# MATH2349 Semester 1, 2018

Code

## *Assignment 3*

*Suraj Kannan - s3668855*

RPubs link : https://rpubs.com/surajkannan/Assignment_3 (https://rpubs.com/surajkannan/Assignment_3)

**Groups** 1. Ancy Antappan - s3691722 2. Suraj Kannan - s3668855 3. Hari Hara Priya Kannan - s3673037

# Required packages

The packages required for this project are loaded as the first step.

Hide

```
library(readr)
library(tidyr)
library(dplyr)
library(outliers)
library(Hmisc)
library(readxl)
```

# Executive Summary

Objective of the study was to create a dataset which could be used to predict the relationship between Number of accidents and the total vehicle ownership in Australian states for the year 2016.

Initially the data is loaded and explored by inspecting the summary. In Tidy tasks, the column names are changed and the values are factorized for merging the datasets. Then, the dataset is checked for missing values and the missing values are handled by replacing the mean values. Then, the dataset is checked for any outliers and visualised using a boxplot. Later, the data is merged and mutated and transformed to obtain the desired output.

# Data

For the analysis two datasets were considered

1.Data describing the number of crashes and variables which explains the scenarion in which crashes occured - https://bitre.gov.au/statistics/safety/fatal_road_crash_database.aspx (https://bitre.gov.au/statistics/safety/fatal_road_crash_database.aspx)

2.Number of vehicle ownership in Australian States http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/9208.012%20months%20ended%2030%20June%202016?OpenDocument (http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/9208.012%20months%20ended%2030%20June%202016?OpenDocument)

Hide

```
Road_fatalities <- read.csv("C:\\Sem2\\Data Preprocessing\\Assignments\\Assignment3\\Road_Fatalities.csv")
head(Road_fatalities)
```

| | Crash.ID <int> | State <fctr> | Mo… <fctr> | Year <int> | Dayweek <fctr> | Time <fctr> | Crash.Type <fctr> | Bus..Involvement <fctr> |
|---|---|---|---|---|---|---|---|---|
| 1 | 20181062 | NSW | April | 2018 | Sunday | 19:00 | Multiple vehicle | No |
| 2 | 20181062 | NSW | April | 2018 | Sunday | 19:00 | Multiple vehicle | No |
| 3 | 20181070 | NSW | April | 2018 | Saturday | 9:05 | Multiple vehicle | No |
| 4 | 20181003 | NSW | April | 2018 | Friday | 16:45 | Multiple vehicle | No |
| 5 | 20181045 | NSW | April | 2018 | Tuesday | 1:30 | Single vehicle | No |
| 6 | 20181107 | NSW | April | 2018 | Sunday | 18:50 | Single vehicle | No |

6 rows | 1-9 of 14 columns

Hide

```
Total_Vehicles_and_Km <- read_excel("Total Vehicles and Km.xlsx")
head(Total_Vehicles_and_Km)
```

| Country <chr> | Total Km travelled in million <dbl> | Number of Vehicles <dbl> |
|---|---|---|
| New South Wales | 70696.00 | 5333001 |
| Victoria | 66850.00 | 4611155 |
| Queensland | 54437.00 | 3791028 |
| South Australia | 16915.00 | 1343791 |
| Western Australia | 29434.00 | 2212489 |
| Tasmania | 4.92 | 457922 |

6 rows

# Understand

Let us first check the dimensions and class of objects in both the datasets

Hide

```
str(Road_fatalities)
```

```
'data.frame':   48982 obs. of  14 variables:
 $ Crash.ID                      : int  20181062 20181062 20181070 20181003 20181045 20181107 20
181107 20181014 20181033 20181101 ...
 $ State                         : Factor w/ 8 levels "ACT","NSW","NT",..: 2 2 2 2 2 2 2 2 2 2
...
 $ Month                         : Factor w/ 12 levels "April","August",..: 1 1 1 1 1 1 1 1 1 1
...
 $ Year                          : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
 $ Dayweek                       : Factor w/ 7 levels "Friday","Monday",..: 4 4 3 1 6 4 4 4 3 1
...
 $ Time                          : Factor w/ 1394 levels "-9","0:00","0:01",..: 654 654 1340 519
89 644 644 445 1012 798 ...
 $ Crash.Type                    : Factor w/ 3 levels "Multiple vehicle",..: 1 1 1 1 3 3 3 3 3 3
...
 $ Bus..Involvement              : Factor w/ 3 levels "-9","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Rigid.Truck..Involvement      : Factor w/ 3 levels "-9","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Articulated.Truck..Involvement.: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Speed.Limit                   : int  100 100 60 80 110 50 50 100 50 100 ...
 $ Road.User                     : Factor w/ 9 levels "-9","9","Bicyclist (includes pillion pass
engers)",..: 4 8 8 4 6 4 8 6 4 4 ...
 $ Gender                        : Factor w/ 4 levels "-9","Female",..: 3 2 3 2 3 3 3 3 3 3 ...
 $ Age                           : int  70 68 94 19 38 24 25 62 24 43 ...
```

Hide

```
str(Total_Vehicles_and_Km)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   8 obs. of  3 variables:
 $ Country                     : chr  "New South Wales" "Victoria" "Queensland" "South Australi
a" ...
 $ Total Km travelled in million: num  70696 66850 54437 16915 29434 ...
 $ Number of Vehicles          : num  5333001 4611155 3791028 1343791 2212489 ...
```

We can observe that the dataset has multiple data types such as factor, int, chr and num.

# Tidy & Manipulate Data I

Our goal is to join the datasets by state. But we can see that in order to do so, the data is not in tidy format. Let us start by inpecting the column we wish to join.

Hide

```
names(Road_fatalities)
```

```
 [1] "Crash.ID"                      "State"
 [3] "Month"                         "Year"
 [5] "Dayweek"                       "Time"
 [7] "Crash.Type"                    "Bus..Involvement"
 [9] "Rigid.Truck..Involvement"      "Articulated.Truck..Involvement."
[11] "Speed.Limit"                   "Road.User"
[13] "Gender"                        "Age"
```

Hide

```
names(Total_Vehicles_and_Km)
```

```
[1] "Country"                       "Total Km travelled in million"
[3] "Number of Vehicles"
```

The states are mentioned under the Country column in the total_vechicles_and_km dataset while it is named as state in the road_fatalirties dataset.

As a first step, we change the column name.

Hide

```
names(Total_Vehicles_and_Km)[1]<-"State"
names(Total_Vehicles_and_Km)
```

```
[1] "State"                         "Total Km travelled in million"
[3] "Number of Vehicles"
```

The “State” variable is character type in the total_vehicles_and_km dataframe by default and this to be converted to factor variable.

Hide

```
Total_Vehicles_and_Km$State<-factor(Total_Vehicles_and_Km$State,levels=c("New South Wales", "Vict
oria","Queensland","South Australia","Western Australia","Tasmania","Northern Territory","Austral
ian Capital Territory"),
                                    labels = c("NSW", "VIC","QLD","SA","WA","TAS","NT","ACT"))
class(Total_Vehicles_and_Km$State)
```

```
[1] "factor"
```

Hide

```
levels(Total_Vehicles_and_Km$State)
```

```
[1] "NSW" "VIC" "QLD" "SA"  "WA"  "TAS" "NT"  "ACT"
```

# Tidy & Manipulate Data II

The “State” variable is character type in the total_vehicles_and_km dataframe by default and this to be converted to factor variable.

Hide

```
Total_Vehicles_and_Km$Country<-factor(Total_Vehicles_and_Km$State,levels=c("New South Wales", "Victoria","Queensland","South Australia","Western Australia","Tasmania","Northern Territory","Australian Capital Territory"),
                                      labels = c("NSW", "VIC","QLD","SA","WA","TAS","NT","ACT"))
class(Total_Vehicles_and_Km$State)
```

```
[1] "factor"
```

Hide

```
levels(Total_Vehicles_and_Km$State)
```

```
[1] "NSW" "VIC" "QLD" "SA"  "WA"  "TAS" "NT"  "ACT"
```

# Scan I

subsetting function was used to eliminate some variables which are outside the scope of the study

Hide

```
Road_Fatalities1<-select(Road_fatalities,-(Dayweek:Articulated.Truck..Involvement.))
```

The study focuses only on the year 2016, hence filter function usesd to filter the data for year 2016

Hide

```
Road_Fatalities1<-filter(Road_Fatalities1,Year==2016)
```

# Checking for missing variables/inconsistencies

Hide

```
colSums(is.na(Road_Fatalities1))
```

```
   Crash.ID       State       Month        Year Speed.Limit   Road.User
          0           0           0           0           0           0
     Gender         Age
          0           0
```

no “NA”’s but all the unavailable information is coded as “-9”, hence changing -9 to NA, then checking for missing variables.

Hide

```
Road_Fatalities1$Speed.Limit[Road_Fatalities1$Speed.Limit == -9] <- NA
Road_Fatalities1$Gender[Road_Fatalities1$Gender == -9] <- NA
```

Hide

```
colSums(is.na(Road_Fatalities1))
```

```
   Crash.ID       State       Month        Year Speed.Limit   Road.User
          0           0           0           0           9           0
     Gender         Age
          1           0
```

We can observe there is 1 NA in Gender and the 9 “NA”’s in Speed.Limit.

For eliminating NA in the Speed.Limit column, we imputed average NA’s with mean value

Hide

```
Road_Fatalities1$Speed.Limit[is.na(Road_Fatalities1$Speed.Limit)]<-mean(Road_Fatalities1$Speed.Li
mit,na.rm=TRUE)
colSums(is.na(Road_Fatalities1))
```

```
   Crash.ID       State       Month        Year Speed.Limit   Road.User
          0           0           0           0           0           0
     Gender         Age
          1           0
```

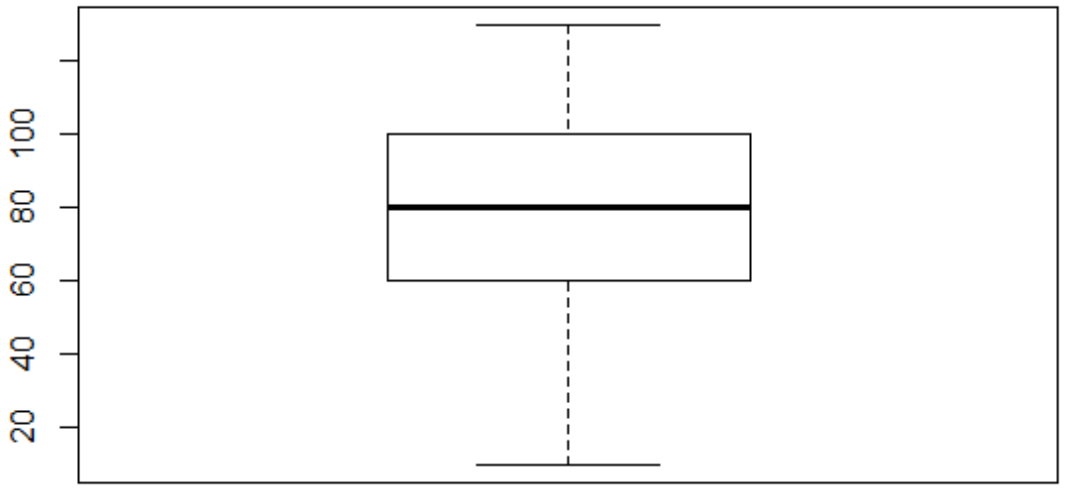
NA succesfully eliminated for speed limit category.

# Scan II

Outliers for Speed.Limit $ Age checked using boxplot function

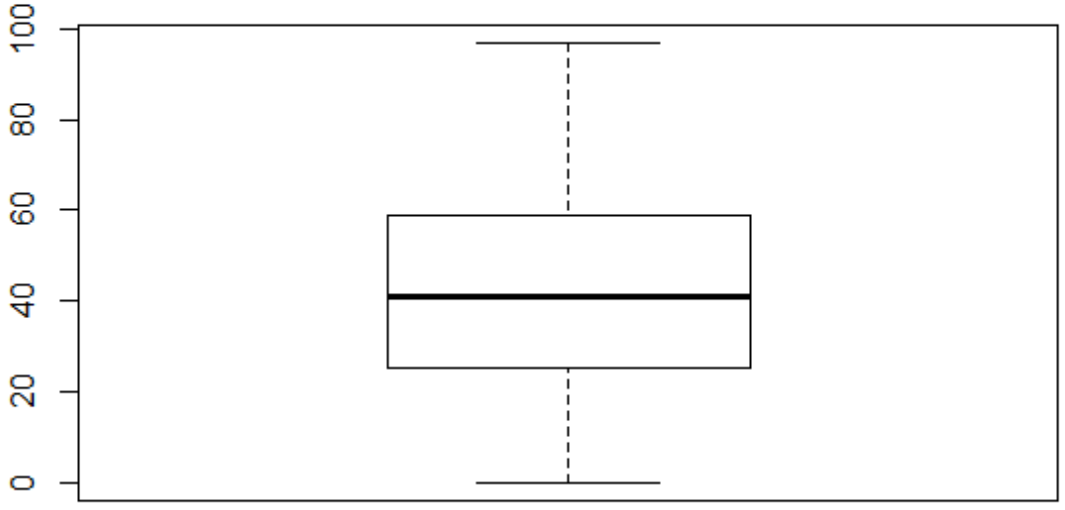
Hide

```
boxplot((Road_Fatalities1$Speed.Limit))
```




Hide

```
boxplot((Road_Fatalities1$Age))
```




From boxplot, it can be predicted that there is no outliers

# Merging

Hide

```
new = summarise(group_by(Road_fatalities,State),count =n())
new
```

| State <fctr> | count <int> |
|---|---|
| ACT | 452 |
| NSW | 15234 |
| NT | 1517 |
| QLD | 9626 |
| SA | 4216 |
| TAS | 1440 |
| VIC | 10759 |
| WA | 5738 |

8 rows

Hide

```
S1<-left_join(new,Total_Vehicles_and_Km,by="State")
S1
```

| State <chr> | count <int> | Total Km travelled in million <dbl> | Number of Vehicles <dbl> | Country <fctr> |
|---|---|---|---|---|
| ACT | 452 | 3746.00 | 284345 | *NA* |
| NSW | 15234 | 70696.00 | 5333001 | *NA* |
| NT | 1517 | 2053.00 | 157944 | *NA* |
| QLD | 9626 | 54437.00 | 3791028 | *NA* |
| SA | 4216 | 16915.00 | 1343791 | *NA* |
| TAS | 1440 | 4.92 | 457922 | *NA* |
| VIC | 10759 | 66850.00 | 4611155 | *NA* |
| WA | 5738 | 29434.00 | 2212489 | *NA* |

8 rows

# Mutate

Percentage of accidents in each state

Hide

```
mutate(S1, percentageofaccidents=round((S1$count/sum(S1$count))*100,2))
```

| State <chr> | co… <int> | Total Km travelled in million <dbl> | Number of Vehicles <dbl> | Coun… <fctr> | percentageofac |
|---|---|---|---|---|---|
| ACT | 452 | 3746.00 | 284345 | *NA* | |
| NSW | 15234 | 70696.00 | 5333001 | *NA* | |
| NT | 1517 | 2053.00 | 157944 | *NA* | |
| QLD | 9626 | 54437.00 | 3791028 | *NA* | |
| SA | 4216 | 16915.00 | 1343791 | *NA* | |
| TAS | 1440 | 4.92 | 457922 | *NA* | |
| VIC | 10759 | 66850.00 | 4611155 | *NA* | |
| WA | 5738 | 29434.00 | 2212489 | *NA* | |

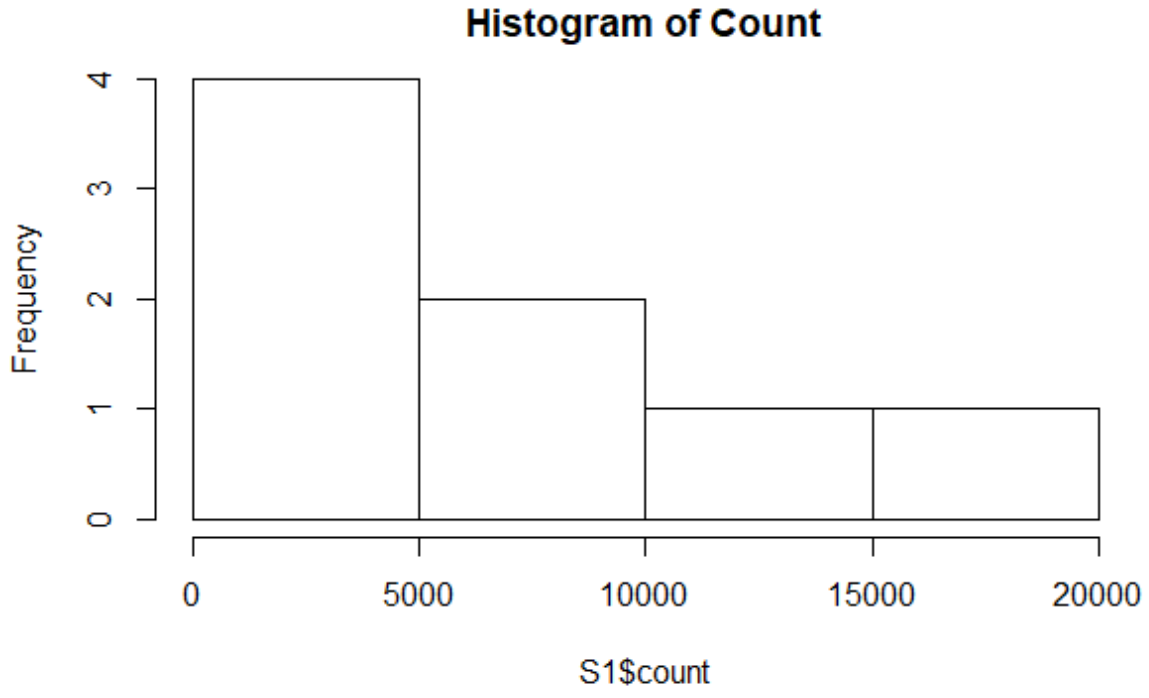
8 rows

# Transform

The histogram is right skewed as the frequency of counts is crowded towards the left. This indicates that the mean counts is greater than the median value. Below is the histogram plotted representing the trsnaformation.

Hide

```
hist(S1$count, main = "Histogram of Count")
```



To normalize the distribution ,log transformation is applied to reduce the skewness and make it normally distributed.

Hide

```
hist(log10(S1$count))
```