list of cities.

3. What is the list of top 10 cities where the crime is committed by people with 25 years+ with 4 yrs. of high school?



Insight:

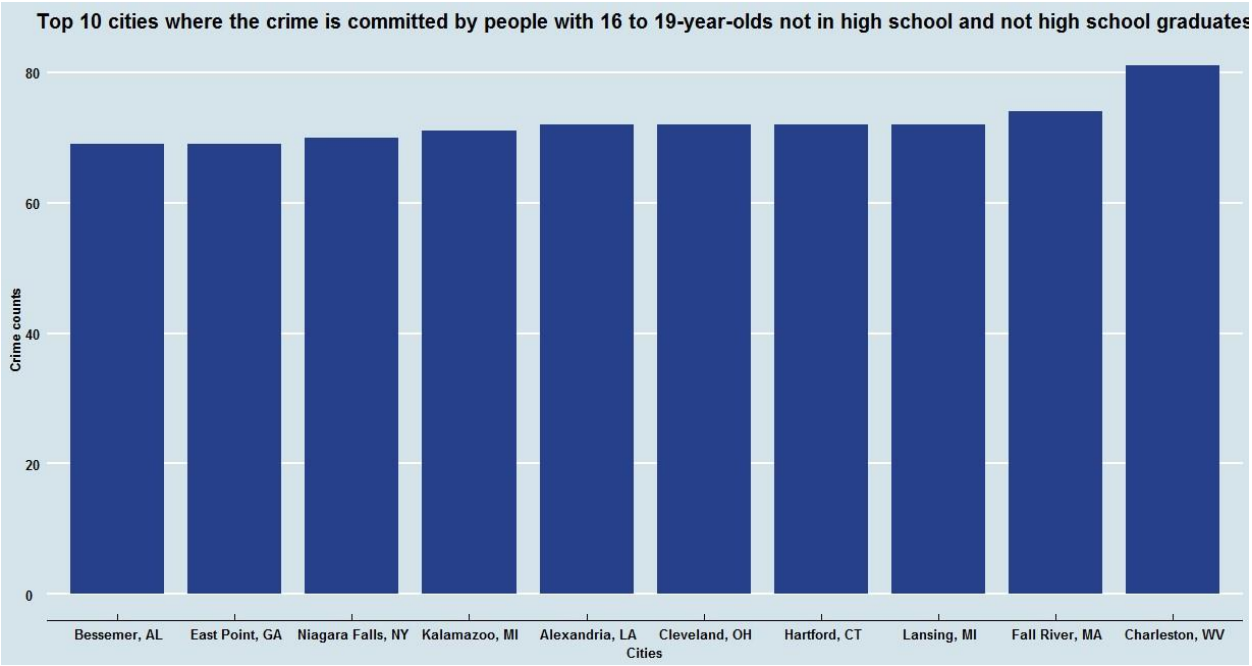
This list helps us in identifying the list of cities, where the focus should be more on the high school education curriculum as most of the crimes are committed by the people with only high school education. Also, we need to identify the available options for higher education in these cities.

R- Code:

```
library(dplyr)
install.packages('ggplot2', dep = TRUE)
library(ggplot2)
library(ggthemes)
data1 <- select(x,Cities,X4)
head(data1)
max_data1 <- data1[order(data1$X4),]
max1 <- tail(max_data1,10)
graph1 <- ggplot(max1,aes(x = reorder(Cities,X4),X4))+theme_economist()
```

```
graph1 + geom_bar(fill = "royalblue4",position = "dodge",stat="identity",width = 0.8,aes())+
  ggtitle("Top 10 Cities where the crime is committed by people of 25 years+ with 4 yrs. of high school")+
  xlab("Cities")+ylab("Crime counts")+theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.title = element_text(face = "bold"))+(theme(axis.text.x = element_text(face = "bold")))+
  (theme(axis.text.y = element_text(face = "bold")))
```

4. What is the list of top 10 cities where the crime is committed by people with 16 to 19-year-olds not in high school and not high school graduates?



Insight:

This data helps us in identifying the cities where most of the crime is committed by committed by people with 16 to 19-year-olds not in high school and not high school graduates. This also indicates that these cities have less opportunities for education although we don't have exact data about that. It helps in the government planning for the improvement of the education facilities in those cities and also taking efforts for enrolling the students in the same.

R-Code:

```
data2 <- select(x,Cities,X5)
head(data2)
max_data2 <- data1[order(data2$X5),]
max2 <- tail(max_data1,10)
graph2 <- ggplot(max1,aes(x = reorder(Cities,X4),X4))+theme_economist()
```

```
graph2 + geom_bar(fill = "royalblue4",position = "dodge",stat="identity",width = 0.8,aes())+
  ggtitle("Top 10 cities where the crime is committed by people with 16 to 19-year-olds
          not in high school and not high school graduates")+xlab("Cities")+ylab("Crime counts")+
  theme(plot.title = element_text(hjust = 0.5))+theme(axis.title = element_text(face = "bold"))+
  (theme(axis.text.x = element_text(face = "bold")))+(theme(axis.text.y = element_text(face = "bold")))
```

3.4.2 Multiple Linear Regression

Multiple regression is an extension of linear regression into relationship between more than two variables. The equation looks like: y = a + b1x1 + b2x2 +...bnxn

There are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The second advantage is the ability to identify outliers, or anomalies.

We create the regression model using the lm() function in R. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

The goal of the model is to establish the relationship between "X1" as a response variable with "X3","X4",”X5”, “X6”and "X7" as predictor variables.

R Code and Output for Mutliple Linear Regression Analysis:

```
> # Create the relationship model.
> model <- lm(X1~X3+X4+X5+X6+X7, data = d1)
>
> # Show the model.
> print(model)

Call:
lm(formula = X1 ~ X3 + X4 + X5 + X6 + X7, data = d1)
```

```
Coefficients:

(Intercept)           X3           X4           X5           X6           X7
    489.649       10.981       -6.089        5.480        0.377        5.500

>
> # Get the Intercept and coefficients as vector elements.
> cat("# # # # The Coefficient Values # # # ","\n")
# # # # The Coefficient Values # # #
>
> a <- coef(model)[1]
> police_funding <- coef(model)[2]
> highschool_25 <- coef(model)[3]
> nostudy_16to19<- coef(model)[4]
> college_18to24 <- coef(model)[5]
> college_25 <- coef(model)[6]
>
> #Printing the coefficients of regression
> print(a)
(Intercept)
   489.6486
> print(police_funding)
      X3
10.98067
> print(highschool_25)
       X4
-6.088529
> print(nostudy_16to19)
      X5
5.480304
> print(college_18to24)
       X6
0.3770443
> print(college_25)
      X7
5.500471
>
```

Now, this model can be used to predict the crime rate in a city in the next years to come if we have the values of the predictive variables starting from X3 to X7.

3.4.3 Conditional Probability

1) What is the probability of choosing a crime from the US cities that is made by a person with25 years and 4 years of high school from Kansas?

Total number of crimes in the US cities = 35898

% of crimes made by a person with age of 25 years and 4 years of high school from Kansas city =59

Total number of crimes in Kansas City = 1419

Total number of crimes made by a person with 25 years and 4 years of high school from Kansas

= 59% * 1419

= 837.21

By Bayes's theorem

P (Crime made by 25 years & 4 years of high school from/Total Crime)

= 837.21/35898

= 0.0233

2)Given that crime has been done by a person with 25 years and 4 years of high school what is the probability that he is from Kansas?

Total number of crimes made my persons with 25 years and 4 years of high school be T.

T = $\sum Total\ crime * \%\ of\ crime\ made\ by\ 24\ of\ years\ and\ 4\ years\ of\ high\ school$

= 20913.6 (Calculated in Excel)

% of crimes made by a person with 25 years and 4 years of high school from Kansas city = 59

Total number of crimes in Kansas City = 1419

Total number of crimes made by a person with 25 years and 4 years of high school from Kansa

= 59% * 1419

= 837.21

By Bayes Theorem,

Probability = Crimes made by person in Kansas with 25 years and 4 years of high school/T

$$= \frac{837.21}{20913.6}$$

$$= 0.0400$$

Investigating whether the crime rates data in the cities in US is distributed normally:

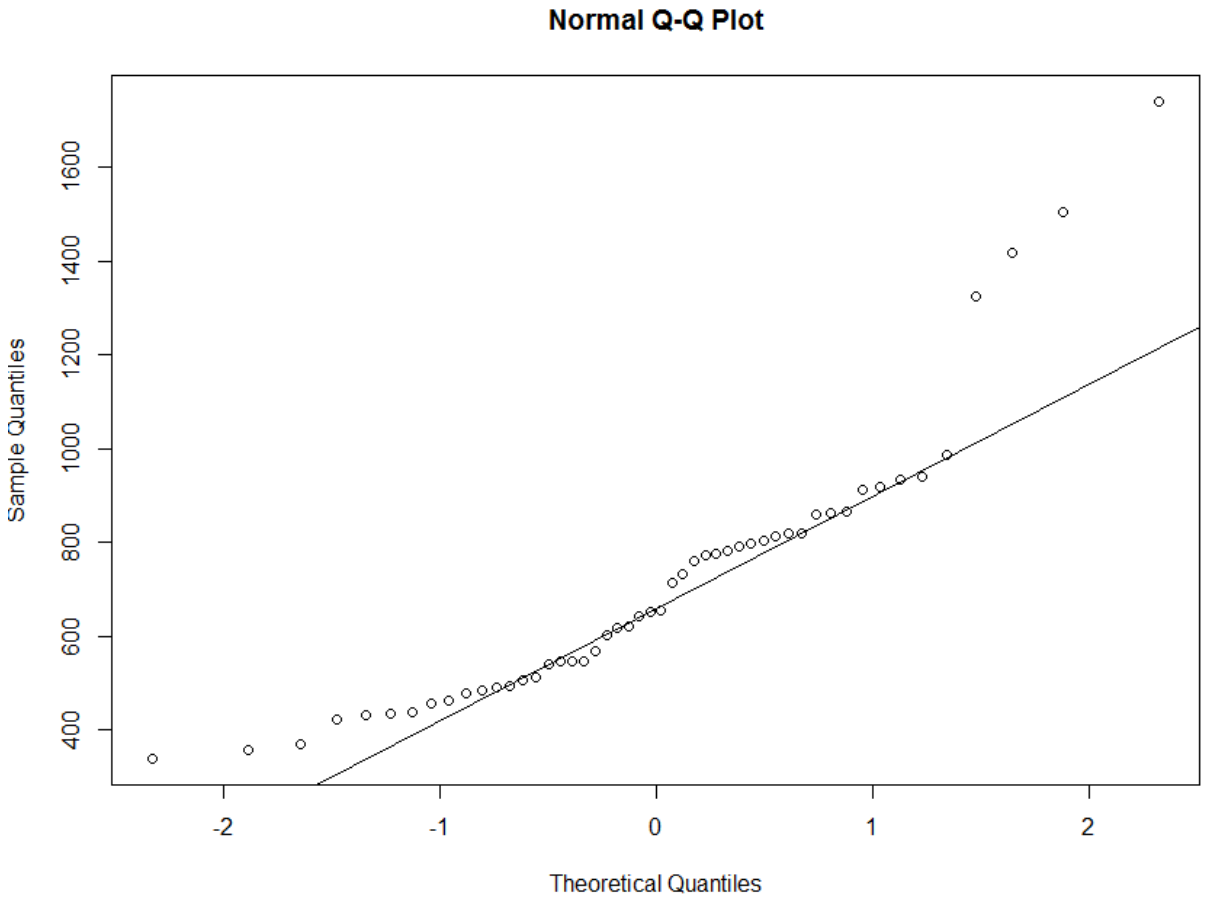
We can find the normality of the crime by drawing a q-q plot for the number of crimes in the US cities.

R-Script:

```
x <- read.csv("asm_project_data.csv")
crime = x$X1

qqnorm(crime)
qqline(crime)
```

q-q plot:



From the above graph we can see that the points don't follow the expected line exactly, but the line is reasonably straight, indicating that there is no problem with the normality assumption.

Skewness and Kurtosis:

Finding the skewness and kurtosis for the total crimes in the US cities.

R-script:

```
n = length(crime)
p = rep(1/n,n)
ey = sum(crime*p)
vy = sum((crime-ey)^2*p)
sdy = sqrt(vy)

#skewness
sk = sum(((crime-ey)/sdy)^3*p)
cat ("Skewness value:",sk)

#kurtosis:
k_value=(((crime-ey)/sdy)^4*p)-3
kurt =(mean(k_value))
cat ("Kurtosis value:",kurt)
```

Output:

```
> #skewness
> sk = sum(((crime-ey)/sdy)^3*p)
> cat ("Skewness value:",sk)
Skewness value: 1.482542
>
> #kurtosis:
> k_value=(((crime-ey)/sdy)^4*p)-3
> kurt =(mean(k_value))
> cat ("Kurtosis value:",kurt)
Kurtosis value: -2.890035
```

Skewness value is 1.4825. Its positive value so the skew is positively skewed. As the value is less than 2 the distribution is markedly different from a normal distribution in its asymmetry.

The main function of the kurtosis is to find whether the values produce outliers are not. Here the Kurtosis value is -2.89. As the value does not exceed + or – 3 it is not an outlier.

3.4.4 ANOVA

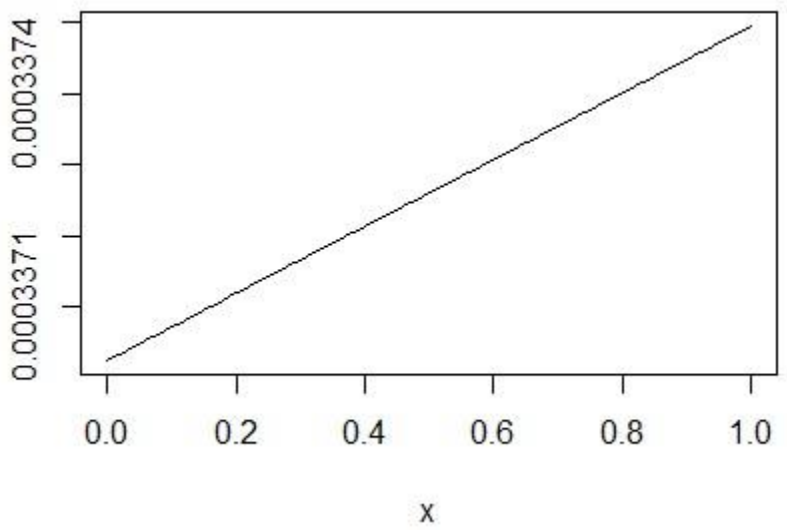
Assumptions for ANOVA:

i) Data should be normally distributed

ii) Data should have equal variances

iii) Samples are randomly and independently selected.

Graph:



Since the graph is not normally distributed, our one of the assumptions have failed. So, we cannot use ANOVA for this dataset.

4 Conclusion and Future Research

4.1 Conclusion

From the analysis of data, we see that the level of education and the reports of crime are closely related in most of the cases, although there are few outliers. This analysis report can be used by the government for understanding the differences in the education levels of people in cities which has a relation with the crime that is reported. They can try to improve the education facilities in those areas such as establishing more and more schools, universities and institutions of higher learning. Also, suitable measures can be taken by the government in promoting awareness among parents that education leads their children to become successful individuals and the chance of they committing crime is very less compared to uneducated individuals. Finally, we conclude or project with the famous quote by Nelson Mandela – "***Education is the most powerful weapon which you can use to change the world."***

4.2 Future Research

In future, we plan to include more and more attributes in our data that is related or has an impact on the crime rate. We will also try to obtain the data from more cities and then observe the trends in the crime rate based on the size of the city, i.e. whether the city is big, medium or small. Because, size of the city has direct relation with the available options for higher education and institutions of higher learning in most of the cases.

We want to come up with the **Predictive and Prescriptive models** in our analysis of the data to better predict the crime happening in future and also to prescribe suitable actions to be taken by government and the social organizations, NGOs,etc.

5 References

[1] Volume Title: Education, Income, and Human Behavior, Chapter Title: On the Relation between Education and Crime, Chapter Author: Isaac Ehrlich, Chapter pages in book: (p. 313 - 338)

[2] http://www.stat.yale.edu/Courses/1997-98/101/condprob.htm

[3] http://www.statmethods.net/stats/regression.html

[4] http://www.stat.columbia.edu/~martin/W2024/R6.pdf

6 Appendix

6.1 Multiple Linear Regression

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is –

```
y = a + b1x1 + b2x2 +...bnxn
```

Following is the description of the parameters used –

y is the response variable.

a, b1, b2...bn are the coefficients.

x1, x2, ...xn are the predictor variables.

6.2 Conditional Probability

The ***conditional probability*** of an event *B* is the probability that the event will occur given the knowledge that an event *A* has already occurred. This probability is written *P(B|A)*, notation for the *probability of B given A*. In the case where events *A* and *B* are *independent* (where event *A* has no effect on the probability of event *B*), the conditional probability of event *B* given event *A* is simply the probability of event *B*, that is *P(B)*.

If events *A* and *B* are not independent, then the probability of the *intersection of A and B* (the probability that both events occur) is defined by
P(A and B) = P(A)P(B|A).

From this definition, the conditional probability *P(B|A)* is easily obtained by dividing by *P(A)*:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Note: This expression is only valid when P(A) is greater than 0.