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

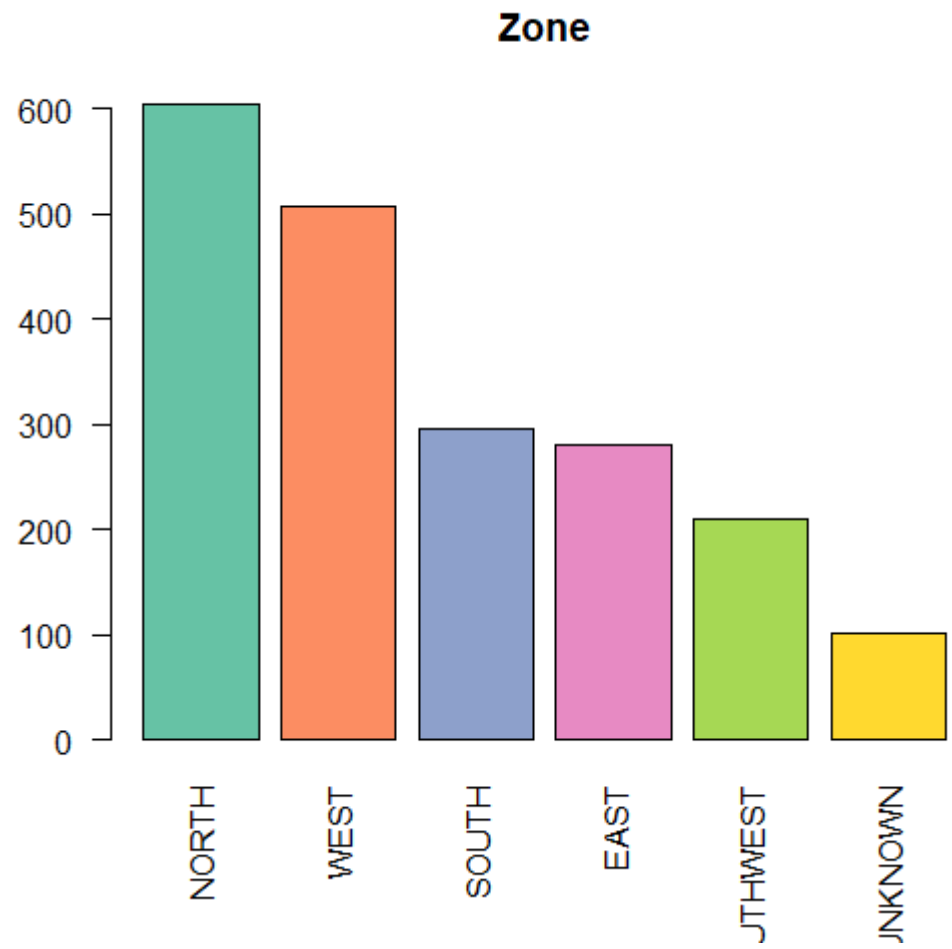
```
count(zoneGroups)
```

```
> # 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

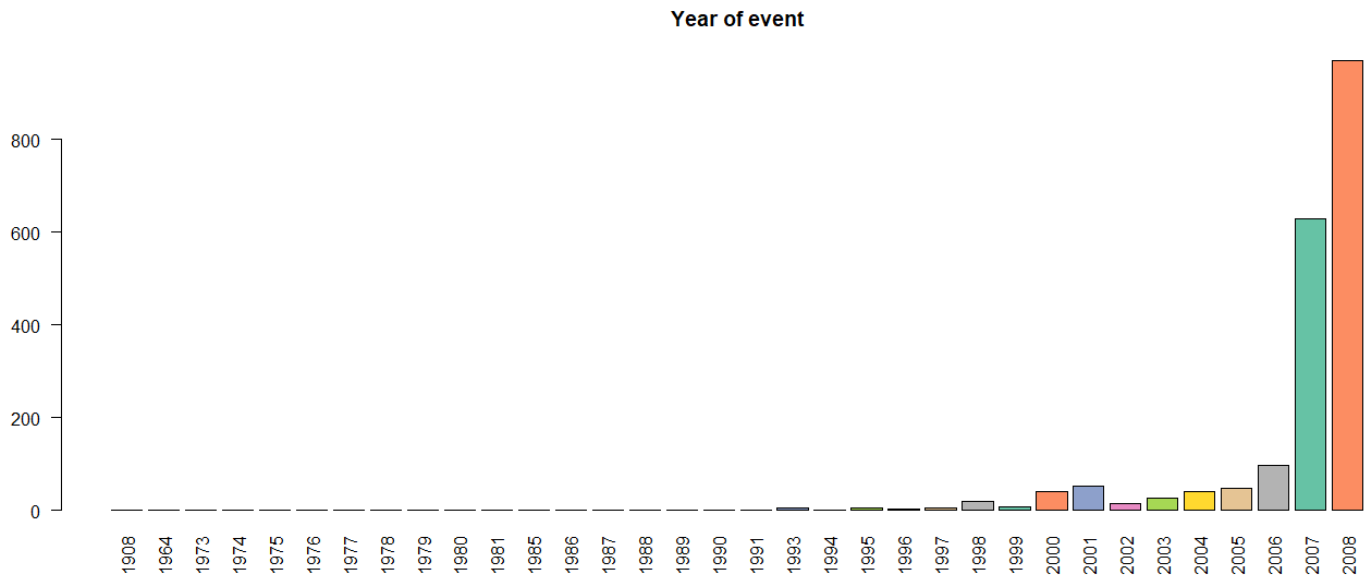
IX. Group the data by year and summarize

```
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
> |
```

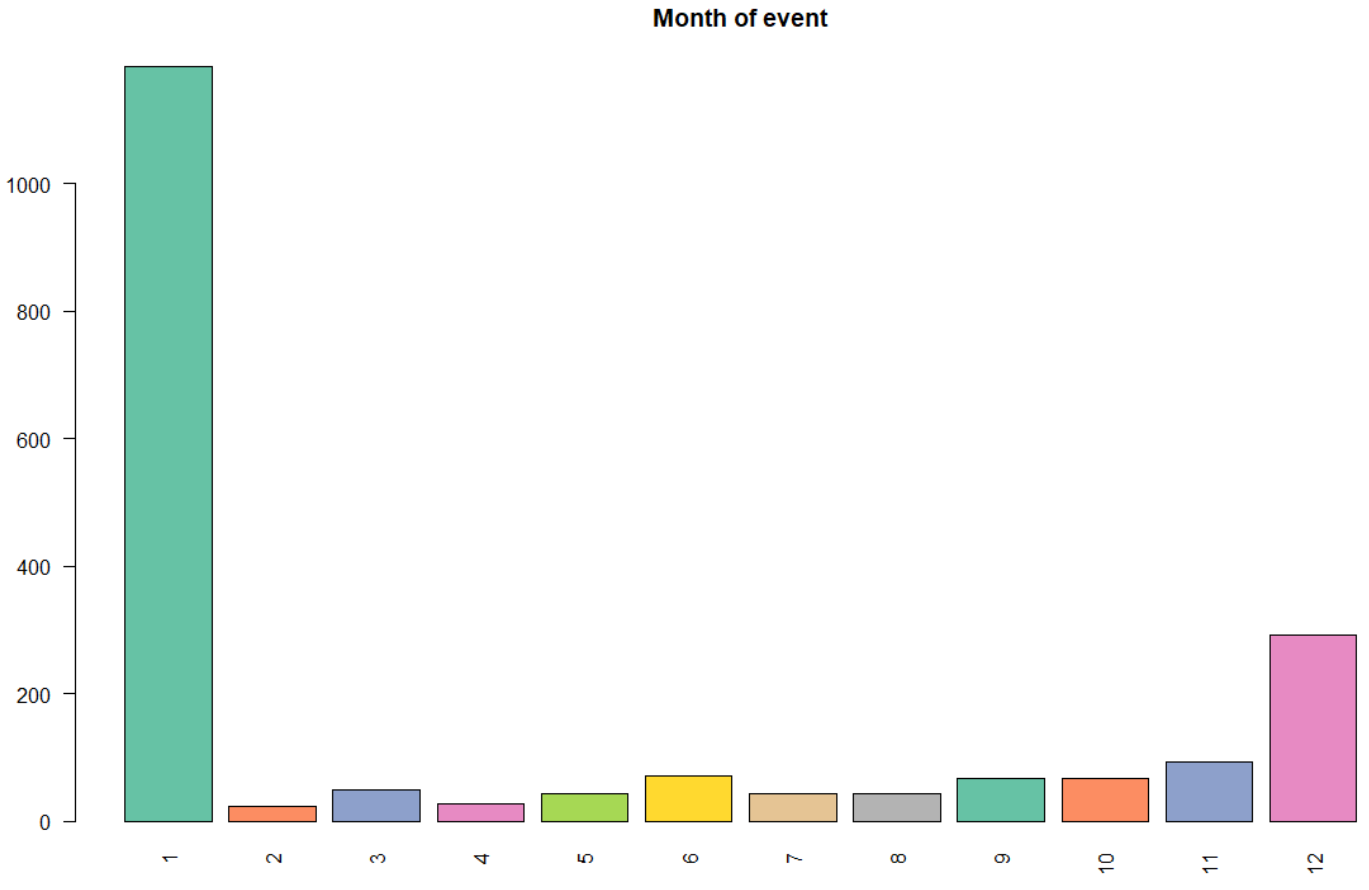
X. Plot a barchart with column Year_of_event that displays the number of offences by year

```
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)
```



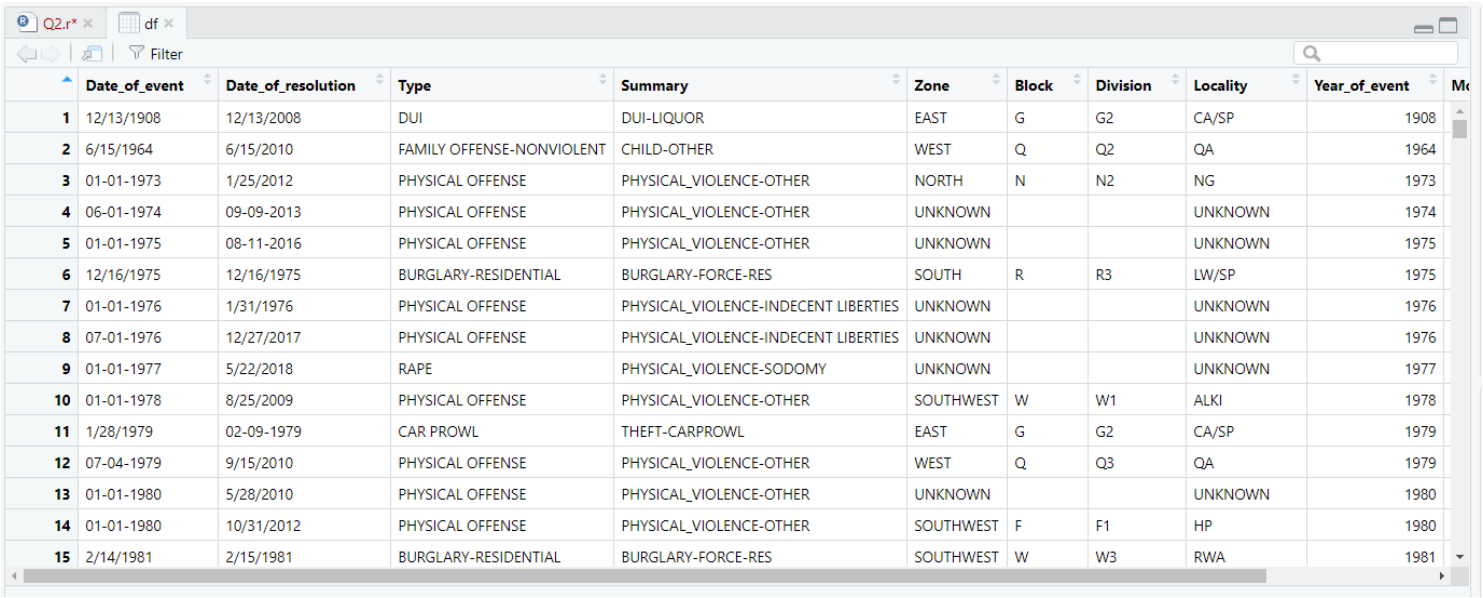
XII. Group and summarize the data by month.

```
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.

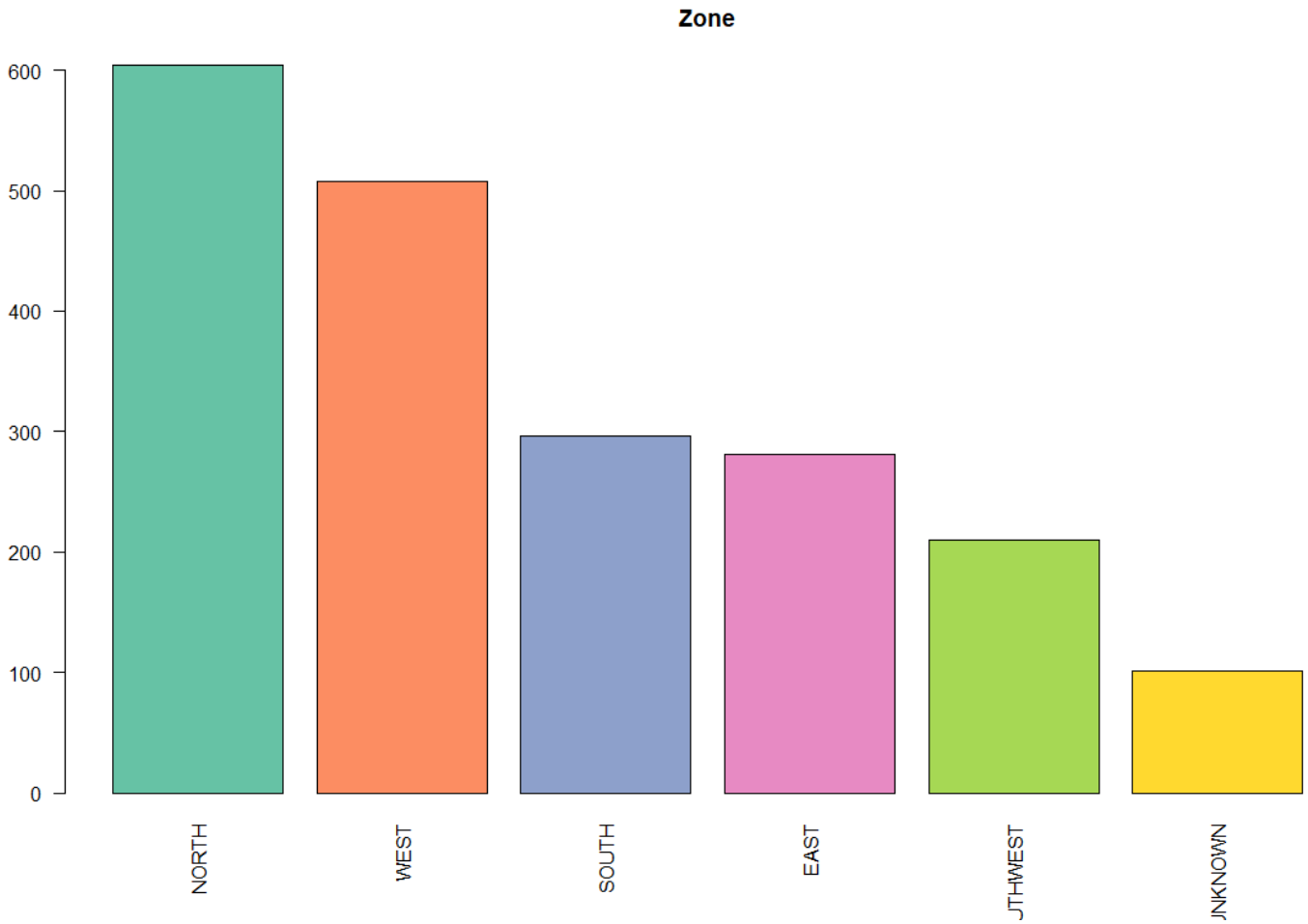
Now we run the same code as in part 6 & 7 above.

```
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)
```



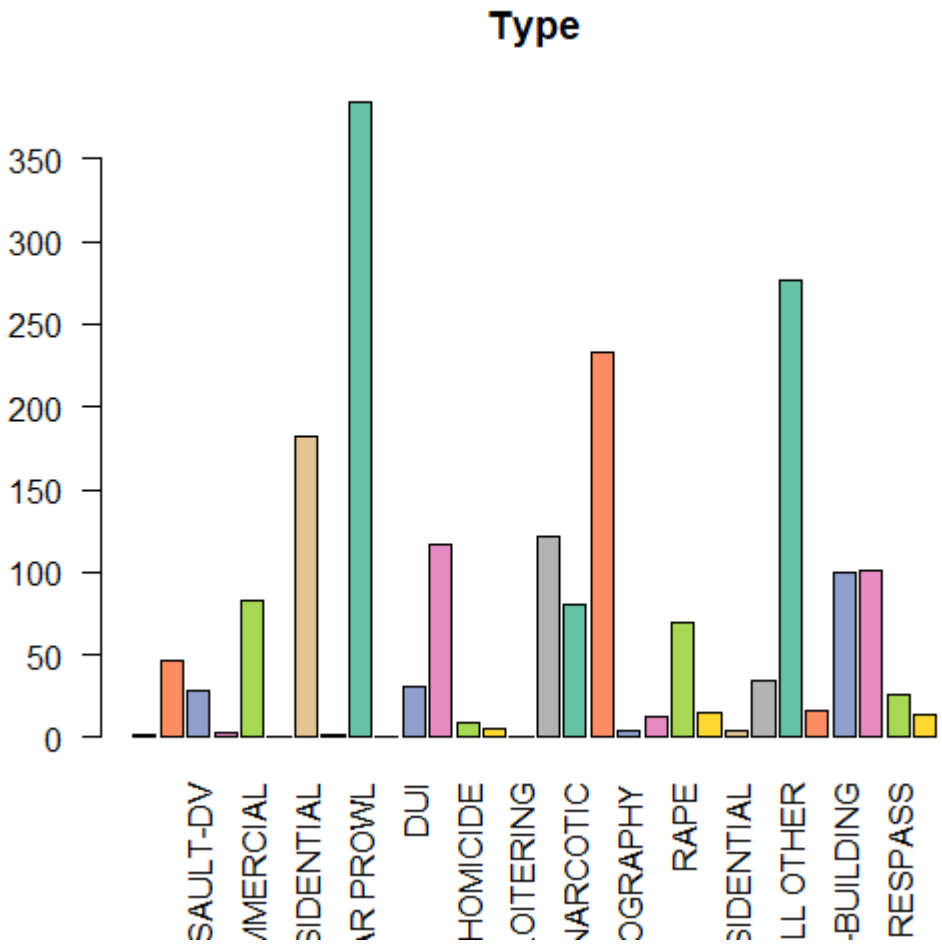
We finally get a proper graph without a meaningless entry in it.

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.

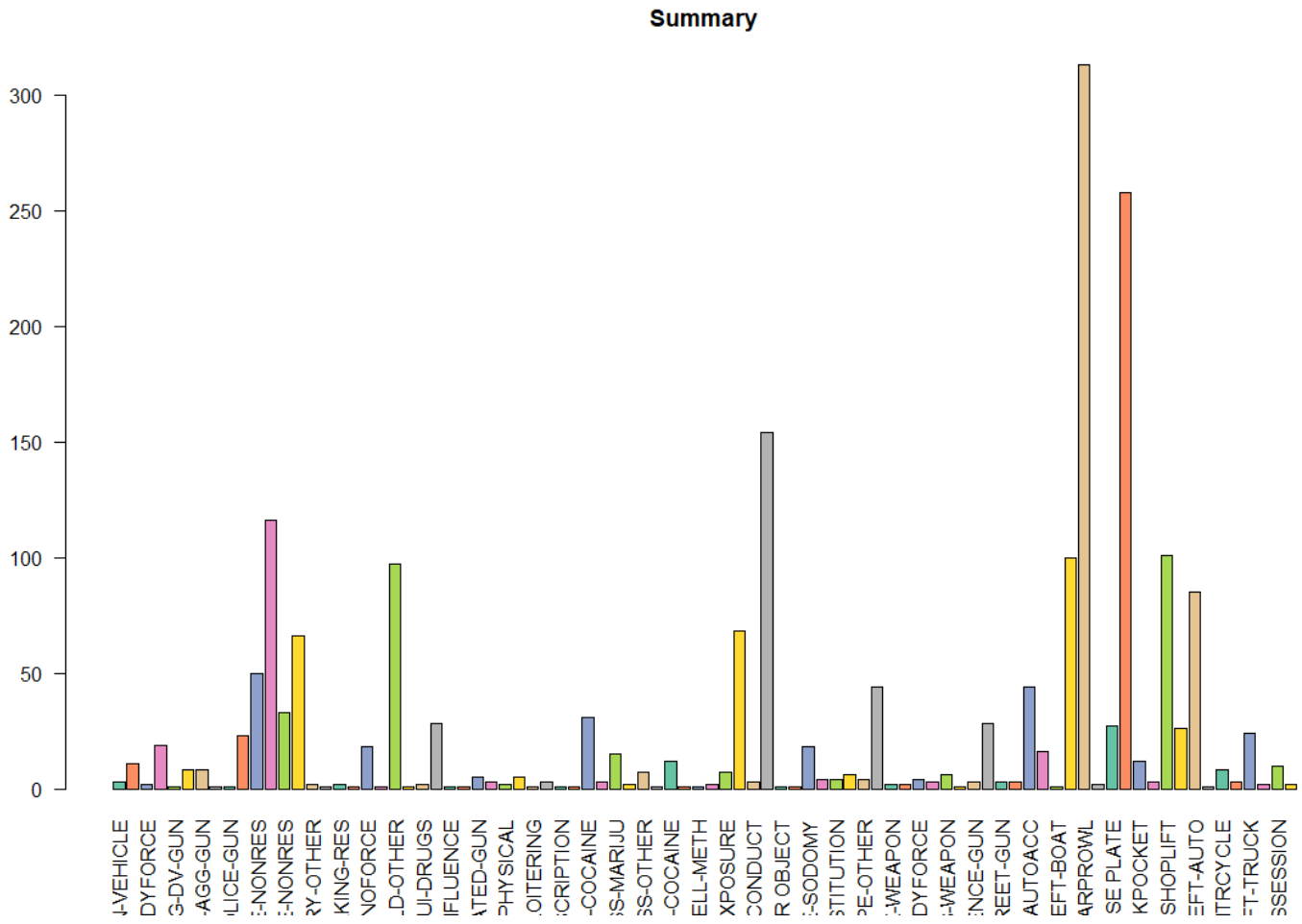
Bar chart that displays the number of offences by Type_of_offence

```
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)
```