MA304 coursework

Dataset Description

How the justice is related to race? Does the police officer behaviour leads to the injuries? Is the racism a premium factor in the crime incidents? These questions alongwith many other queries are needed to be answered from the dataset provided. It has many details related to officer, injuries, incident location, injuries caused, time and date of incident etc. We will discuss the multiple variables in the dataset to make a concise analysis to answer the questions mentioned above.

Overview of data

We can get an overview of the our dataset with the help of `head` function. It gives us important information about the data types of variables in the dataset which can help to determine what variables we should keep. The information from this very initial step can help for EDA analysis.

```
df <- read.csv("~/Documents/R_data_Visualizations/37-00049_UOF-P_2016_prepped.csv",na.strings = c(""))
```

```
#df <- clean_names(df)
#datatable(kable(head(df,5),booktabs=T,format='html') %>%
 #  kable_styling(latex_options = "striped"))

datatable(df)
```

EDA Analysis

- Step1: Getting shape of dataset

```
dim(df)
```

```
## [1] 2384   47
```

- Step2: Removing extra column

This step involves several step from getting duplicates in the data to the actual check of normalization of the data. In the 1st step we will remove the 1st column since it has already same variable names as the 1st row.

- Step3: Converting variable types

We observe there are `character` variables which can converted to factors and `double` variables which can be converted to numeric for plotting so we use the R package named `commonutiladdins` to convert them to desired data type. After converting the result is given as

```
df <-  lapply(df, as.factor) %>% data.frame()

df$OFFICER_YEARS_ON_FORCE <- as.numeric(as.character(df$OFFICER_YEARS_ON_FORCE))

df$STREET_NUMBER <- as.numeric(as.character(df$OFFICER_YEARS_ON_FORCE))
df$SECTOR <- as.numeric(as.character(df$SECTOR))
df$LOCATION_LATITUDE <- as.numeric(as.character(df$LOCATION_LATITUDE))
df$LOCATION_LONGITUDE <- as.numeric(as.character(df$LOCATION_LONGITUDE))
```

```
datatable(diagnose(df))
```

From the above we can notice that there are columns with lots of missing values we can remove them with the help of code below

- Step4: Removing columns with `Nan` values

```
df <-  df %>% select(!matches("USED"))%>%
  select(-c(LOCATION_CITY,LOCATION_STATE,NUMBER_EC_CYCLES, OFFICER_ID, SUBJECT_ID,BEAT, UOF_NUMBER))
```

- Step5: Converting data time format to separate columns

Before the data visualization and normality check, we observe there are variables with date format so we will use `stringr` package to mutate new columns with separate day, date and hour for incidents. It will help us to analyse the data further in data visualization section.

```
df$INCIDENT_DATE <- as.Date(df$INCIDENT_DATE, format = "%m/%d/%Y")
df$INCIDENT_DATE <- gsub("00","20",df$INCIDENT_DATE)
df$INCIDENT_DATE <- as.Date(df$INCIDENT_DATE, format = "%Y-%m-%d")
df$INCIDENT_TIME <- format(strptime(df$INCIDENT_TIME, "%I:%M:%S %p"), "%H:%M:%S")
df$INCIDENT_MONTH <- months(as.Date(df$INCIDENT_DATE))
df$INC_MONTH <-format(df$INCIDENT_DATE,"%m")
df$INCIDENT_HOUR <- as.numeric(substr(df$INCIDENT_TIME, 0, 2))
df$INCIDENT_DAY <- wday(df$INCIDENT_DATE)
df$INC_HOUR <- substr(df$INCIDENT_TIME, 0, 2)
df$INC_DATE <- substr(df$INCIDENT_DATE, 9, 10)

## Create group of datas:

df_year <-  df %>%
  group_by(INCIDENT_DATE,INCIDENT_MONTH,INCIDENT_DAY) %>%
  summarize(count = n())

df_month <-  df %>%
  group_by(INC_MONTH) %>%
  summarize(count = n())

df_day <-  df %>%
  group_by(INCIDENT_DAY,INCIDENT_HOUR) %>%
```

```
  summarize(count = n())

df$INC_HOUR <- substr(df$INCIDENT_TIME, 0, 2)

df   %>% group_by(INC_HOUR) %>%
  summarize(avg =n()) -> df_hour_n
```

- Step6: Central tendency check

For the EDA analysis we have used the package `dlookr` and `Dataexplorer` of R. Following table provides us the information about the central tendency of our dataset.

```
diagnose_numeric(df)
```

```
## # A tibble: 7 x 10
##   variables              min    Q1   mean median    Q3   max  zero minus outlier
##   <chr>                <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <int> <int>   <int>
## 1 OFFICER_YEARS_ON_FO~   0     3     8.05    6    10    36       3     0     240
## 2 SECTOR               110   210   389.    350   610   750       0     0       0
## 3 STREET_NUMBER          0     3     8.05    6    10    36       3     0     240
## 4 LOCATION_LATITUDE     32.6  32.7  32.8    32.8  32.9  33.0     0     0       0
## 5 LOCATION_LONGITUDE   -97.0 -96.8 -96.8   -96.8 -96.8 -96.6     0  2328      44
## 6 INCIDENT_HOUR          0     5    13.0    16    20    23     142     0       0
## 7 INCIDENT_DAY           1     2     4.05    4     6     7       0     0       0
```

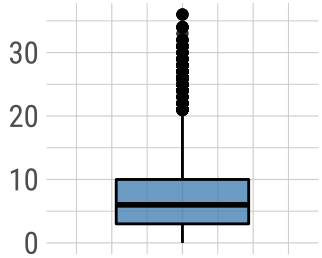
```
# target="SUBJECT_GENDER"
#   op_file="report.html",
#   op_dir=getwd())
```

We can check the outliers of the dataset with the help of boxplots for numeric variables.

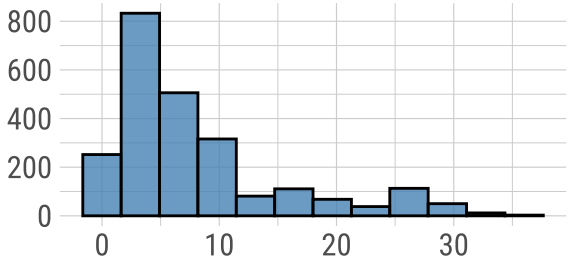
```
df %>%
  plot_outlier(diagnose_outlier(df) %>%
                 filter(outliers_ratio >= 0.5) %>%
                 select(variables) %>%
                 unlist())
```

Outlier Diagnosis Plot (OFFICER_YEARS_ON_FORCE)

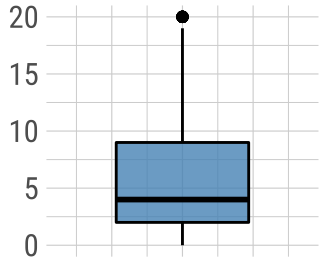
With outliers



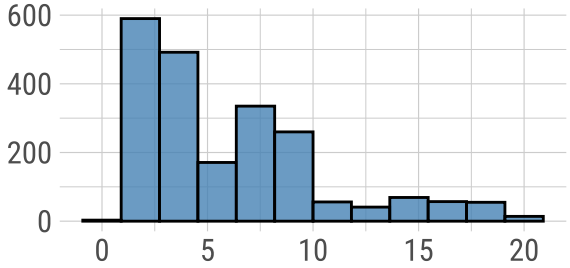
With outliers



Without outliers

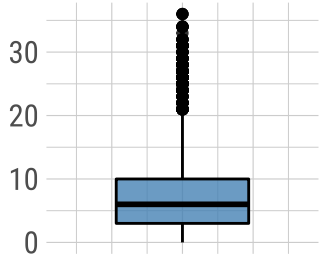


Without outliers

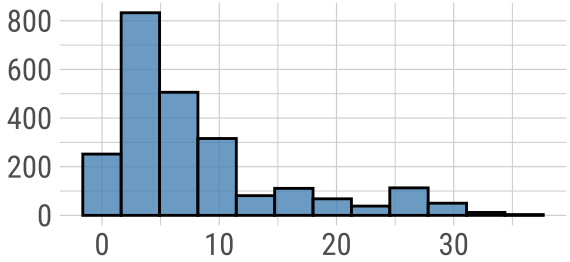


Outlier Diagnosis Plot (STREET_NUMBER)

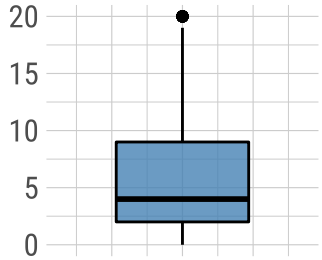
With outliers



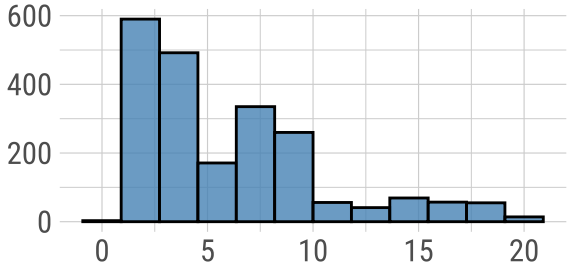
With outliers

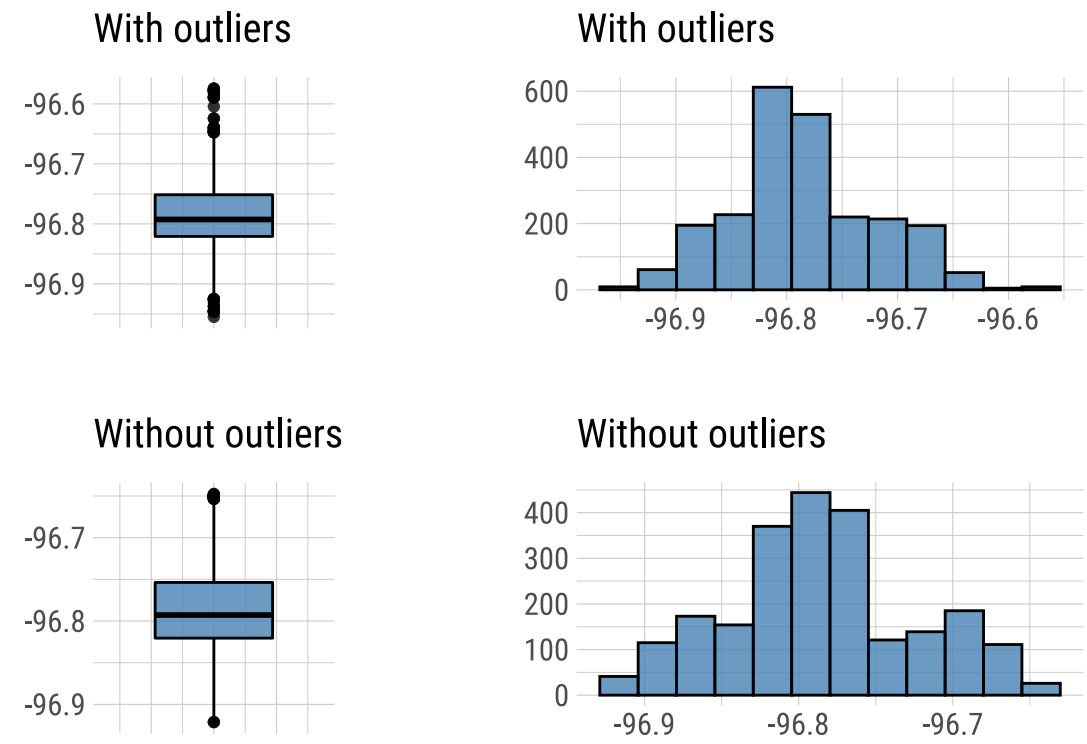


Without outliers



Without outliers





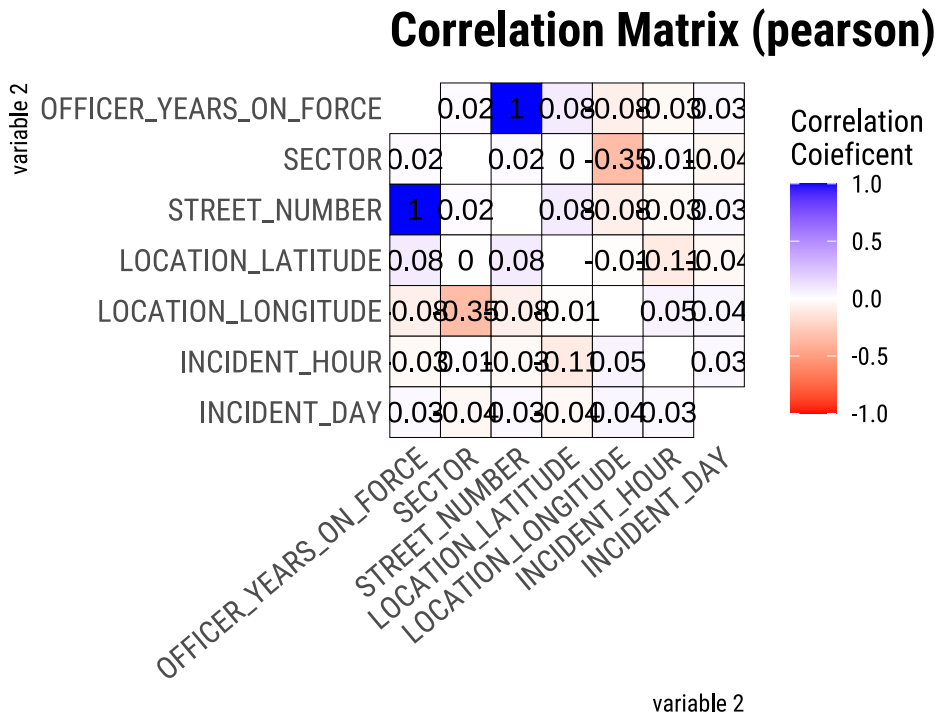
- Step7: Corelation plot

The corelation map given above can be extracted in a tabular form

```
correlate(df)
```

```
## # A tibble: 42 x 3
##    var1                   var2                   coef_corr
##    <fct>                  <fct>                      <dbl>
##  1 SECTOR                 OFFICER_YEARS_ON_FORCE   0.0182
##  2 STREET_NUMBER          OFFICER_YEARS_ON_FORCE   1
##  3 LOCATION_LATITUDE      OFFICER_YEARS_ON_FORCE   0.0839
##  4 LOCATION_LONGITUDE     OFFICER_YEARS_ON_FORCE  -0.0821
##  5 INCIDENT_HOUR          OFFICER_YEARS_ON_FORCE  -0.0329
##  6 INCIDENT_DAY           OFFICER_YEARS_ON_FORCE   0.0286
##  7 OFFICER_YEARS_ON_FORCE SECTOR                   0.0182
##  8 STREET_NUMBER          SECTOR                   0.0182
##  9 LOCATION_LATITUDE      SECTOR                   0.00273
## 10 LOCATION_LONGITUDE     SECTOR                  -0.349
## # ... with 32 more rows
```

```
df %>%
  correlate() %>%
  plot()
```



- Step 8: Normality check We can check the normality of our dataset which will help us to reject or accept our null hypothesis which is given individually for each numeric variable. For instance, for the variable 2, 17, 1, 24, 7, 7, 7, 9, 4, 8, 7, 6, 8, 4, 5, 9, 12, 3, 4, 7, 2, 3, 1, 3, 21, 10, 28, 31, 31, 28, 2, 30, 8, 2, 1, 4, 15, 3, 3, 8, 4, 27, 10, 7, 9, 7, 8, 1, 2, 7, 10, 2, 3, 24, 26, 24, 26, 4, 11, 11, 1, 6, 24, 13, 3, 1, 7, 19, 1, 3, 7, 1, 27, 11, 2, 3, 27, 2, 2, 4, 2, 8, 3, 29, 28, 2, 27, 13, 6, 3, 13, 2, 32, 2, 3, 25, 3, 6, 16, 11, 10, 1, 1, 17, 7, 7, 27, 5, 8, 6, 15, 3, 2, 9, 3, 3, 2, 21, 1, 15, 4, 3, 9, 10, 4, 9, 8, 16, 3, 8, 3, 2, 1, 6, 6, 4, 9, 6, 9, 2, 14, 1, 9, 3, 3, 9, 10, 1, 4, 10, 1, 8, 1, 9, 3, 1, 1, 3, 1, 1, 4, 1, 3, 3, 3, 12, 3, 7, 3, 2, 11, 16, 2, 6, 1, 2, 19, 2, 3, 4, 1, 1, 34, 4, 3, 10, 7, 7, 1, 10, 4, 3, 3, 2, 4, 8, 7, 27, 8, 14, 27, 2, 10, 26, 4, 17, 31, 11, 3, 4, 11, 12, 8, 6, 2, 3, 3, 14, 1, 4, 3, 3, 8, 28, 1, 1, 25, 8, 6, 7, 2, 1, 31, 7, 1, 2, 11, 3, 4, 1, 7, 9, 2, 6, 3, 1, 3, 2, 7, 1, 10, 7, 6, 19, 16, 11, 1, 7, 9, 26, 6, 2, 3, 3, 3, 26, 26, 10, 2, 3, 1, 3, 3, 26, 27, 8, 8, 6, 1, 29, 3, 3, 9, 18, 9, 9, 3, 14, 17, 8, 8, 26, 3, 19, 1, 3, 18, 10, 1, 4, 25, 14, 17, 3, 2, 21, 11, 7, 1, 3, 28, 6, 7, 7, 9, 9, 27, 2, 5, 10, 10, 9, 9, 7, 4, 9, 9, 27, 32, 8, 4, 8, 21, 6, 7, 6, 21, 31, 3, 1, 1, 3, 2, 3, 1, 4, 22, 3, 3, 2, 10, 3, 12, 1, 4, 10, 12, 9, 16, 8, 3, 1, 1, 3, 14, 7, 20, 6, 3, 3, 2, 3, 10, 8, 2, 9, 9, 3, 15, 3, 28, 26, 2, 3, 3, 16, 1, 6, 3, 3, 3, 11, 19, 7, 4, 2, 8, 1, 21, 4, 2, 23, 4, 6, 9, 6, 1, 3, 8, 3, 7, 6, 17, 28, 7, 19, 3, 3, 3, 14, 24, 4, 14, 9, 2, 1, 1, 14, 2, 3, 10, 6, 26, 7, 3, 1, 9, 2, 3, 8, 3, 11, 3, 3, 6, 10, 32, 1, 15, 9, 9, 7, 1, 2, 9, 1, 4, 13, 6, 7, 21, 3, 1, 7, 7, 8, 2, 4, 8, 5, 10, 2, 11, 2, 18, 3, 10, 21, 10, 9, 19, 3, 2, 2, 4, 3, 1, 7, 18, 3, 26, 8, 4, 8, 1, 10, 8, 1, 28, 3, 6, 15, 8, 3, 10, 2, 2, 7, 2, 3, 2, 4, 8, 3, 4, 3, 8, 3, 16, 26, 10, 4, 8, 7, 1, 13, 1, 1, 9, 3, 6, 18, 7, 14, 2, 21, 12, 16, 7, 9, 8, 7, 26, 8, 6, 1, 24, 10, 3, 8, 11, 8, 10, 8, 9, 18, 3, 25, 2, 11, 15, 4, 8, 3, 10, 3, 11, 19, 2, 6, 1, 27, 26, 9, 15, 1, 7, 3, 3, 2, 6, 1, 12, 2, 7, 8, 2, 8, 3, 3, 26, 9, 2, 4, 4, 26, 9, 3, 7, 1, 31, 6, 3, 12, 20, 27, 7, 1, 5, 8, 28, 2, 2, 16, 2, 27, 10, 1, 6, 9, 6, 7, 3, 10, 9, 2, 1, 1, 18, 1, 3, 2, 7, 3, 7, 3, 7, 27, 8, 29, 2, 1, 3, 10, 22, 9, 4, 9, 1, 1, 2, 15, 3, 11, 3, 3, 3, 2, 1, 28, 26, 25, 3, 2, 1, 8, 8, 2, 3, 1, 9, 2, 9, 2, 3, 1, 2, 25, 8, 3, 3, 26, 6, 21, 22, 7, 10, 3, 27, 0, 3, 10, 8, 14, 2, 27, 26, 25, 6, 11, 14, 3, 28, 11, 7, 13, 3, 4, 8, 2, 1, 2, 9, 6, 21, 2, 2, 18, 3, 1, 8, 1, 4, 1, 15, 8, 2, 9, 2, 9, 34, 3, 2, 8, 2, 3, 1, 2, 26, 3, 3, 7, 2, 25, 20, 10, 1, 10, 26, 2, 1, 11, 3, 2, 2, 24, 2, 7, 3, 14, 1, 3, 7, 3, 13, 27, 1, 28, 2, 27, 5, 4, 25, 3, 13, 19, 1, 9, 3, 1, 7, 12, 9, 6, 14, 1, 3, 20, 4, 8, 3, 3, 19, 27, 25, 4, 3, 7, 2, 2, 25, 3, 3, 10, 2, 15, 3, 1, 2, 4, 3, 2, 14, 2, 3, 3, 1, 6, 32, 16, 4, 7, 3, 2, 8, 11, 2, 19, 3, 8, 21, 4, 19, 19, 2, 3, 9, 3, 2, 7, 9, 14, 1, 2, 1, 8, 3, 19, 1, 7, 4, 2, 2, 7, 3, 9, 1, 7, 9, 13, 6, 6, 11, 5, 8, 7, 6, 3, 11, 14, 1, 9, 4, 6, 7, 9, 1, 1, 11, 3, 2, 16, 7, 1, 7, 6, 1, 24, 19, 14, 7, 9, 1, 14, 29, 2, 3, 10, 9, 7, 1, 9, 3, 4, 2, 27, 6, 3, 7, 10, 1, 3, 11, 6, 8, 2, 7, 2, 29, 10, 1, 22, 3, 3, 3, 2, 7, 1, 3, 26, 2, 11, 3, 3, 7, 2, 1, 3, 2, 2, 3, 8, 3, 28, 27, 2, 1, 3, 1, 12, 2, 29, 2, 27, 8, 19, 7, 16, 2, 3, 3, 9, 1, 8, 3, 16, 7, 8, 7, 2, 18, 2, 1, 2, 8, 8, 2, 3, 3, 1, 3, 7, 4, 2, 2, 11, 15, 3, 4, 3, 2, 9, 2, 14, 2, 9, 4, 3, 25, 8, 6, 2, 3, 17, 16, 2, 2, 8, 7, 17, 9, 3, 13, 30, 1, 9, 3, 7, 6, 9, 7, 10, 13, 6, 14, 2, 2, 19, 4, 2, 17,

3, 4, 7, 14, 1, 26, 7, 8, 26, 6, 6, 1, 8, 9, 2, 4, 8, 2, 4, 14, 16, 15, 2, 12, 9, 4, 2, 10, 2, 4, 3, 9, 25, 7, 4, 11, 9, 31, 3, 3, 2, 2, 10, 1, 17, 4, 2, 23, 6, 11, 10, 4, 9, 5, 7, 3, 3, 9, 3, 6, 3, 7, 10, 3, 2, 2, 17, 10, 1, 1, 3, 21, 2, 3, 9, 22, 3, 7, 4, 2, 1, 29, 1, 2, 25, 9, 14, 3, 1, 3, 4, 11, 8, 10, 7, 22, 19, 7, 4, 11, 11, 3, 8, 7, 7, 4, 3, 16, 13, 3, 26, 27, 9, 2, 1, 18, 1, 7, 2, 4, 19, 31, 10, 9, 26, 1, 7, 10, 27, 3, 7, 1, 8, 16, 6, 30, 1, 8, 3, 1, 19, 1, 7, 21, 25, 2, 1, 1, 4, 2, 3, 3, 18, 2, 9, 13, 31, 25, 8, 25, 2, 1, 4, 1, 10, 16, 5, 7, 6, 2, 10, 4, 3, 6, 10, 10, 8, 4, 2, 3, 3, 6, 12, 4, 4, 2, 1, 4, 8, 9, 9, 1, 3, 9, 1, 9, 8, 16, 2, 2, 3, 2, 9, 9, 11, 1, 7, 10, 2, 9, 12, 9, 2, 7, 6, 6, 2, 16, 7, 2, 10, 10, 9, 2, 26, 6, 25, 23, 8, 1, 8, 6, 10, 10, 3, 4, 9, 15, 12, 1, 1, 3, 6, 26, 4, 9, 4, 7, 7, 26, 20, 4, 9, 3, 17, 3, 3, 8, 4, 7, 7, 26, 18, 24, 10, 4, 6, 10, 4, 3, 9, 3, 7, 7, 3, 27, 7, 4, 3, 7, 7, 10, 11, 2, 4, 15, 20, 20, 3, 2, 4, 3, 3, 3, 1, 15, 4, 4, 6, 23, 8, 14, 1, 23, 2, 10, 2, 31, 10, 3, 7, 8, 33, 26, 1, 9, 28, 7, 4, 3, 9, 2, 18, 26, 8, 1, 7, 20, 4, 6, 9, 19, 2, 2, 7, 9, 1, 2, 4, 10, 11, 2, 16, 7, 3, 4, 2, 3, 9, 4, 27, 3, 1, 3, 8, 8, 3, 7, 4, 3, 9, 8, 9, 4, 12, 9, 6, 2, 8, 10, 15, 7, 4, 10, 1, 1, 2, 11, 26, 2, 7, 3, 3, 10, 7, 1, 2, 4, 2, 6, 8, 2, 2, 24, 1, 9, 10, 1, 3, 6, 10, 3, 2, 16, 19, 4, 3, 25, 4, 4, 3, 3, 5, 1, 2, 4, 2, 3, 16, 2, 1, 2, 8, 6, 3, 6, 15, 3, 2, 1, 9, 17, 26, 3, 3, 2, 1, 9, 1, 4, 24, 21, 6, 3, 1, 11, 19, 8, 3, 1, 7, 3, 1, 8, 6, 8, 26, 12, 7, 10, 3, 16, 1, 7, 31, 23, 2, 4, 2, 21, 1, 2, 1, 8, 6, 4, 25, 8, 2, 34, 6, 4, 8, 14, 1, 3, 7, 8, 3, 4, 13, 3, 1, 2, 7, 1, 8, 24, 2, 8, 21, 7, 4, 6, 1, 26, 4, 16, 4, 7, 2, 6, 21, 1, 24, 25, 27, 31, 4, 3, 9, 1, 6, 7, 26, 9, 24, 3, 9, 2, 4, 3, 1, 10, 20, 9, 1, 21, 18, 2, 9, 1, 3, 16, 3, 36, 10, 2, 4, 5, 4, 9, 2, 9, 1, 4, 2, 31, 6, 16, 7, 8, 2, 9, 18, 3, 4, 2, 6, 11, 6, 6, 4, 1, 1, 9, 6, 16, 9, 4, 5, 7, 4, 1, 9, 7, 4, 11, 6, 9, 2, 3, 9, 7, 8, 4, 26, 3, 3, 8, 1, 1, 8, 34, 2, 18, 9, 9, 4, 4, 7, 11, 3, 9, 2, 7, 9, 3, 3, 3, 10, 15, 6, 8, 7, 11, 7, 21, 9, 1, 9, 4, 1, 7, 3, 3, 22, 10, 6, 2, 7, 9, 19, 9, 24, 6, 3, 4, 24, 2, 3, 2, 2, 26, 1, 11, 9, 1, 6, 2, 3, 1, 16, 1, 7, 4, 3, 1, 16, 2, 3, 8, 1, 4, 7, 7, 8, 4, 1, 2, 2, 9, 2, 6, 1, 2, 20, 2, 6, 2, 3, 3, 24, 12, 3, 2, 2, 2, 9, 3, 2, 6, 2, 4, 4, 4, 26, 3, 4, 11, 4, 2, 7, 7, 2, 8, 4, 13, 5, 2, 26, 6, 10, 3, 4, 6, 6, 1, 6, 1, 3, 15, 8, 3, 3, 4, 26, 28, 3, 2, 3, 2, 4, 4, 3, 4, 3, 26, 7, 1, 2, 3, 2, 1, 9, 15, 2, 4, 17, 10, 10, 34, 1, 18, 11, 8, 7, 22, 2, 2, 1, 3, 2, 1, 2, 2, 11, 9, 25, 7, 6, 24, 13, 26, 6, 1, 6, 1, 2, 3, 2, 4, 10, 3, 4, 7, 3, 1, 3, 7, 16, 2, 26, 2, 8, 1, 3, 26, 2, 2, 3, 3, 1, 14, 3, 7, 6, 3, 17, 6, 6, 8, 2, 6, 31, 34, 13, 9, 4, 8, 27, 4, 31, 8, 7, 15, 7, 4, 15, 2, 4, 1, 21, 1, 2, 28, 1, 16, 7, 9, 6, 6, 7, 12, 2, 1, 7, 26, 6, 4, 9, 3, 3, 3, 2, 2, 2, 7, 9, 27, 10, 8, 2, 14, 2, 8, 10, 21, 1, 24, 4, 7, 7, 3, 18, 6, 15, 1, 4, 9, 10, 9, 4, 7, 2, 8, 21, 18, 4, 24, 1, 6, 9, 8, 8, 3, 8, 2, 6, 2, 2, 7, 1, 18, 4, 2, 9, 1, 14, 1, 8, 2, 6, 9, 7, 6, 1, 12, 13, 1, 7, 7, 2, 17, 7, 8, 7, 4, 11, 1, 16, 2, 24, 9, 11, 9, 10, 16, 1, 1, 1, 1, 7, 7, 1, 4, 1, 3, 26, 1, 8, 1, 8, 7, 6, 0, 20, 9, 2, 7, 31, 6, 3, 32, 8, 22, 2, 10, 2, 6, 7, 9, 1, 9, 26, 2, 10, 2, 9, 7, 2, 30, 2, 8, 19, 6, 13, 30, 2, 21, 7, 2, 2, 8, 7, 13, 1, 10, 9, 6, 6, 3, 10, 1, 4, 11, 5, 3, 2, 13, 1, 2, 3, 9, 6, 2, 10, 1, 2, 26, 14, 1, 9, 3, 2, 27, 22, 10, 3, 9, 1, 6, 7, 15, 4, 6, 4, 10, 4, 11, 6, 10, 3, 6, 1, 2, 7, 16, 3, 8, 7, 3, 2, 1, 4, 16, 11, 6, 26, 7, 19, 9, 6, 1, 10, 11, 6, 2, 9, 1, 19, 2, 8, 8, 9, 7, 7, 2, 2, 1, 4, 29, 8, 7, 27, 6, 36, 10, 25, 2, 2, 9, 7, 2, 2, 18, 6, 1, 5, 7, 1, 2, 2, 24, 6, 18, 7, 14, 2, 2, 20, 2, 2, 3, 4, 6, 9, 10, 4, 15, 2, 25, 2, 6, 1, 7, 25, 10, 8, 6, 17, 6, 4, 14, 4, 20, 2, 3, 6, 1, 2, 6, 18, 7, 6, 8, 7, 6, 3, 11, 2, 1, 2, 6, 14, 14, 6, 2, 30, 36, 6, 10, 1, 3, 2, 1, 7, 27, 3, 26, 2, 3, 9, 7, 7, 16, 26, 4, 31, 6, 3, 6, 3, 6, 2, 2, 2, 9, 2, 10, 10, 11, 26, 10, 4, 11, 2, 27, 1, 19, 19, 2, 3, 2, 5, 7, 1, 9, 7, 2, 1, 2, 6, 2, 2, 1, 2, 3, 6, 8, 4, 8, 3, 2, 2, 10, 2, 2, 9, 2, 4, 25, 2, 8, 19, 1, 2, 14, 1, 1, 14, 2, 2, 5, 15, 8, 14, 2, 2, 2, 2, 3, 8, 7, 4, 4, 6, 5, 2, 3, 6, 2, 8, 2, 6, 6, 2, 4, 7, 8, 29, 9, 14, 3, 4, 28, 6, 3, 2, 3, 5, 2, 2, 8, 10, 18, 1, 2, 13, 2, 4, 5, 1, 20, 2, 6, 9, 15, 7, 2, 29, 2, 7, 14, 9, 2, 2, 2, 7, 2, 16, 7, 9, 10, 8, 4, 4, 6, 15, 3, 2, 1, 11, 11, 2, 2, 1, 18, 14, 7, 17, 2, 2, 5, 0, 2, 9, 7, 17, 7, 6, 9 we make a null hypothesis that it is not normal. If we get p-value < 0.05 we can reject our null hypothesis at 5% siginificance level.

```
normality(df)
```

```
## # A tibble: 7 x 4
##   vars                   statistic  p_value sample
##   <chr>                      <dbl>    <dbl>  <dbl>
## 1 OFFICER_YEARS_ON_FORCE     0.811 1.91e-46   2383
## 2 SECTOR                     0.912 3.56e-35   2383
## 3 STREET_NUMBER              0.811 1.91e-46   2383
## 4 LOCATION_LATITUDE          0.955 4.33e-26   2383
## 5 LOCATION_LONGITUDE         0.984 1.35e-15   2383
## 6 INCIDENT_HOUR              0.890 3.38e-38   2383
## 7 INCIDENT_DAY               0.900 6.97e-37   2383
```

The p-value for all the numeric variables is less than 0.05 at significance level of 5% so we consider that our

data is normal after the steps given above.

- Step 10: Skewness check

We have not included the skewness plots in our dataset till now which is given below with a code snippet.

```
find_skewness(df)
```

```
## [1]  6 21 26
```

```
find_skewness(df, index = FALSE)
```

```
## [1] "OFFICER_YEARS_ON_FORCE" "STREET_NUMBER"          "LOCATION_LATITUDE"
```

```
find_skewness(df, value = TRUE)
```

```
## OFFICER_YEARS_ON_FORCE                 SECTOR          STREET_NUMBER
##                  1.484                  0.268                  1.484
##      LOCATION_LATITUDE     LOCATION_LONGITUDE          INCIDENT_HOUR
##                  0.594                  0.287                 -0.462
```

So the major numeric variables of on duty years is skewed which need to analysed to remove skewness. Other 2 variables can be reject for skewnss removal since they do not weigh much in the analysis.

We can remove the skewness by

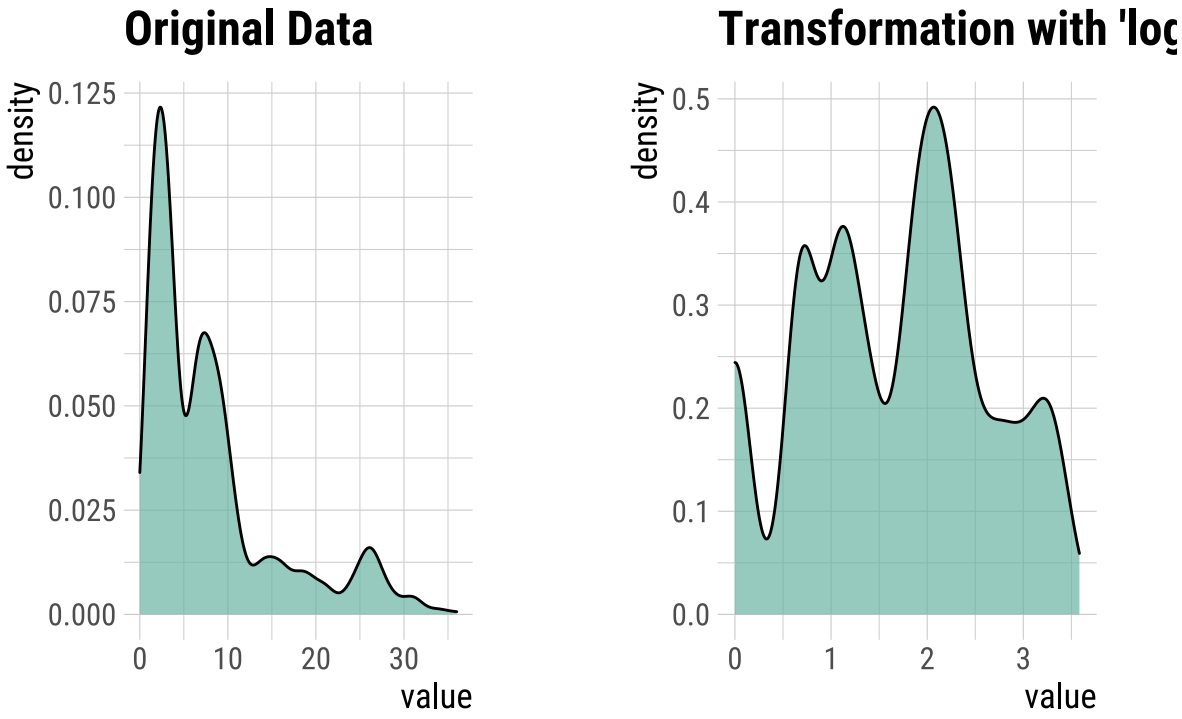
```
OFFICER_YEARS_ON_FORCE = transform(df$OFFICER_YEARS_ON_FORCE, method = "log")

summary(OFFICER_YEARS_ON_FORCE)
```

```
## * Resolving Skewness with log
##
## * Information of Transformation (before vs after)
##              Original Transformation
## n        2383.0000000   2383.0000000
## na          0.0000000      0.0000000
## mean        8.0490978           -Inf
## sd          7.5624813            NaN
## se_mean     0.1549181            NaN
## IQR         7.0000000      1.2039728
## skewness    1.4852096            NaN
## kurtosis    1.4770790            NaN
## p00         0.0000000           -Inf
## p01         1.0000000      0.0000000
## p05         1.0000000      0.0000000
## p10         1.0000000      0.0000000
## p20         2.0000000      0.6931472
## p25         3.0000000      1.0986123
## p30         3.0000000      1.0986123
## p40         4.0000000      1.3862944
## p50         6.0000000      1.7917595
```

```
## p60          7.0000000      1.9459101
## p70          9.0000000      2.1972246
## p75         10.0000000      2.3025851
## p80         11.0000000      2.3978953
## p90         21.0000000      3.0445224
## p95         26.0000000      3.2580965
## p99         31.0000000      3.4339872
## p100        36.0000000      3.5835189
```

plotting the variables after skewness removal

```
plot(OFFICER_YEARS_ON_FORCE)
```



Alongwith with individual steps described above we can create an automatic EDA report from the data.

Data Visualizations

The data visualization are greatly helpful for insights into the data. Following graphs will explain mainly the subject characteristics to the other parameters. We will explain each graph in a short sentences before each graph.

The below figure shows that the incidents were greatly reduced in the mornings hours.

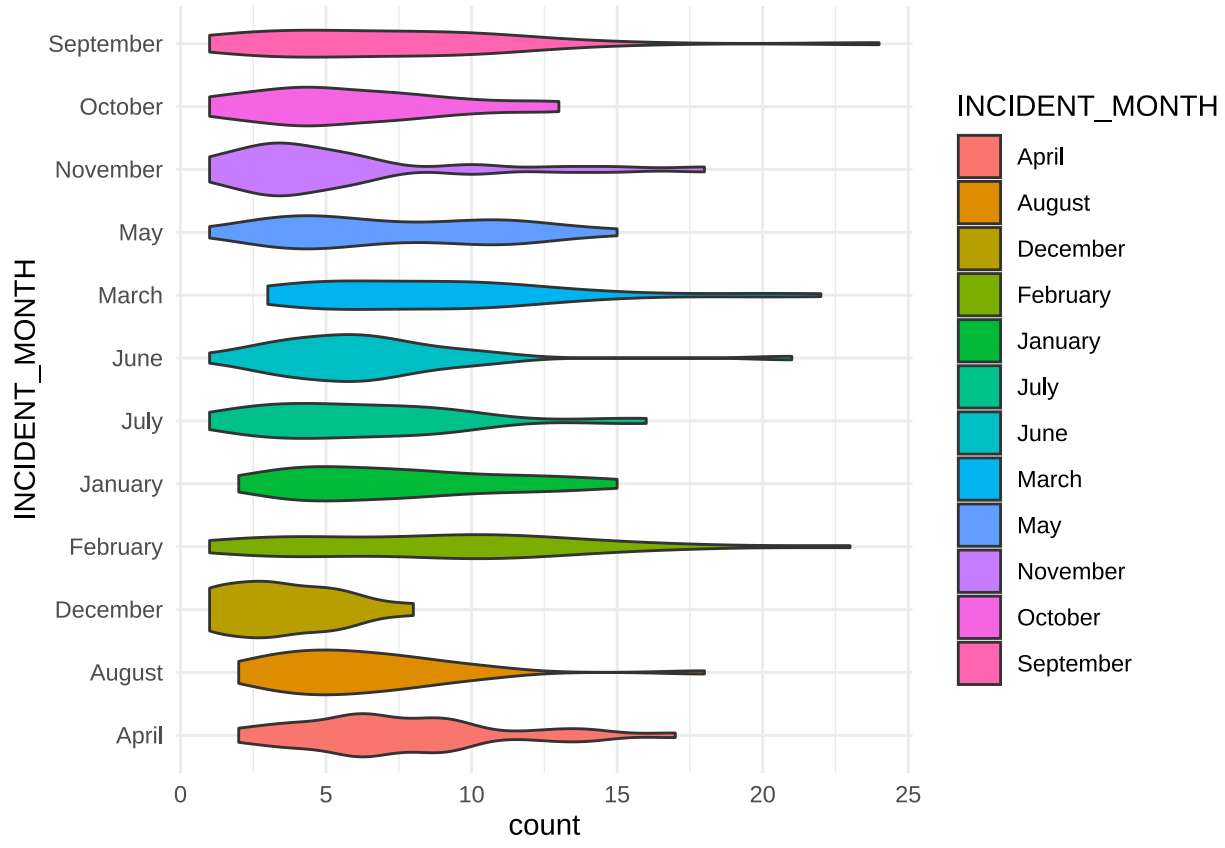
```
p <- df_day %>%
 filter(!is.na(INCIDENT_HOUR)) %>%
 ggplot() +
 aes(x = INCIDENT_HOUR, y = count, size = count) +
 geom_point(shape = "circle open",
```

```
 colour = "#B22222") +
 geom_smooth(span = 0.75) +
 ggthemes::theme_base()


ggplotly(p)
```

Incidents frequency at different hours

Following violin plots show that the incidents are greatly reducted in the month of feb as compared to other months. Moreover the variations from the mean can be greatly observed in the winter months.

```
ggplot(df_year) +
 aes(x = count, y = INCIDENT_MONTH, fill = INCIDENT_MONTH) +
 geom_violin(adjust = 1L,
 scale = "area") +
 scale_fill_hue(direction = 1) +
 theme_minimal()
```



Following graph describes the fact that Black Males are most likely to be involved in incidents.

```
p <- df %>%
 filter(!(SUBJECT_RACE %in% "NULL")) %>%
 filter(SUBJECT_GENDER %in% c("Female", "Male")) %>%
 filter(!(INCIDENT_REASON %in%
```

```
 "NULL")) %>%
 filter(!(REASON_FOR_FORCE %in% "NULL")) %>%
 filter(!is.na(INCIDENT_HOUR)) %>%
 filter(!is.na(INC_HOUR)) %>%
 ggplot() +
 aes(x = SUBJECT_RACE, fill = SUBJECT_RACE) +
 geom_bar() +
 scale_fill_hue(direction = 1) +
 ggthemes::theme_base() +
 theme(legend.position = "bottom") +
 facet_wrap(vars(SUBJECT_GENDER), ncol = 2L)

ggplotly(p)
```

Whenever there is chance of Arrest, the police officer will most likely use force as shown in graph below.

```
p <- df %>%
 filter(!(SUBJECT_RACE %in% "NULL")) %>%
 filter(SUBJECT_GENDER %in% c("Female", "Male")) %>%
 filter(!(INCIDENT_REASON %in%
 "NULL")) %>%
 filter(!(REASON_FOR_FORCE %in% "NULL")) %>%
 filter(!is.na(INCIDENT_HOUR)) %>%
 filter(!is.na(INC_HOUR)) %>%
 ggplot() +
 aes(x = REASON_FOR_FORCE, fill = SUBJECT_GENDER, weight = OFFICER_YEARS_ON_FORCE) +
 geom_bar() +
 scale_fill_hue(direction = 1) +
 ggthemes::theme_few() +
 theme(legend.position = "bottom") +
 facet_wrap(vars(OFFICER_INJURY),
 scales = "free", ncol = 1L)

ggplotly(p)
```

Both the young officers and old officers will undergo injury during incidents.

```
p <- df %>%
 filter(!(SUBJECT_RACE %in% "NULL")) %>%
 filter(SUBJECT_GENDER %in% c("Female", "Male")) %>%
 filter(!(INCIDENT_REASON %in%
 "NULL")) %>%
 filter(!(REASON_FOR_FORCE %in% "NULL")) %>%
 filter(!is.na(INCIDENT_HOUR)) %>%
 filter(!is.na(INC_HOUR)) %>%
 ggplot() +
 aes(x = OFFICER_YEARS_ON_FORCE, fill = OFFICER_INJURY, weight = OFFICER_YEARS_ON_FORCE) +
 geom_density(adjust = 1L) +
 scale_fill_hue(direction = 1) +
 ggthemes::theme_base() +
 theme(legend.position = "bottom")

ggplotly(p)
```

The below graph show that there is no clear corelation between officer race and subject race during incidents. It can be equal for all cases although ratio of males is much higher to be involved in incidents.

```
p <- df %>%
 filter(!(SUBJECT_RACE %in% "NULL")) %>%
 filter(SUBJECT_GENDER %in% c("Female", "Male")) %>%
 filter(!(INCIDENT_REASON %in%
 "NULL")) %>%
 filter(!(REASON_FOR_FORCE %in% "NULL")) %>%
 filter(!is.na(INCIDENT_HOUR)) %>%
 filter(!is.na(INC_HOUR)) %>%
 ggplot() +
 aes(x = SUBJECT_RACE, y = OFFICER_RACE, fill = OFFICER_GENDER) +
 geom_tile(size = 0.5) +
 scale_fill_hue(direction = 1) +
 ggthemes::theme_base() +
 theme(legend.position = "bottom")

ggplotly(p)
```

We can check from the below graph that if the injury caused to the officer is related to the subject arrest. It was considered as a hypothesis that injury caused to the officer by subject in incident may lead to arrest and this hypothesis seems clealy rejected.

```
p <- ggplot(df) +
 aes(x = SUBJECT_WAS_ARRESTED) +
 geom_bar(fill = "#112446") +
 labs(title = "Officer injury relation with the Subject arrest") +
 ggthemes::theme_base() +
 facet_wrap(vars(OFFICER_INJURY))

ggplotly(p)
```

Summary

Data Analysis has been conducted for the Dallas, USA Police enquity dataset. Several EDA steps are conducted to clean the data which was verified in the normality and skewness tests. Afterwards the data visualization show that Black males are mostly to be involved in incidents as compared to Hispanics and White males. The number of asians in incidents are on 2nd rank. The Gender of officers will most likely to have no effect on the arrests. Similar trend is found for the race of officer in relation to race of subject.