Final Project

Terrence Nemayire

December 9, 2019

1. Introduction

This paper is on the recorded 2006 crimes committed in the various cities of USA.Source data that will be used has been obtained from : https://data.world/ucr/crime-in-us-2006-offenses. The dataset being an excel file containing different variable of crim,e type. With crimes patterns and population growth, models will be created that can be used to predict or assists in decision making process. To embark on this research various tools and softwares listed below will be needed to help in answering the Research question

2. Research Question:

Does the size of population affect the type of crime or crimes committed in particular towns in the United States of America? and the baseline hypothesis is a) Population size affects type of crimes committed b) and the null hpothesis being- Population size doesnt affect crimes commited

3. Methodology and Tools

To carry out the research analysis, the following tools will need to be present and installed on a standard Computer withe the following minimu requirements -Hardware (4GhZ cPU, 40g HDD, 4G RAM) -Minimum Requirements to be installed -Windows Operating System 7/8/10 (32/64bit) -Microsoft Office Package 2007 and above (Ms Excel and Ms Word) -R-Studio Version 1.1.463 - https://cran.rstudio.com/ -R version 3.6.1 -MiKTeX 2.9 Setup -In R â “Studio have update packages of the following; Knitr, Yaml, Htmltools, caTools, Bitops, Rmarkdown, ggplot, ggplot2, LaTex -GitHub account

4.1 Analysis -Loading data

The purpose of this analysis is to find out if there is a relationship between population and specific crime types that occur in various citiies. At this stage l calling the excel raw data file into R NB The files was already cleaned from the original source and they will be no need for a Key or data dictionary since all the variables are self explanatory. Once the file has been loaded into a dataset in R. the second stage will be to load the dataset into a dataframe that can be manipulated easily by inbuilt libraries. The last part would be to view the summary report of the dataframe for further analysis that will be used in answering the Question. Code block below.

```
library(readxl)
crimes <- read_excel("project_data/crimefile.xlsx")
crimeframe <- crimes
```

```
[,c("City","population","violent_crime","murder","rape","robbery","assault")]
crimeframe
```

```
## # A tibble: 5,499 x 7
##    City       population violent_crime murder  rape robbery assault
##    <chr>           <dbl>         <dbl>  <dbl> <dbl>   <dbl>   <dbl>
##  1 Abbeville        2990            11      0     1       0      10
##  2 Adamsville       4889            44      0     2      13      29
##  3 Alabaster       27766            27      0     0       8      19
##  4 Aliceville       2487            33      1     1       2      29
##  5 Andalusia        8770            49      0     8       8      33
##  6 Anniston        23956           521     14    29     161     317
##  7 Ardmore          1116             1      0     0       0       1
##  8 Ashford          1949             3      0     0       0       3
##  9 Ashville         2451             5      0     1       1       3
## 10 Atmore           7598            69      0     2      11      56
## # ... with 5,489 more rows
```

```
summary (crimeframe)
```

```
##      City            population      violent_crime         murder      
##  Length:5499        Min.   :     18   Min.   :    0.0   Min.   :  0.000  
##  Class :character   1st Qu.:   2474   1st Qu.:    3.0   1st Qu.:  0.000  
##  Mode  :character   Median :   6500   Median :   13.0   Median :  0.000  
##                     Mean   :  24266   Mean   :  136.9   Mean   :  1.749  
##                     3rd Qu.:  18230   3rd Qu.:   51.0   3rd Qu.:  0.000  
##                     Max.   :8165001   Max.   :52086.0   Max.   :596.000  
##                     NA's   :1         NA's   :2                          
##       rape             robbery           assault       
##  Min.   :   0.000   Min.   :    0.00   Min.   :    0.00  
##  1st Qu.:   0.000   1st Qu.:    0.00   1st Qu.:    2.00  
##  Median :   1.000   Median :    2.00   Median :    9.00  
##  Mean   :   7.931   Mean   :   51.38   Mean   :   82.56  
##  3rd Qu.:   5.000   3rd Qu.:   11.00   3rd Qu.:   33.00  
##  Max.   :1071.000   Max.   :23511.00   Max.   :26908.00  
##  NA's   :1                             NA's   :2
```

4.2 Analysis - Finding coerrelation between the variables

After analysis of the dataframe, the various coefficients of the variables, alot had a result of NA as a coefficient. These results of NA indicates that the variables in question are not linearly related to the other variables, or the data from the dataset is noty sufficient enough to prove a meaning to the level of signficance and for multiple regression, the NA coeeficients depicts that the variables do not add much value to the models or affects the response variable (Y). to find coefficient, l have to create anew dataframe colled corcrimeframe eliminating column on cities.

```
corcrimeframe <- crimeframe
[,c("population","violent_crime","murder","rape","robbery","assault")]
cor(corcrimeframe)
```

```
##                population violent_crime   murder rape  robbery assault
## population              1            NA       NA   NA       NA      NA
## violent_crime          NA             1       NA   NA       NA      NA
## murder                 NA            NA 1.000000   NA 0.827243      NA
## rape                   NA            NA       NA    1       NA      NA
## robbery                NA            NA 0.827243   NA 1.000000      NA
## assault                NA            NA       NA   NA       NA       1
```

4.3 Anlysis - Linear Regression Calculatioin

Linear regression in this case is used to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, l can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known. the Resultant formulae will be in the form of Y = AX + B Now l will calculate the Linear regression model from these variables,

NB: The response variable (Y) is murder cases reported and the population is the predictor variable (X) murder =Intercept + (beta * population)

```
linearMod <- lm(murder ~ population, data=corcrimeframe)
print(linearMod)
```

```
##
## Call:
## lm(formula = murder ~ population, data = corcrimeframe)
##
## Coefficients:
## (Intercept)   population
##  -3.250e-01    8.547e-05
```

```
summary (linearMod)
```

```
##
## Call:
## lm(formula = murder ~ population, data = corcrimeframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -240.93   -0.68   -0.04    0.20  342.73
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.250e-01  1.262e-01  -2.576     0.01 *
## population   8.547e-05  8.765e-07  97.510   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.221 on 5496 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.6336
## F-statistic:  9508 on 1 and 5496 DF,  p-value: < 2.2e-16
```

The linear regression model for the two variables (murder and population) is as follow murder =-3.250e-01 + 8.547e-05(population) hence the mathematical model is: Y = -3.250e-01 + 8.547e-05X (Hence with this model l can try to predict the possible number of murders if population in a city increases or decreases), since from the coefficents table, a strong signficant relationship exists. Using the inbuilt summary function in R, l can find the p values of the model to determine model's significance statistically. The p value statistical signficance of 2.2e-16 is above the pre-determined significance level of 0.05 hence the population variable has more signficance in this linear regression model, hence to say the more the population the more the number of recorded murder cases and the model can be used accurately to predict.

Below is a scatter plot diagram based on the Pearson Coeefficient Model. The product-moment correlation coefficient is a measure of the strength of the linear relationship between the two variables (Population and Murder)

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

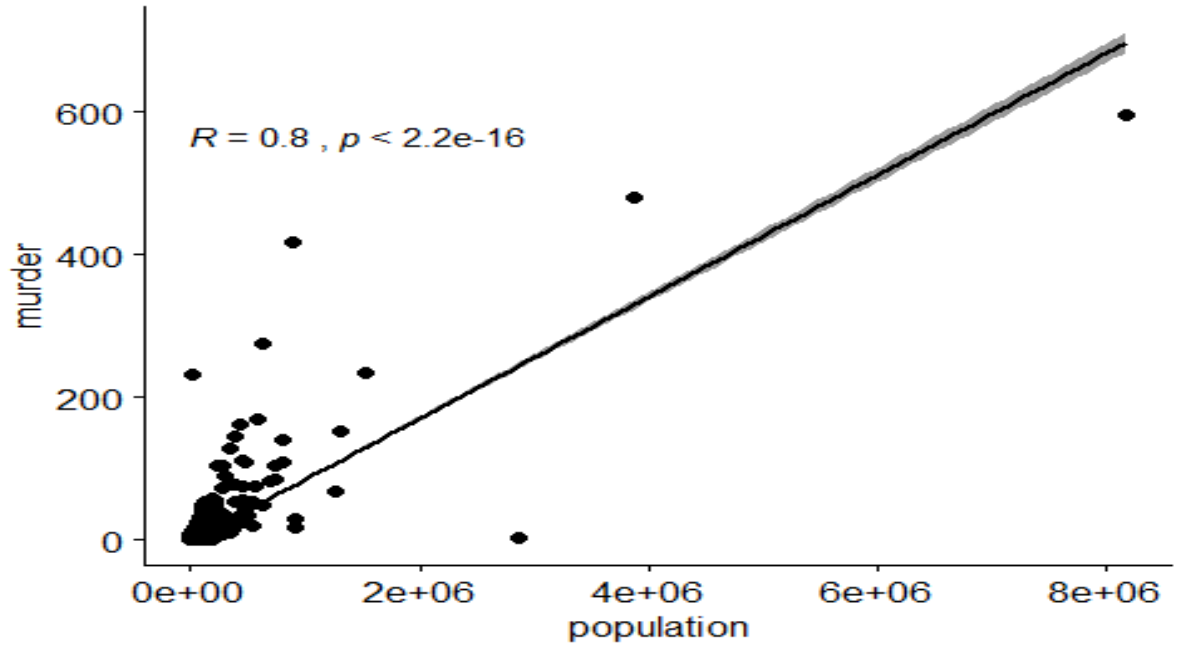
```
## Loading required package: magrittr
```

```
ggscatter(corcrimeframe, x = "population", y = "murder",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_cor).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



```
corgraph <- cor.test(corcrimeframe$population, corcrimeframe$murder)
                     method = ("pearson")
corgraph

##
##  Pearson's product-moment correlation
##
## data:  corcrimeframe$population and corcrimeframe$murder
## t = 97.51, df = 5496, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7861630 0.8055371
## sample estimates:
##       cor
## 0.7960539
```

4.4 Analysis -Multiple Linear Regression

Multiple linear Regression, is a statistical technique that uses several variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the independent variables and response-dependent variable. From the question above, I intend to answer the question, Is there a relation between the number of murder case against changes from other variables (population, assult cases, rape,robbery and violent crimes). The expected regression model that best suits is one which is proven statistically to have a high degree of freedom value, stigma error rate, p value using the t test and the Adjusted R-Squared value. The model equation will be as y = a + b1x1 + b2x2 +...bnxn, where x1, x2..xn repreesnt the predictor variable, b1, b2..bn being the variable coefficients and the value of a being the intercept (constant)

a) MultiModel_1 -building the first multi regression model using all the variables, Y is murder and X predictor variants being other 6 variables, (population, violent_crime, murder, rape, robbery ,assault)

b) MultiModel_2- building the second multi regression model

```
MultiModel_1 <- lm(murder ~
population+violent_crime+murder+rape+robbery+assault , data=corcrimeframe)

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned

summary (MultiModel_1)

##
## Call:
## lm(formula = murder ~ population + violent_crime + murder + rape +
##     robbery + assault, data = corcrimeframe)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.686  -0.324  -0.206   0.135 231.820
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.922e-01  7.300e-02   4.003 6.34e-05 ***
## population    -5.310e-05  1.609e-06 -33.009  < 2e-16 ***
## violent_crime  1.616e-02  2.198e-04  73.527  < 2e-16 ***
## rape           7.540e-03  4.030e-03   1.871   0.0614 .
## robbery        1.038e-02  6.702e-04  15.493  < 2e-16 ***
## assault       -7.410e-04  5.309e-04  -1.396   0.1628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.096 on 5489 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8883, Adjusted R-squared:  0.8882
## F-statistic:  8727 on 5 and 5489 DF,  p-value: < 2.2e-16

MultiModel_2 <- lm(murder ~ population+violent_crime+murder+robbery,
data=corcrimeframe)  #

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned

summary (MultiModel_2)

##
## Call:
## lm(formula = murder ~ population + violent_crime + murder + robbery,
##     data = corcrimeframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.998  -0.337  -0.218   0.130 231.860
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.090e-01  7.072e-02   4.369 1.27e-05 ***
## population    -5.310e-05  1.597e-06 -33.248  < 2e-16 ***
## violent_crime  1.629e-02  1.640e-04  99.369  < 2e-16 ***
## robbery        9.704e-03  4.284e-04  22.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.098 on 5492 degrees of freedom
##   (3 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.8881, Adjusted R-squared:  0.8881
## F-statistic: 1.453e+04 on 3 and 5492 DF,  p-value: < 2.2e-16

## .
```

5 Results

a) Model_2 Y= 0.29 + (-0.00)population + (0.016)violent_crime + (0.01)rape + (0.01)robbery + (-0,00)assault From the first model, predictor variables (Rape and Assault and negligable and has no signficance in the model that can answer the resarch question, hence the second model l will remove the 2 negligable variables and improve model.

b) Model_2 Y= 0.31 + (-0.00)population + (0.016)violent_crime +(0.01)robbery From the second model, all the variables have a p value that is signaficanT based on the t-test and the variables have a high signficant effect to the response variable. The Adjusted R-squared test IN the new model has a 89% accuracy rate for the model using the (population, violent_crime and robbery variables) hence l can safely conclude to say Model 2 can statistically be used for further analysis or prediction.

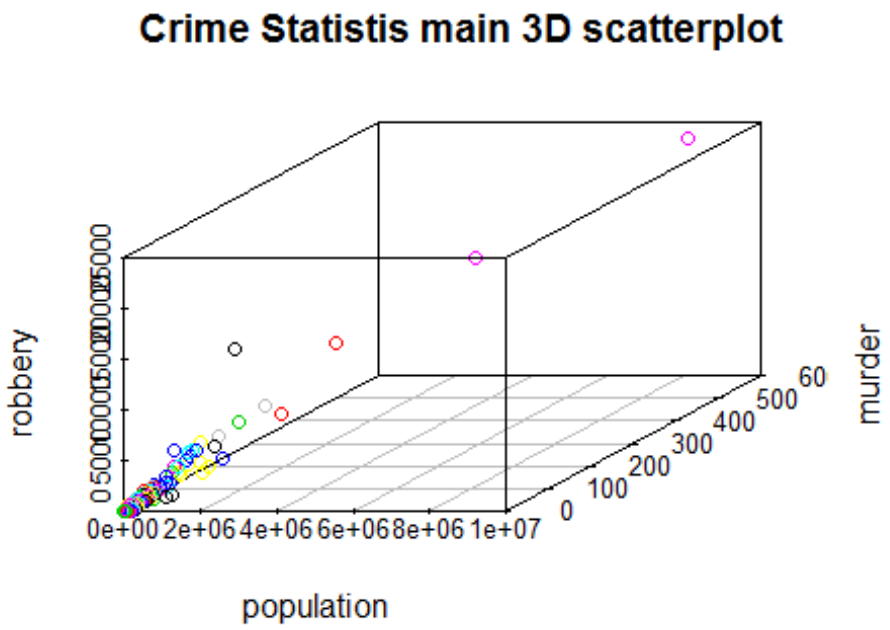
```
print(MultiModel_1)

##
## Call:
## lm(formula = murder ~ population + violent_crime + murder + rape +
##     robbery + assault, data = corcrimeframe)
##
## Coefficients:
##   (Intercept)     population  violent_crime           rape        robbery
##     0.2921880     -0.0000531      0.0161600      0.0075399      0.0103834
##       assault
##    -0.0007410

print(MultiModel_2)

##
## Call:
## lm(formula = murder ~ population + violent_crime + murder + robbery,
##     data = corcrimeframe)
##
## Coefficients:
##   (Intercept)     population  violent_crime        robbery
##     0.3089815     -0.0000531      0.0162939      0.0097039

library(scatterplot3d)
attach(crimes)
scatterplot3d(population,murder,robbery,violent_crime,main ="Crime Statistis
main 3D scatterplot")
```



6 Conclusion

The experiment has been conducted successfully to establish the strength of relationships between the various variables in the dataset. Regression models have been formulated for future use in prediction and aid in decison making process in crime analysis. From the dataset used for this experiment, l can statistically say that murder crimes are related to the size of the population in a city and also other variables have little of low signficance levels in determing murder crimes.