Data Science - Special Assignment

Name - Tarun Luthra
College Roll No. - 2018 CSC 1023
Exam Roll No. - 18068570014
Semester - 6
Year - 3
Subject - Data Science
Exam - Data Science - External Practical
Set - 3

In case any of the codes or images/screenshots present in the PDF file is not legible or readable, you may also refer to this Github Repository.
It contains everything done during this assignment - https://github.com/tarunluthra123/Data-Science-Special-Assignment

Question 1

Use the “USArrests” built-in dataset to plot beautiful graphs and find meaningful insight about the dataset. You may draft questions yourself and summarize the results. You will be marked for [10]
a. Creativity
b. Presentation
c. Originality
d. Summarization of results

Getting to know the dataset

Let us start by viewing and understanding the dataset.

```
data(USArrests)
help(USArrests)
```

USArrests {datasets} R Documentation

Violent Crime Rates by US State

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Usage

```
USArrests
```

Format

A data frame with 50 observations on 4 variables.

[,1] Murder	numeric Murder arrests (per 100,000)
[,2] Assault	numeric Assault arrests (per 100,000)
[,3] UrbanPop	numeric Percent urban population
[,4] Rape	numeric Rape arrests (per 100,000)

Note

`USArrests` contains the data as in McNeil's monograph. For the `UrbanPop` percentages, a review of the table (No. 21) in the Statistical Abstracts 1975 reveals a transcription error for Maryland (and that McNeil used the same “round to even” rule that R's `round()` uses), as found by Daniel S Coven (Arizona).

See the example below on how to correct the error and improve accuracy for the ‘<n>.5’ percentages.

Source

World Almanac and Book of facts 1975. (Crime rates).

Statistical Abstracts of the United States 1975, p.20, (Urban rates), possibly available at http://books.google.ch/books?id=zl9qAAAAMAAJ&pg=PA20

```
names(USArrests)
```

```
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

```
dim(USArrests)
```

```
[1] 50  4
```

```
View(USArrests)
```

Q2.r | Q1.r* | USArrests | df

Filter

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3
Kansas	6.0	115	66	18.0
Kentucky	9.7	109	52	16.3
Louisiana	15.4	249	66	22.2
Maine	2.1	83	51	7.8
Maryland	11.3	300	67	27.8

Showing 1 to 21 of 50 entries, 4 total columns

So this is a dataset about the arrest reports in USA. There are four attributes and we can try to find the relations between them. We can also try to find the least crime filled states or the most crime filled states. The dataset provides us with a list of 50 rows (1 for each state) and its statistics for 4 types of crimes - Murder, Assault, UrbanPop and Rape. Let us try to have some closer look at the dataset by obtaining its summary.

```
summary(USArrests)
```

```
> summary(USArrests)
     Murder          Assault         UrbanPop          Rape       
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00  
> 
```

Let us also try to see if the four crimes have any correlation to each other.

```
cor(USArrests)
```

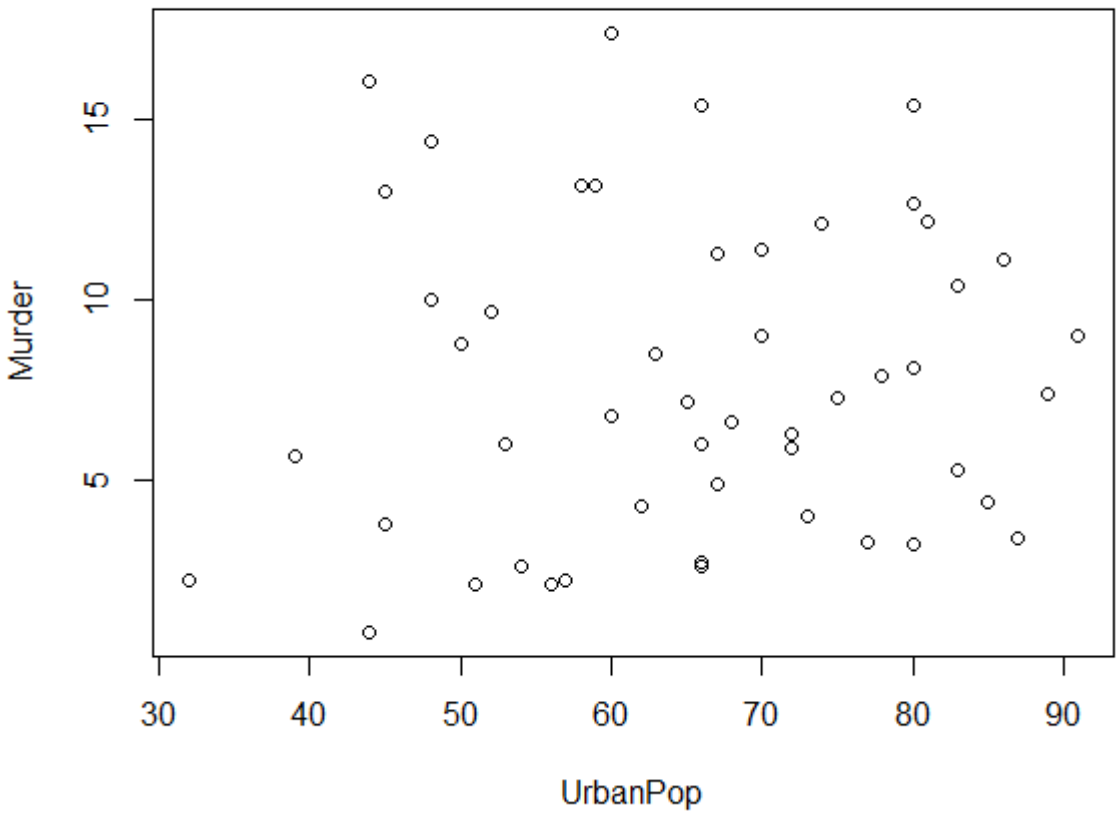
```
> cor(USArrests)
             Murder   Assault   UrbanPop      Rape
Murder   1.00000000 0.8018733 0.06957262 0.5635788
Assault  0.80187331 1.0000000 0.25887170 0.6652412
UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
Rape     0.56357883 0.6652412 0.41134124 1.0000000
>
```

Through the above data we infer that **Assault** are the most frequent crimes that are happening as it has the highest averages. Further **Rape** crimes are the least likely to happen (which is a good thing).

Visualising data

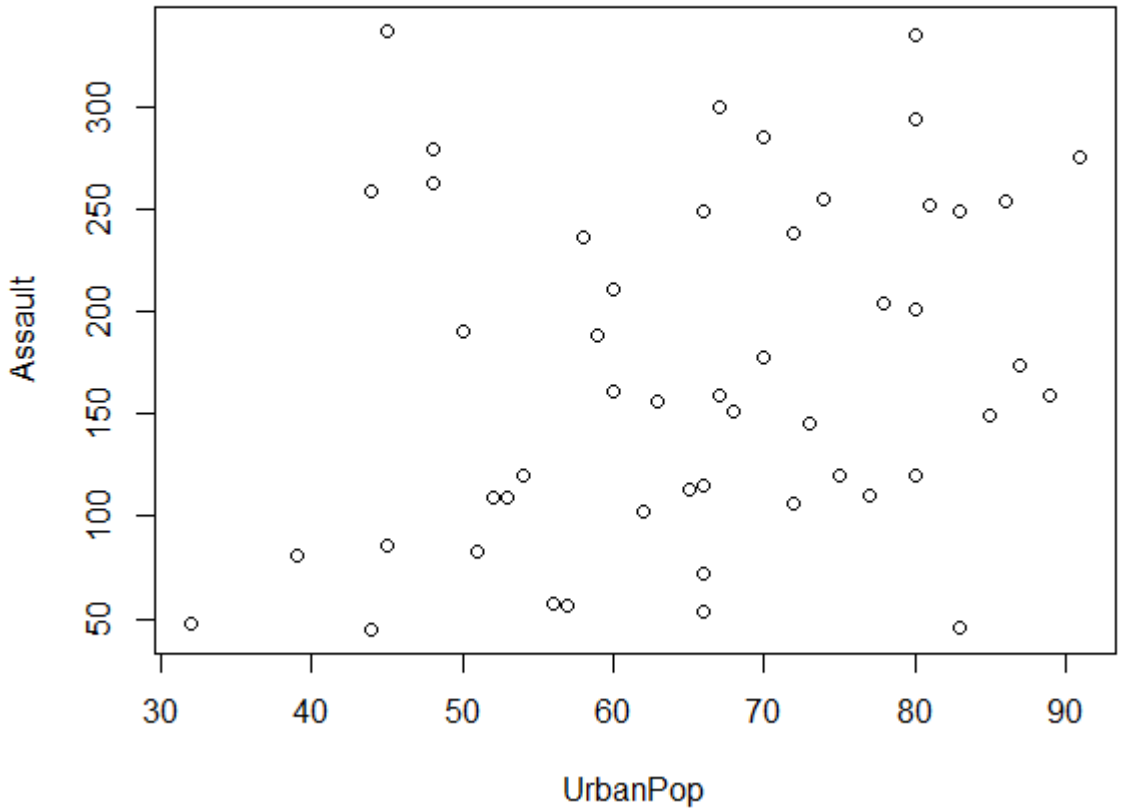
Running plot for these combinations, Murder and Assault do not appear to have a relation to UrbanPop. The distribution of plot points are scattered to the point that they do not appear to correlate to UrbanPop.

```
with(USArrests, plot(UrbanPop, Murder))
```



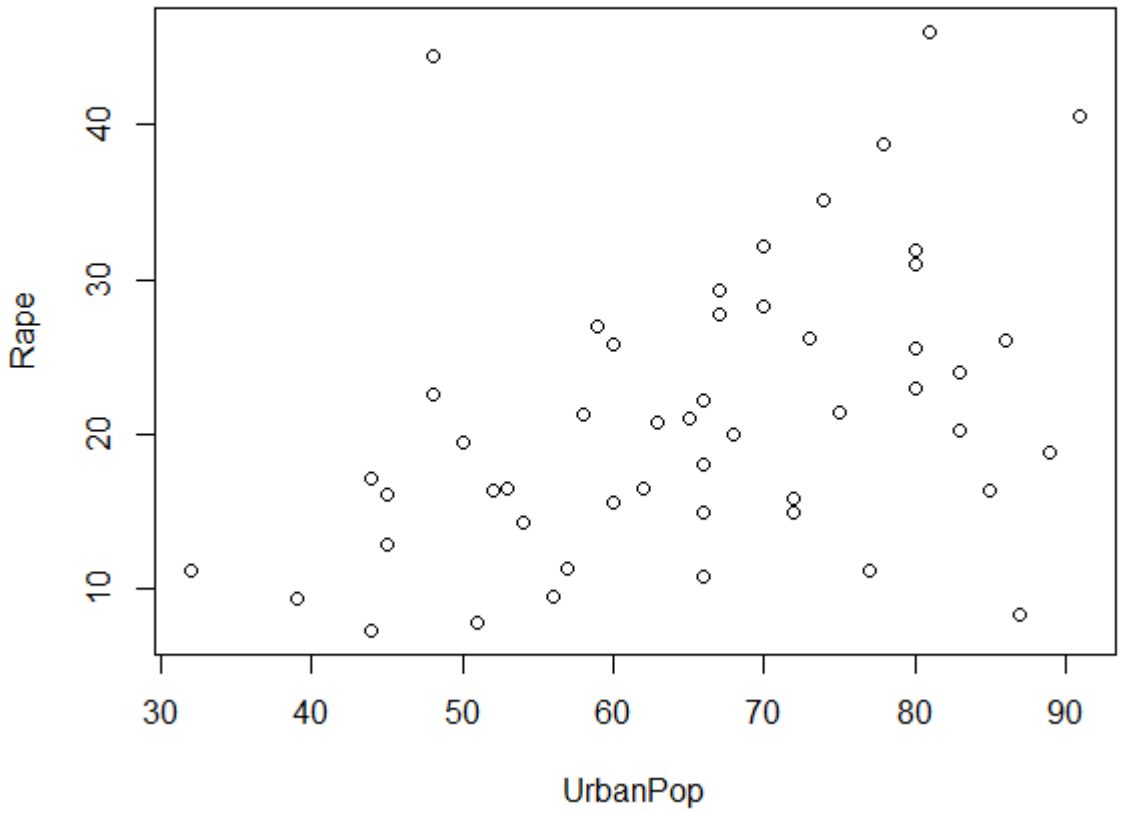
We see that most murders are likely to occur with higher Urban population. Although this isn't a fair measure since the plot seems to be very scattered. Let us try another one.

```
with(USArrests, plot(UrbanPop, Assault))
```



It is quite evident that more Assaults are likely to occur with more Urban Population.

```
with(USArrests, plot(UrbanPop, Rape))
```



Rape crimes are more likely to occur in states with average urban population. This goes to contrary to standard belief that more Urban population leads to more rapes.

Which states has most and least assault, murder, and rape arrests?

Let us try to figure out which states are more safer than the others so that if we ever plan a trip to US, we know where to steer clear off.

Most and Least assault

```
x <- which(USArrests$Assault == max(USArrests$Assault))
rownames(USArrests)[x]
```

```
[1] "North Carolina"
```

```
x <- which(USArrests$Assault == min(USArrests$Assault))
rownames(USArrests)[x]
```

```
[1] "North Dakota"
```

Most and Least murder

```
x <- which(USArrests$Murder == max(USArrests$Murder))
rownames(USArrests)[x]
```

```
[1] "Georgia"
```

```
x <- which(USArrests$Murder == min(USArrests$Murder))
rownames(USArrests)[x]
```

```
[1] "North Dakota"
```

Most and least rape

```
x <- which(USArrests$Rape == max(USArrests$Rape))
rownames(USArrests)[x]
```

```
[1] "Nevada"
```

```
x <- which(USArrests$Rape == min(USArrests$Rape))
rownames(USArrests)[x]
```

```
[1] "North Dakota"
```

States which have assault arrests more than median of the country.

```
assault.median = median(USArrests$Assault)
assault.median
```

```
[1] 159
```

```
subset(USArrests, Assault > assault.median, select= c(UrbanPop, Assault))
```

```
> subset(USArrests, Assault > assault.median, select= c(UrbanPop, Assault))
               UrbanPop Assault
Alabama              58     236
Alaska               48     263
Arizona              80     294
Arkansas             50     190
California           91     276
Colorado             78     204
Delaware             72     238
Florida              80     335
Georgia              60     211
Illinois             83     249
Louisiana            66     249
Maryland             67     300
Michigan             74     255
Mississippi          44     259
Missouri             70     178
Nevada               81     252
New Mexico           70     285
New York             86     254
North Carolina       45     337
Rhode Island         87     174
South Carolina       48     279
Tennessee            59     188
Texas                80     201
Wyoming              60     161
>
```

States that are in the bottom 25% of murder

These are the safer states that I would prefer to go to.

```
bottomQuartileMurderRate <- quantile(USArrests$Murder)[2]
bottomQuartileMurderRate
```

```
25%
4.075
```

```
subset(USArrests, Murder < bottomQuartileMurderRate, select= c(UrbanPop, Murder))
```

```
> subset(USArrests, Murder < bottomQuartileMurderRate, select= c(UrbanPop, Murder))
              UrbanPop Murder
Connecticut         77    3.3
Idaho               54    2.6
Iowa                57    2.2
Maine               51    2.1
Minnesota           66    2.7
New Hampshire       56    2.1
North Dakota        44    0.8
Rhode Island        87    3.4
South Dakota        45    3.8
Utah                80    3.2
Vermont             32    2.2
Washington          73    4.0
Wisconsin           66    2.6
>
```

States which are in the top 25% of the murder.

Better stay away from these states for our own safety.

```
topQuartileMurderRate <- quantile(USArrests$Murder)[4]
topQuartileMurderRate
```

```
##   75%
## 11.25
```

```
subset(USArrests, Murder > topQuartileMurderRate, select= c(UrbanPop, Murder))
```

```
> subset(USArrests, Murder > topQuartileMurderRate, select= c(UrbanPop, Murder))
               UrbanPop Murder
Alabama              58   13.2
Florida              80   15.4
Georgia              60   17.4
Louisiana            66   15.4
Maryland             67   11.3
Michigan             74   12.1
Mississippi          44   16.1
Nevada               81   12.2
New Mexico           70   11.4
North Carolina       45   13.0
South Carolina       48   14.4
Tennessee            59   13.2
Texas                80   12.7
>
```

Question 2

2. Download the file Set3.csv and write the correct code for each of the following : [15]
   I. Read the contents of the file
   II. Count of the number of records
   III. View the data in a tabular format
   IV. Filter the offences on the basis of Locality (QA only) and Type_of_offence (PHYSICAL OFFENSE only).
   V. Group the offences by Zone.
   VI. Get a count of the number of records for each group
   VII. Using ggplot(), plot a barchart displaying the number of offences in each Zone. (use all the possible parameters)
   VIII. Create a new column called Year_of_event containing the only the year of the event
   IX. Group the data by year and summarize
   X. Plot a barchart with column Year_of_event that displays the number of offences by year
   XI. Create another bar chart that displays the number of offences by month instead of year
   XII. Group and summarize the data by month.
   XIII. Rename the columns to make them more user friendly and view the results
   XIV. What's the need of filtering the data? Show examples using appropriate commands
   XV. What other charts can you plot for XI ? Which one will leverage more information and why?(elaborate in comments)

Library Imports

```
library(dplyr)
library(ggplot2)
library(lubridate)
library(RColorBrewer)
```

I. Read the contents of the file

```
df <- read.csv('Set3.csv')
```

II. Count of the number of records

```
count(df)
```

```
> # II. Count of the number of records
> count(df)
     n
1 2000
>
```

III. View the data in a tabular format

```
View(df)
```

Q2.r* × df ×

Filter

	Date_of_event	Date_of_resolution	Type_of_offence	Summary_of_
1	12/13/1908	12/13/2008	DUI	DUI-LIQUC
2	6/15/1964	6/15/2010	FAMILY OFFENSE-NONVIOLENT	CHILD-OTH
3	01-01-1973	1/25/2012	PHYSICAL OFFENSE	PHYSICAL_
4	06-01-1974	09-09-2013	PHYSICAL OFFENSE	PHYSICAL_
5	01-01-1975	08-11-2016	PHYSICAL OFFENSE	PHYSICAL_
6	12/16/1975	12/16/1975	BURGLARY-RESIDENTIAL	BURGLARY
7	01-01-1976	1/31/1976	PHYSICAL OFFENSE	PHYSICAL_
8	07-01-1976	12/27/2017	PHYSICAL OFFENSE	PHYSICAL_
9	01-01-1977	5/22/2018	RAPE	PHYSICAL_
10	01-01-1978	8/25/2009	PHYSICAL OFFENSE	PHYSICAL_

Showing 1 to 11 of 2,000 entries, 8 total columns

Console Jobs ×

C:/Users/Tarun Luthra/Desktop/Data Science/

```
> library(dplyr)
> # I. Read the contents of the file
> df <- read.csv('Set3.csv')
> # II. Count of the number of records
> count(df)
     n
1 2000
> # III. View the data in a tabular format
> View(df)
>
```

IV. Filter the offences on the basis of Locality (QA only) and Type_of_offence (PHYSICAL OFFENSE only).

```
filteredDf <- df %>% filter(Locality == 'QA' & Type_of_offence=='PHYSICAL OFFENSE')
View(filteredDf)
```

Q2.r | filteredDf | df

Filter

	Date_of_event	Date_of_resolution	Type_of_offence	Summary_of_offence	Zone	Block	Division	Locality
1	07-04-1979	9/15/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q3	QA
2	05-06-2004	8/13/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	WEST	Q	Q2	QA
3	06-07-2006	06-07-2006	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q3	QA
4	01-01-2007	7/31/2009	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q2	QA
5	07-10-2007	10/20/2008	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q2	QA
6	7/18/2007	7/18/2007	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	WEST	Q	Q3	QA
7	11-04-2007	3/31/2008	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	WEST	Q	Q3	QA
8	01-02-2008	01-02-2008	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT EXPOSURE	WEST	Q	Q3	QA

V. Group the offences by Zone.

```
zoneGroups <- df %>% group_by(Zone)
zoneGroups
```

```
> # V. Group the offences by Zone.
> zoneGroups <- df %>% group_by(Zone)
> zoneGroups
# A tibble: 2,000 x 8
# Groups:   Zone [6]
   Date_of_event Date_of_resolut~ Type_of_offence
   <chr>         <chr>            <chr>
 1 12/13/1908    12/13/2008       DUI
 2 6/15/1964     6/15/2010        FAMILY OFFENSE~
 3 01-01-1973    1/25/2012        PHYSICAL OFFEN~
 4 06-01-1974    09-09-2013       PHYSICAL OFFEN~
 5 01-01-1975    08-11-2016       PHYSICAL OFFEN~
 6 12/16/1975    12/16/1975       BURGLARY-RESID~
 7 01-01-1976    1/31/1976        PHYSICAL OFFEN~
 8 07-01-1976    12/27/2017       PHYSICAL OFFEN~
 9 01-01-1977    5/22/2018        RAPE
10 01-01-1978    8/25/2009        PHYSICAL OFFEN~
# ... with 1,990 more rows, and 5 more variables:
#   Summary_of_offence <chr>, Zone <chr>, Block <chr>,
#   Division <chr>, Locality <chr>
```

VI. Get a count of the number of records for each group

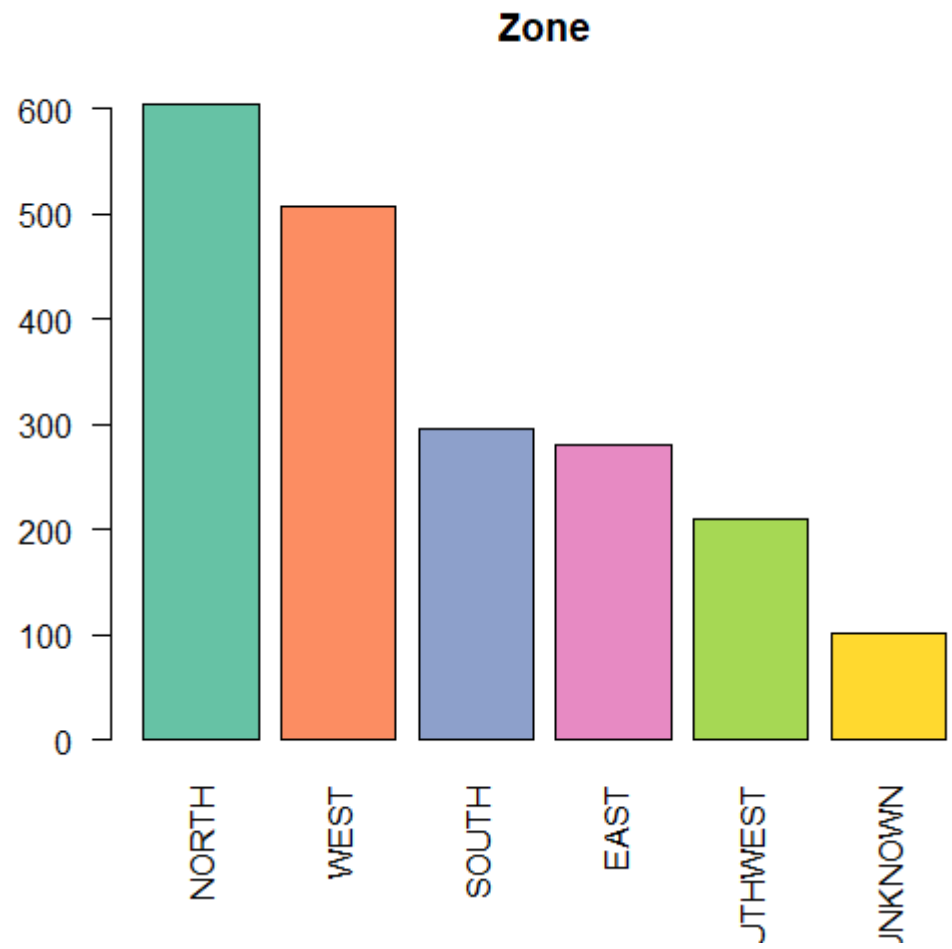
```
count(zoneGroups)
```

```
> # VI. Get a count of the number of records for each group
> count(zoneGroups)
# A tibble: 6 x 2
# Groups:   Zone [6]
  Zone           n
  <chr>      <int>
1 EAST         281
2 NORTH        604
3 SOUTH        296
4 SOUTHWEST    210
5 UNKNOWN      101
6 WEST         508
> 
```

Note: The entries with Zone = 'UNKNOWN' were **intentionally** left in the dataframe and as can be seen above, they become a part of the analysis. This causes the following graph in Part 7 to get affected as well. The reason for this is explained in detail in **Part 14**. Kindly refer to it before thinking of this as a mistake.

VII. Using ggplot(), plot a barchart displaying the number of offences in each Zone. (use all the possible parameters)

```
coul <- brewer.pal(8, "Set2")
barplot(sort(table(zoneGroups$Zone), decreasing = T), las = 2, main = "Zone", col=coul)
```



Note that an alternative plot could also be constructed with the following function using the ggplot()

```
# Alternative method to generate plot
ggplot(count(zoneGroups), aes(x=Zone,y=n)) +
  geom_bar(stat="identity" )
```

However since the two plots are very like and the plot generated through barplot() is neater and cleaner, I have used that as my primary method. Following bar plots can also utilize the ggplot() method however only barplot() code is written for that.

VIII. Create a new column called Year_of_event containing the only the year of the event

```
df$Year_of_event <- df %>% with(year(mdy(Date_of_event)))
View(df)
```

Q2.r* | filteredDf | df

Filter

	Date_of_event	Date_of_resolution	Type_of_offence	Summary_of_offence	Zone	Block	Division	Locality	Year_of_event
1	12/13/1908	12/13/2008	DUI	DUI-LIQUOR	EAST	G	G2	CA/SP	1908
2	6/15/1964	6/15/2010	FAMILY OFFENSE-NONVIOLENT	CHILD-OTHER	WEST	Q	Q2	QA	1964
3	01-01-1973	1/25/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	NORTH	N	N2	NG	1973
4	06-01-1974	09-09-2013	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1974
5	01-01-1975	08-11-2016	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1975
6	12/16/1975	12/16/1975	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTH	R	R3	LW/SP	1975
7	01-01-1976	1/31/1976	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	UNKNOWN			UNKNOWN	1976
8	07-01-1976	12/27/2017	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	UNKNOWN			UNKNOWN	1976
9	01-01-1977	5/22/2018	RAPE	PHYSICAL_VIOLENCE-SODOMY	UNKNOWN			UNKNOWN	1977
10	01-01-1978	8/25/2009	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	W	W1	ALKI	1978
11	1/28/1979	02-09-1979	CAR PROWL	THEFT-CARPROWL	EAST	G	G2	CA/SP	1979
12	07-04-1979	9/15/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q3	QA	1979
13	01-01-1980	5/28/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1980
14	01-01-1980	10/31/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	F	F1	HP	1980
15	2/14/1981	2/15/1981	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTHWEST	W	W3	RWA	1981

Showing 1 to 15 of 2,000 entries, 9 total columns

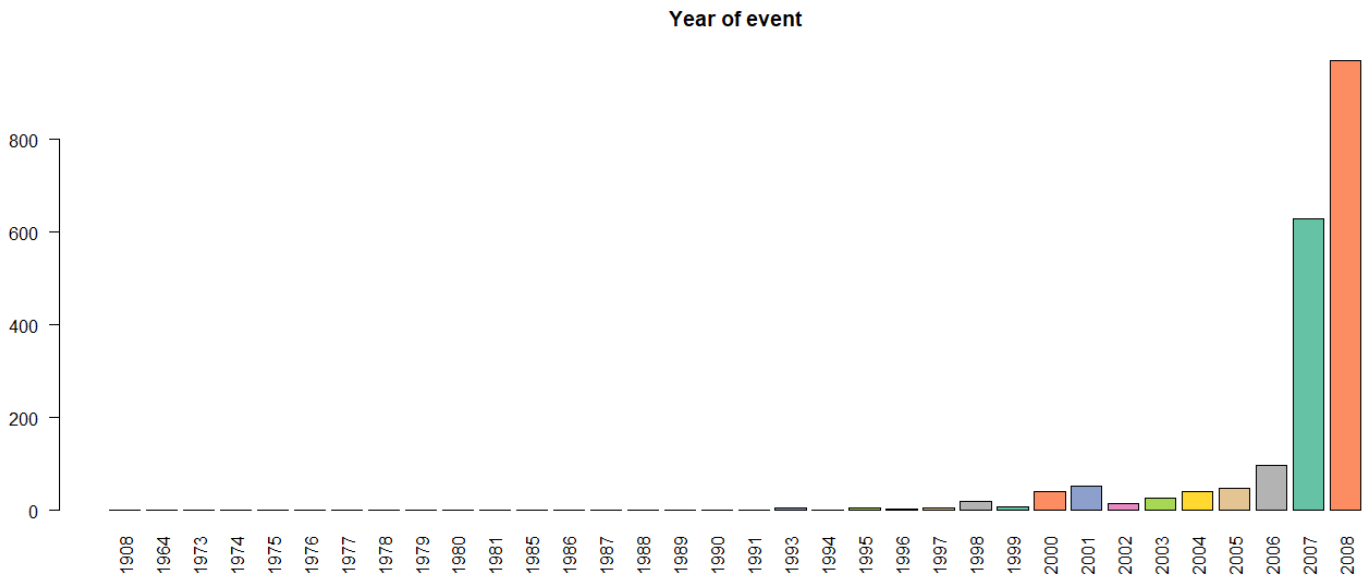
IX. Group the data by year and summarize

```
yearGroup <- df %>% group_by(Year_of_event)
summary(yearGroup)
```

```
> # IX. Group the data by year and summarize
> yearGroup <- df %>% group_by(Year_of_event)
> summary(yearGroup)
 Date_of_event      Date_of_resolution Type_of_offence    Summary_of_offence     Zone              Block
 Length:2000        Length:2000        Length:2000        Length:2000        Length:2000        Length:2000
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


   Division           Locality         Year_of_event
 Length:2000        Length:2000        Min.   :1908
 Class :character   Class :character   1st Qu.:2007
 Mode  :character   Mode  :character   Median :2007
                                       Mean   :2006
                                       3rd Qu.:2008
                                       Max.   :2008
> |
```

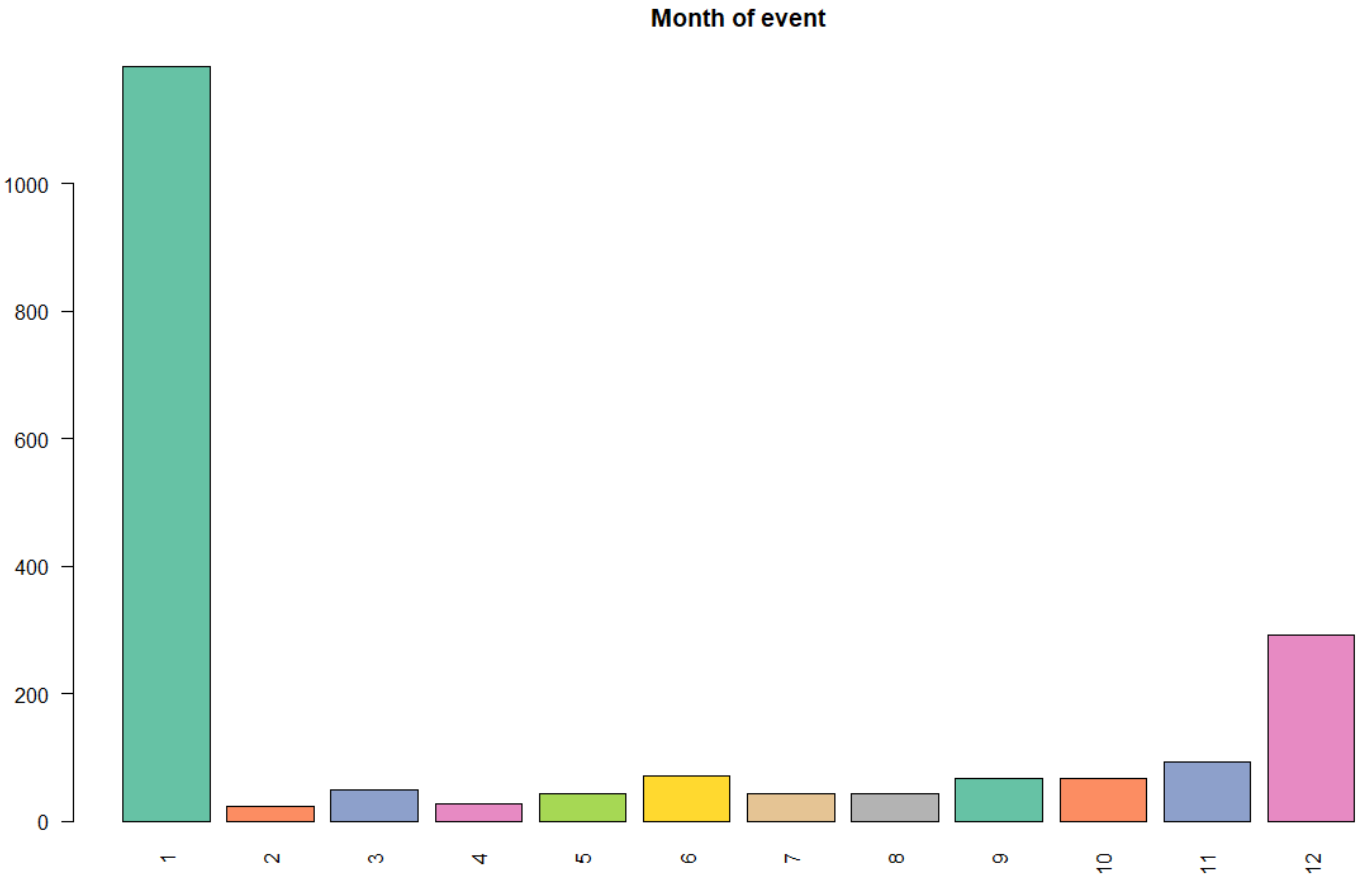
X. Plot a barchart with column Year_of_event that displays the number of offences by year

```
barplot(table(yearGroup$Year_of_event),
        las = 2, main = "Year of event", col=coul)
```



XI. Create another bar chart that displays the number of offences by month instead of year

```
df$Month_of_event <- df %>% with(month(mdy(Date_of_event)))

monthGroup <- df %>% group_by(Month_of_event)

barplot(table(monthGroup$Month_of_event),
        las = 2, main = "Month of event" ,col=coul)
```



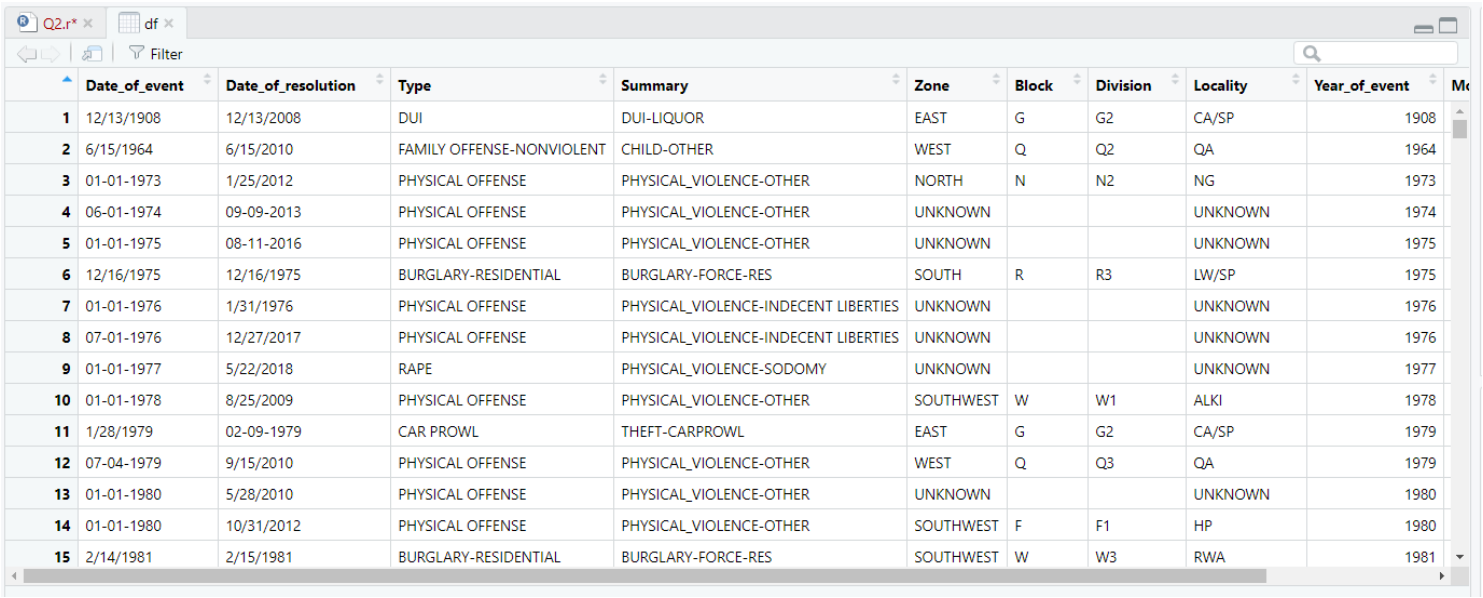
XII. Group and summarize the data by month.

```
df$Month_of_event <- df %>% with(month(mdy(Date_of_event)))

monthGroup <- df %>% group_by(Month_of_event)

summary(monthGroup)
```

```
> summary(monthGroup)
 Date_of_event      Date_of_resolution Type_of_offence    Summary_of_offence     Zone              Block
 Length:2000        Length:2000        Length:2000        Length:2000        Length:2000        Length:2000
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


   Division           Locality         Year_of_event  Month_of_event
 Length:2000        Length:2000        Min.   :1908   Min.   : 1.000
 Class :character   Class :character   1st Qu.:2007   1st Qu.: 1.000
 Mode  :character   Mode  :character   Median :2007   Median : 1.000
                                       Mean   :2006   Mean   : 4.275
                                       3rd Qu.:2008   3rd Qu.: 9.000
                                       Max.   :2008   Max.   :12.000
> |
```

XIII. Rename the columns to make them more user friendly and view the results

```
names(df)[names(df) == "Summary_of_offence"] <- "Summary"
names(df)[names(df) == "Type_of_offence"] <- "Type"
View(df)
```



Q2.r* | df | Filter

	Date_of_event	Date_of_resolution	Type	Summary	Zone	Block	Division	Locality	Year_of_event	Mo
1	12/13/1908	12/13/2008	DUI	DUI-LIQUOR	EAST	G	G2	CA/SP	1908	
2	6/15/1964	6/15/2010	FAMILY OFFENSE-NONVIOLENT	CHILD-OTHER	WEST	Q	Q2	QA	1964	
3	01-01-1973	1/25/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	NORTH	N	N2	NG	1973	
4	06-01-1974	09-09-2013	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1974	
5	01-01-1975	08-11-2016	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1975	
6	12/16/1975	12/16/1975	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTH	R	R3	LW/SP	1975	
7	01-01-1976	1/31/1976	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	UNKNOWN			UNKNOWN	1976	
8	07-01-1976	12/27/2017	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-INDECENT LIBERTIES	UNKNOWN			UNKNOWN	1976	
9	01-01-1977	5/22/2018	RAPE	PHYSICAL_VIOLENCE-SODOMY	UNKNOWN			UNKNOWN	1977	
10	01-01-1978	8/25/2009	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	W	W1	ALKI	1978	
11	1/28/1979	02-09-1979	CAR PROWL	THEFT-CARPROWL	EAST	G	G2	CA/SP	1979	
12	07-04-1979	9/15/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q3	QA	1979	
13	01-01-1980	5/28/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	UNKNOWN			UNKNOWN	1980	
14	01-01-1980	10/31/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	F	F1	HP	1980	
15	2/14/1981	2/15/1981	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTHWEST	W	W3	RWA	1981	

XIV. What's the need of filtering the data? Show examples using appropriate commands

Filtering data allows us to clean up our data and remove the entries that provide no information to our analysis.
Further, keeping these entries in our dataset could harm our results potentially and ultimately lead to a wrong/faulty analysis.

Consider the given dataframe for this problem.
In the above analysis, part 6 & 7. when we created zoneGroups by grouping our data entries based on their zone, we noticed that one of the zones was defined as **"UNKNOWN"**.
This occured due to the fact that several entries in our dataset do not have any dataset defined.
This gives us a faulty barplot as well which contains one bar for **UNKNOWN** in it as can be seen above.

This can be corrected however by filtering our data properly and removing these entries before plotting the data.

Start by filtering out the data from the dataframe.

```
df <- df %>% filter(!is.na(Zone) & Zone != 'UNKNOWN')
```

Q2.r* | df | zoneGroups | filteredDf

Filter

	Date_of_event	Date_of_resolution	Type_of_offence	Summary_of_offence	Zone	Block	Division	Locality
1	12/13/1908	12/13/2008	DUI	DUI-LIQUOR	EAST	G	G2	CA/SP
2	6/15/1964	6/15/2010	FAMILY OFFENSE-NONVIOLENT	CHILD-OTHER	WEST	Q	Q2	QA
3	01-01-1973	1/25/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	NORTH	N	N2	NG
4	12/16/1975	12/16/1975	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTH	R	R3	LW/SP
5	01-01-1978	8/25/2009	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	W	W1	ALKI
6	1/28/1979	02-09-1979	CAR PROWL	THEFT-CARPROWL	EAST	G	G2	CA/SP
7	07-04-1979	9/15/2010	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	WEST	Q	Q3	QA
8	01-01-1980	10/31/2012	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	F	F1	HP
9	2/14/1981	2/15/1981	BURGLARY-RESIDENTIAL	BURGLARY-FORCE-RES	SOUTHWEST	W	W3	RWA
10	8/22/1981	8/22/1981	HOMICIDE	HOMICIDE-PREMEDITATED-WEAPON	SOUTH	S	S2	BD
11	9/13/1985	10/13/2014	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	EAST	G	G1	FH
12	1/19/1986	12/19/2016	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTH	S	S2	BD
13	9/29/1988	9/29/1988	MOTOR VEHICLE THEFT	VEH-THEFT-AUTO	WEST	M	M2	SC
14	01-01-1989	1/30/2013	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	SOUTHWEST	W	W2	AJ
15	01-01-1991	4/27/2011	RAPE	PHYSICAL_VIOLENCE-SODOMY	SOUTHWEST	F	F3	SP
16	01-01-1991	7/31/2009	FAMILY OFFENSE-NONVIOLENT	CHILD-ABUSED-NOFORCE	SOUTHWEST	W	W1	*NA*
17	01-01-1993	4/25/2017	PHYSICAL OFFENSE	PHYSICAL_VIOLENCE-OTHER	EAST	G	G2	CA/SP
18	01-01-1993	10-02-2014	THEFT-ALL OTHER	THEFT-OTH	EAST	E	E2	CH
19	02-09-1993	02-09-2008	RAPE	RAPE-WEAPON	SOUTHWEST	W	W3	RWA
20	07-09-1993	12/14/2017	RAPE	RAPE-STRONGARM	NORTH	L	L3	LAKECITY
21	10-08-1993	10-08-1993	HOMICIDE	HOMICIDE-PREMEDITATED-GUN	SOUTH	R	R2	CR

Showing 1 to 22 of 1,899 entries, 8 total columns

We notice that all entries with Zone = 'UNKNOWN' have been removed.

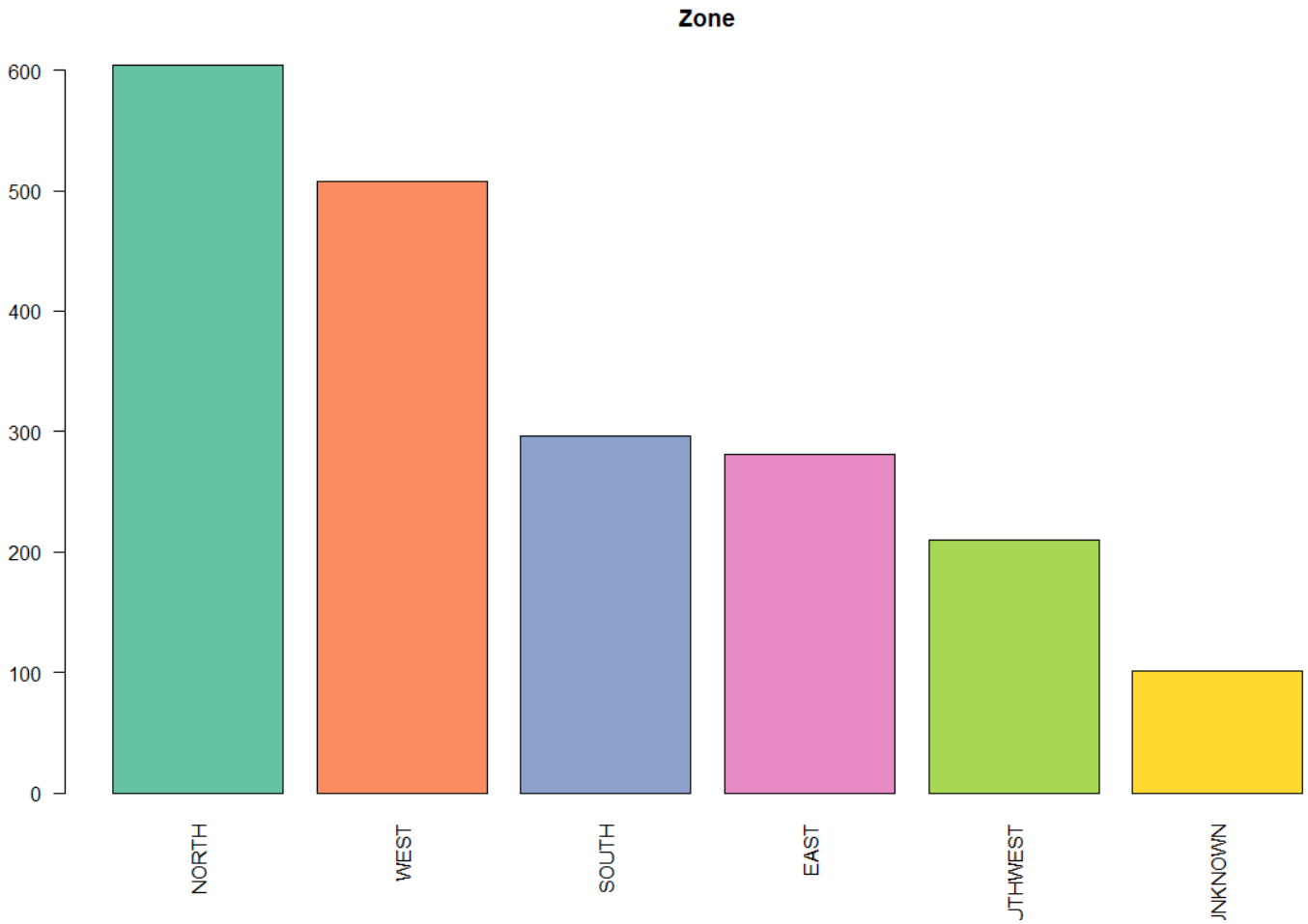
Now we run the same code as in part 6 & 7 above.

```
zoneGroups <- df %>% group_by(Zone)
count(zoneGroups)
```

```
> count(zoneGroups)
# A tibble: 5 x 2
# Groups:   Zone [5]
  Zone          n
  <chr>     <int>
1 EAST        281
2 NORTH       604
3 SOUTH       296
4 SOUTHWEST   210
5 WEST        508
>
```

Note that the entries with UNKNOWN are now gone.

```
barplot(sort(table(zoneGroups$Zone), decreasing = T),
        las = 2, main = "Zone.", col=coul)
```



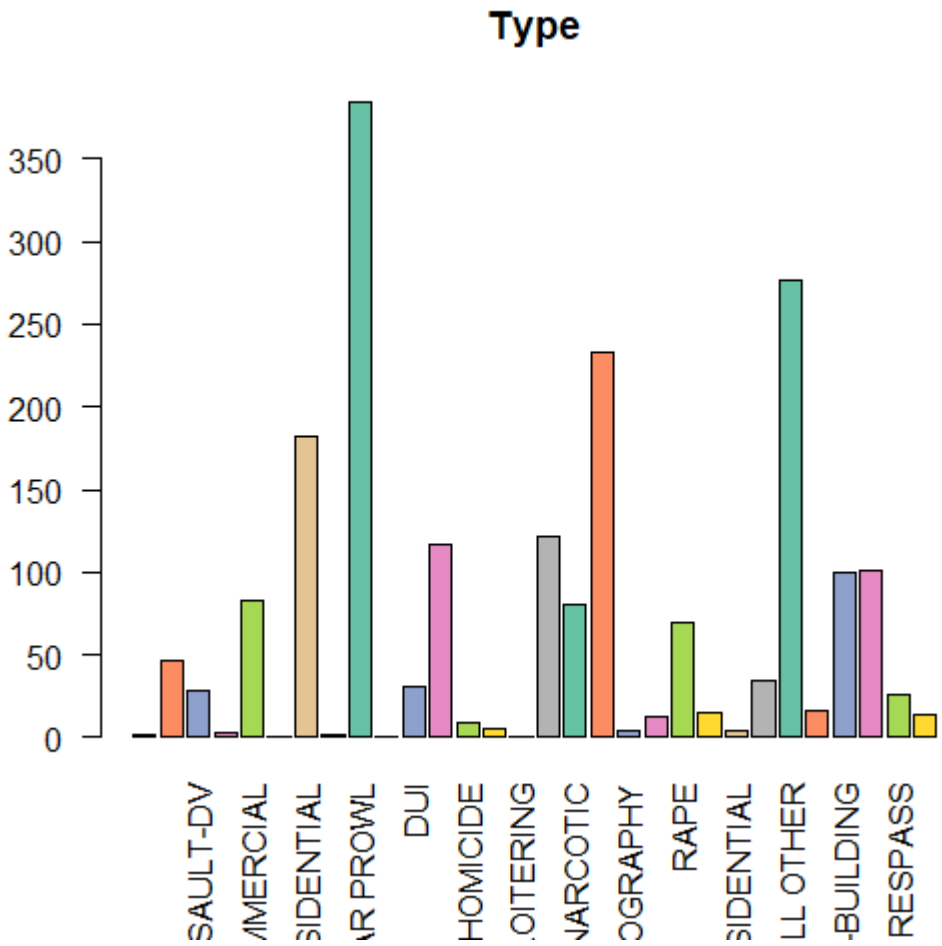
We finally get a proper graph without a meaningless entry in it.

XV. What other charts can you plot for XI ? Which one will leverage more information and why?(elaborate in comments)

There are several possibilities. Let us try a few of them out.

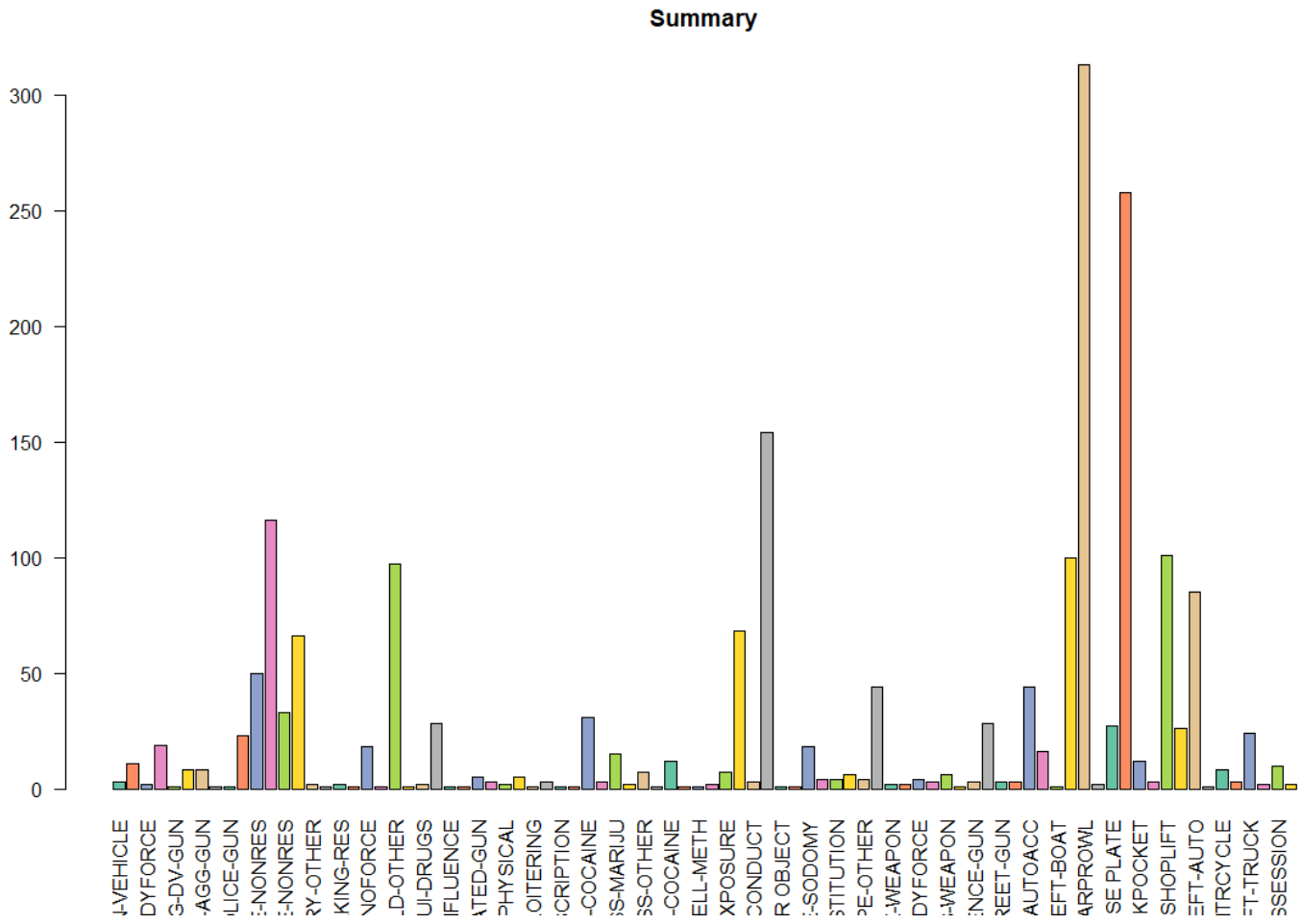
Bar chart that displays the number of offences by Type_of_offence

```
barplot(table(df$Summary_of_offence), las = 2, main = "Summary",col=coul)
```



Bar chart that displays the number of offences by Summary_of_offence

```
barplot(table(df$Summary_of_offence), las = 2, main = "Summary",col=coul)
```



Bar chart that displays the number of offences by Block

```
barplot(table(df$Block[df$Block!=""]), las=2, main="Block", col=coul)
```