***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)) {

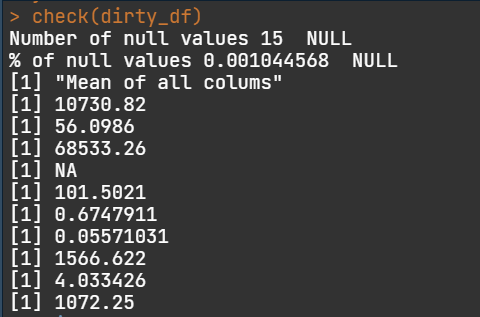
print(sum(is.na(dirty\_df[,i])))

}

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

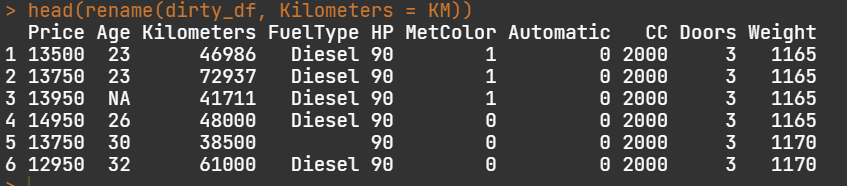
Output

Null Values of dataset and mean of every column

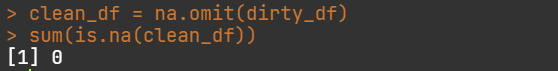


Fourth column is categorical data so it can’t be `meaned`

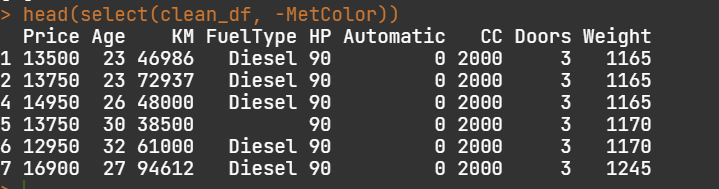
Renaming a column



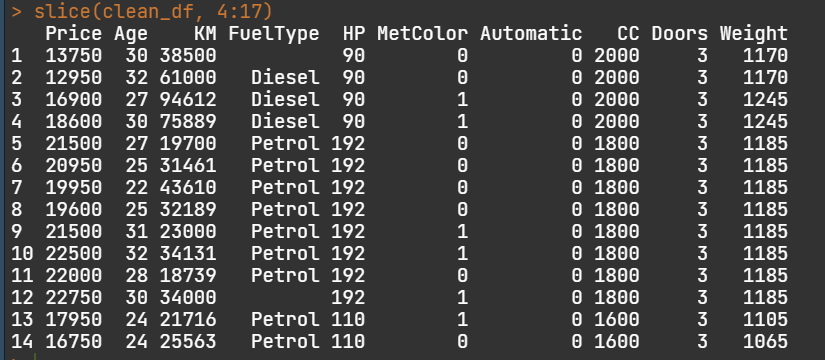
Omitting the NA values



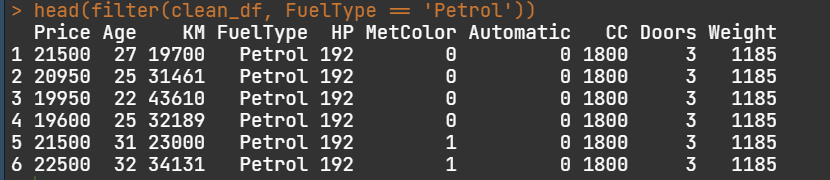
Removing a column (MetColor) from dataset



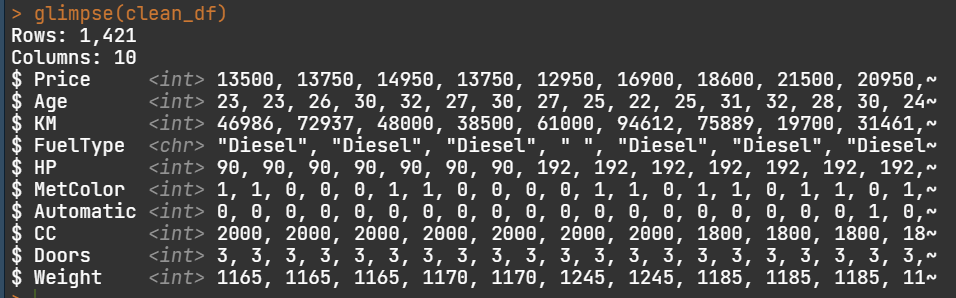
Taking a slice of dataset



Filtering the dataset to get all petrol vehicles



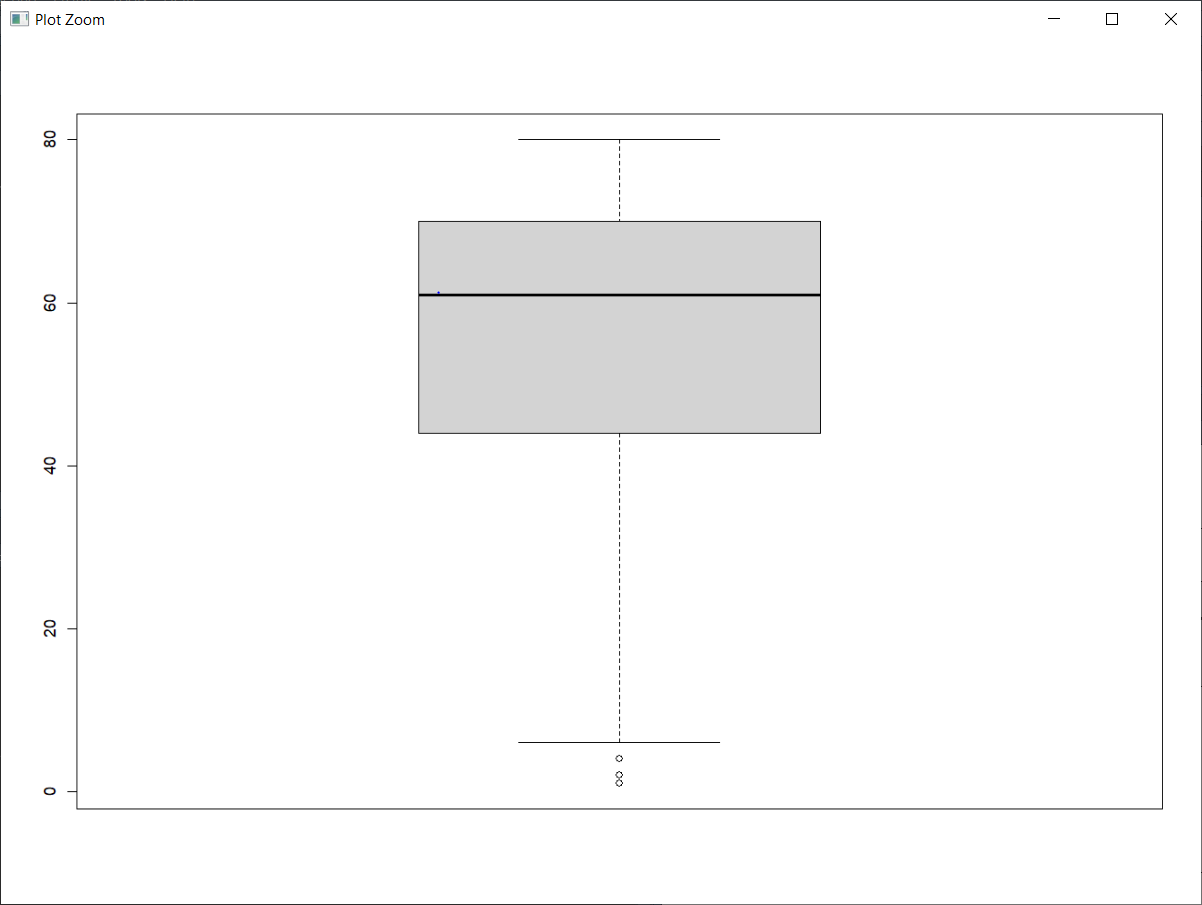
Taking a glimpse of our dataset



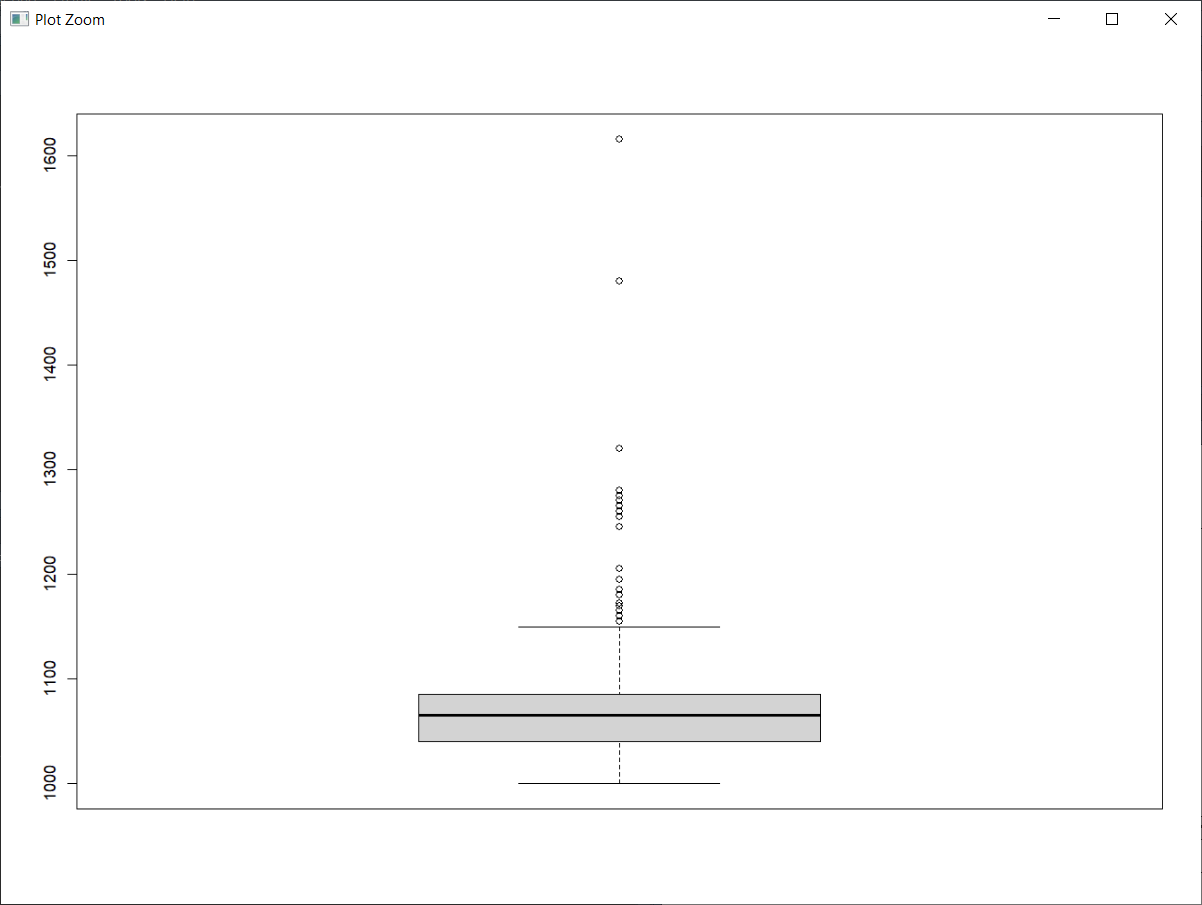
Boxplot of Price of vehicles



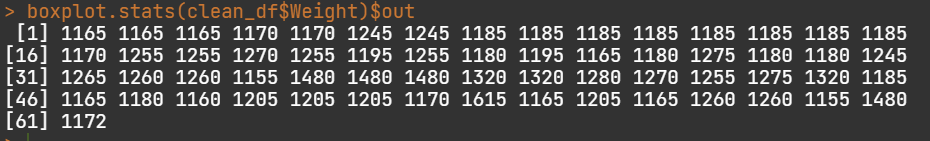
Boxplot of Age of Vehicles



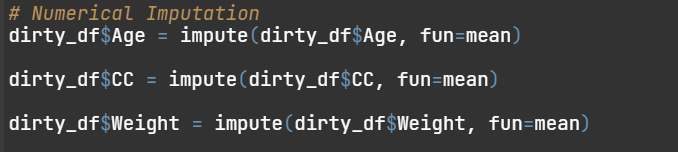
Boxplot of Weight of vehicles

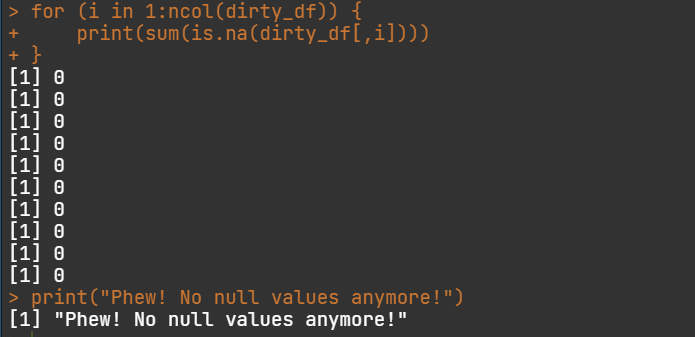


Getting all outliers of weight



Impute the columns with NA values





The dataset is now clean