DMPM Assignment 3

Name: Rushikesh Jyoti

Division: A

Roll no: 27

SRN: 201901139

Question: Preprocess and clean the given dataset

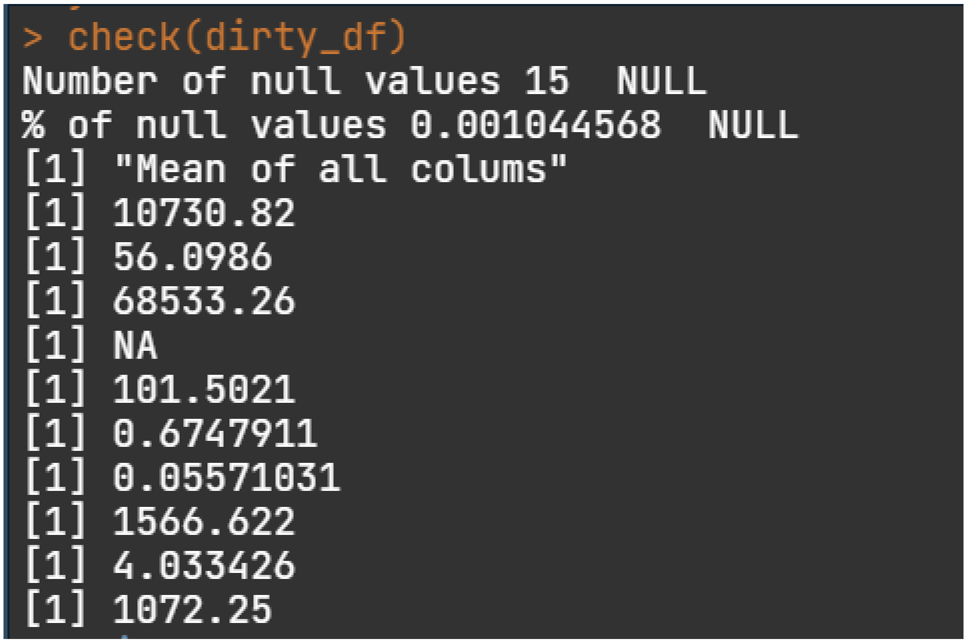
Code

```
# install.packages("tidyverse")

# install.packages("Hmisc")


library(tidyverse)

library(dplyr)

library(Hmisc)


df = read.csv('ToyotaCorolla.csv')

dirty_df = read.csv('ToyotaCorolla - Dirty.csv')




check = function(dataset) {

  print(cat("Number of null values", sum(is.na(dataset)), " "))

  print(cat("% of null values", mean(is.na(dataset)), " "))


  print("Mean of all colums")

  for (i in 1:ncol(dataset)) {

    print(mean(dataset[,i], na.rm = TRUE))
    }
}

check(dirty_df)

head(rename(dirty_df, Kilometers = KM))

clean_df = na.omit(dirty_df)

head(select(clean_df, -MetColor))

head(arrange(clean_df, Age))

slice(clean_df, 4:17)

head(filter(clean_df, FuelType == 'Petrol'))

glimpse(clean_df)

boxplot(clean_df$Price)

boxplot(clean_df$Age)

boxplot(clean_df$Weight)

print("Outliers of Weight are ")
boxplot.stats(clean_df$Weight)$out
```

```
# Numerical Imputation
dirty_df$Age = impute(dirty_df$Age, fun=mean)


dirty_df$CC = impute(dirty_df$CC, fun=mean)


dirty_df$Weight = impute(dirty_df$Weight, fun=mean)


for (i in 1:ncol(dirty_df)) {
  print(sum(is.na(dirty_df[,i])))
}


print("Phew! No null values anymore!")
```

Output

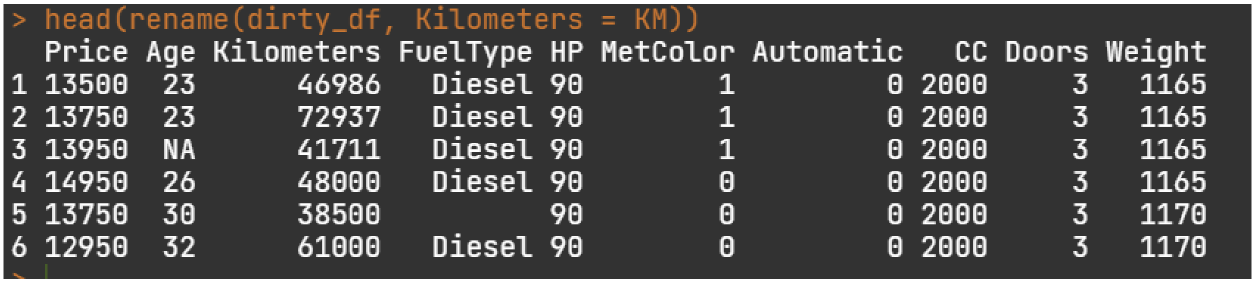
Null Values of dataset and mean of every column



```
> check(dirty_df)
Number of null values 15  NULL
% of null values 0.001044568  NULL
[1] "Mean of all colums"
[1] 10730.82
[1] 56.0986
[1] 68533.26
[1] NA
[1] 101.5021
[1] 0.6747911
[1] 0.05571031
[1] 1566.622
[1] 4.033426
[1] 1072.25
```

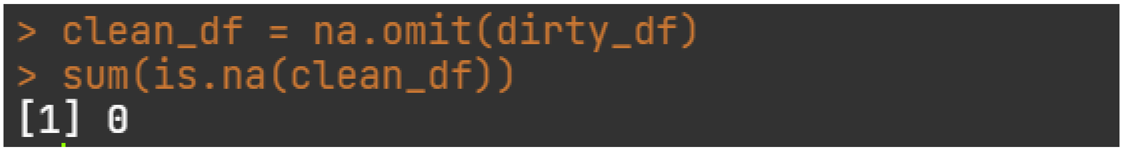
Fourth column is categorical data so it can’t be `meaned`

Renaming a column



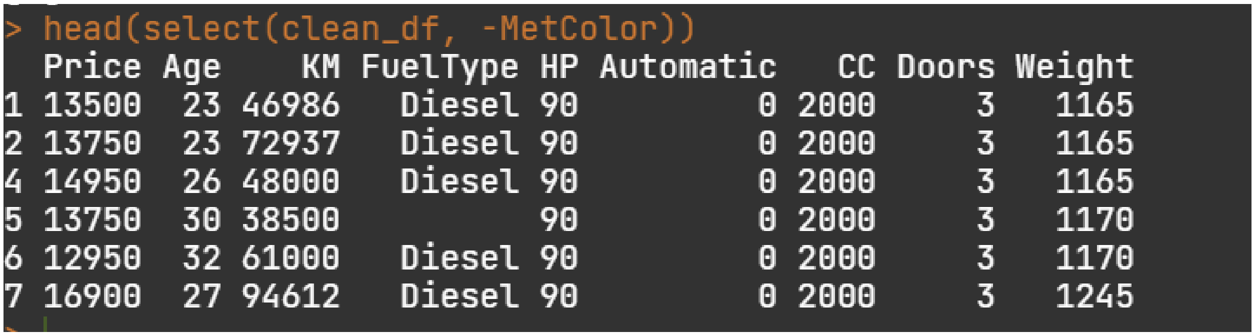
```
> head(rename(dirty_df, Kilometers = KM))
  Price Age Kilometers FuelType HP MetColor Automatic   CC Doors Weight
1 13500  23      46986   Diesel 90        1         0 2000     3   1165
2 13750  23      72937   Diesel 90        1         0 2000     3   1165
3 13950  NA      41711   Diesel 90        1         0 2000     3   1165
4 14950  26      48000   Diesel 90        0         0 2000     3   1165
5 13750  30      38500          90        0         0 2000     3   1170
6 12950  32      61000   Diesel 90        0         0 2000     3   1170
```

Omitting the NA values



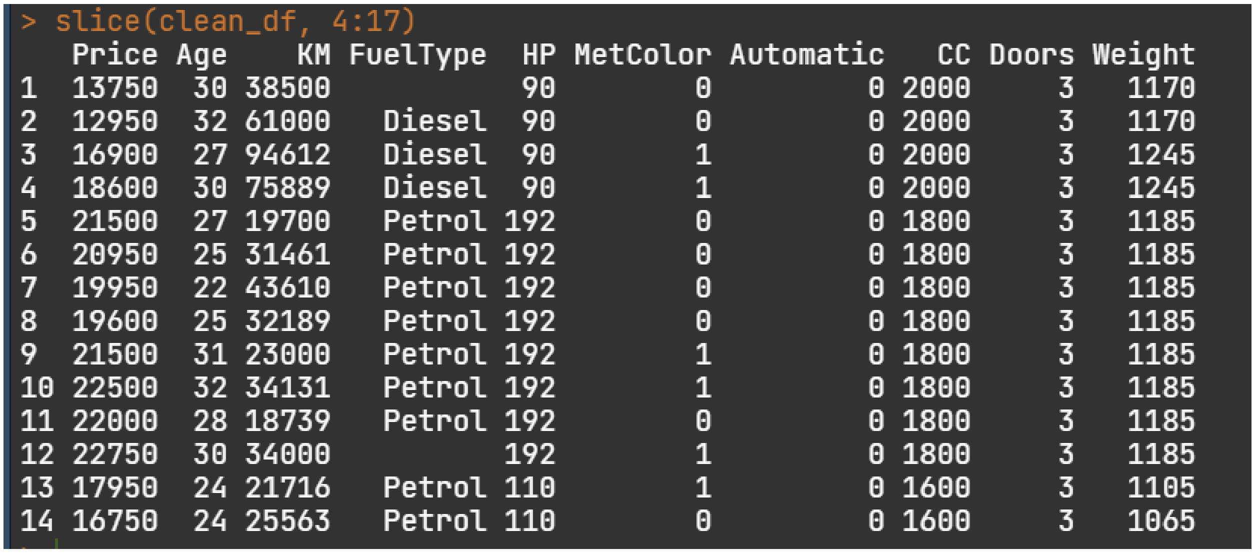
```
> clean_df = na.omit(dirty_df)
> sum(is.na(clean_df))
[1] 0
```

Removing a column (MetColor) from dataset



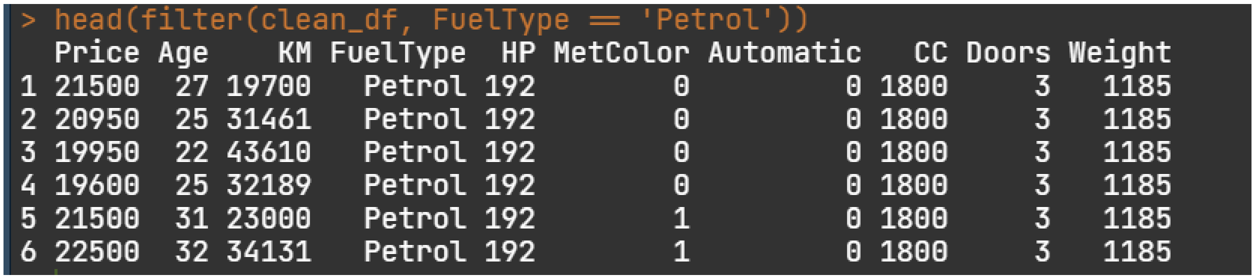
```
> head(select(clean_df, -MetColor))
  Price Age    KM FuelType HP Automatic   CC Doors Weight
1 13500  23 46986   Diesel 90         0 2000     3   1165
2 13750  23 72937   Diesel 90         0 2000     3   1165
4 14950  26 48000   Diesel 90         0 2000     3   1165
5 13750  30 38500          90         0 2000     3   1170
6 12950  32 61000   Diesel 90         0 2000     3   1170
7 16900  27 94612   Diesel 90         0 2000     3   1245
```

Taking a slice of dataset



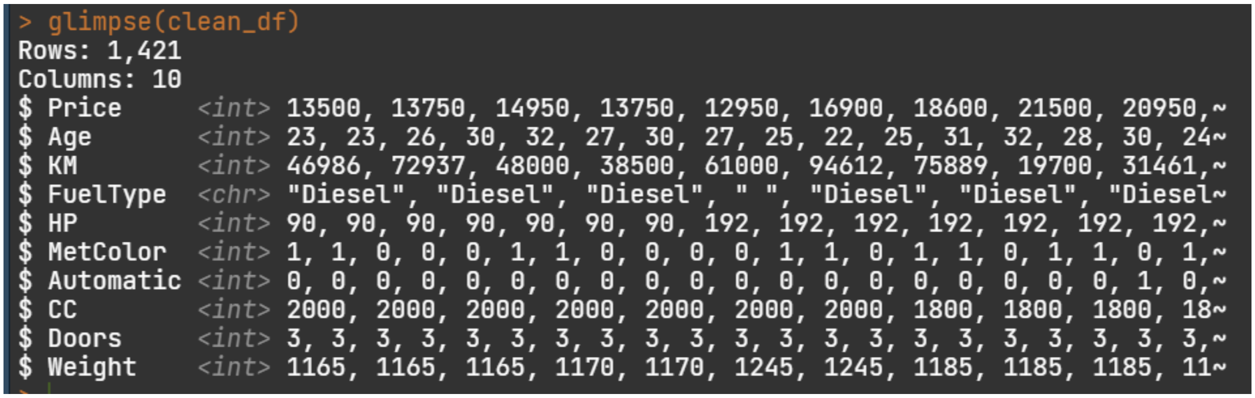
```
> slice(clean_df, 4:17)
   Price Age    KM FuelType  HP MetColor Automatic   CC Doors Weight
1  13750  30 38500           90        0         0 2000     3   1170
2  12950  32 61000   Diesel  90        0         0 2000     3   1170
3  16900  27 94612   Diesel  90        1         0 2000     3   1245
4  18600  30 75889   Diesel  90        1         0 2000     3   1245
5  21500  27 19700   Petrol 192        0         0 1800     3   1185
6  20950  25 31461   Petrol 192        0         0 1800     3   1185
7  19950  22 43610   Petrol 192        0         0 1800     3   1185
8  19600  25 32189   Petrol 192        0         0 1800     3   1185
9  21500  31 23000   Petrol 192        1         0 1800     3   1185
10 22500  32 34131   Petrol 192        1         0 1800     3   1185
11 22000  28 18739   Petrol 192        0         0 1800     3   1185
12 22750  30 34000          192        1         0 1800     3   1185
13 17950  24 21716   Petrol 110        1         0 1600     3   1105
14 16750  24 25563   Petrol 110        0         0 1600     3   1065
```

Filtering the dataset to get all petrol vehicles



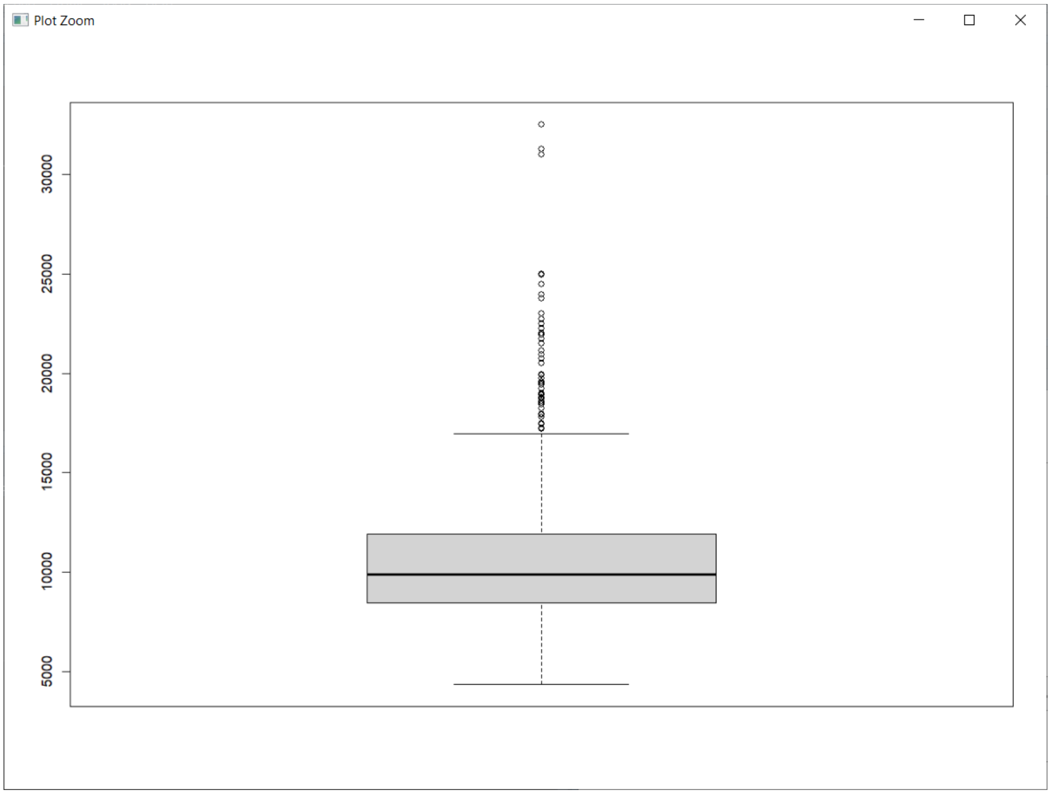
```
> head(filter(clean_df, FuelType == 'Petrol'))
  Price Age    KM FuelType  HP MetColor Automatic   CC Doors Weight
1 21500  27 19700   Petrol 192        0         0 1800     3   1185
2 20950  25 31461   Petrol 192        0         0 1800     3   1185
3 19950  22 43610   Petrol 192        0         0 1800     3   1185
4 19600  25 32189   Petrol 192        0         0 1800     3   1185
5 21500  31 23000   Petrol 192        1         0 1800     3   1185
6 22500  32 34131   Petrol 192        1         0 1800     3   1185
```

Taking a glimpse of our dataset

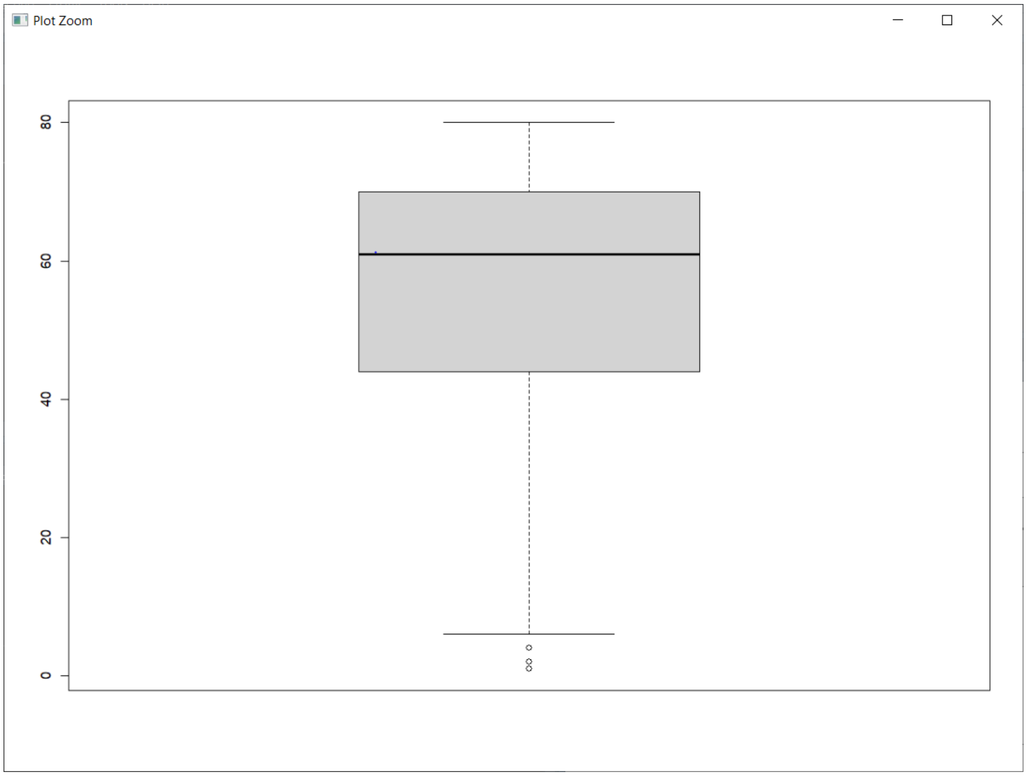


```
> glimpse(clean_df)
Rows: 1,421
Columns: 10
$ Price     <int> 13500, 13750, 14950, 13750, 12950, 16900, 18600, 21500, 20950,~
$ Age       <int> 23, 23, 26, 30, 32, 27, 30, 27, 25, 22, 25, 31, 32, 28, 30, 24~
$ KM        <int> 46986, 72937, 48000, 38500, 61000, 94612, 75889, 19700, 31461,~
$ FuelType  <chr> "Diesel", "Diesel", "Diesel", " ", "Diesel", "Diesel", "Diesel~
$ HP        <int> 90, 90, 90, 90, 90, 90, 90, 192, 192, 192, 192, 192, 192, 192,~
$ MetColor  <int> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1,~
$ Automatic <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
$ CC        <int> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 1800, 1800, 1800, 18~
$ Doors     <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
$ Weight    <int> 1165, 1165, 1165, 1170, 1170, 1245, 1245, 1185, 1185, 1185, 11~
```

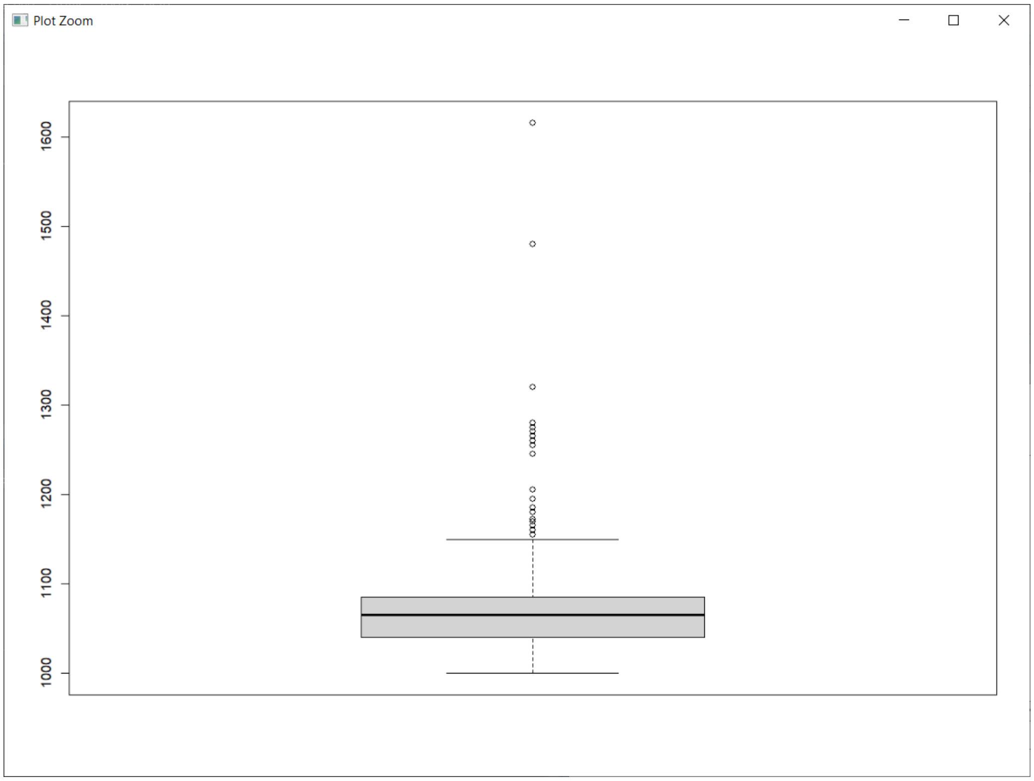
Boxplot of Price of vehicles



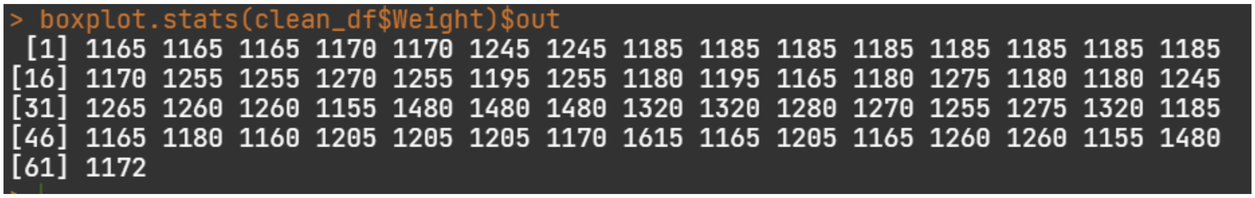
Boxplot of Age of Vehicles



Boxplot of Weight of vehicles

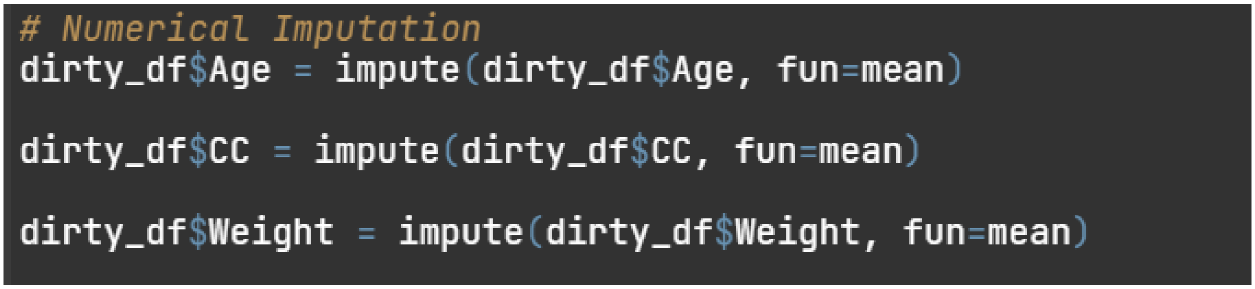


Getting all outliers of weight

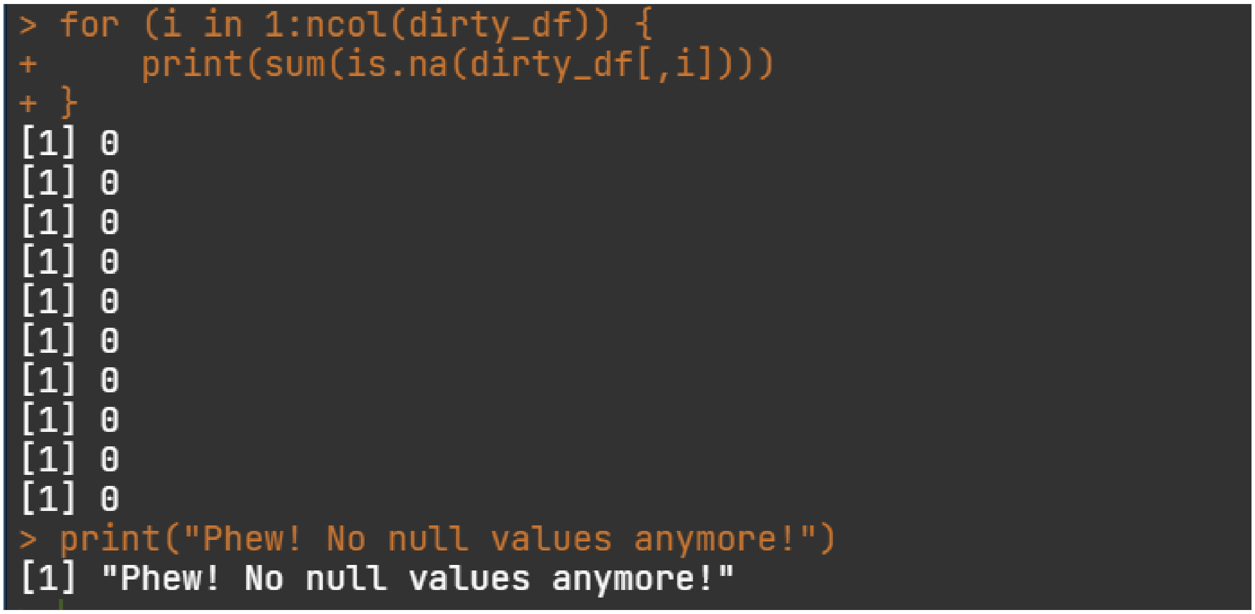


```
> boxplot.stats(clean_df$Weight)$out
 [1] 1165 1165 1165 1170 1170 1245 1245 1185 1185 1185 1185 1185 1185 1185 1185
[16] 1170 1255 1255 1270 1255 1195 1255 1180 1195 1165 1180 1275 1180 1180 1245
[31] 1265 1260 1260 1155 1480 1480 1480 1320 1320 1280 1270 1255 1275 1320 1185
[46] 1165 1180 1160 1205 1205 1205 1170 1615 1165 1205 1165 1260 1260 1155 1480
[61] 1172
```

Impute the columns with NA values



```
# Numerical Imputation
dirty_df$Age = impute(dirty_df$Age, fun=mean)

dirty_df$CC = impute(dirty_df$CC, fun=mean)

dirty_df$Weight = impute(dirty_df$Weight, fun=mean)
```



```
> for (i in 1:ncol(dirty_df)) {
+     print(sum(is.na(dirty_df[,i])))
+ }
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
> print("Phew! No null values anymore!")
[1] "Phew! No null values anymore!"
```

The dataset is now clean