Individual Assignment

TECHNOLOGY PARK MALAYSIA

CT051-3-M-DM

Data Management

Assignment Part-2

APUMF2204DSBA(DE)(PR)

Student's TP: TP067696

Student's Name: Mr. Muhammad Arif Bin Jamaluddin

Lecturer's Name: Dr. MURUGANANTHAN VELAYUTHAM

Abstract

The topic of this assignment is mass shooting in USA where gun violence datasets from 2013-2018 is used to study about data management. This assignment required to prepared and explored the datasets selected which in these cases is gun violence whereby several justifications of pre-processing, transformation and feature engineering method is implemented using SAS studio. The main contents of this assignment part 2 is method section where it consists of initial data exploration, data pre-processing, exploratory data analysis (EDA) and hypothesis. Next is followed up by discussion where it summarizes the finding of all the method parts. Introduction and related work briefly explain on the background information about gun violence in United States whereby it explained the reason why mass shooting happens and statistics of mass shooting recorded in the past in United States. The purpose of this data analysis is to study the topic by doing statistical approach on the datasets to find the relationship and correlation between the variables.

Keyword; Gun violence, SAS studio, mass shooting, USA, data management

Table of Contents

Abstract.....2

Introduction.....4

Related work.....5

Method.....7

Discussion.....35

Conclusion.....37

References.....37

Introduction

Mass shooting has been a recent phenomenon in the United States of America where it happens very frequently especially in school, public areas or where there is a mass gathering of people.

Before going into further explanation on the statistics, what is the definition of mass shootings? Mass shootings can be defined as when a mass murderer that kills more than four people in an event excluding himself, usually in a single location or a place by Federal Bureau of Investigation (FBI) in the 1980s (Krouse and Richardson, 2015). These mass public shootings are rare events—they constitute less than 15 percent of all mass killings in the United States and are responsible for less than 0.5 percent of all homicides (Duwe, 2020). Although this statistic shows that the number is insignificant but it still poses threat to the American society.

There are a lot of statistics and research papers recorded by the US government where for example in 24 May 2022 where recently happen this year, an 18-year-old male teenager shoot and killed 21 people at Robb Elementary school in Uvalde, Texas. The statistics shows that in 2022, there is already 200 mass shootings incidents happen in USA where 27 of it comes from mass shooting incidents happen in school.

The trend of mass shootings is increasing in USA from 149 mass shootings and 27 deaths per day in 1999 to 248 mass shootings and 51 deaths per day in 2021 (Federico G de Cosío, 2022). The negative impacts of mass shootings can be seen on the economic cost for all type of gun violence where each year the US government spend $12.7 billion USD or the local government of each states spend roughly $34.8 million USD to deal with aftermath of gun violence (Everytown. The Economic Cost of Gun Violence, 2021). Staggering amount of $280 billion USD is the annual expenditure for taxpayers, survivors, families, employers, and communities to deal with the impact of gun violence or mass shootings.

There are different methods to defined mass shooting where the findings can be summarized by the traits of the mass shootings for example offender behaviors, types of firearms used and community level shootings.

It can be proved that by using data science and analytical tools such as statistics, gun shooting related fatality can be backed up using data science evidence to regulate gun ownership in United States of America to prevent more from this incident to happen. The flow of this papers starts with related work where it discussed about the research papers discussing the datasets

about mass shooting incidents in United States of America. Then, method section contained detail exploration of the dataset, pre-processing, feature engineering, EDA and hypothesis about the mass shooting in United States of America where SAS codes are implemented to study about this topic especially on exploring the data and evaluates it.

In the domain of data science most data are unstructured where it is incomplete such as missing values and outliers where several methods can be applied to deals with this noisy data. Then discussion part where the summary of findings can be discussed about the methodology part and eventually discussion to concluded everything for this assignment paper.

Related work

In this section discussed about research papers regarding mass shootings incidents in USA based on the data collected by the USA government (statistics and reports). Mass shootings definition according to Federal Bureau Investigation (FBI) is usually with the purpose to address criminal profiling procedures (Ressler, Burgess, and Douglas, 1988). There are diverse definitions of mass shootings from the news or media outlets that reports on these mass shootings incidents where it can confuse the public on the relationship between gun law and mass shootings trends (Elsass, H. Jaymi, Jaclyn Schildkraut, Mark C. Staffor, 2016).

In USA, the homicide rate caused by gun violence is 26 times higher compare to developed countries (GunPolicy.org,2022) and it shows that in several states in USA with fragile and loose gun law restrictions and highest gun ownership per person shows greater rates of mass shootings incidents (Paul M. Reeping et al,2019). Because of this event, this sparks debate among the general public to have a better and much more strict gun control legislation to prevent same tragedies from happening again. Thus, perhaps with tighter gun control the probability of mass shootings will likely less to happen.

The victims of these mass shootings can be statistically analyzed based on the death age by groups, ethnicity, and the demographic or the regions in the USA where the shootings take places. It shows that 64% of all death rates is people from age range between 15-24 and 25-34 and also most of deaths are among males. The cumulative death by ethnic group shows that the highest is black/African American, followed by white, Hispanic, Asian/pacific islander and American Indian/ Alaskan native. The trend can be seen for cumulative deaths by month where

January and July are the peaks month correlated with mass shootings and after that it starts to decline from august to December.

Mass shootings patterns or behavior can be identified by this characteristic such as it was executed by people who carry firearms or gun illegally, that person shows some sign of not normal behaviors such as mental illness and had history of domestic violence at home. There are several definitions what constitute mass shootings for example the casualty threshold (deaths or injuries caused by the firearms), location of the mass shootings takes place for example school, workplace, residential area, malls and motivation of the shooters. For example, there are several threshold definitions on mass shooting such as two or more injured victims is considered threshold by (Lott and Landes,2000). Whereas other definition used 6 victim casualty thresholds by (Kleck, 2016).

There are differences between mass shootings that involves domestic and gang violence than indiscriminate killings in public events such school shootings where several researchers and analysts state that it should categorized separately. Definitions proposed by (Duwe, Kovandzic, and Moody,2002) states that mass communal shootings exclude that both the sufferer and convicts were participating in illegal activities for example crime, gang members and drug dealings.

Furthermore, mass shootings events happen in public area where the offender choose the victims randomly proposed by (Gius,2015) (Luca, Malhotra, and Poliquin, 2020). But then again, this approach currently has flaw such as its overlooking significant percentage of gun violence that comes from family or assault linked murder (Fox and Levin, 2015). It is a very subjective in a way that how the offender choosing the victims indiscriminately or the incidents is related to crime or gang activities. Thus, accurate information from most of the police reports or news about this mass shootings events or not included whether what is the shooter's main motivation or is there any correlation with gangs' members.

From 2009 until 2020, there were 1363 casualties of mass shooting where 947 people were killed and 240 were injured. Based on this numbers at least the number of children and teens death are 362 and 21 law enforcements are killed and the rest 35 are wounded. There is interesting pattern noticed where most of shooters acted alone and overall 32% percent of the mass shooters which roughly at 92 people eventually end up killing themselves by suicide. The rest of the perpetrator end up killed by the police which are around 24 and the remaining 145 were arrested into custody by the police.

Although it is known that mass shootings that occurs in public especially school gathers more attention media, but most of it happens at home around 61% and the rest happen at public places such as school, bars, restaurants and malls. Unfortunately, there is a lack of research regarding gun violence in USA and also the government does not provide sufficient funding and thus this leads to donations and private funding to conduct research on understanding more the cause of mass shooting in USA.

Thus, the objectives of this assignment are to explore mass shootings/ gun violence datasets using SAS codes to understand the correlation between dependent and independent variables in the datasets obtained.

Method

The datasets used for this assignment obtained from Kaggle where the title is obtained from Gun Violence Data where it shows all the statistics and record of 260k gun violence accidents in USA from 2013-2018. The aim of this datasets is to make learned predictions about upcoming trends regarding the gun violence using statistical method. The detail exploration and explanation are conducted using SAS code.

Initial Data Exploration

The first step is using visual exploration to understand what is in a dataset and the characteristics of the data. Thus, using SAS software the codes below is generated by importing excel into SAS Studio under WORK.IMPORT1 function and the code is executed to obtain information as shown in figure 1 and 2. In this assignment the imported data obtained originally contained huge amount of observations data about mass shooting from 2013 until 2018.

```
/* Generated Code (IMPORT) */
/* Source File: Gun Violence data 2013-2014.xlsx */
/* Source Path: /home/u61522473/sasuser.v94/Assignment */
/* Code generated on: 05/07/2022 22:21 */

%web_drop_table(WORK.IMPORT1);


FILENAME REFFILE '/home/u61522473/sasuser.v94/Assignment/Gun Violence data 2013-2014.xlsx';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLSX
    OUT=WORK.IMPORT1;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

Line 1, Column 1  UTF-
```

Figure 1

Based on figure 2 it shows that the data has 52133 observations 29 variables/attributes.

CODE LOG RESULTS OUTPUT DATA

Table of Contents

The CONTENTS Procedure

Data Set Name	WORK.IMPORT1	Observations	52133
Member Type	DATA	Variables	29
Engine	V9	Indexes	0
Created	05/07/2022 22:24:41	Observation Length	12760
Last Modified	05/07/2022 22:24:41	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information

Data Set Page Size	131072
Number of Data Set Pages	5214
First Data Page	1

Figure 2

Categorize the data into categorical and continuous. Table 1 shows the categorical and continuous data type for this dataset where it's divided into categorical and continuous variables.

Table 1

Categorical	**Continuous**
1. State 2. City_or_country 3. Address 4. Incident_url 5. Source_url 6. Gun_type 7. incident_characteritics 8. location_description 9. notes 10. participant_gender 11. participant_name 12. participant_relationship 13. participant_status 14. participant_type 15. sources 16. participant_age_group 17. participant_age 18. gun_stolen	1. incident_id 2. date 3. n_killed 4. n_injured 5. incident_url_fields_missing 6. congressional_district 7. latitude 8. longitude 9. n_guns_involved 10. state_house_district 11. state_senate_district

Table 2 below shows the data description of the variables from the gun violence datasets obtained from Kaggle from 2013 and 2014 and the data contained 29 attributes.

Table 2

Variables/attributes	Descriptions
incident_id	Incident identification number
date	Date of the shooting took place
state	The state where the shooting took place
city_or_county	The city where the shooting took place
address	The address of the event
n_killed	The number of victims killed
n_injured	The number of victims injured
incident_url	Incident sources url
source_url	Sources of the website link about the detailed event
incident_url_fields_missing	The incident url fields missing
congressional_district	Legislative districts

gun_stolen	Gun stolen
gun_type	Type of gun used
incident_characteristics	The severity of the casualties
latitude	Latitude value of the event
location_description	The description of the event location
longitude	Longitude value of the event
n_guns_involved	The number of guns involved
notes	Detail description of the event
participant_age	The age of the participant
participant_age_group	The age of the participant in group manner
participant_gender	The participant gender
participant_name	The participant's name
participant_relationship	Relationship between offender and victim
participant_status	The offender status after the event
participant_type	The participant type
sources	The sources of the news
state_house_district	State house district codes
state_senate_district	State senate district codes

The next part is to identify the data into characters or numerical type as shown in table 3 below. Based on table 3 it shows that there are 18 Char type data and 11 Num type data and figure 3 is used to execute the SAS code to get the output in table 3.

```
ods noproctitle;
ods select attributes position;

PROC datasets;
    CONTENTS DATA=WORK.IMPORT1 order=varnum;
run;
```

Figure 3

Table 3

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	incident_id	Num	8	BEST.		incident_id

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
2	date	Num	8	MMDDYY10.		date
3	state	Char	20	$20.	$20.	state
4	city_or_county	Char	42	$42.	$42.	city_or_county
5	address	Char	97	$97.	$97.	address
6	n_killed	Num	8	BEST.		n_killed
7	n_injured	Num	8	BEST.		n_injured
8	incident_url	Char	50	$50.	$50.	incident_url
9	source_url	Char	255	$255.	$255.	source_url
10	incident_url_fields_missing	Num	8	BEST.		incident_url_fields_missing
11	congressional_district	Num	8	BEST.		congressional_district
12	gun_stolen	Char	3808	$3808.	$3808.	gun_stolen
13	gun_type	Char	3808	$3808.	$3808.	gun_type
14	incident_characteristics	Char	618	$618.	$618.	incident_characteristics
15	latitude	Num	8	BEST.		latitude
16	location_description	Char	56	$56.	$56.	location_description
17	longitude	Num	8	BEST.		longitude
18	n_guns_involved	Num	8	BEST.		n_guns_involved
19	notes	Char	273	$273.	$273.	notes
20	participant_age	Char	148	$148.	$148.	participant_age
21	participant_age_group	Char	318	$318.	$318.	participant_age_group
22	participant_gender	Char	224	$224.	$224.	participant_gender
23	participant_name	Char	401	$401.	$401.	participant_name
24	participant_relationship	Char	182	$182.	$182.	participant_relationship

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
25	participant_status	Char	328	$328.	$328.	participant_status
26	participant_type	Char	408	$408.	$408.	participant_type
27	sources	Char	1631	$1631.	$1631.	sources
28	state_house_district	Num	8	BEST.		state_house_district
29	state_senate_district	Num	8	BEST.		state_senate_district

The next part is the discussion of descriptive statistics where it consisted of summary statistics, univariate and bivariate analysis, missing values and outlier treatment for the missing values data. In this section it contained the identification values of the summarising properties for each attribute including frequency and spread e.g., value ranges of the attributes, frequency of values, distributions, medians, means, variances, and percentiles.

For example, to display mean, standard deviation, minimum and maximum value for the variables of n_killed and n_injured the code in SAS is used as below,

```
27 Proc MEANS Data=WORK.IMPORT1;
28 var n_killed n_injured;
29 run;
30
```

Figure 4

The MEANS Procedure

Table 4

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
n_killed	n_killed	52132	0.2469500	0.5141859	0	11.0000000

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
n_injured	n_injured	52132	0.4600054	0.7535712	0	19.0000000

The values of mean, standard deviation, minimum and maximum for n_killed and n_injured are shown above in table 4. For example, to display the mean, standard deviation, minimum and maximum value for all numerical variables the SAS code is used as below at figure 5,

```
31 Proc MEANS Data=WORK.IMPORT1;
32 run;
33
```

Figure 5

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
incident_id	incident_id	52132	175217.69	69672.85	92114.00	1074613.00
date	date	52132	19908.90	104.6464114	19359.00	20088.00
n_killed	n_killed	52132	0.2469500	0.5141859	0	11.0000000
n_injured	n_injured	52132	0.4600054	0.7535712	0	19.0000000
incident_url_fields_missing	incident_url_fields_missing	52132	0	0	0	0
congressional_district	congressional_district	51944	8.0201756	8.5784115	0	53.0000000
latitude	latitude	51972	37.3192984	4.9555418	19.4554000	71.3001000
longitude	longitude	51972	-88.8682308	13.9477790	-165.4440000	-67.7617000
n_guns_involved	n_guns_involved	5917	1.3601487	5.0477357	1.0000000	280.0000000
state_house_district	state_house_district	45180	55.4715139	39.4884873	1.0000000	217.0000000
state_senate_district	state_senate_district	46492	20.3678052	14.0108171	1.0000000	94.0000000

Figure 6

In Figure 6 above, all the mean, standard deviation, minimum and maximum value for all numerical variables is shown by the output. For example, to display MIN, MAX and SUM for the variable n_killed and n_injured is shown in SAS code below in figure 7.

```
34 Proc MEANS Data=WORK.IMPORT1 min max sum;
35 var n_killed n_injured;
36 run;
37
```

Figure 7

The MEANS Procedure

Table 5

Variable	Label	Minimum	Maximum	Sum
n_killed	n_killed	0	11.0000000	12874.00
n_injured	n_injured	0	19.0000000	23981.00

For example, to display the number of missing values in all variables the SAS code is used shown below.

```
38 Proc MEANS NMISS Data=WORK.IMPORT1;
39 run;
40
```

Figure 8

The MEANS Procedure

Table 6

Variable	Label	N Miss
incident_id	incident_id	1
date	date	1
n_killed	n_killed	1
n_injured	n_injured	1
incident_url_fields_missing	incident_url_fields_missing	1
congressional_district	congressional_district	189
latitude	latitude	161
longitude	longitude	161
n_guns_involved	n_guns_involved	46216
state_house_district	state_house_district	6953
state_senate_district	state_senate_district	5641

It shows that variable of n_guns_involved shows the highest N Miss value at 46216 and all of the previous section discussed is data from 2013 and 2014. For this section comparison is made between data from 2014 and 2015 about the gun violence datasets and investigate the difference between different years and its relationship with the variables. The same SAS code is used and the output for 2014 and 2015 is shown below and figure 9 and 10 shows the results of the observations of both years.

Data Set Name	WORK.IMPORT	Observations	51854
Member Type	DATA	Variables	29
Engine	V9	Indexes	0
Created	07/07/2022 23:06:10	Observation Length	12600
Last Modified	07/07/2022 23:06:10	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		

Figure 9

The CONTENTS Procedure

Data Set Name	WORK.IMPORT1	Observations	53579
Member Type	DATA	Variables	29
Engine	V9	Indexes	0
Created	07/07/2022 23:07:01	Observation Length	18872
Last Modified	07/07/2022 23:07:01	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			

Figure 10

The number of observations in year 2014 of gun violence data is 51854 meanwhile for 2015 is 53579. Figure 11 shows the SAS code to get the output of the variables in figure 12 and 13.

```
ods noproctitle;
ods graphics /imagemap=on;

proc means Data=WORK.IMPORT chartype mean std min max median n nmiss vardef=df
      qrange qmethod=os;
   var  n_killed n_injured latitude longitude n_guns_involved;
quit;
```

Figure 11

Variable	Label	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Quartile Range
n_killed	n_killed	0.2421654	0.4989407	0	8.0000000	0	51853	1	0
n_injured	n_injured	0.4436002	0.7013601	0	16.0000000	0	51853	1	1.0000000
latitude	latitude	37.3176407	4.9587200	19.4554000	71.3001000	38.1917000	51695	159	7.1763000
longitude	longitude	-88.8624531	13.9395607	-165.4440000	-67.7617000	-85.5315000	51695	159	13.5207000
n_guns_involved	n_guns_involved	1.3627299	5.0904244	1.0000000	280.0000000	1.0000000	5817	46037	0

Figure 12 of 2014 datasets of gun violence

Variable	Label	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Quartile Range
n_killed	n_killed	0.2516658	0.5122075	0	16.0000000	0	53579	0	0
n_injured	n_injured	0.5033129	0.7150228	0	19.0000000	0	53579	0	1.0000000
latitude	latitude	37.6959244	5.2043980	19.4331000	67.2524000	38.6413000	53148	431	7.5741500
longitude	longitude	-89.4720828	14.6001496	-166.0970000	-67.8424000	-86.1503000	53148	431	14.7165500
n_guns_involved	n_guns_involved	1.7738892	9.6111950	1.0000000	400.0000000	1.0000000	20773	32806	0

Figure 13 of 2015 datasets of gun violence

It can be seen that n_killed, n_injured, latitude, longitude and n_guns_involved values in 2015 data shows much higher values of mean, standard deviation, minimum, maximum, median, N, N Miss and quartile range. The only similarities the values of n_killed, n_injured, n_guns_involved for both datasets are for minimum, median, and quartile range. The only odd data from between this two can be seen on n_guns_involved at median for 2014 datasets where it is lower than the one in 2015 which is 20773 and also the same for N Miss where in 2014 the value is 46037 which is much lower than 2015 at 32806. The next part is univariate analysis where analysis is done to know if the variable type is categorical or continuous. For continuous variable it is used to understand central tendency and the spread variable using data visualization tools to highlight the missing and outlier values. The first part is the explanation

of the data visualization using boxplot and the latter part is histogram. The variables selected are n_killed, n_injured and n_guns_involved for the boxplot.

```
ods graphics/ reset width=5in height=3 in imagemap;

proc sgplot DATA=WORK.IMPORT;
    vbox n_killed / fillattrs=(color=cxFF6060);
    yaxis grid;
run;

ods graphics / reset;
```

Figure 14

The SAS code in figure 14 is executed to get the boxplot as shown below.

Figure 15

The figure below is SAS code for n_injured variables 2014 datasets

```
ods graphics/ reset width=5in height=3 in imagemap;

proc sgplot DATA=WORK.IMPORT;
    vbox n_injured / fillattrs=(color=cxFF6060);
    yaxis grid;
run;

ods graphics / reset;
```

Figure 16

Figure 17

```
ods graphics/ reset width=5in height=3 in imagemap;

proc sgplot DATA=WORK.IMPORT;
    vbox n_guns_involved / fillattrs=(color=cxFF6060);
    yaxis grid;
run;

ods graphics / reset;
```

Figure 18

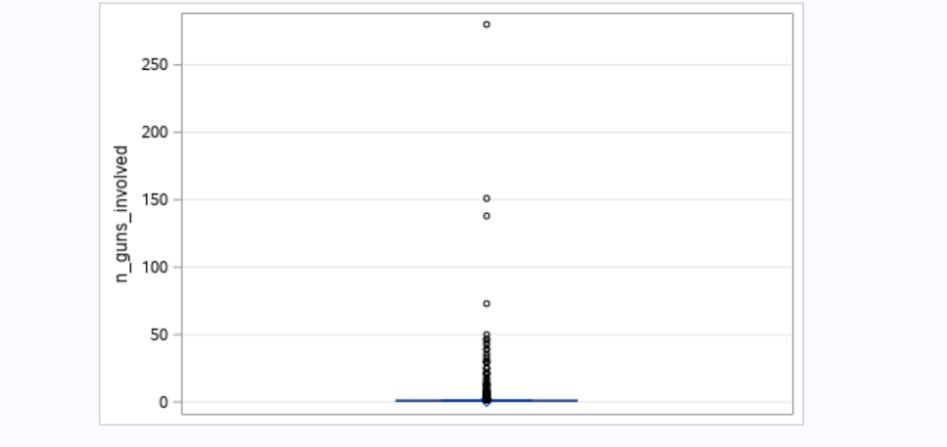


Figure 19

Figure 14-19 shows the SAS code and the boxplot of 2014 datasets of gun violence data. The next part is the SAS code and boxplot for 2015 datasets and comparison is done if there are any differences on the shape of boxplot. Figure 20 until 22 shows the boxplot of 2015 gun violence datasets.

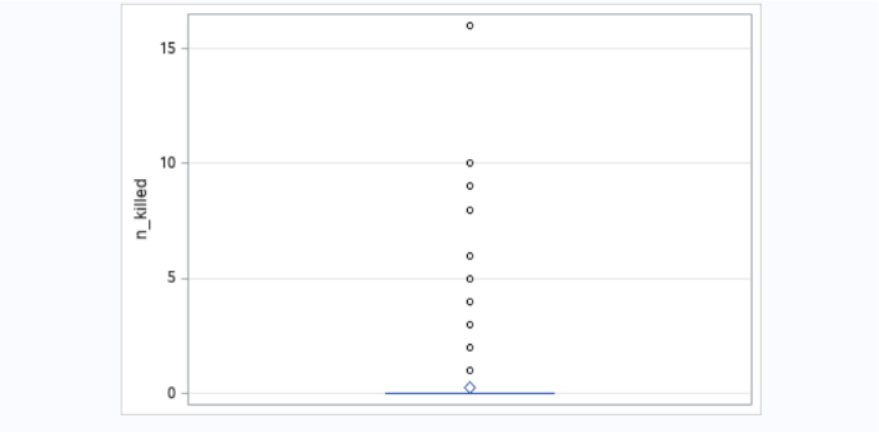


Figure 20

Figure 21

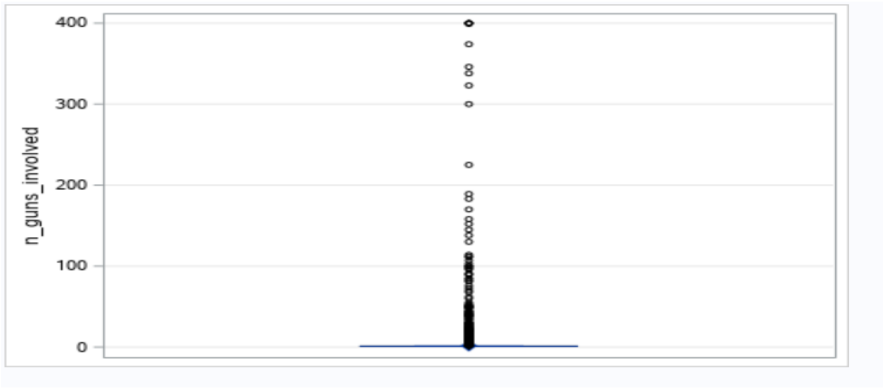
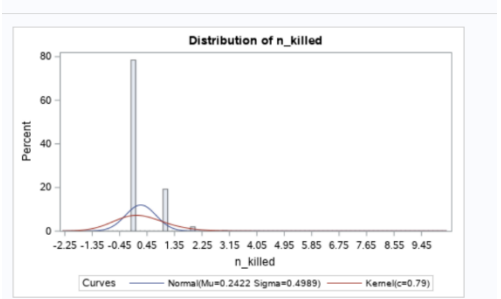


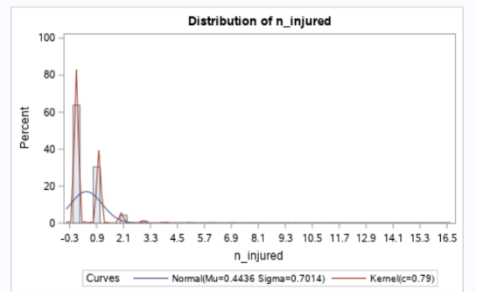
Figure 22

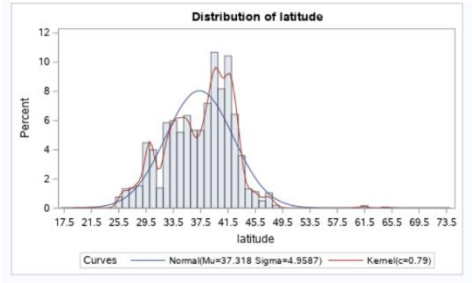
The next section is discussion of histogram to see the distributions of the variables. The main objective of histogram is to graphically sum up the distribution of univariate data set and again the same step where comparison is made for histogram about datasets on 2014- and 2015-gun violence data.

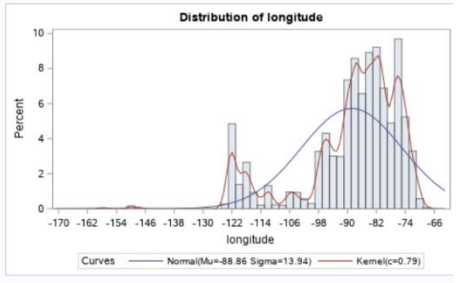
```
ods noproctitle;
ods graphics / reset width=5in height=3 in imagemap;

proc univariate data=WORK.IMPORT;
    ods select Histogram;
    var n_killed n_injured latitude
        longitude n_guns_involved;
    histogram n_killed n_injured latitude
        longitude n_guns_involved / normal kernel;
run;
```

Figure 23









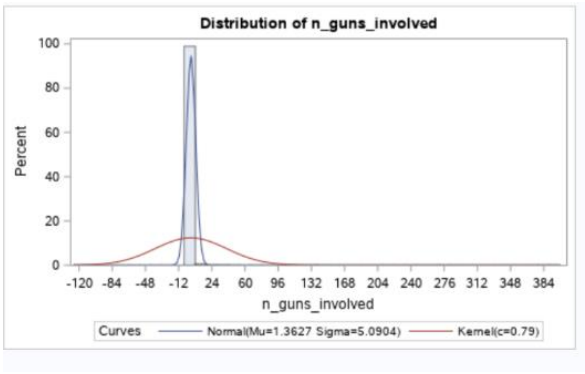
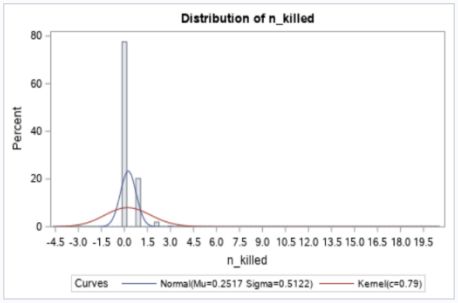
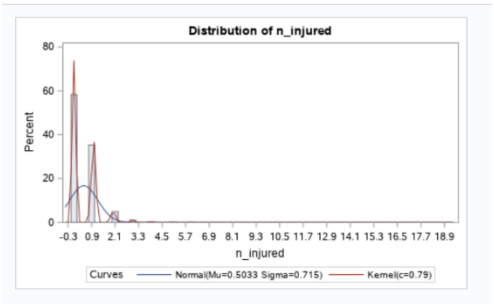
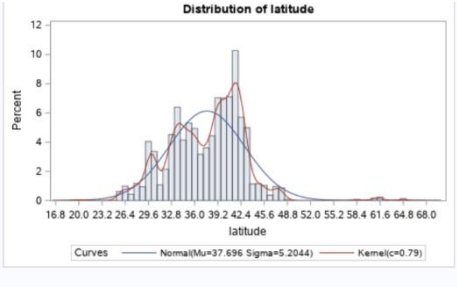
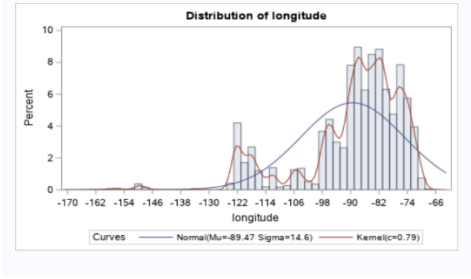


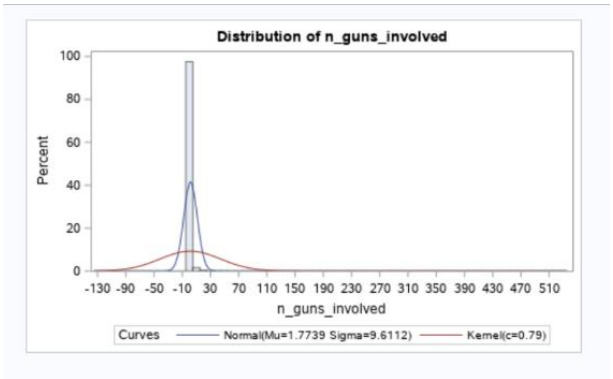
Figure 23 shows the SAS code for 2014-gun violence datasets and below it is the distribution of the histogram variables for the 2014-gun violence datasets. The same method is applied to the 2015 datasets and the histogram for the 2015-gun violence datasets is shown below.











Data Pre-processing

In this part it is required to investigate the method to handle the incomplete, noisy and inconsistent data and the applied technique is further explained in this section here. In this section the method of handling the noisy data my removing the features or instances is not allowed in this assignment and thus another technique should be explored.

Before going into the details of the technique data pre-processing is a part of data preparation where its transforming raw data into useful data. In a traditional data pre-processing is an introduction step before data mining process and distinct ways are used in data pre-processing. These tools are sampling, transformation, denoising, imputation, normalization and feature extraction. In this assignment normalization and imputation method are used for the data pre-processing.

1. **Imputation**

The SAS code below shows the imputation SAS code where the median data is used for the variables. This part here is the section for the 2014-gun violence datasets and it can be viewed on figure 27 and 30 the missing data frequencies values

```
data gun_data;
set WORK.IMPORT;
if n_killed='' then n_killed=0;
if n_injured='' then n_injured=0;
if n_guns_involved='' then n_guns_involved=1;
run;
```

Figure 24

```
ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
run;
```

Figure 25

```
proc freq data=work.gun_data;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: .,A,B, etc = Missing";
    format n_killed n_injured n_guns_involved _nmissprint.;
    tables  n_killed n_injured n_guns_involved / missing nocum;

run;
```

Figure 26

Missing Data Frequencies
Legend: .,A,B, etc = Missing

n_killed		
n_killed	Frequency	Percent
Non-missing	51854	100.00

n_injured		
n_injured	Frequency	Percent
Non-missing	51854	100.00

n_guns_involved		
n_guns_involved	Frequency	Percent
Non-missing	51854	100.00

Figure 27

```
proc freq data=work.gun_data noprint;
    table n_killed * n_injured * n_guns_involved / missing out=work._MissingData_;
    format n_killed n_injured n_guns_involved _nmissprint.;
run;
```

Figure 28

```
proc print data=work._MissingData_ noobs label;
     title3 "Missing Data Patterns across variables";
     title h=2 "Legend: .,A,B, etc = Missing";
     format n_killed n_injured n_guns_involved _nmissprint.;
     label count="Frequency" percent="Percent";
run;
```

Figure 29

Legend: .,A,B, etc = Missing

n_killed	n_injured	n_guns_involved	Frequency	Percent
Non-missing	Non-missing	Non-missing	51854	100

Figure 30

```
title3;
/* clean up */
proc delete data=work._MissingData_;
run;
```

Figure 31

This part here discussing the section of 2015-gun violence datasets for imputation SAS code for the missing values. The same method for SAS code for 2015- gun violence is applied here from the SAS code used in 2014 gun-violence datasets as shown in figure 24,25,26,28,29 and 31. Figure 32 and 33 indicates the missing data frequencies for the 2015 datasets.

Missing Data Frequencies

Legend: .,A,B, etc = Missing

n_killed		
n_killed	Frequency	Percent
Non-missing	53579	100.00

n_injured		
n_injured	Frequency	Percent
Non-missing	53579	100.00

n_guns_involved		
n_guns_involved	Frequency	Percent
Non-missing	53579	100.00

Figure 32

Legend: .,A,B, etc = Missing

n_killed	n_injured	n_guns_involved	Frequency	Percent
Non-missing	Non-missing	Non-missing	53579	100

Figure 33

The tables indicates that by using this imputation method there will be nor more missing values indicated by the imputation method.

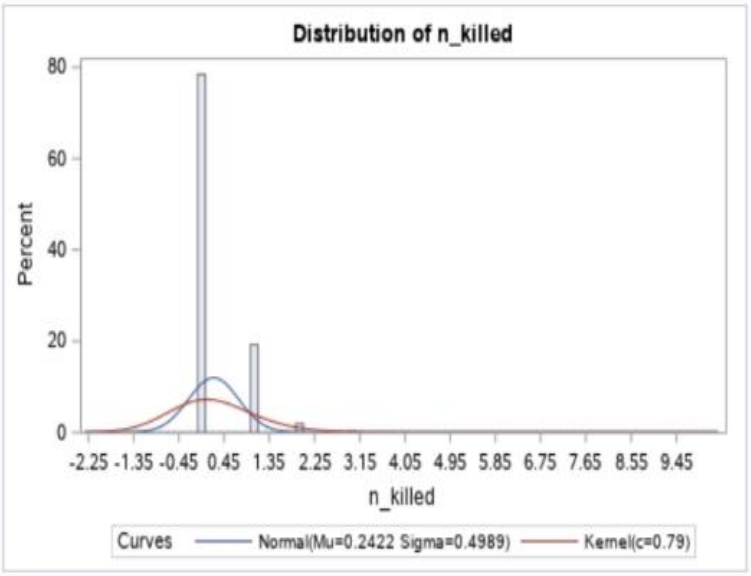
2. Normalization

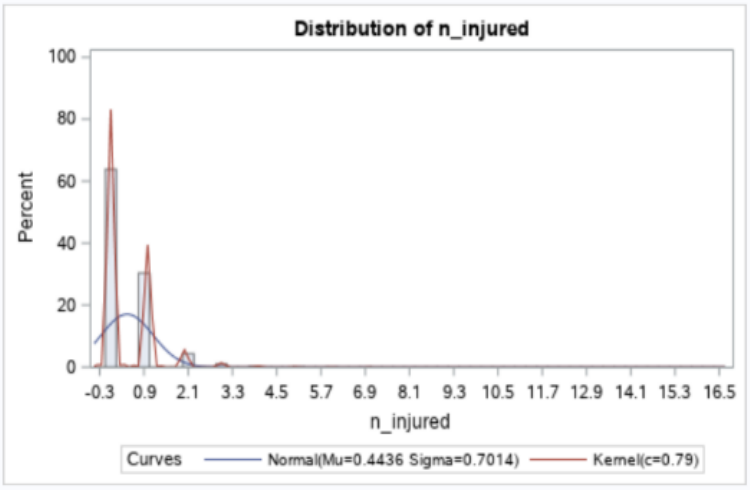
In figure 34 below it shows the code to normalized the data and as for this code it is applied for both 2014 and 2015.

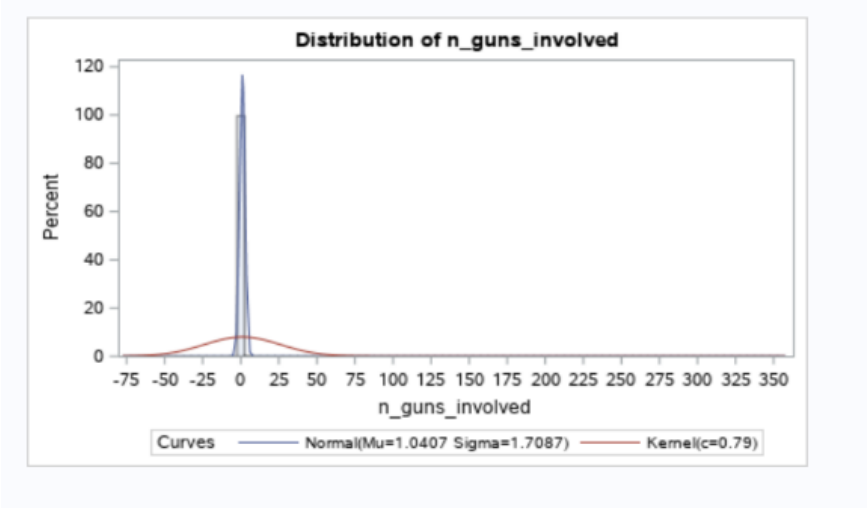
```
ods noproctitle;

proc stdize data=work.gun_data method=std nomiss out=work.Stdizedata;
   var n_killed n_injured n_guns_involved;
run;

ods noproctitle;
ods graphics / reset width=5in height=3 in imagemap;

proc univariate data=work.gun_data;
   ods select Histogram;
   var n_killed n_injured n_guns_involved;
   histogram n_killed n_injured n_guns_involved / normal kernel;
run;
```

Figure 34

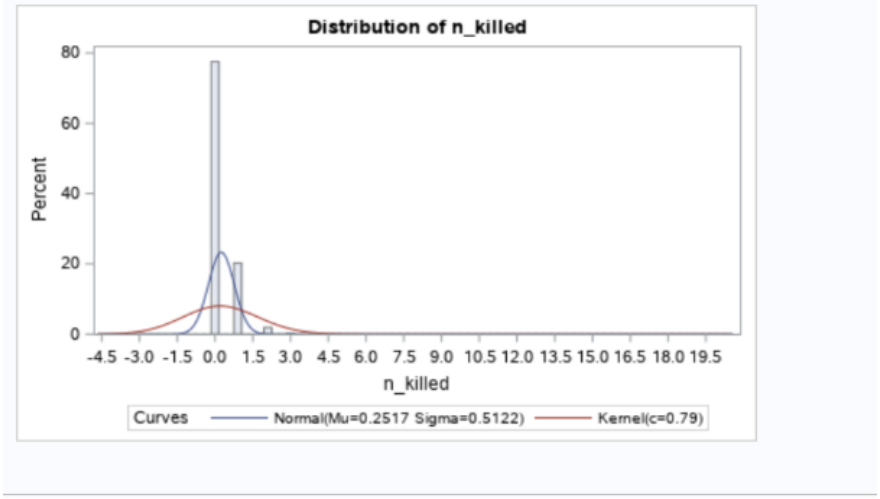
The three figures below show the 2014- gun violence datasets histogram where three variables are selected such as n_killed, n_injured and n_guns_involved.

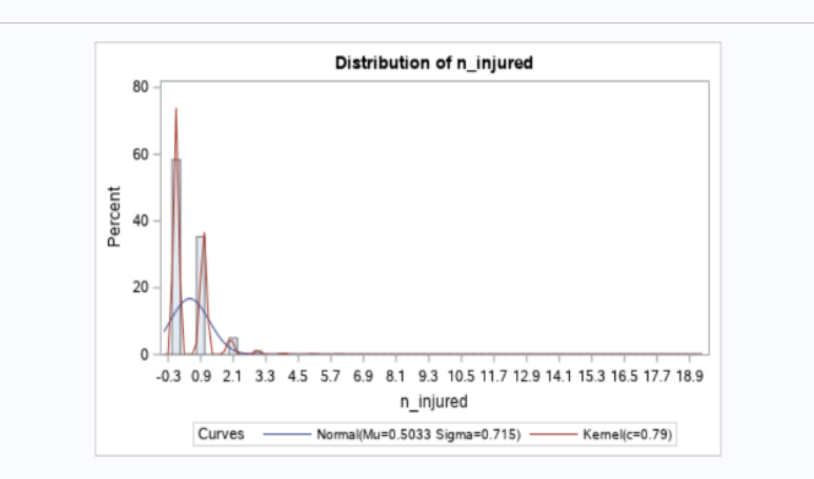

Distribution of n_killed
Percent
80
60
40
20
0
-2.25 -1.35 -0.45 0.45 1.35 2.25 3.15 4.05 4.95 5.85 6.75 7.65 8.55 9.45
n_killed
Curves
Normal(Mu=0.2422 Sigma=0.4989)
Kernel(c=0.79)

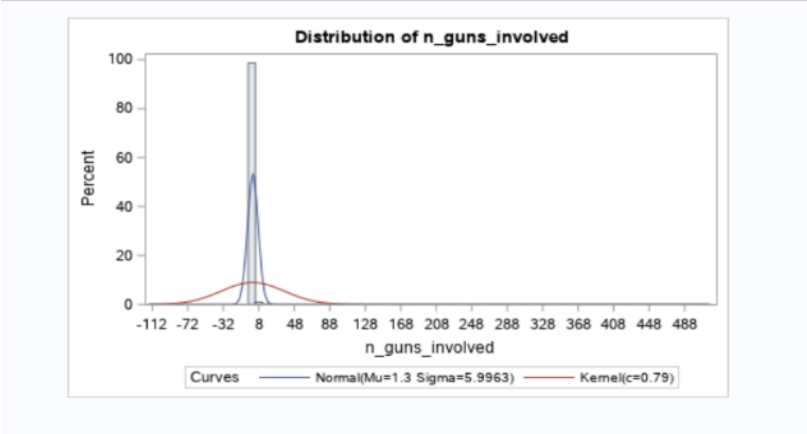

Distribution of n_injured
Percent
100
80
60
40
20
0
-0.3 0.9 2.1 3.3 4.5 5.7 6.9 8.1 9.3 10.5 11.7 12.9 14.1 15.3 16.5
n_injured
Curves
Normal(Mu=0.4436 Sigma=0.7014)
Kernel(c=0.79)



The three figure below on the other hand is the histogram visualization for the 2015-gun violence datasets.







Feature engineering

It is a process of selecting, manipulating, and converting raw data into features that can be implemented in supervised learning. It is a machine learning technique that leverage data to create new variables that are not available in training set. The main goal is to simplify and fastening up the data transformations and the same time increases model accuracy.

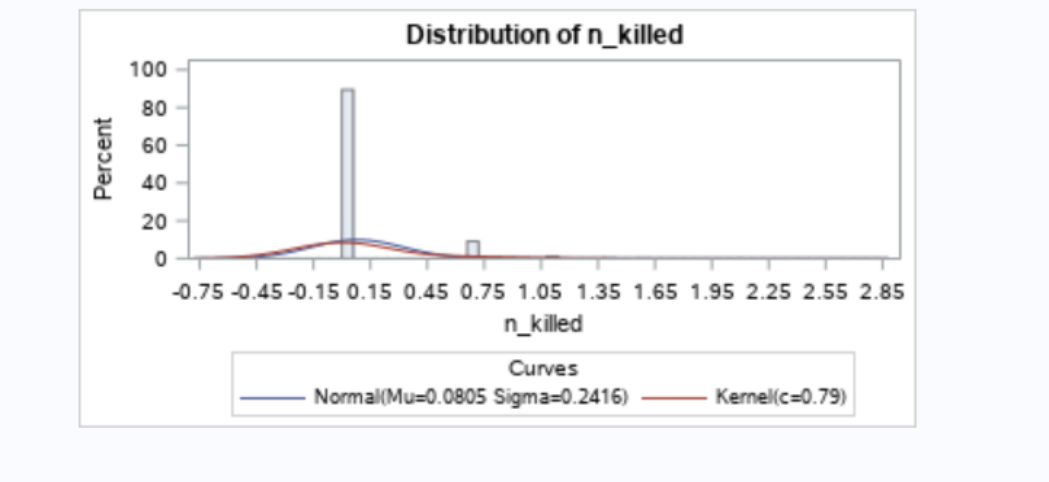
Logarithm transformation is one of the most commonly used mathematical transformations in feature engineering whereby it is used to handle skewed data and hence after the transformation, the data distribution becomes really close to normal. The code in figure 35

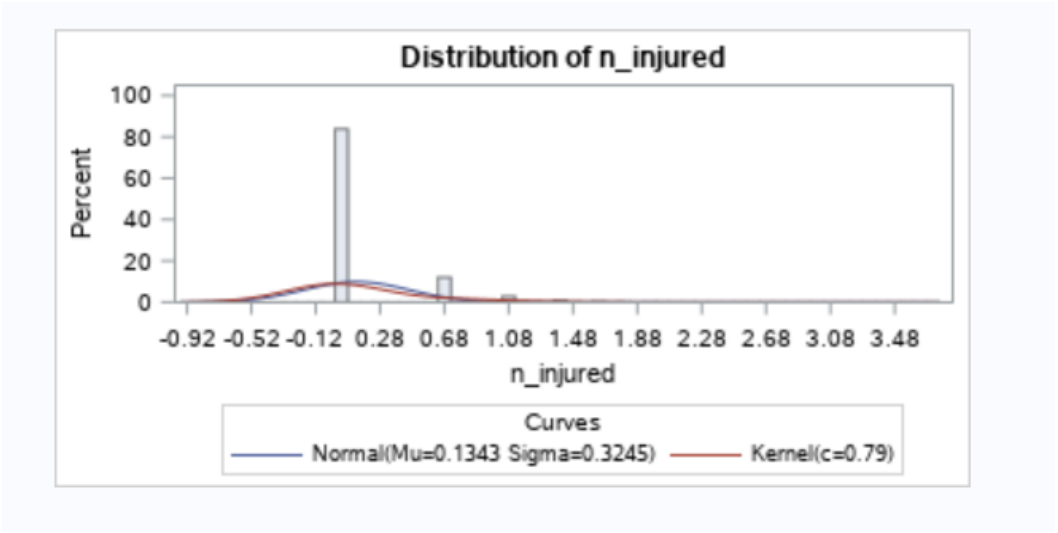
below shows the SAS code use for log transformation and hence the following part shows the histogram distribution of the 2014- and 2015-gun violence datasets where it is becoming less skewed compare without feature engineering/ log transformation.

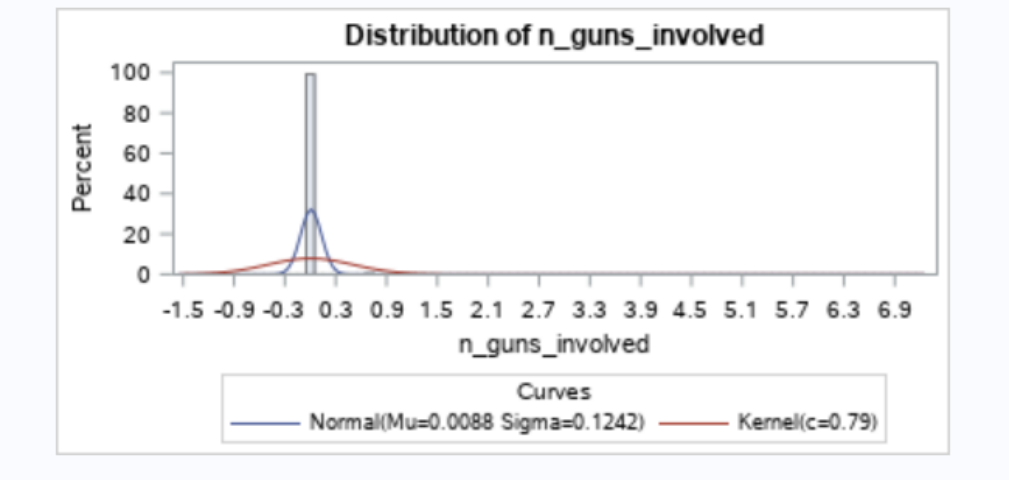
```
data work.transformation;
    set work.gun_data;
    n_killed=log(n_killed);
    n_injured=log(n_injured);
    n_guns_involved=log(n_guns_involved);
run;
```

Figure 35

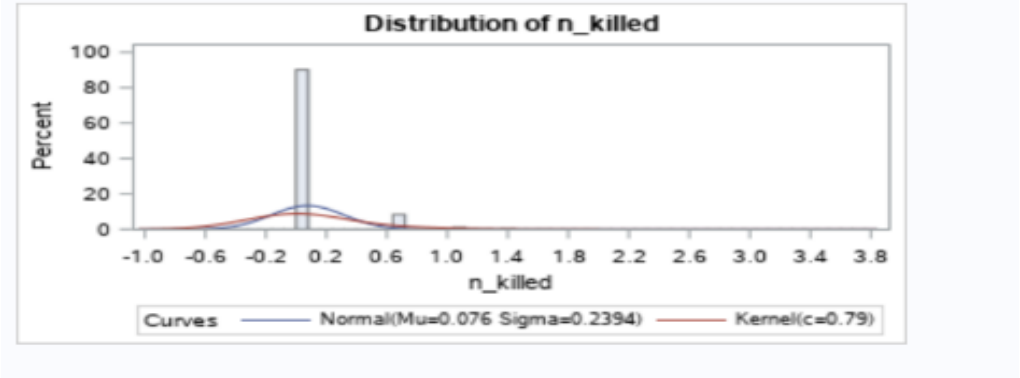
The three figure below shows the histogram distribution of the variables after undergoing log transformation

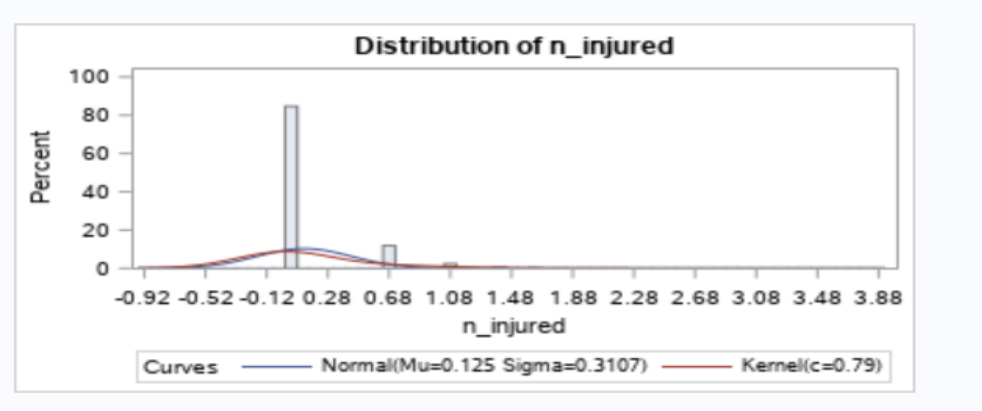


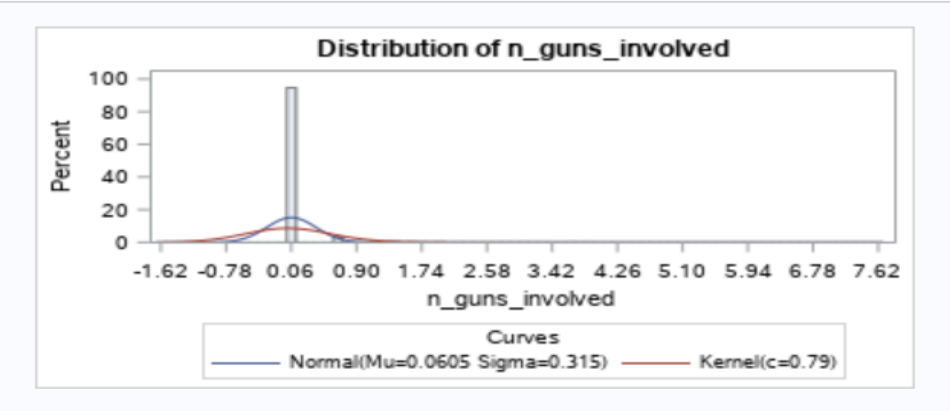




The three figures below show the log transformation histogram distribution for the 2015 datasets







Exploratory Data Analysis (EDA)

Exploratory data analysis main function is to help look at data before making assumptions where it can be use to detect any errors such as understanding better and well about the data patterns, identify outliers, and find relationship between the variables. There are several exploratory data analysis tools such as clustering and dimension reduction, univariate, bivariate, multivariate visualizations, k- means clustering and predictive models.

Figure 36 shows the SAS code for the scatter plot with the heatmap and correlation with the regression line at it shows that there is a relationship between n_guns_involved and n_killed based on the scatter plot in figure 37 as the number of guns involved increases the number of killed people is increasing as well.

```
ods graphics / reset width=4 in height=2 in imagemap;

proc sgplot data=work.transformation;
    heatmap x=n_killed y=n_guns_involved / name='HeatMap';
    reg x=n_killed y=n_guns_involved / nomarkers;
    gradlegend 'HeatMap';
    keylegend / linelength=20 fillheight=2.5pct fillaspect=golden;
run;

ods graphics / reset;
```

Figure 36

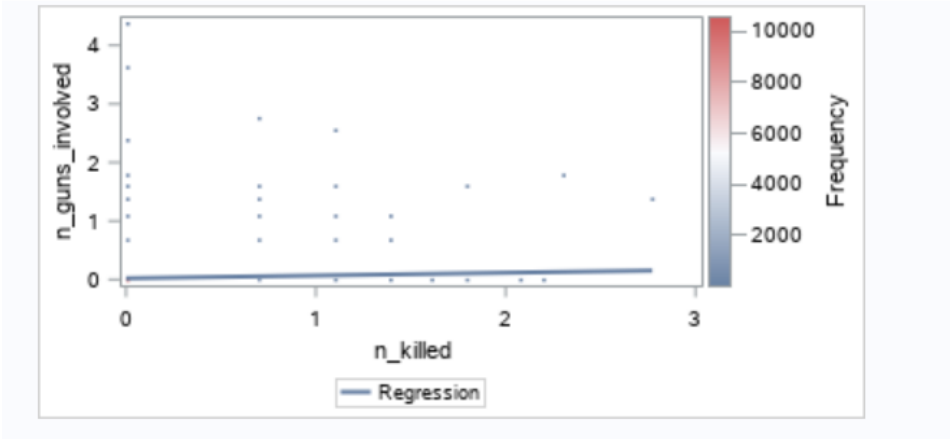


Figure 37

Figure 38 shows the SAS code that explained figure 39 whereby it shows that there is a relationship between n_guns_involved and n_injured where by as the number of guns involved increases the number of people injured increases.

```
ods graphics / reset width=4.2in height=2.5in imagemap;

proc sgplot data=work.transformation;
     reg x=n_guns_involved y=n_injured / nomarkers;
     scatter x=n_guns_involved y=n_injured /;
     xaxis grid;
     yaxis grid;
run;

ods graphics / reset;
```

Figure 38

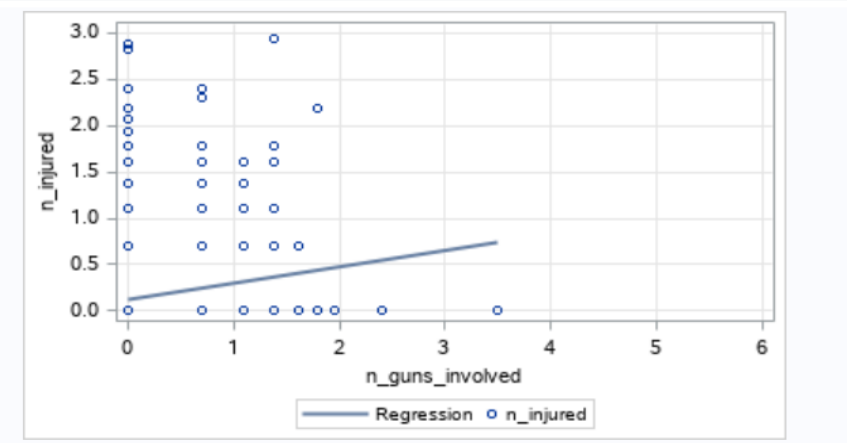


Figure 39

Figure 40 shows the SAS code for the pearson correlation whereby the target variable used is number of guns involved and table 7 shows the coefficients relationship between n_guns_involved and other numerical variables from the datasets.

```
ods noproctitle;
ods graphics/ imagemap=on;

proc corr data=work.gun_data pearson nosimple noprob plots=none;
  var n_guns_involved;
  with incident_id date n_killed n_injured
       incident_url_fields_missing congressional_district
       latitude longitude state_house_district
       state_senate_district;
run;
```

Figure 40

10 With Variables:	incident_id date n_killed n_injured incident_url_fields_missing congressional_district latitude longitude state_house_district state_senate_district
1 Variables:	n_guns_involved

Table 7

Pearson Correlation Coefficients Number of Observations	
	n_guns_involved
incident_id	0.01420 53579
date	0.01012 53579
n_killed	-0.01946 53579
n_injured	-0.03039 53579
incident_url_fields_missing	. 53579
congressional_district	0.02078 53114
latitude	-0.00354 53148
longitude	-0.01529 53148
state_house_district	-0.01041 46941
state_senate_district	-0.00069 48004

From the table 7 it shows that there is high negative correlation between the target variables which number of guns involved with the number of people killed at -0.01946 and number of people injured at -0.03039 and for the rest there is no relationship between the rest variables.

Hypothesis

Thus, it can be said that from the cleaned and transform datasets several hypotheses can be derived from the visualization done on the previous section with data visualization using SAS studio. The list below 5 interpretations of hypotheses of this dataset using the SAS studio Codes.

- The variable is grouped into categorical and continuous
- The detail explanation of descriptive statistics
- Imputation can be used to get missing data values
- Normalization is used to standardized the data
- Log transformation is used to make the datasets distribution less skewed

Discussion

This part explained the detail explanation of all the process starting from the initial data exploration, data pre-processing, exploratory data analysis and hypothesis about the summary of the findings. These parts consist of details explanation about the specific traits or values of certian attributes from the datasets imported into the SAS, extracting relevant insights or information from the distributions, finding out any correlation between the variables and more. In this assignment it the datasets used is from Kaggle about gun violence where it has range of data from 2013-2018 but for the purpose of this assignment only 2014- and 2015-gun violence datasets are used for data analysis.

For initial data exploration the output of the SAS code shows 52133 observations and 29 attributes or variables. Then the data is categorized into categorical and continuous where there is 18 Char data type and 11 Num data type. After categorizing the data, descriptive statistics can be done to show and summarize basic features of the gun violence datasets where it is presented to describe the data sample and the measurements. For example, WORK.IMPORT1;

are used for most of the analysis to get the desired output such as the values of mean, distributions, medians, variances, percentiles, standard deviation, N Miss.

Next there is a comparison between two different years of the gun violence datasets used where observations number of 2014 datasets are 51854 whereas 2015 is 53579 and it can be concluded in the initial data exploration the 2015-gun violence datasets shows much higher values for most of the variables such n_killed, n_injured, latitude, longitude and n_guns_involved for its mean, standard deviation, minimum, maximum, median, N, N Miss and quartile range. The next section is the data visualization using boxplot and histogram whereby both 2014 and 2015 datasets are already explained in the initial data exploration part. After that, histogram is used to study the distributions of the variables and the variables selected are n_killed, n_injured, latitude, longitude and n_guns_involved because its numeric variable. The mean value distribution for 2014 mu is 0.2422 meanwhile for 2015 is 0.2517 for number of people killed and it shows the higher value of mu indicates the data is more spread out for the 2015 one. For the 2015 number of people injured mu value is 0.5033 which is higher than 2014 at 0.4436 meanwhile it is also higher in 2015 where the mu value is 1.7739 whereas in 2014 is 1.3627.

Next is data pre-processing where two methods proposed that is imputation and normalization and the main function of data pre-processing is to handle noisy data. Imputation is done for both 2014- and 2015-gun violence datasets where only three variables are selected and in 2015 the number of frequencies is 53579 and 2014 is 51854. The second part of data pre-processing where the datasets is normalized and standardized. After that is feature engineering where log transformation is applied for both 2014 and 2015 datasets. It is found that mu value is changing and this indicates that the data becomes less skewed and it is found that mu in 2014 for n_killed is higher at 0.0805 than 2015 at 0.076 for histogram distribution. For the number of people injured the mu value in 2014 is 0.1343 meanwhile in 2015 is 0.125 and number of guns involved in 2014 is 0.0088 while in 2015 is 0.0605. This shows that after log transformation the data is more spread in 2014 than 2015.

For exploratory data analysis scatter plot is used to show the correlation between the variables and as number of guns increases the number of people killed also increases. This same goes with number of people injured and number of guns involved where it shows linear trend or relationship.

Conclusion

It can be concluded that gun violence datasets obtain from Kaggle about Gun Violence in United States from 2013-2018 shows ample and significant discussion for this assignment. It can be concluded there is a significance difference on the output results of SAS code for both year 2014 and 2015. The results of the analysis shows that 2015 has significant results as compare to 2014 and can be deduced that as number of observations increases it provides more insights on the data analysis. From the variables point of view as the number of guns involved increases the number of death and injuries increases and thus much stricter gun control is needed in USA to reduce more death and injuries related to gun violence especially mass shootings.

References

Krouse, William J., and Daniel J. Richardson, Mass Murder with Firearms: Incidents and Victims, 1999–2013, Washington, D.C.: Congressional Research Service, R44126, 2015.

Duwe, Grant, "Patterns and Prevalence of Lethal Mass Violence," Criminology and Public Policy, Vol. 19, No. 1, 2020, pp. 17–35.

CDC WONDER Online Database. Data are from the Multiple Cause of Death Files, 1999-2022. Accessed at http://wonder.cdc.gov/mcd-icd10.html on Jun 20, 202 Maps developed in SPSS 28 by Federico G de Cosío

Everytown. The Economic Cost of Gun Violence. Everytown Research & Policy. Feb 17, 2021. Consulted on June 24, 2022. Available at: https://everytownresearch.org/report/the-economic-cost-of-gun-violence/.

Ressler, Robert K., Ann W. Burgess, and John E. Douglas, Sexual Homicide: Patterns and Motives, New York: Simon and Schuster, 1988.

Elsass, H. Jaymi, Jaclyn Schildkraut, and Mark C. Stafford, "Studying School Shootings: Challenges and Considerations for Research," American Journal of Criminal Justice, Vol. 41, No. 3, 2016, pp. 444–464

Everytown analysis of the most recent year of gun homicides by country (2013 to 2019), GunPolicy.org (accessed January 7, 2022).

Paul M. Reeping et al., "State Gun Laws, Gun Ownership, and Mass Shootings in the US: Cross Sectional Time Series," BMJ 364 (March 2019): 1542

Lott, John R., Jr., and William Landes, "Multiple Victim Public Shootings," 2000 (unpublished). As of June 29, 2017: http://www.shack.co.nz/pistoltaupo/SSRN_ID272929_code010610560.pdf

Kleck, Gary, "Large-Capacity Magazines and the Casualty Counts in Mass Shootings: The Plausibility of Linkages," Justice Research and Policy, Vol. 17, No. 1, 2016, pp. 28–47.

Duwe, Grant, Tomislav Kovandzic, and Carlisle E. Moody, "The Impact of Right-to-Carry Concealed Firearm Laws on Mass Public Shootings," Homicide Studies, Vol. 6, No. 4, 2002, pp. 271–296.

Gius, Mark, "The Impact of State and Federal Assault Weapons Bans on Public Mass Shootings," Applied Economics Letters, Vol. 22, No. 4, 2015c, pp. 281–284.

Luca, Michael, Deepak Malhotra, and Christopher Poliquin, "The Impact of Mass Shootings on Gun Policy," Journal of Public Economics, Vol. 181, January 2020.

Fox, James A., and Jack Levin, Extreme Killing: Understanding Serial and Mass Murder, 3rd ed., Thousand Oaks, Calif.: Sage Publications, 2015.