

CP3403 Data Mining

ASSESSMENT TASK [INSERT NUMBER] COLLEGE OF [INSERT COLLEGE]



TASK COVER SHEET

Students									
Please sign, date and attach cover sheet to front of assessment task for all hard copy submissions									
SUBJECT CODE	CP-3403								
STUDENT FAMILY NAME	Student Given Name	JCU Student Number							
i. Liu	JieYuan	1	3	8	3	0	6	2	8
ii. Koh	Derrick Shao Wei	1	3	6	6	3	3	7	0
iii. Zhang	JiaYu	1	3	8	5	1	0	4	9
iv. Tin	Myint Kyaw	1	3	8	7	6	1	8	6
ASSESSMENT TITLE	Final Project Submission								
DUE DATE	28/05/2021								
LECTURER NAME	Eric Tham								
TUTOR NAME	Eric Tham								
<p align="center"><u>Student Declaration</u></p> <ol style="list-style-type: none"> 1. This assignment is our original work and no part has been copied/ reproduced from any other person's work or from any other source, except where acknowledgement has been made (see <i>Learning, Teaching and Assessment Policy 5.1</i>). 2. This work has not been submitted for any other course/subject (see <i>Learning, Teaching and Assessment Policy 5.9</i>). 3. This assignment has not been written for us. 4. We hold a copy of this assignment and can produce a copy if requested. 5. This work may be used for the purposes of moderation and identifying plagiarism. 6. We give permission for a copy of this marked assignment to be retained by the College for benchmarking and course review and accreditation purposes. <p>Learning, Teaching and Assessment Policy 5.1. A student who submits work containing plagiarised material for assessment will be subject to the provisions of the Student Academic Misconduct Requirements.</p> <p>Note definition of plagiarism and self plagiarism in Learning, Teaching and Assessment Policy</p>									
<p><u>Student signature(s)</u></p> <p>i.....Liu JieYuan..... Submission date ...28...../...05...../ 2019 iv.....Derrick..... Submission date ...28...../...05 / 2019</p> <p>ii.....Zhang JiaYu..... Submission date ...28...../...05...../ 2019 vTin Myint Kyaw..... Submission date ...28...../...05...../ 2019</p>									

Baltimore Crime Patterns Analysis

Abstract

Data mining is an out of the world approach to how we view complex entities. Through the process of mining data, many other relationships among the entities will surface and provide crucial information. Our team has a vision that if proper usage of data mining is applied on crime cases, there could potentially be an increase in crime prevention for the police department. Hence, our goal for this assignment will be data mining on the crime patterns in a town. The objectives of this report include general patterns on how crimes were committed in Baltimore - during which time are crimes more frequent and where are the crime hot spots. Equipped with these information and patterns, we hope that it can provide assistance to the police department in extensive prevention for all these crimes. That is what we want to achieve.

The data set: <https://www.kaggle.com/sohier/crime-in-baltimore>

Table of Contents

ABSTRACT	1
1. INTRODUCTION	4
1.1 - BUSINESS SCENARIO	4
2. METHOD	5
2.1 - K-MEANS.....	5
2.2 - DBSCAN	5
2.3 - RSTUDIO	5
3. DATA PRE-PROCESSING AND DESCRIPTION	6
3.1 - DATASET USED FOR THE REPORT	6
3.2 - DATASET DESCRIPTION.....	6
3.3 - DATA CLEANING	7
3.4 - DATA TRANSFORMATION.....	10
3.5 - DATA REDUCTION	11
4. DATA MODELLING	13
4.1 - FINDING CRIME HOTSPOTS.....	13
4.1.1 <i>K-means clustering</i>	14
4.1.2 <i>DBSCAN</i>	15
4.2 DATA VISUALIZATION	18
4.2.1 <i>Which Neighborhoods Have More Crime rates?</i>	18
4.2.2 <i>Weapon Used in Crimes</i>	19
4.2.3 <i>Crimes in Every Month</i>	20
4.2.4 <i>Indoor vs. Outdoor Cases</i>	21
4.2.5 <i>Crimes Occurrence in Different Time of Days</i>	21
5. CONCLUSION	24
REFERENCES.....	25
APPENDIX.....	25

1. Introduction

First, we choose the area that has a problem we want to tackle, upon deciding to enter the crime prevention sector, we sourced for our data set regarding crimes. Sourcing an appropriate dataset with relevant columns is important because there were some datasets without relevant information like time or type of crime. Throughout this report, there is our data description, data cleaning, data reduction and data preprocessing so that we have a good set of 5000 crimes data to work with.

Following the method, we took advantage of K-means, DBSCAN to get the results and graphs we needed to analyze the important relationships of crimes. In the modelling stage, we used the Longitude and Latitude from the dataset to plot the area with the highest crime rate. Allowing us to categorize the different neighborhoods and the amount of crime that happened. This report will also include the number of crimes in accordance with the weapon used to commit them to allow the police to take note while frisking suspicious personnel. Throughout this report, we have concluded the neighborhood and the time of day where crimes are more rampant. Which is good In assisting the police department for their manpower dissemination.

1.1 - Business Scenario

Our business scenario will be a police department engaging our data mining service to tackle the crime rate in their city. Our team provides recommendations and solutions to organizations that are looking to improve their efficiencies on what they do, and this time, we are working with the police department. The police department has provided us with a dataset of crimes happening in their city from the year 2012 and 2017, information is in the next paragraph. With this data, our team will use Weka and RStudio to further understand the further relationships among these crimes.

2. Method

2.1 - K-means

K-means aims to partition n observations into k clusters where each cluster belongs to the nearest mean (the center of the cluster) (Wikipedia, n.d.). In this report, the K-means clustering algorithm is used to confirm area groups which have not been clustered in the data set.

2.2 - DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering non-parametric algorithm. It groups together points that are closely packed together, marking as outliers points that lie alone in low-density regions (Wikipedia, n.d.). It separates the clusters of high density from low density clusters (Lutins, 2017). Therefore, it can also be used to find out the noisy data.

2.3 - RStudio

For the weapon column, we categorise them using counts and arrange them in descending order. Additionally, a percentage column is added. This allows us to see the percentage for each weapon as shown in table 1 more clearly.

Formula: $\text{Percentage} = \text{Count}/5000$

Also, get the Result of Crimes in Every Month:

First mutate the dataset, adding a new column which splits the month from the CrimeDate. After that, change the month from integer to character. Finally, use ggplot to plot the bar chart as figure 13, which will display the crimes rates for every month.

And get the data about the Crimes Happened in Different Time of Days:

Firstly, split the hour from the CrimeTime column. Then add a new column showing the crimes that happened in which hour. Next, build a table about the crimes that happened in different hours by grouping it and lastly, summarizing it and arranging them to attain table 2. Hence, the statistics of data would be obtained. In order to get the trend, we plot a line graph by ggplot and then obtain a graph showing the trend about crimes with time goes by, named it as figure 15.

Additionally, obtaining the mean, median, min and max number of the crimes that happened in different hours.

In order to compare the number of indoor crimes and outdoor crimes, ggplot from Rstudio was used to plot a bar chart shown in Figure 16.

3. Data Pre-Processing and Description

3.1 - Dataset Used for the Report

The dataset that will be used in this report is obtained from Kaggle - <https://www.kaggle.com/sohier/crime-in-baltimore>, which includes the crime data recorded by the Baltimore Police Department.

3.2 - Dataset Description

The original dataset has plenty of data, which contains 276530 rows. The dataset is about the crimes that happened between 2012 to 2017.

The following attributes contained in the file:

CrimeDate - Date. The date that crimes happened.

CrimeTime - Time. The specific time when crimes happened.

CrimeCode - String.

Location - String. The specific location where crimes happened.

Description - Nominal. The rough direction where crimes happened.

Inside/Outside - Nominal.

Weapon - Nominal. The weapon that is used by criminals.

Post - Numeric. Postal code that crimes happened

District - String.

Neighborhood - String.

Longitude - Numeric.

Latitude - Numeric.

Location 1 - The combination of longitude and latitude.

Premise - Nominal.

Total Incidents - Numeric.

3.3 - Data Cleaning

The original dataset we obtained from Kaggle was too big and contained 276529 rows. Due to the size of the dataset, we ran into problems with our devices not being able to effectively handle the dataset. Therefore, we decided to focus on a specific part of the dataset and remove the rest. The dataset originally contained the crime data from 2012 to 2017, with the data for 2017 being incomplete. Therefore, only the data for crimes that were committed in 2016 were kept and the rest were removed from the dataset. However, the data for 2016 still contained too much data in order for our devices to handle -- there were 48749 rows of data for crimes committed in 2016, so R was used to randomly select 5000 data, as follows:

First, the original csv file is loaded into Rstudio, then it was renamed as 'df' in order to make future identification easier. After that, prepare to use two R libraries called 'datasets' and 'dplyr' respectively to manipulate the dataset.

The whole dataset has plenty of rows, therefore it is impractical to find whether there are missing values. Therefore, using `filter()` and then `complete.cases()` were used to filter the missing values out.

```
df <- read.csv("BPD_Part_1_Victim_Based_Crime_Data.csv")  
  
View(df)  
  
library(datasets)  
  
library(dplyr)'  
  
df<-filter(df,complete.cases(df))  
  
View(df1)
```

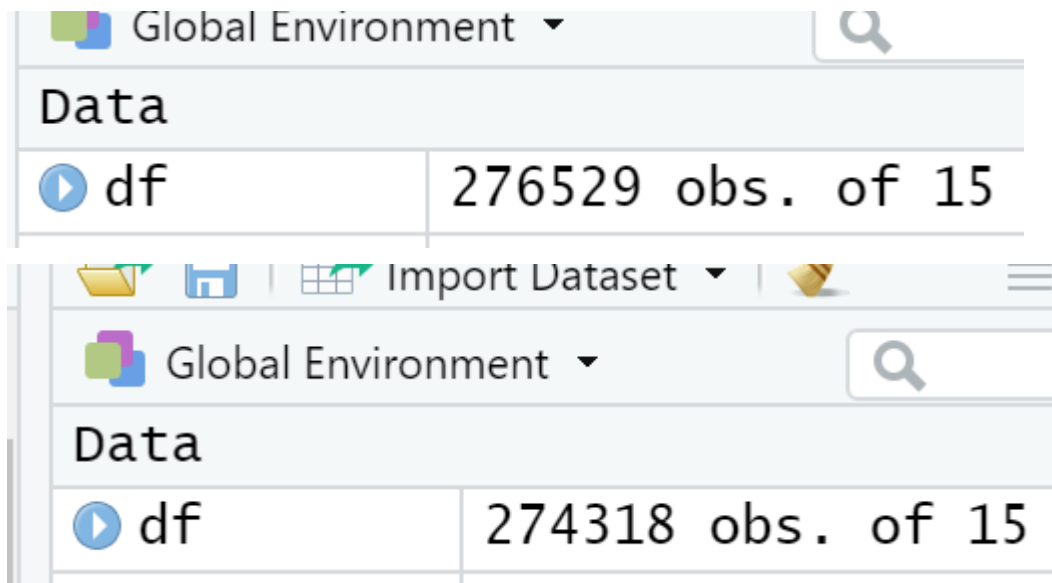


Figure 1: The original dataset and the dataset removed missing value

Next, there are 14 columns, but some columns are not needed for our purposes, such as longitude, latitude (as Location.1 is enough), Total.incidents (all of them are 1), Post, Premise. Therefore, these columns are removed. The rest of the information is kept.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            5000 non-null   int64
1   CrimeDate             5000 non-null   object
2   CrimeTime             5000 non-null   object
3   Location              4999 non-null   object
4   Description           5000 non-null   object
5   Inside.Outside        4938 non-null   object
6   Weapon               1740 non-null   object
7   District             5000 non-null   object
8   Neighborhood          4988 non-null   object
9   Longitude            5000 non-null   float64
10  Latitude             5000 non-null   float64
11  Location.1           5000 non-null   object
dtypes: float64(2), int64(1), object(9)
memory usage: 468.9+ KB
```

Except for 1740 columns, the values were missing in the “Weapon” columns, so we filled all the missing values as “Missing”. And for missing values in the “location” column,

we filled it as “Unknown”; and “X” in the missing values of “inside or outside” column; and filled in missing values in “Neighbourhood” column as “Unknown”.

```
In [15]: data["Weapon"].fillna("Missing")
```

```
Out[15]: 0      HANDS
1      Missing
2      Missing
3      Missing
4      KNIFE
...
4995   FIREARM
4996   Missing
4997   HANDS
4998   Missing
4999   Missing
Name: Weapon, Length: 5000, dtype: object
```

```
In [62]: data["Weapon"] = data["Weapon"].fillna("Missing")
data["Location"] = data["Location"].fillna("Unknown")
data["Inside.Outside"] = data["Inside.Outside"].fillna("X")
data["Neighborhood"] = data["Neighborhood"].fillna("Unknown")
```

Figure 2: Deal with the missing value of ‘Weapon’, ‘Location’, ‘Inside.Outside’ and ‘Neighborhood’.

```
: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            5000 non-null   int64
1   CrimeDate             5000 non-null   object
2   CrimeTime            5000 non-null   object
3   Location              5000 non-null   object
4   Description           5000 non-null   object
5   Inside.Outside        5000 non-null   object
6   Weapon               5000 non-null   object
7   District              5000 non-null   object
8   Neighborhood          5000 non-null   object
9   Longitude             5000 non-null   float64
10  Latitude              5000 non-null   float64
11  Location.1            5000 non-null   object
dtypes: float64(2), int64(1), object(9)
memory usage: 468.9+ KB
```

Now we filled in all the missing values.

3.4 - Data Transformation

After the dataset has been cleaned, it is easy to find that the format of the CrimeDate is ‘%m/%d%Y’, which is quite complex. Besides, the Rstudio will collapse while using the CrimeDate column. In this case, the format of the date should be changed to ‘%Y-%m-%d’.

```
df$CrimeDate<-as.Date(df$CrimeDate,format='%m/%d/%Y')
```

```
df$CrimeDate<-as.Date(df$CrimeDate,format='%Y-%m-%d')
```

	CrimeDate		CrimeDate
1	09/02/2017	1	2017-09-02
2	09/02/2017	2	2017-09-02
3	09/02/2017	3	2017-09-02
4	09/02/2017	4	2017-09-02
5	09/02/2017	5	2017-09-02

Figure 3: The original date format and changed date format.

In addition, in the column named Inside. Outside, sometimes the outside crimes are represented as ‘O’, and the rest is present as ‘outside’. To make it the same, use Rstudio to select the column which is equal to ‘Outside’. After that change all of that to ‘O’. Similarly, change all the ‘Inside’ into ‘I’.

```
sampldf$Inside.Outside[sampldf$Inside.Outside=='Outside']<-'O'
```

```
sampldf$Inside.Outside[sampldf$Inside.Outside=='Inside']<-'I'
```

Inside.Outside
Outside
O
O

Figure 4: Original record without same format.

3.5 - Data Reduction

There is a lot of data in the dataset. But we decided to focus on 2016, the date in other years will be removed from the new dataset.

First, we build a new dataframe called df1, then use mutate() to add a new column which shows the year of every crime, named CrimeYear. After that, the filter() function will select the data which happened in 2016. Finally, delete the column named CrimeYear to make the dataset clean.

```
df1<-df%>%mutate(CrimeYear=as.integer(format(df$CrimeDate,'%Y')))  
  
View(df1)  
  
df2<-filter(df1,df1$CrimeYear==2016)  
  
View(df2)  
  
df2$CrimeYear=NULL
```

After selecting the crimes that happened in 2016, there are still more than 48000 rows, the scale of the dataset is still large. To make it smaller, sampling is a good choice. In order to keep the result accurate, the size will be set as **5000**. Using sample_n(), to randomly pick up 5000 data from the former dataset, then named it as 'sampledf'.

```
sampledf<-sample_n(df2,size=5000)
```

As the data in the column named CrimeCode is not relevant to our objective, the column will be deleted.

```
sampld$CrimeCode<-NULL
```

df2	48548 obs. of 15 varia...
sampldf	5000 obs. of 14 variab...

Figure 5: The preprocessed dataset named *sampldf* and the original dataset.

Finally, export the processed data:

```
write.csv(sampldf, 'CP3403DATASET_PROJECT.csv')
```

4. Data Modelling

4.1 - Finding Crime Hotspots

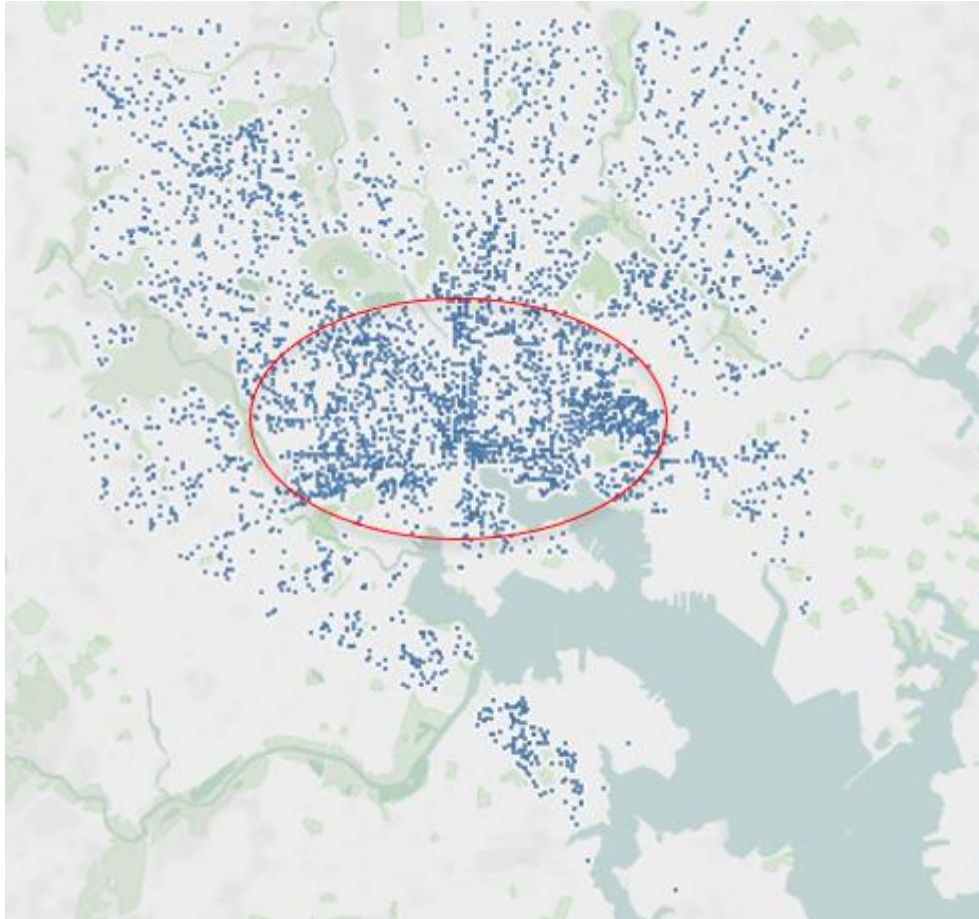


Figure 6: *Crime Hotspots*

From the figure above, it is clear that crimes are more likely to happen in the area within the red circle. Since the density of the spots is larger than other areas.

However, in order to make it more specific, we need to use the k-means algorithm to make the area into 6 clusters. Which we will use to split the dataset into northwest, northeast, west, middle, east, and south

4.1.1 K-means clustering

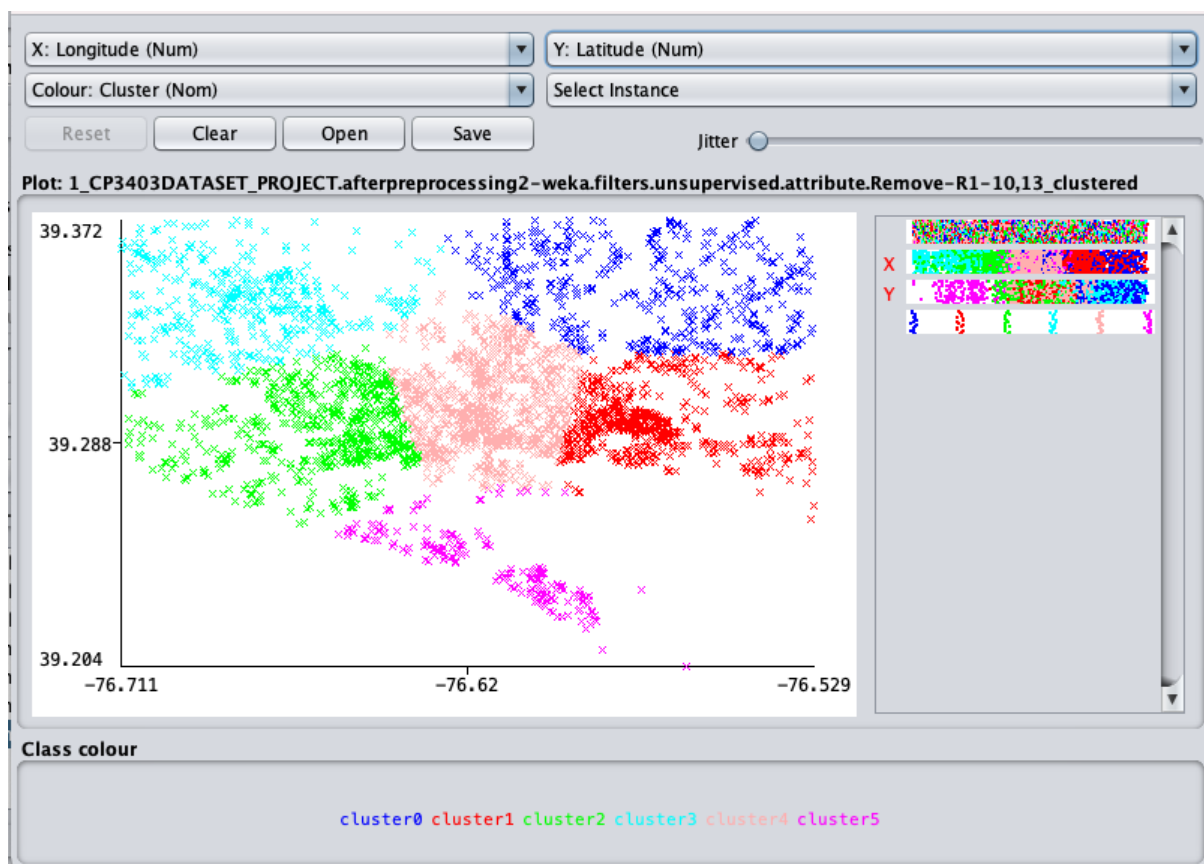


Figure 7: The crime spots are split into 6 clusters.

Final cluster centroids:

Attribute	Full Data (1250.0)	Cluster#					
		0 (162.0)	1 (241.0)	2 (248.0)	3 (155.0)	4 (385.0)	5 (59.0)
Longitude	-76.6175	-76.5727	-76.5712	-76.6592	-76.6794	-76.6143	-76.6122
Latitude	39.3065	39.3464	39.2956	39.2927	39.3429	39.301	39.2402

Figure 9: Final Cluster Centroids

Clustered Instances

0	490 (13%)
1	699 (19%)
2	742 (20%)
3	482 (13%)
4	1107 (30%)
5	230 (6%)

Figure 10: Clustered Instances

The numbers indicate the areas of crime spots and their frequencies. The fourth cluster has the most frequency of crimes. Furthermore, area 1 and 2 has 20% of total crimes. Therefore, the police department should assign more manpower patrol to area 1, 2 and 4, which represents the middle, west, and east parts respectively.

4.1.2 DBSCAN

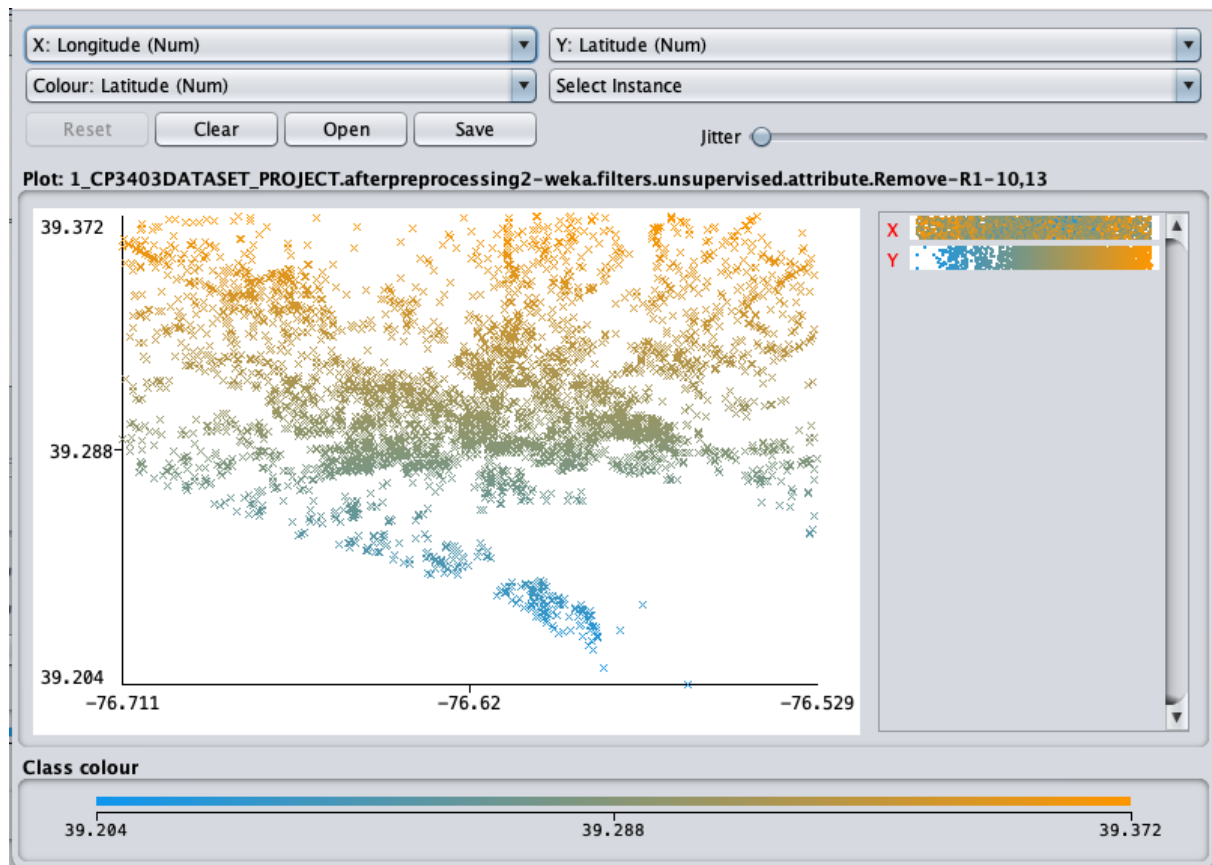


Figure 11

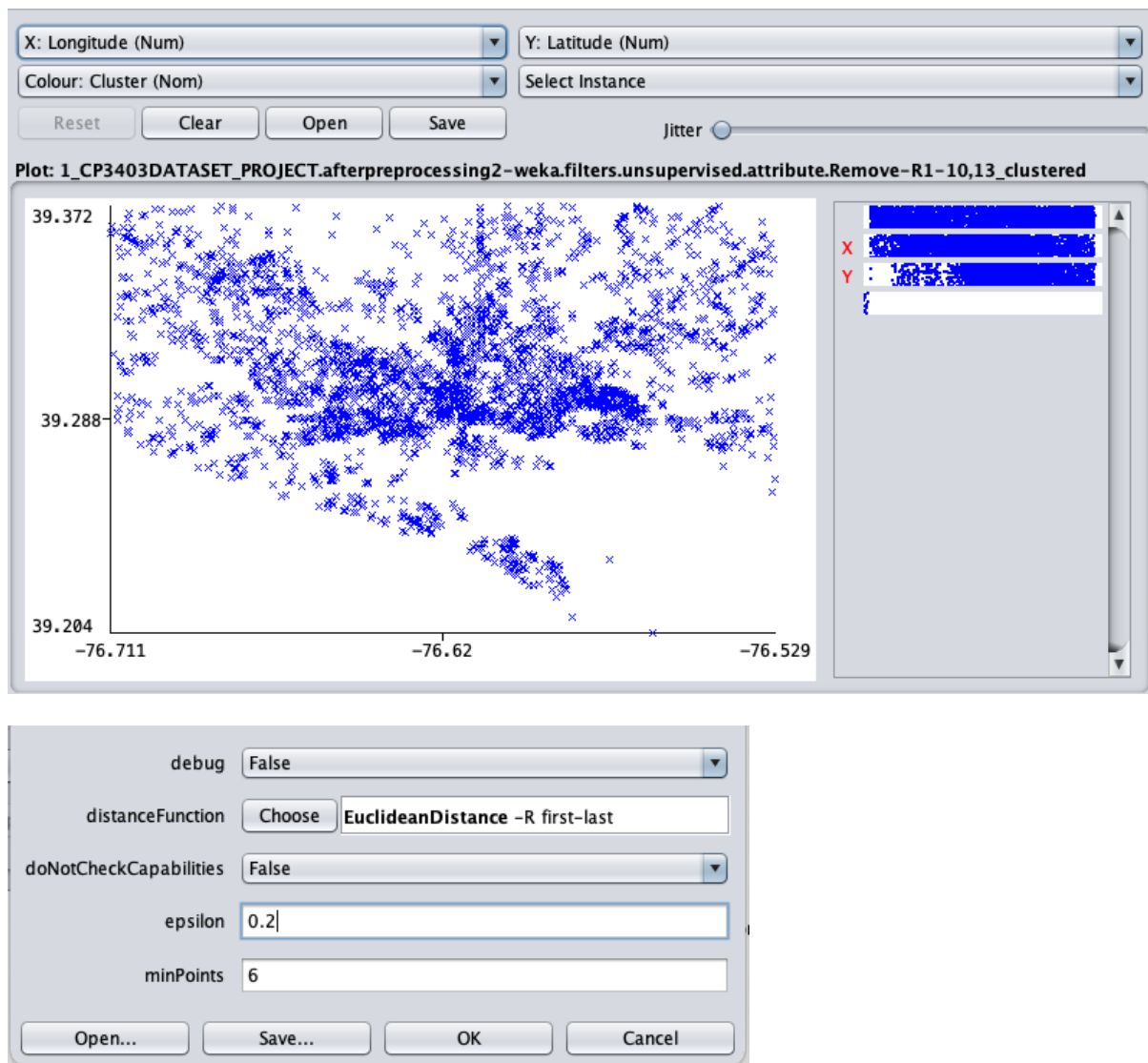


Figure 13: DBSCAN Results with the setting 1

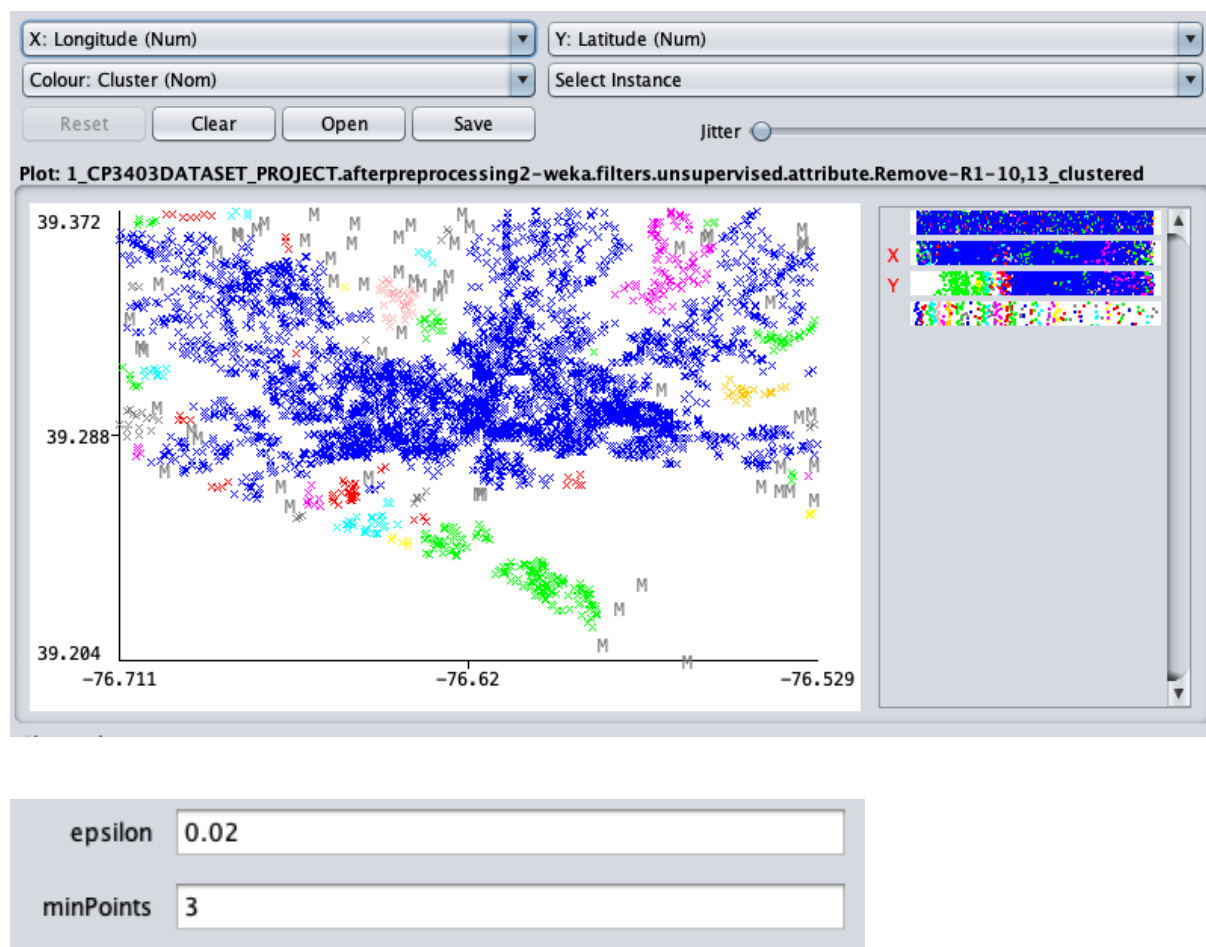


Figure 14: DBSCAN settings 2

Then, DBSCAN is used to groups together points that are packed together, and separates the low density clusters from high density clusters. How the points be clustered is determined by options: the epsilon and minpoints.

minPoints -- minimum number of DataObjects required in an epsilon-range-query

epsilon -- radius of the epsilon-range-queries

First, we set the epsilon as 0.2, and minpoints as 6, the cluster is shown in Figure 12, then we set the epsilon as 0.02 and minpoint as 3, the outcome shown in Figure 14.

Clearly, the second graph shows the clusters and points out the missing value, whereas the first graph only has one cluster.

From the graph, it is suggested that the police resources should be assigned to the clusters that have colors, and there could be less police assigned to the areas that are pointed

as “M”. And when police are going on patrol, they should go to the clustered areas by DBSCAN within their assigned areas by K-means more frequently.

4.2 Data Visualization

4.2.1 Which Neighborhoods Have More Crime rates?

In order to decide where more police should be assigned to control the crime rate, the report needs to analyze the crime in different locations and split the region data into smaller areas. The following figure shows the top 15 neighborhoods that the crimes were committed in, ranked from the largest to smallest.

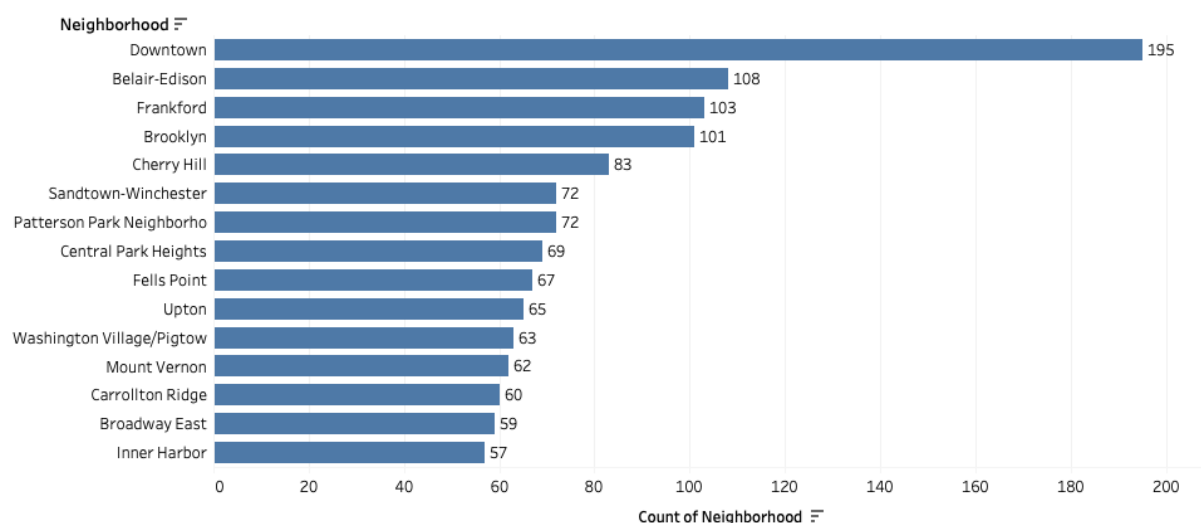


Figure 16: Number of crimes in each neighbourhood

As evident from Figure 16, the Downtown neighborhood has the most crime, compared to other neighborhoods. Also, the difference in crime rates between the Downtown neighbourhood and the neighbourhood with the second highest number of crimes - Belair-Edison - is almost twofold. Therefore, it is suggested that more police should be assigned to the Downtown area to cope with the high crime rate.

4.2.2 Weapon Used in Crimes

figure

1

	Weapon	Count	Percentage
1	Missing	3260	0.6520
2	HANDS	792	0.1584
3	FIREARM	470	0.0940
4	OTHER	300	0.0600
5	KNIFE	178	0.0356

Table 1: The count of different types of weapons and their use rate shown in the table.

According to Table 1, which described the results about the weapons used in the crimes. From the table, we can see that 65.2% of data concerning weapon use is missing, so we may not use this data to make accurate conclusion. The missing values could primarily mean that either the crime recorded in the sample is committed without the use of weapons (theft), or the police could not determine the type of weapon used. However, when comparing the rest items, many criminals use their hands as weapons, firearms were used in only 9% of the crimes and the other criminals prefer to use knives.

On the contrary, the result does not mean that the police should let their guard down because the criminals only used their hands when committing crimes. In this case, because of a lot of missing records, the exact usage trend of weapons could not be confirmed. Therefore, the police or the related workers should be fully prepared and stay vigilant to avoid severe injuries.

4.2.3 Crimes in Every Month

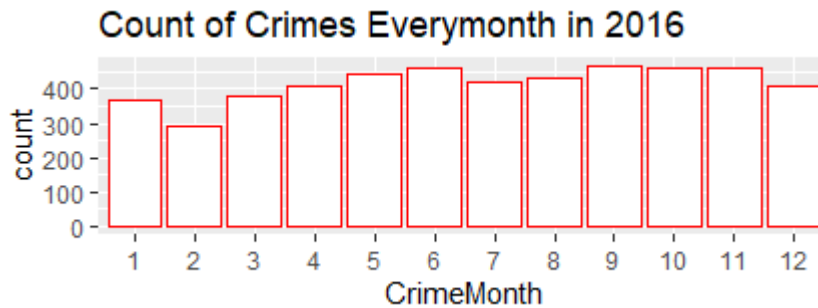


Figure 17: *Crime rates in different months*

In the bar chart depicted in Figure 17, there are about 350 crimes happening every month. February has the least crimes, which is about 280. June, September, October, and November have the most crime cases, which is up to about 450.

There is a decline in crime rate from January to February and the trend reaches its lowest point. It might be caused by the oncoming winter; as the cold weather might decrease the crimes. Also, the month of February has the least crimes out of all the months, which could cause inaccuracy on the analysis. After that, the crime rate increases steadily until June, which has the most crimes. Followed by a sudden drop, which could be caused by the oncoming summer, the higher temperature might also decrease the crimes. Two months later, the crimes reached a higher point again.

In this case, it is concluded that the crimes might be affected by seasons.

4.2.4 Indoor vs. Outdoor Cases

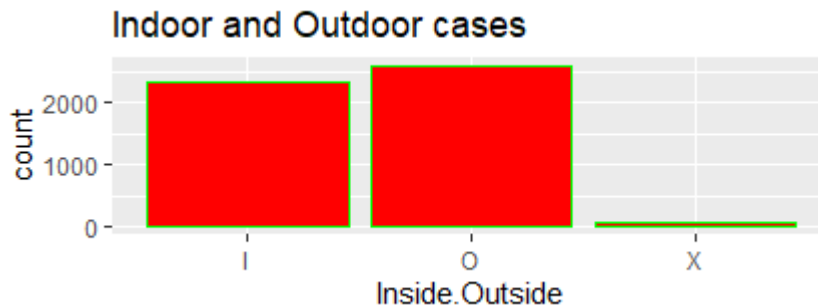


Figure 18: The number of indoor and outdoor cases in 2016.

As shown on figure 14 bar chart , there are about 2300 indoor crimes and 2600 outdoor crimes, the rest are unclear cases.

Comparing the number of crimes indoors and outdoors, the difference is not much but the result still points towards a higher crime rate in the outdoors. Therefore, more notices should be given out by the police alerting residents to be more careful while outdoors without letting their guard down when they are at home too. Keeping windows and doors locked when residences are sleeping or not at home is a good way to go about that.

4.2.5 Crimes Occurrence in Different Time of Days

	CrimeHour	Count		CrimeHour	Count
1	0	214	13	12	234
2	1	202	14	13	230
3	2	152	15	14	249
4	3	102	16	15	247
5	4	83	17	16	261
6	5	70	18	17	309
7	6	85	19	18	346
8	7	151	20	19	278
9	8	189	21	20	275
10	9	178	22	21	282
11	10	175	23	22	267
12	11	195	24	23	226

Table 2: The total cases of crimes happened in different times of the day

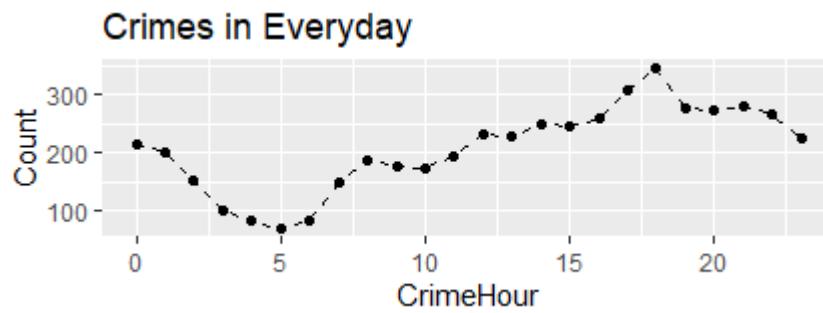


Figure 15: Plot for the total cases of crimes happened in different times.

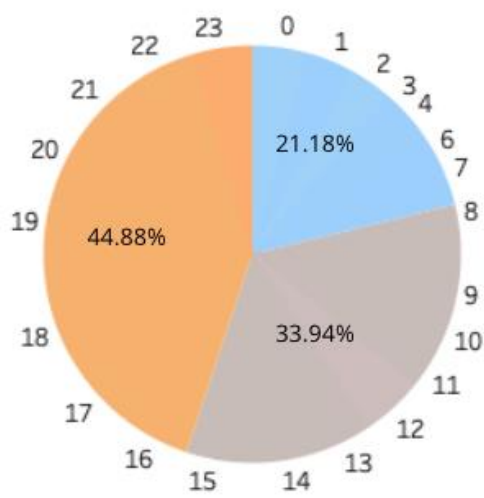


Figure 16: Pie chart depicting the amount of crime over various times of the day

```

> CrimesHourmedian<-median(df4$Count)
> CrimesHourmedian
[1] 220
> CrimesHourmean<-round(mean(df4$Count),2)
> CrimesHourmean
[1] 208.33
> CrimesMaxhour<-max(df4$Count)
> CrimesMaxhour
[1] 346
> CrimesMinhour<-min(df4$Count)
> CrimesMinhour
[1] 70

```

Figure 17: The median, mean, min and max of crimes between 12 a.m to 12 p.m.

As the general working hours is around 8 hours per day, we split the day into 3 parts. The crimes committed during 12am to 8am represent 21.18% of all crimes while 33.94% of crimes committed falls under 9am to 3pm. Lastly, the time period where crimes are most frequent is from 4pm to 11pm at 44.88%, almost twice the frequency compared to 12am to 8am.

From the line graph shown in Figure 15 above, the peak time of committing the crime is at 6pm, which has 346 crimes happening, and there are only 70 crimes happening at 5am., which has the least crimes.

Besides, the mean of the crimes of all the time of a day is 208.3 while the median is 220. In this case, it could be considered that crimes usually happened between 11am and 12am. Therefore, more police resources should be assigned in the afternoon and during nighttime. Besides, the police should strengthen the patrol during this period, and remind people to pay attention to safety and be more mindful of their own safety during this period and improve the awareness of self-protection. Also, they can save some effort in the morning compared to nighttime.

5. Conclusion

This report explored the crime in the city of Baltimore in the year 2016, from a public dataset obtained from the Kaggle. The key findings obtained from modeling of the dataset is: Downtown area is the neighbourhood with the most crime, 6:00 pm is the time period with the most crimes and crime rates in Baltimore are linked to the time of the year with lower crime rates in winter months and summer months.

References

Lutins E., (2017). *DBSCAN: What is it? When to Use it? How to use it*. Retrieved from:

<https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818>

Wikipedia. (n.d.). Retrieved from: https://en.wikipedia.org/wiki/K-means_clustering

Wikipedia. (n.d.). Retrieved from: <https://en.wikipedia.org/wiki/DBSCAN>

Appendix

```
df<-read.csv('1_CP3403DATASET_PROJECT.afterpreprocessing2.csv')

View(df)

library(dplyr)

library(ggplot2)

df$CrimeDate<-as.Date(df$CrimeDate,format='%d/%m/%Y')

df$CrimeDate<-as.Date(df$CrimeDate,format='%Y-%m-%d')

#Add a new column about the month crimes happened

df1<-df%>%mutate(CrimeMonth=as.integer(format(df$CrimeDate,'%m')) )

View(df1)

df1$CrimeMonth<-as.character(df1$CrimeMonth) #Integer to character
```

```
#Bar chart about Count of Crimes Every Month in 2016
```

```
plmonth<-ggplot(data=df1) +  
  
  geom_bar(color='red',fill='white',aes(x=CrimeMonth))+  
  
  labs(title='Count of Crimes Everymonth in 2016')+  
  
  xlim(as.character(c(1:12)))
```

```
Plmonth  
.....
```

```
#Split the hour from the CrimeTime column
```

```
library(hms)  
  
df2<-df1  
  
df2$CrimeTime<-as_hms(df2$CrimeTime)  
  
df2$CrimeTime<-as.POSIXct(df2$CrimeTime)  
  
df3<-df2%>%mutate(CrimeHour=as.integer(format(df2$CrimeTime,'%H')))  
  
df3%>%summarise(sum(df3$CrimeHour))  
  
write.csv(df3,'HOUR.csv')  
  
df3$CrimeHour<-as.character(df3$CrimeHour)
```

```
#Build a table about the time of days that crimes happened
```

```
df4<-df3%>%group_by(CrimeHour)%>%summarise(Count=n())
```

```
df4$CrimeHour<-as.integer(df4$CrimeHour,'%H')
```

```
df4<-arrange(df4,-desc(CrimeHour))
```

```
write.csv(df4,'df4.csv')
```

```
#The median, mean, max and min of the hour crimes happened
```

```
CrimesHourmedian<-median(df4$Count)
```

```
CrimesHourmedian
```

```
CrimesHourmean<-round(mean(df4$Count),2)
```

```
CrimesHourmean
```

```
CrimesMaxhour<-max(df4$Count)
```

```
CrimesMaxhour
```

```
CrimesMinhour<-min(df4$Count)
```

```
CrimesMinhour
```

```
#Line graph about crimes happened in different time of days
```

```
p2<-ggplot(data=df4,aes(x=CrimeHour,y=Count)) +
```

```
  geom_line(linetype=2)+
```

```
  geom_point()+
```

```
  ggtitle('Crimes in Everyday')
```

```
p2
```

```
#Bar chart of the statistics about indoor and outdoor crimes
```

```
p3<-ggplot(data=df1) +
```

```
  geom_bar(color='green',fill='red',aes(x=Inside.Outside))+
```

```
  labs(title='Indoor and Outdoor cases')
```

```
p3
```

```
#Build a table about weapon usage
```

```
df5<-df1%>%group_by(Weapon)%>%summarise(Count=n())
```

```
df5<-arrange(df5,desc(Count))
```

```
df5<-df5%>%mutate(Percentage=Count/5000)
```