Analyis of Police officer deaths in the United States of America 1900-2016

Daniel Gorman
BSc (Honors) in Computing
National College of Ireland
Dublin, Ireland
Daniel.gorman@student.ncirl.ie

***Abstract*—** United States of America has been long associated with guns, its buried deep in American history from long before the country was formed. American law enforcement is known worldwide for the operations they have undertaken to stop crime and corruption on both their own and foreign soil. But with these operations comes a cost. That cost has been the loss of thousands of officers over their history. Hundreds of well trained dogs died also some with their partners and some alone. It's not just gunfire that has killed these brave officers, it has come in various way including a large proportion of officers dying from heart attacks, although from the datasets available it's hard to say if it was the constant pressure of the job and fear of being shot at any time like the 11,000. The following report will highlight my findings from the research I've conducted.

Keywords— United States of Ameirca, Police, Law enforcement, Prohibition, War on Drugs, death.

I. Introduction

In today's age with so many news mediums and social media platforms. It has become common to hear of police brutality and unjust killings by law enforcement. United States of America law enforcement has come under scrutiny because of the increased coverage they have received with the rise of social media. The killings of Eric Garner and Michael Brown drove police killings of black people to become the top news story of 2014 per the Associated Press annual poll of U.S editors and news directors [1]. US Law enforcement has been charged with the killings of over 2,000 people from January 2015 to December 2016 [2], with many of the killings unjustified and unarmed civilians. Law enforcement officers are dealt an almost impossible task of been protectors of the citizens with guns been allowed to be owned by citizens, making everyday interactions between citizens and authorities concerning, resulting in many officer's lives being lost.

Since the formation of the US Law enforcement, over 20,000 officers have died while on duty. The deaths have come in various ways e.g. Gun shot, heart attacks. With this report I intent to analyses the deaths of law enforcement officers since the beginning of the 20th Century. Examining many operations undertaken by law enforcement including prohibition which lead to the subsequent rise of organized crime in the US. Another event is the 'War on drugs' introduced by President Ronald Regan in 1971. By analyzing these prominent events in US history, I intend to discover did police killings rise dramatically and maintain steadily because of these operations undertaken by law enforcement. I also want to find out if there is a coloration between gun ownership and police officers been killed.

II. Data Collection

A. Police Deaths

The Police Deaths dataset was acquired from Kaggle [3]. The dataset was constructed based on Law enforcement officers killed in the United States of America since the formation of the Law enforcement in 1791 till 2016. The dataset contains the name of each officer, his/her rank, the year of death, cause of death and the location of the death. The dataset was originally scraped from 'Officer down memorial page' [4] by FiveThrtyEight a website knowns for opinion poll analysis and blogging. Thousands of records were discarded due to the location of the death occurring outside of the 50 states, for this study I have included District of Columbia as an unrecognized state and United States Dept agents killed as it would relate to officers killed overseas most likely in countries against the war on drugs. After filtering and sub setting of the dataset, I was left with 20,000 records of police officers killed between 1900-2016. I choose this dataset as it was from a reputable user from Kaggle, FiveThirtyEight. After some research, I found that number of records in the dataset matched that of those fallen officers which verified the integrity of the dataset [5]. Some work done with this dataset is located on Kaggle, User 'Donyoe' performed a series of analysis on the dataset like "Cause of death", "Police deaths by state". There hasn't been any work of note documented previously on this dataset or relating datasets other than Albert P. Cardarelli journal entry "An analysis of police killed by criminal action 1961-1963" [6].

B. Gun Stats

Building on from the police death dataset, I have sourced the gun ownership figures by state as of 2007 sourced from

Demographic Data [7]. Before cleaning the dataset, it contained the name of each state and the percentage of gun ownership in each state, along with the other crime statistics e.g. Gun murders per 100,000 and Violent crime. To make the data richer I added the state by state populations from the 1900-2010 census [8]. It allows for easier reading as gun ownership stats are in a percentage leaving readers very uninformed on the population and gun ownership in each state. After cleaning we're left with 52 columns of data e.g. the states of United States and for this report I have included the District of Columbia which isn't recognized as an official state. The dataset now contains the name of each state and its abbreviation name, the population of the state from 1900-2010 in 10 year intervals and the gun owners as of 2007. There have been various articles written about the material in this dataset located on the source website e.g. correlation of 'gun ownership and gun death'. A study, published in the American Journal of public health, found that between 1996 and 2010 almost 1 officer per 10,000 was being murdered by use of gun in states with high gun ownership (<50%) [9].

Table 1: Police Deaths

Name	***Type***	***Description***
Dept	String	Department name the officer was assigned to before death
Cause	String	The cause of death e.g. Gunfire
Year	INT	Year of death
State	String	Location the death occurred abbreviated
Rank	String	Job title of officer
Day	String	Day of death
Date	INT	Date of death

Table 2: Gun Stats

Name	Type	Description
State_full	String	Full name of state
State	String	Name of state abbreviated
Pop_2010	INT	Population of state for year 2010
Guns_2007	INT	Number of guns owned in each state 2007
Gun Murder per 100k	INT	Gun murder rate per 100k inhabitants

III. METHODOLOGY

Throughout this section I will discuss the implementation of the Knowledge Discovery in Databases (KDD) at a high level. The KDD follows a series of processes including, Data selection, Data Preprocessing, Transformation, Data Mining, Interpretation and evaluation of patterns into knowledge.

A. Data Selection

Data selection began with researching significant periods of American history to find periods of times strict laws or operations were put in place to oppose criminal activity. After much research, I decided to focus on era of prohibition (1920-1933) and Americas war on drugs (1971-2000). The reason behind choosing this time frame is 1971-2000 was considered the peak years on the war on drugs denoted by the vast amount of federal spending and if I choose a larger time scale the results would be influenced by the events of 9/11 and aftermath. The other dataset chosen as part of this report is the gun ownership and population figures by state. After extensive research for a dataset based on gun ownership by state. I had to create a dataset based on variables from other datasets. I took the name and abbreviation of each state and coupled it together with the gun ownership estimates of 2007, as based on numerous surveys as there is no official database for gun ownership available to the public. I then added the population of each state taken from the census as it is the most accurate available. My motivation for binding these two datasets together was that I could run tests based on the population of states and gun owner's vs officers killed to find if there is correlation between high gun ownership and high population states and officers killed throughout the history.

B. Data Preprocessing

Data preprocessing is the process of removing unneeded variables and rows. During this stage outliers and meaningless data is removed to pave the way for a more accurate analysis. Also during this stage, we decide on the best strategy for dealing with missing data fields. Cleaning of the dataset took place after I had selected the periods of history I was focusing on. Some preprocessing techniques included the removing of unneeded columns like the Canine column which doesn't fit the goal of the analysis. these columns included ‘cause’, ‘person’, ‘eow’, and ‘canine’. The reason behind ‘person and eow have been described above. I will discuss in more depth in the next section how these processes were carried out. Also at this stage data integration will take place with the merging of the two selected datasets above.

C. Transformation

Transformation began by creating new variable called ‘rank’, this variable was filled by an existing variable called ‘person’. Using a function, I could remove the name of the police officer and store only the rank of the officer in the newly created variable called ‘rank’. The ranks were stored as a factor before being transformed to table for easier analyzing. The same process was used to create a variable for days of the week which were extracted from the end of watch (eow) variable. The column ‘cause’ was removed due to having a similar variable called ‘cause_short’ which has been since renamed to ‘cause’. Canine was not needed as I didn’t need it for the analysis I am conducting.

D. Data Mining

Data mining is the act of searching for patterns within a dataset, MapReduce will be used to draw patterns from different periods of history to find if there has been certain characteristics throughout history that has caused police officers to die in certain states. MapReduce is a twostep process, Map and Reduce. The job of the Mapper is to perform filtering and sorting before the Reducer can perform a summary operation. This stage will be discussed in great depth in the next section.

E. Interpreation/Knowledge

This is the last stage of the KDD cycle, it consists of two stages, interpreting what the resulting data mining stages means and what we’ve learned from it. If executed accurately, significant, rich knowledge could be drawn if the results are interpreted correctly. This will be discussed in more detail in the results section below were all my findings will be displayed.

IV. Implementation

A. Architecture

I have built my application workflow around the KDD process for a structured approach for the analysis and competition of this report. Throughout the following section I will describe the techniques and approaches I’ve followed to complete the analysis using various tools such as MapReduce with Python, R programming language to run a series of test and resulting visual graphs. The report was produced using the architecture below;

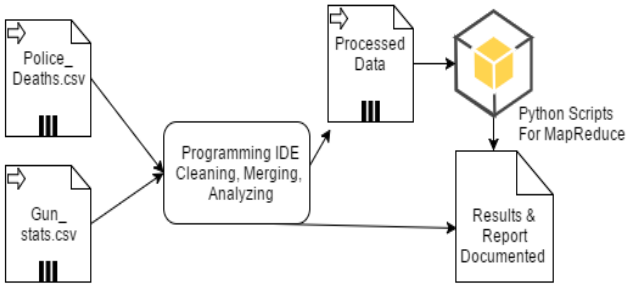

Figure 1: Architecture Diagram

B. Data Selection and Pre-processing

The process began with the data selection and cleaning of the datasets, I downloaded the files, then converted them to .CSV files as it can be problematic importing excel sheets other than CSV files into RStudio. After reading the files into RStudio and setting the file to factors. Installation of packages was next for some cleaning of the files and visuals. After cleaning took place, I merged the two .CSV files through a merge function storing the new larger file in a data frame called “df”. As there was no variable for rank of the fallen officer I had the idea of extracting the rank from the “persons” variable and storing it in a new variable called rank. This was accomplished using two functions [10], one setup for the removal of a string mentioned and the 2nd function for storing it in a variable using the strings provided. The same functions were reused to extract the days of the week from the data column which I found more beneficial to use days of the week instead of the date when analyzing this historical dataset. Next was removing unused variables and changing any new variables created to factors e.g. newly created rank and day of the week column.

Next was the creation of subsets from the main dataset. I created various subsets including subsets for prohibition era, war on drugs era and modern era with the drug culture in USA swiftly changing with the introduction of many drug laws across America deeming certain drugs no longer a felony. The different subsets were broken up for comparisons later to determine if certain states, rank of officers are in more danger than others. This is where the 2nd dataset is implemented for analyzing the modern era subset to find if there is a correlation between high population states, gun ownership Vs police deaths. Using the newly cleaned data, producing visuals to be shown in the results was implemented efficiently. Using a variety of visuals from the library tidyverse [11]. To make comparisons between different era’s clearer to readers and to see if post that ‘era’ the trend of officers dying slowly decline or steadied, this was accomplished using filters on the subsets to only show relevant results. As for the prohibition era ending in 1933 and the great depression (1929-1939) we expect crime to quickly rise endangering more police officers than ever before with the overlapping ban on alcohol and resulting rise in the mafia and black market, this will be discussed in next section.

C. *MapReduce*

The implementation of MapReduce proved a strenuous task, after cleaning of the datasets were complete, I wanted to compare the deaths of officers over 3 different periods by state to find if over time officers were dying in the same states throughout history or if they were changing over time with the rise of criminal activity and other factors. The 3 splits from the processed dataset are from 3 different periods in history all ranging from 13-16 years, below are the selected years;

1) Prohibition 1920-1933
2)Post Prohibition and the great depression 1934-1950
3)War on Drugs 1971-1984

Began by making 3 subsets in RStudio for the different periods, the reason for doing this was the states in the CSV file, 'state' had no numeric value only the name of the state for each observation. After creating the subsets in RStudio, I then created a data frame with 'state' as a factor with the numeric variable 'deaths', with the use of tidyverse library, I was able to extract the occurrence of each state into the numeric variable 'deaths'. The process was repeated for other 2 splits before outputting the 3 new data frames to CSV files. Code snippet below shows the data frame being created.

```r

#Creating subset for Prohibiton Era
during<-subset(df, year >= 1920 &  year <= 1933)

#Creating data.frame for numeric value of each state
split_1 <- during %>%
  group_by(state) %>%
  summarise(Deaths = n())

#Write to CSV file
write.csv(split_1, "split_1.csv" )
```

Figure 2:Subset being created after filtering numeric data

Project.R* × split_1 ×

Filter

	state	Deaths
1	AK	2
2	AL	56
3	AR	51
4	AZ	13
5	CA	182

Figure 3: Result of the filter

Next, was to initialize the MapReduce environment. The environment for the usage of MapReduce was setup using Python IDLE to for easy access and editing of code, also used was Command Prompt (CMD) to process the CSV files and run python commands which are attached to this folder in the form a .bat file.

The Mapper was setup to take the top 10 records by numeric value sorted by highest to lowest from each split before outputting the top 10 from each to a new CSV file created named Out.CSV, then the 2nd and 3rd split were run using the same commands, adding more record to the outputted CSV file. Once that was completed, the CSV file was filled with the 30 records, 10 form each period. The reducer was then processed against the new Out.CSV file to find the top 20 states officers deaths have occurred sorted highest to lowest with. The following figures show the code used to produce and setup the Mapper and Reducer through python.

```python
#!/usr/bin/env python
import sys

# Mapper to return local top 10 Officer killed by State
# Data source: https://www.kaggle.com/fivethirtyeight/police-officer-deaths-in-the-us
# Data header: "Prohibiton" "State" "Deaths"

# Initialise a list to store the top N records as a collection of touples (deaths, record)
myList = []
n = 10  # Number of top N records

for line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()
        # split data values into list
        data = line.split(",") # CSV file

        # convert deaths (currently a string) to int
        try:
                Deaths = int (data[2])
        except ValueError:
                # ignore/discard this line
                continue

        # add (deaths, record) touple to list
        myList.append( (Deaths, line) )
        # sort list in reverse order
        myList.sort(reverse=True)

        # keep only first N records
        if len(myList) > n:
                myList = myList[:n]

# Print top N records
for (k,v) in myList:
        print(v)
```

Figure 4: topTenStatesMapper.py

```python
#!/usr/bin/env python
import sys

# Reducer to return overall top N records
# Data source: https://www.kaggle.com/fivethirtyeight/police-officer-deaths-in-the-us
# Data header: "Prohibiton" "State" "Deaths"

# Initialise a list to store the top N records as a collection of touples (deaths, record)
myList = []
n = 20  # Number of top N records

for line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()
        # split data values into list
        data = line.split(",")

        # convert deaths (currently a string) to int
        try:
                Deaths = int(data[2])
        except ValueError:
                # ignore/discard this line
                continue

        # add (deaths, record) touple to list
        myList.append( (Deaths, line) )
        # sort list in reverse order
        myList.sort(reverse=True)

        # keep only first N records
        if len(myList) > n:
                myList = myList[:n]

# Print top N records
for (k,v) in myList:
        print(v)
```

Figure 5: topTenStatesReducer.py

The two figures displayed above show the code used to extract the top 10 from each import and exporting it to the CSV before the reducer takes the overall top 20 states from the three different periods, the two python files were compiled a combined four times to find the desired outcome. With this then I could then analyze to find patterns in the data between the periods in history. The results of MapReduce will be discussed more in the next section

D. *Analysis Testing*

Testing was implemented by a series of statistical test, ranging from summary, means to correlation tests between the two datasets. Correlation was tested to see if there is correlation between high gunownership and deaths among officers.

Setting up a t-test required setting the null and alternative hypothesis, I set out to test that the mean of officers dying by state during the reign of George Bush (2001-2008) was equal to that of under Barrack Obama (2009-2016), the hypothesis was set as follows;

$$H_0: \mu = \text{Deaths}$$
$$H_1: \mu \neq \text{Deaths}$$

To setup the required environment, I setup two new data frames one for each president time in charge. I then merged the two datasets together under the state variable. Now the new data frame "tess" contained 3 variables including state and the two numeric values associated with the president's time in charge. During this time a scatterplot was plotted to represent the data of both time frames visually to see if the data was changing overtime, using the function abline to show if G.Bush death toll was equal to that of B.Obama. It will give a good representation of what to expect after the paired t-test has been performed

E. Visuals

Creating visuals is a necessary process as many readers don't want read through pages of text and not see the data visually represented in some shape or form. I used numerous combinations of graphs and plots to represent the results, using basic plots like histograms, ggplots and bar charts to show an array of results like ranks of officer's deaths by year, state etc. This was possible by using the renowned ggplot2 library. Filtering and subsets were used to breakdown the plots to represent the relevant data

V. Results

From the investigation of my datasets I found several results that shocked me, I was amazed by the sheer number of officers who have died while on duty (21,809) excluding canine dogs, a lot of these deaths have come in spells when law enforcement has been set out to target illegal activity notably the importing and sale of alcohol during prohibition and the war on drugs. Before prohibition was enforced a total of 4,520 officers died dating back to 1791, during the 13 year stretch of prohibition (1920-1933) 3,759 officers died, many causes of death rose during this period like automobile crashes and accidents most likely due to more automobiles on the road and purists of traffickers. The period of prohibition could be seen as a terrible decision for many reasons such as the sheer number of officers that died during the period and the subsequent rise in of the mafia across America as the ban on alcohol gave the mafia another source of income on the black market. The graph below shows the sharp increase from the formation of law enforcement in 1791 until the end of prohibition 1931

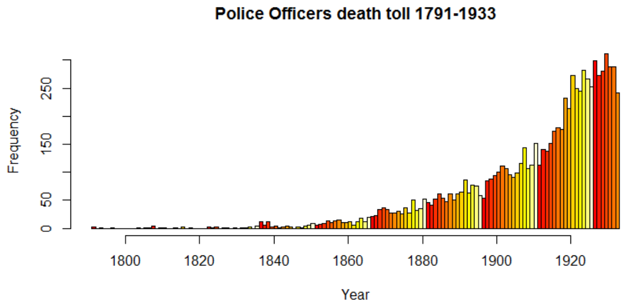

Figure 6: Death toll officers until 1933

This histogram shows the sharp rise in deaths among police officers at the turn of the 20th century then once again when the ban on alcohol is enforced. At the peak of officer's death per year they were reaching upwards of 300 on a few occasions many in the latter part of the 1920s. The great depression that began in 1929 could have been a factor with desperation creeping in among citizens. Unsurprisingly most officers killed during the period (1920-1933) were patrolman and general police officers. Among my findings, I found the states that rose the most with the introduction of prohibition were states to the north-east in America, states like Ohio, Pennsylvania, Illinois all rose to within the top 7 states in deaths among officers. The graph below displays the top 7 states including United States federal agents killed displayed as "US";

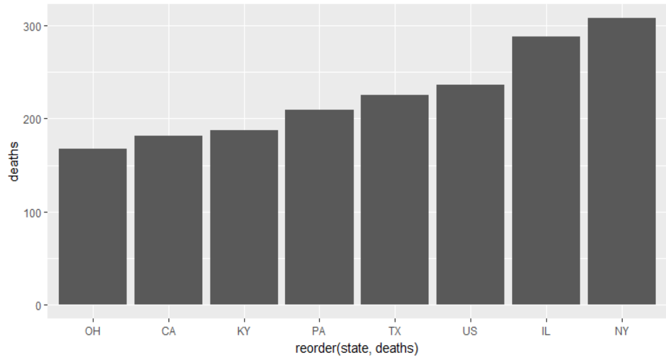

Figure 7: Death by state during Prohibition (1920-1933)

The graphs show's states during the prohibition that deaths were greater than 140. As can be seen many of the states are bordering countries like Mexico and Canada, the furthest inland state is Kentucky. Federal agent's deaths increased as would be expected with a federal law being enforced country wide. Compared to the graph below we can notice some states officer deaths dropping post prohibition, once again this isn't the sole factor as there is no explanation behind the deaths of the police officers. But we can take from the sudden spikes and drops in periods of history that these events played a prominent role in the deaths of officers. I can only believe the reason New Jersey doesn't feature among these states is due to the power Enoch Johnson had during this period, Johnson was well known for not enforcing the prohibition law and making money from the sale of alcohol in Atlantic City in turn saving many local officers from certain death while boosting the local economy with visitors coming for various activates like gambling and alcohol.

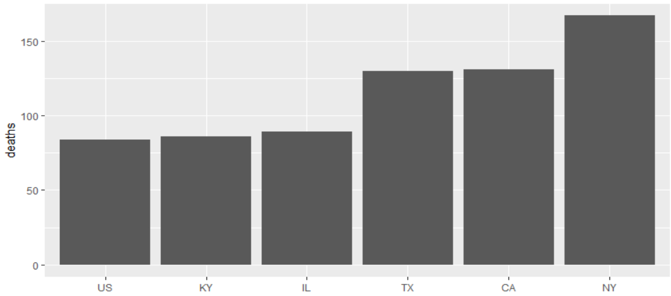

Figure 8: Death by state post Prohibition (1934-1944)

New York by this stage had it established itself as the capital of the mafia country wide, prohibition was enough to place it atop of the unpopular charts of the crime capital of the country. States like Illinois and Ohio dropped off dramatically. Illinois dropped over 120% and Ohio doesn't even feature among the top states. The most interesting state is that of California which throughout both periods held high levels of deaths for unknown reasons.

There were few results that I was surprised to see, the number of officers that have died to due to a heart attack on duty. 4.4% (971) of fallen officers were diagnosed as dying from a heart attack. 52.5% (11589) died from gun fire which was very unsurprising bearing in mind the gun laws and gun culture that exist in United States throughout their history.

A. MapReduce Results

From the outset of this project I wanted to compare different era of American history and to see if the same states time and time again recorded high number of police deaths. Using MapReduce I could setup an environment to test this question, approach and implementation to MapReduce were discussed in the section above, essentially, I took the top ten states from three different periods in American history which I believed police would be most at risk. I then stored these in a csv file before using a reducer to find the top twenty states of police deaths. From my interpretation of the MapReduce results, I found California seems to be following an increasing trend, with their rapidly rising population through the 20th century as seen in the 2nd dataset, it has led to increased criminal activity and a higher rate of officers dying. New York State was the only state to appear in the top ten in all three periods but did show signs of regression as time progressed. It should be no surprise to readers that Florida showed up in the top twenty states during the war on drugs with over 150 officers dying during the period with the obvious factor being increased activity around stopping drugs being imported and sold on the streets. US federal agents weren't assigned to any state, to accommodate this I have created an extra 'state' named US. There are many different results to take from the output of the MapReduce file. Some show correlation of high deaths e.g. New York, Texas, Kentucky. While others only show up when the black market is flourishing like Ohio and Illinois. With the clear majority of deaths involving criminal activity e.g. pursing a suspect and crashing, terrorist attack or being killed by gunfire. It obvious to take from background research these aren't freak accidents that some states rise and fall off once certain laws or operations have been implemented and then rescinded the like prohibition act.

B. Paired t-test

Set out to test if police deaths were equal under George Bush (2001-2008) compared to that of Barrack Obama. Before a test was conducted, I examined the before and after results on a scatterplot to get an idea of the result we should expect, incorporating a 45 Degree angle to show the before is equal to after.

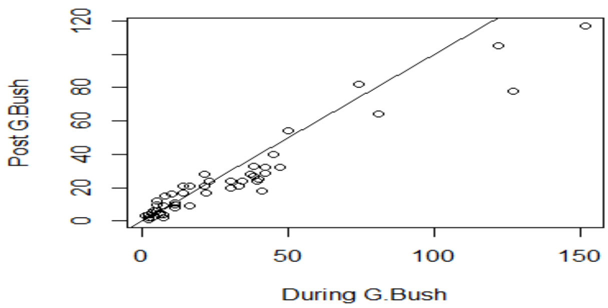

Figure 9: Deaths by state during G.Bush and B.Obama presidency

We can see from this scatterplot most plots fall on the line indicating little change between the two president's reigns, furthermore the dotted points are slightly uneven above and below line the further we progress, one cause of this is due to the 9/11 attacks that would drive up the deaths for New York under George Bush. After the t-test was executed in RStudio at an alpha value of 0.05 the obtained t-statistic of 3.05 and p-value of 0.0018, we can reject the null hypothesis in favor of the alternative hypothesis that states that the mean of officer's deaths under George Bush were not equal to the mean of officers under Barrack Obama

C. Corrleation

After testing the correlation from the two datasets we found the following;

The p-value of the test is 2.2^{-16}, which is less than the significance level alpha = 0.05. We can conclude that Deaths and Gun ownership levels are significantly correlated with a correlation coefficient of 0.87 and p-value of 2.2 ^{-16}.

The meaning behind this states that states with high gun ownership are strongly correlated with states with high rates of death among officers. In my opinion this is right, more guns in almost every case will lead to higher rates of death.

VI. CONCLUSIONS

From undertaking this assignment, I learnt a lot in regards to research and manipulating data with the use of R. I felt I got a better grasp of the power of R than I did in previous assignments. I had a clear idea of what I wanted to do with the dataset which helped shape the code and functions I would be using. In general I found out many interesting libraries like tidyverse and gmodels which came in handy towards the end of the assignment. It certainly made me apply myself with the limited time available with other submissions due at similar times,

If I was to undertake this project again I would certainly do things differently from the beginning. I felt I neglected the project and the quantity of work it involved until very late in the semester. I would maybe pick a different topic as I picked police deaths in the USA, it made for a very weak 2nd dataset as it was very hard to merge a different dataset with it. No countries I researched had open datasets on police deaths, although it would make for a very interesting project if Europe made a dataset available, with the different gun laws and culture it would make for good topic.

I would have spent more time with MapReduce and tried to implement Hadoop for more marks, approaching the question I was unsure of the work that MapReduce would require and if the work was satisfactory. Most certainly if I had more time and the right datasets were available I would do the project again but this time I was limited with research time, which I feel hampered my progress late on when it came to building a linear regression models, I had attempted to build a linear model but I didn’t fully understand the visuals hence why I left it out of the document but have the code at the bottom of my RScript.

I have commented all my code related to figures and enclosed the RScript in the ZIP folder.

VII. REFERENCES

[1] CRARY, D. (2016). *Police killings of blacks voted top story of 2014*. [online] AP. Available at: http://bigstory.ap.org/article/ad250438af4e4fae95d7e41f537661ef/ap-poll-police-killings-blacks-voted-top-story-2014 [Accessed 1 Dec. 2016].

[2] the Guardian. (2016). *The Counted: tracking people killed by police in the United States | US News | The Guardian*. [online] Available at: https://www.theguardian.com/us-news/series/counted-us-police-killings [Accessed 5 Dec. 2016].

[3] Kaggle.com. (2016). *FiveThirtyEight | Kaggle*. [online] Available at: https://www.kaggle.com/fivethirtyeight [Accessed 1 Dec. 2016].

[4] Nleomf.org. (2016). *National Law Enforcement Officers Memorial Fund: Law Enforcement Facts*. [online] Available at: http://www.nleomf.org/facts/enforcement/ [Accessed 4 Dec. 2016].

[5] Odmp.org. (2016). *The Officer Down Memorial Page (ODMP)*. [online] Available at: https://www.odmp.org/ [Accessed 4 Dec. 2016].

[6] Jstor.org. (2016). *An Analysis of Police Killed by Criminal Action: 1961-1963 on JSTOR*. [online] Available at: http://www.jstor.org/stable/1141788?seq=1#page_scan_tab_contents [Accessed 8 Dec. 2016].

[7] Demographic Data. (2016). *Gun Ownership Statistics by State - Demographic Data*. [online] Available at: http://demographicdata.org/facts-and-figures/gun-ownership-statistics/ [Accessed 7 Dec. 2016].

[8] Stats.indiana.edu. (2016). *State Census Counts: STATS Indiana*. [online] Available at: http://www.stats.indiana.edu/population/PopTotals/historic_counts_states.asp [Accessed 7 Dec. 2016].

[9] Washington Post. (2016). *More police officers die on the job in states with more guns*. [online] Available at: https://www.washingtonpost.com/news/wonk/wp/2016/07/08/more-police-officers-die-on-the-job-in-states-with-more-guns/ [Accessed 6 Dec. 2016].

[10] R, H. (2016). *How to create new column in dataframe based on partial string matching other column in R*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/19747384/how-to-create-new-column-in-dataframe-based-on-partial-string-matching-other-col [Accessed 13 Dec. 2016].

[11] Wickham, H. (2016). *Easily Install and Load 'Tidyverse' Packages [R package tidyverse version 1.0.0]*. [online] Cran.r-project.org. Available at: https://cran.r-project.org/web/packages/tidyverse/index.html [Accessed 15 Dec. 2016].