IST 772 Project

Warren Fernandes

12/10/2021

Diagnostics

```
# Loading required libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(imputeTS)
```

```
## Warning: package 'imputeTS' was built under R version 4.0.4
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(dplyr)
library(BayesFactor)
```

```
## Warning: package 'BayesFactor' was built under R version 4.0.5
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 4.0.5
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richa
rddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

```
library(RColorBrewer)
# Change the setwd() to where you have stored the data set
#setwd("~/Desktop/Week11772")

autoData <- read_csv("Automobile_data.csv")
```

```
## Rows: 205 Columns: 26
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (16): normalized-losses, make, fuel-type, aspiration, num-of-doors, body...
## dbl (10): symboling, wheel-base, length, width, height, curb-weight, engine-...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(autoData)
```

```
##    symboling       normalized-losses      make            fuel-type
##  Min.   :-2.0000   Length:205         Length:205         Length:205
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.8341
##  3rd Qu.: 2.0000
##  Max.   : 3.0000
##   aspiration        num-of-doors        body-style        drive-wheels
##  Length:205         Length:205         Length:205         Length:205
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  engine-location      wheel-base         length          width
##  Length:205         Min.   : 86.60   Min.   :141.1   Min.   :60.30
##  Class :character   1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10
##  Mode  :character   Median : 97.00   Median :173.2   Median :65.50
##                     Mean   : 98.76   Mean   :174.0   Mean   :65.91
##                     3rd Qu.:102.40   3rd Qu.:183.1   3rd Qu.:66.90
##                     Max.   :120.90   Max.   :208.1   Max.   :72.30
##      height       curb-weight   engine-type        num-of-cylinders
##  Min.   :47.80   Min.   :1488   Length:205         Length:205
##  1st Qu.:52.00   1st Qu.:2145   Class :character   Class :character
##  Median :54.10   Median :2414   Mode  :character   Mode  :character
##  Mean   :53.72   Mean   :2556
##  3rd Qu.:55.50   3rd Qu.:2935
##  Max.   :59.80   Max.   :4066
##   engine-size    fuel-system           bore              stroke
##  Min.   : 61.0   Length:205         Length:205         Length:205
##  1st Qu.: 97.0   Class :character   Class :character   Class :character
##  Median :120.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :126.9
##  3rd Qu.:141.0
##  Max.   :326.0
##  compression-ratio  horsepower          peak-rpm           city-mpg
##  Min.   : 7.00     Length:205         Length:205         Min.   :13.00
##  1st Qu.: 8.60     Class :character   Class :character   1st Qu.:19.00
##  Median : 9.00     Mode  :character   Mode  :character   Median :24.00
##  Mean   :10.14                                           Mean   :25.22
##  3rd Qu.: 9.40                                           3rd Qu.:30.00
##  Max.   :23.00                                           Max.   :49.00
##   highway-mpg       price
##  Min.   :16.00   Length:205
##  1st Qu.:25.00   Class :character
##  Median :30.00   Mode  :character
##  Mean   :30.75
##  3rd Qu.:34.00
##  Max.   :54.00
```

```
str(autoData)
```

```
## spec_tbl_df [205 x 26] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ symboling        : num [1:205] 3 3 1 2 2 2 1 1 1 0 ...
##  $ normalized-losses: chr [1:205] "?" "?" "?" "164" ...
##  $ make             : chr [1:205] "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
##  $ fuel-type        : chr [1:205] "gas" "gas" "gas" "gas" ...
##  $ aspiration       : chr [1:205] "std" "std" "std" "std" ...
##  $ num-of-doors     : chr [1:205] "two" "two" "two" "four" ...
##  $ body-style       : chr [1:205] "convertible" "convertible" "hatchback" "sedan" ...
##  $ drive-wheels     : chr [1:205] "rwd" "rwd" "rwd" "fwd" ...
##  $ engine-location  : chr [1:205] "front" "front" "front" "front" ...
##  $ wheel-base       : num [1:205] 88.6 88.6 94.5 99.8 99.4 ...
##  $ length           : num [1:205] 169 169 171 177 177 ...
##  $ width            : num [1:205] 64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ height           : num [1:205] 48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curb-weight      : num [1:205] 2548 2548 2823 2337 2824 ...
##  $ engine-type      : chr [1:205] "dohc" "dohc" "ohcv" "ohc" ...
##  $ num-of-cylinders : chr [1:205] "four" "four" "six" "four" ...
##  $ engine-size      : num [1:205] 130 130 152 109 136 136 136 136 131 131 ...
##  $ fuel-system      : chr [1:205] "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ bore             : chr [1:205] "3.47" "3.47" "2.68" "3.19" ...
##  $ stroke           : chr [1:205] "2.68" "2.68" "3.47" "3.4" ...
##  $ compression-ratio: num [1:205] 9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower       : chr [1:205] "111" "111" "154" "102" ...
##  $ peak-rpm         : chr [1:205] "5000" "5000" "5000" "5500" ...
##  $ city-mpg         : num [1:205] 21 21 19 24 18 19 19 19 17 16 ...
##  $ highway-mpg      : num [1:205] 27 27 26 30 22 25 25 25 20 22 ...
##  $ price            : chr [1:205] "13495" "16500" "16500" "13950" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   symboling = col_double(),
##   ..   `normalized-losses` = col_character(),
##   ..   make = col_character(),
##   ..   `fuel-type` = col_character(),
##   ..   aspiration = col_character(),
##   ..   `num-of-doors` = col_character(),
##   ..   `body-style` = col_character(),
##   ..   `drive-wheels` = col_character(),
##   ..   `engine-location` = col_character(),
##   ..   `wheel-base` = col_double(),
##   ..   length = col_double(),
##   ..   width = col_double(),
##   ..   height = col_double(),
##   ..   `curb-weight` = col_double(),
##   ..   `engine-type` = col_character(),
##   ..   `num-of-cylinders` = col_character(),
##   ..   `engine-size` = col_double(),
##   ..   `fuel-system` = col_character(),
##   ..   bore = col_character(),
##   ..   stroke = col_character(),
##   ..   `compression-ratio` = col_double(),
##   ..   horsepower = col_character(),
##   ..   `peak-rpm` = col_character(),
```

```
##   ..   `city-mpg` = col_double(),
##   ..   `highway-mpg` = col_double(),
##   ..   price = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Set Information:

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process “symboling”. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

Attributes -

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

```
# Checking for NA values
sapply(autoData, function(x) sum(is.na(x)))
```

```
##         symboling normalized-losses              make         fuel-type
##                 0                 0                 0                 0
##        aspiration      num-of-doors        body-style      drive-wheels
##                 0                 0                 0                 0
##   engine-location        wheel-base            length             width
##                 0                 0                 0                 0
##            height       curb-weight       engine-type  num-of-cylinders
##                 0                 0                 0                 0
##       engine-size       fuel-system              bore            stroke
##                 0                 0                 0                 0
## compression-ratio        horsepower          peak-rpm          city-mpg
##                 0                 0                 0                 0
##       highway-mpg             price
##                 0                 0
```

```
# There seem to be no NA values so let's look at the data to find missing values.
#View(autoData)
autoData <- autoData %>% mutate_all(~replace(., . == '?', NA))
head(autoData)
```

```
## # A tibble: 6 x 26
##   symboling `normalized-losses` make        `fuel-type` aspiration `num-of-doors`
##       <dbl> <chr>               <chr>       <chr>       <chr>      <chr>
## 1         3 <NA>                alfa-romero gas         std        two
## 2         3 <NA>                alfa-romero gas         std        two
## 3         1 <NA>                alfa-romero gas         std        two
## 4         2 164                 audi        gas         std        four
## 5         2 164                 audi        gas         std        four
## 6         2 <NA>                audi        gas         std        two
## # ... with 20 more variables: body-style <chr>, drive-wheels <chr>,
## #   engine-location <chr>, wheel-base <dbl>, length <dbl>, width <dbl>,
## #   height <dbl>, curb-weight <dbl>, engine-type <chr>, num-of-cylinders <chr>,
## #   engine-size <dbl>, fuel-system <chr>, bore <chr>, stroke <chr>,
## #   compression-ratio <dbl>, horsepower <chr>, peak-rpm <chr>, city-mpg <dbl>,
## #   highway-mpg <dbl>, price <chr>
```

```
# Missing values with '?' value have been replaced by NA.

sapply(autoData, function(x) sum(is.na(x)))
```

```
##         symboling normalized-losses              make         fuel-type
##                 0                41                 0                 0
##        aspiration      num-of-doors        body-style      drive-wheels
##                 0                 2                 0                 0
##   engine-location        wheel-base            length             width
##                 0                 0                 0                 0
##            height       curb-weight       engine-type  num-of-cylinders
##                 0                 0                 0                 0
##       engine-size       fuel-system              bore            stroke
##                 0                 0                 4                 4
## compression-ratio        horsepower          peak-rpm          city-mpg
##                 0                 2                 2                 0
##       highway-mpg             price
##                 0                 4
```

```
# We won't be using symboling and mormalized-losses data for our statistical analysis. The summa
ry information on different cars would ideally provide enough attributes for a thorough staistic
al analysis. The rest of the NA values will be filled using interpolation.
```

```
drop_cols <-  c('symboling', 'normalized-losses')
autoData <- autoData[,!(names(autoData) %in% drop_cols)]
head(autoData,10)
```

```
## # A tibble: 10 x 24
##    make        `fuel-type` aspiration `num-of-doors` `body-style` `drive-wheels`
##    <chr>       <chr>       <chr>      <chr>          <chr>        <chr>
##  1 alfa-romero gas         std        two            convertible  rwd
##  2 alfa-romero gas         std        two            convertible  rwd
##  3 alfa-romero gas         std        two            hatchback    rwd
##  4 audi        gas         std        four           sedan        fwd
##  5 audi        gas         std        four           sedan        4wd
##  6 audi        gas         std        two            sedan        fwd
##  7 audi        gas         std        four           sedan        fwd
##  8 audi        gas         std        four           wagon        fwd
##  9 audi        gas         turbo      four           sedan        fwd
## 10 audi        gas         turbo      two            hatchback    4wd
## # ... with 18 more variables: engine-location <chr>, wheel-base <dbl>,
## #   length <dbl>, width <dbl>, height <dbl>, curb-weight <dbl>,
## #   engine-type <chr>, num-of-cylinders <chr>, engine-size <dbl>,
## #   fuel-system <chr>, bore <chr>, stroke <chr>, compression-ratio <dbl>,
## #   horsepower <chr>, peak-rpm <chr>, city-mpg <dbl>, highway-mpg <dbl>,
## #   price <chr>
```

```
# Setting character variables to either factors or numeric variables
factor_cols <- c('fuel-type', 'aspiration', 'num-of-doors',  'body-style', 'drive-wheels', 'engi
ne-location', 'engine-type', 'num-of-cylinders', 'fuel-system')
autoData[factor_cols] <- lapply(autoData[factor_cols], as.factor)

num_cols <- c('bore', 'stroke', 'horsepower', 'peak-rpm', 'price')
autoData[num_cols] <- lapply(autoData[num_cols], as.numeric)

str(autoData)
```

```
## tibble [205 x 24] (S3: tbl_df/tbl/data.frame)
##  $ make             : chr [1:205] "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
##  $ fuel-type        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ aspiration       : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2 ...
##  $ num-of-doors     : Factor w/ 2 levels "four","two": 2 2 2 1 1 2 1 1 1 2 ...
##  $ body-style       : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5 4 3 ...
##  $ drive-wheels     : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
##  $ engine-location  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1 1 ...
##  $ wheel-base       : num [1:205] 88.6 88.6 94.5 99.8 99.4 ...
##  $ length           : num [1:205] 169 169 171 177 177 ...
##  $ width            : num [1:205] 64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ height           : num [1:205] 48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curb-weight      : num [1:205] 2548 2548 2823 2337 2824 ...
##  $ engine-type      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4 4 4 ...
##  $ num-of-cylinders : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2 2 2 ...
##  $ engine-size      : num [1:205] 130 130 152 109 136 136 136 136 131 131 ...
##  $ fuel-system      : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ bore             : num [1:205] 3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
##  $ stroke           : num [1:205] 2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compression-ratio: num [1:205] 9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower       : num [1:205] 111 111 154 102 115 110 110 110 140 160 ...
##  $ peak-rpm         : num [1:205] 5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
##  $ city-mpg         : num [1:205] 21 21 19 24 18 19 19 19 17 16 ...
##  $ highway-mpg      : num [1:205] 27 27 26 30 22 25 25 25 20 22 ...
##  $ price            : num [1:205] 13495 16500 16500 13950 17450 ...
```

After processing the data, we replace the NA values before moving on to the analysis

```
# Dealing with NA values
autoData$bore <- na_interpolation(autoData$bore)
autoData$stroke <- na_interpolation(autoData$stroke)
autoData$horsepower <- na_interpolation(autoData$horsepower)
autoData$`peak-rpm` <- na_interpolation(autoData$`peak-rpm`)
autoData$price <- na_interpolation(autoData$price)
autoData$`num-of-doors` <- autoData$`num-of-doors` %>% replace_na("four")
# The NA value is for sedan cars which mostly have four doors.
sapply(autoData, function(x) sum(is.na(x)))
```

```
##              make         fuel-type        aspiration      num-of-doors
##                 0                 0                 0                 0
##        body-style      drive-wheels   engine-location        wheel-base
##                 0                 0                 0                 0
##            length             width            height       curb-weight
##                 0                 0                 0                 0
##       engine-type  num-of-cylinders       engine-size       fuel-system
##                 0                 0                 0                 0
##              bore            stroke compression-ratio        horsepower
##                 0                 0                 0                 0
##          peak-rpm          city-mpg       highway-mpg             price
##                 0                 0                 0                 0
```

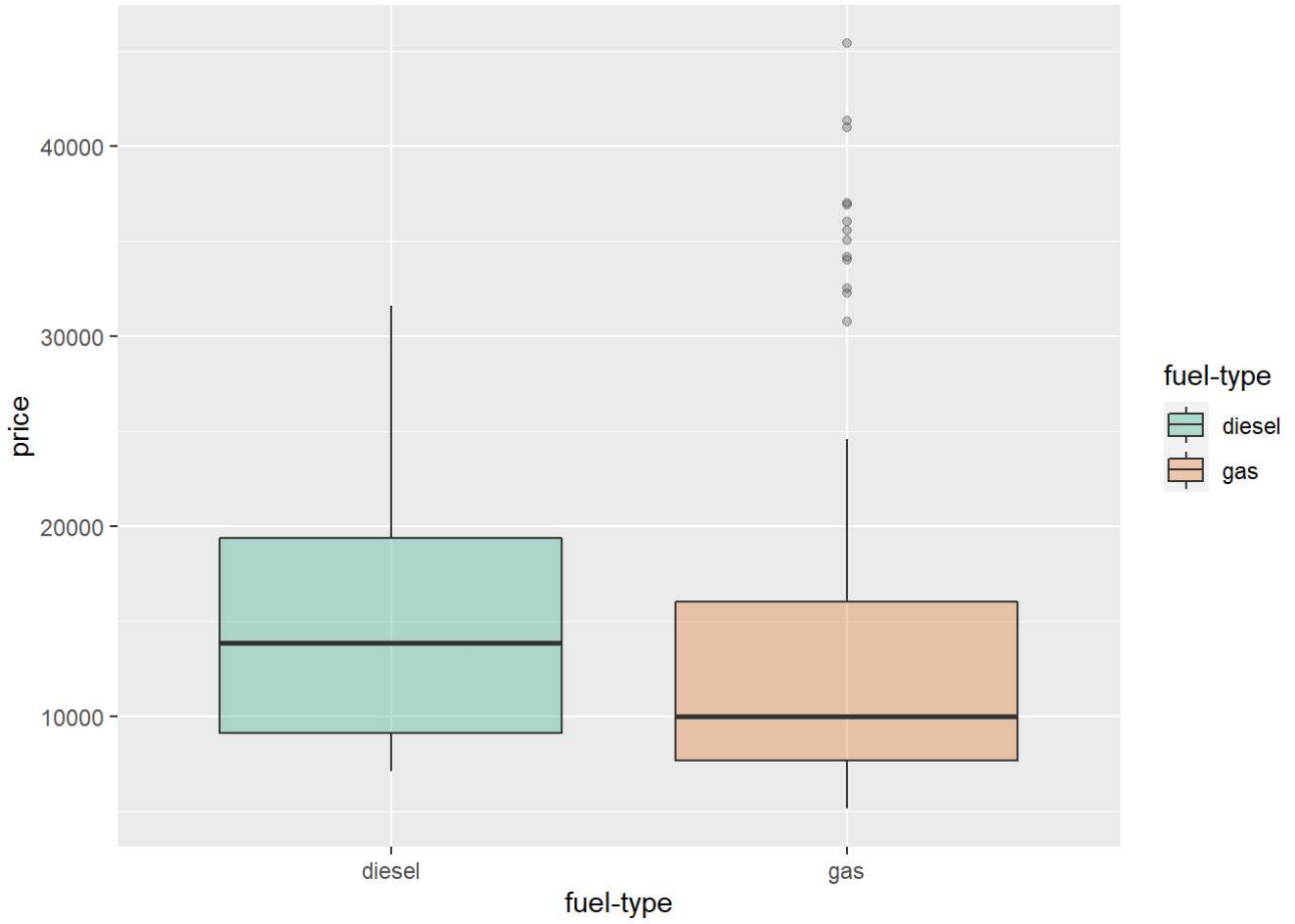
```
View(autoData)
# There are no missing values in our numeric variables now.
```

EDA

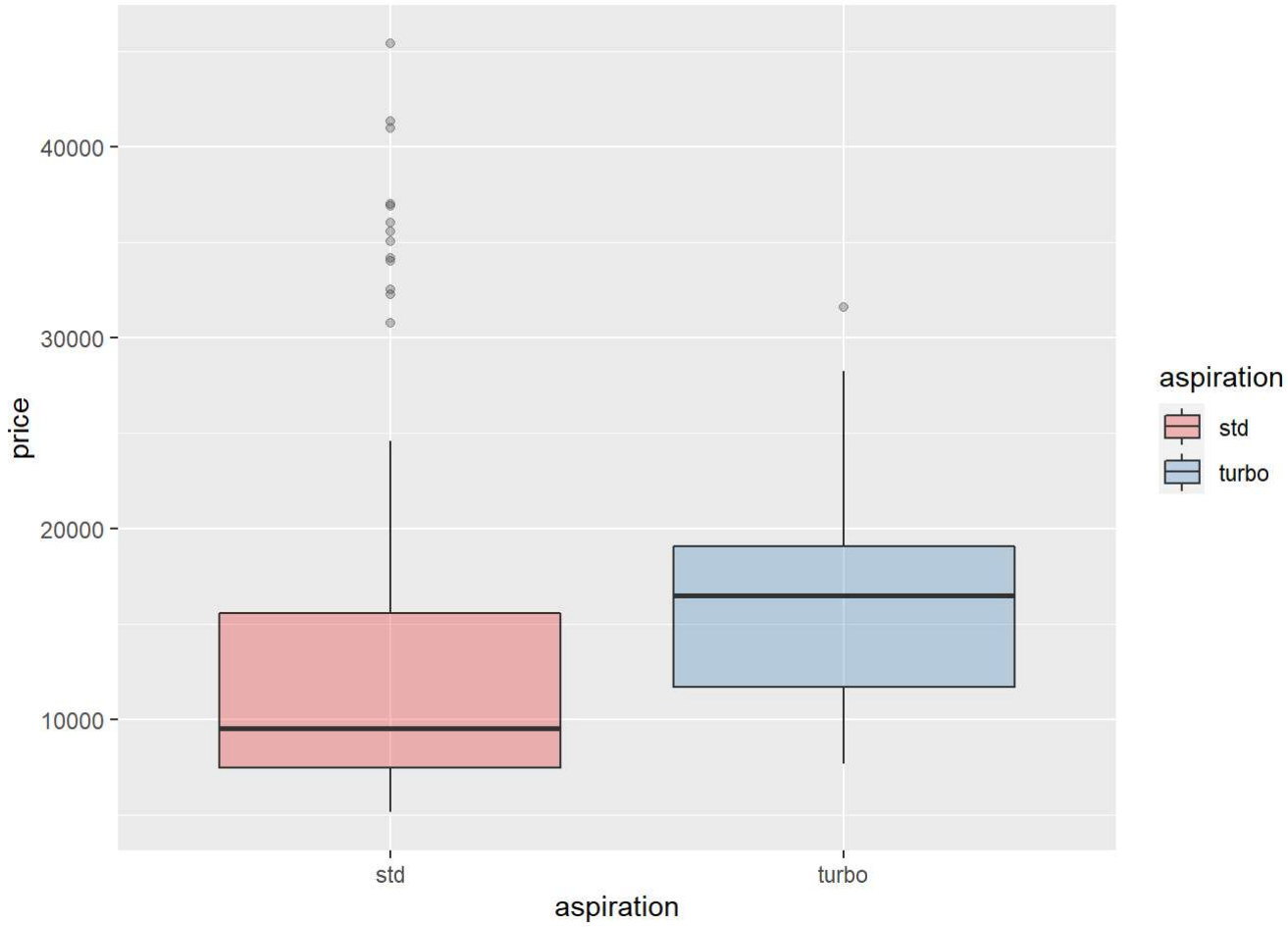
```
table(autoData$`fuel-type`)
```

```
##
## diesel    gas
##     20    185
```

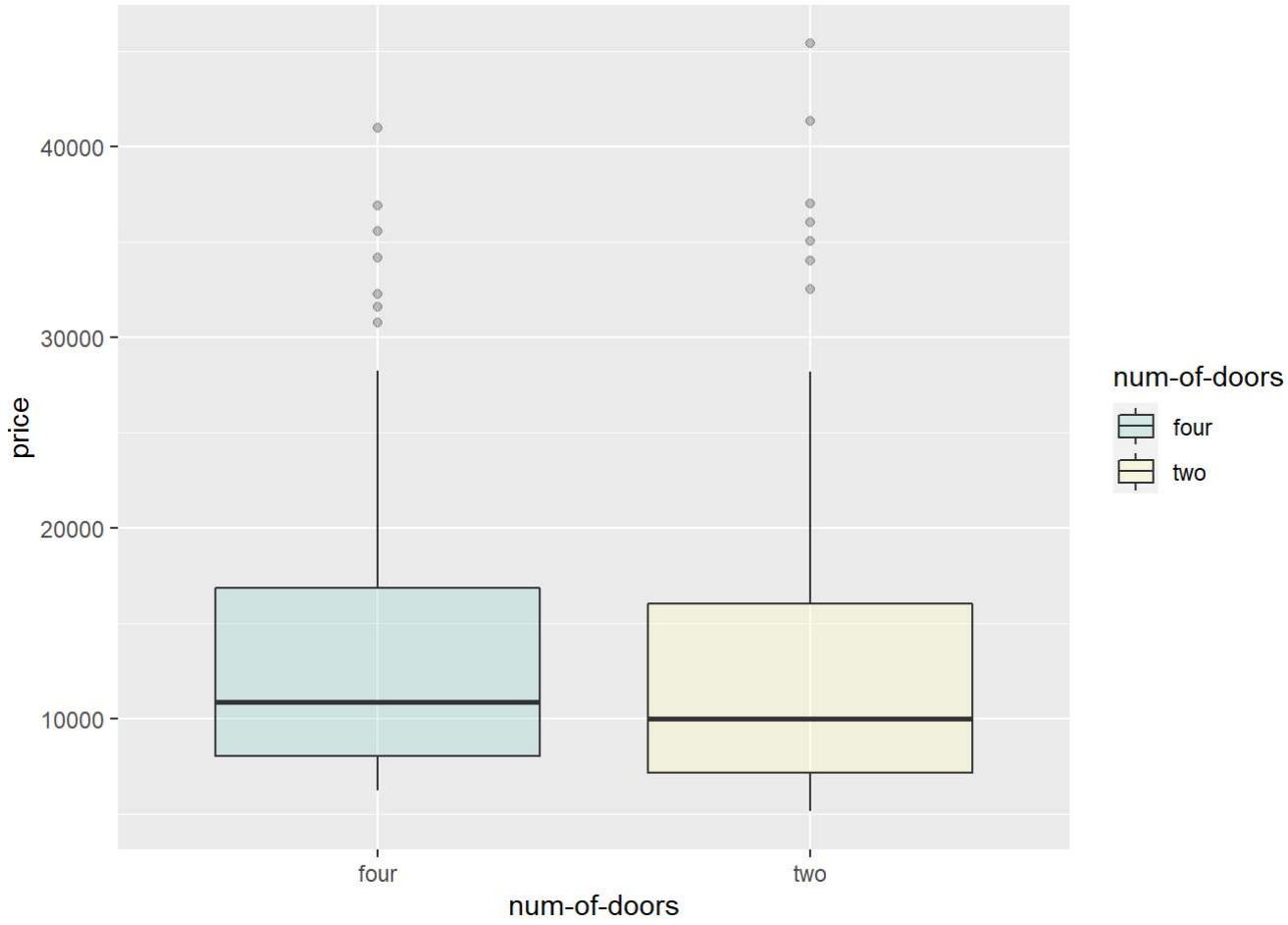
```
ggplot(autoData, aes(x=`fuel-type`, y=price, fill=`fuel-type`)) + geom_boxplot(alpha=0.3) + scal
e_fill_brewer(palette="Dark2")
```




```
# Mean price of gas is less than the mean price of diesel.There are many outliers for gas prices above $30000.
#The lowest value of diesel is slightly less than $10000 whereas the highest value of diesel is  approximately $30000. The lowest value of gas is less than $10000 whereas he highest value of diesel is approximately $25000

ggplot(autoData, aes(x=aspiration, y=price, fill=aspiration)) + geom_boxplot(alpha=0.3) + scale_fill_brewer(palette="Set1")
```

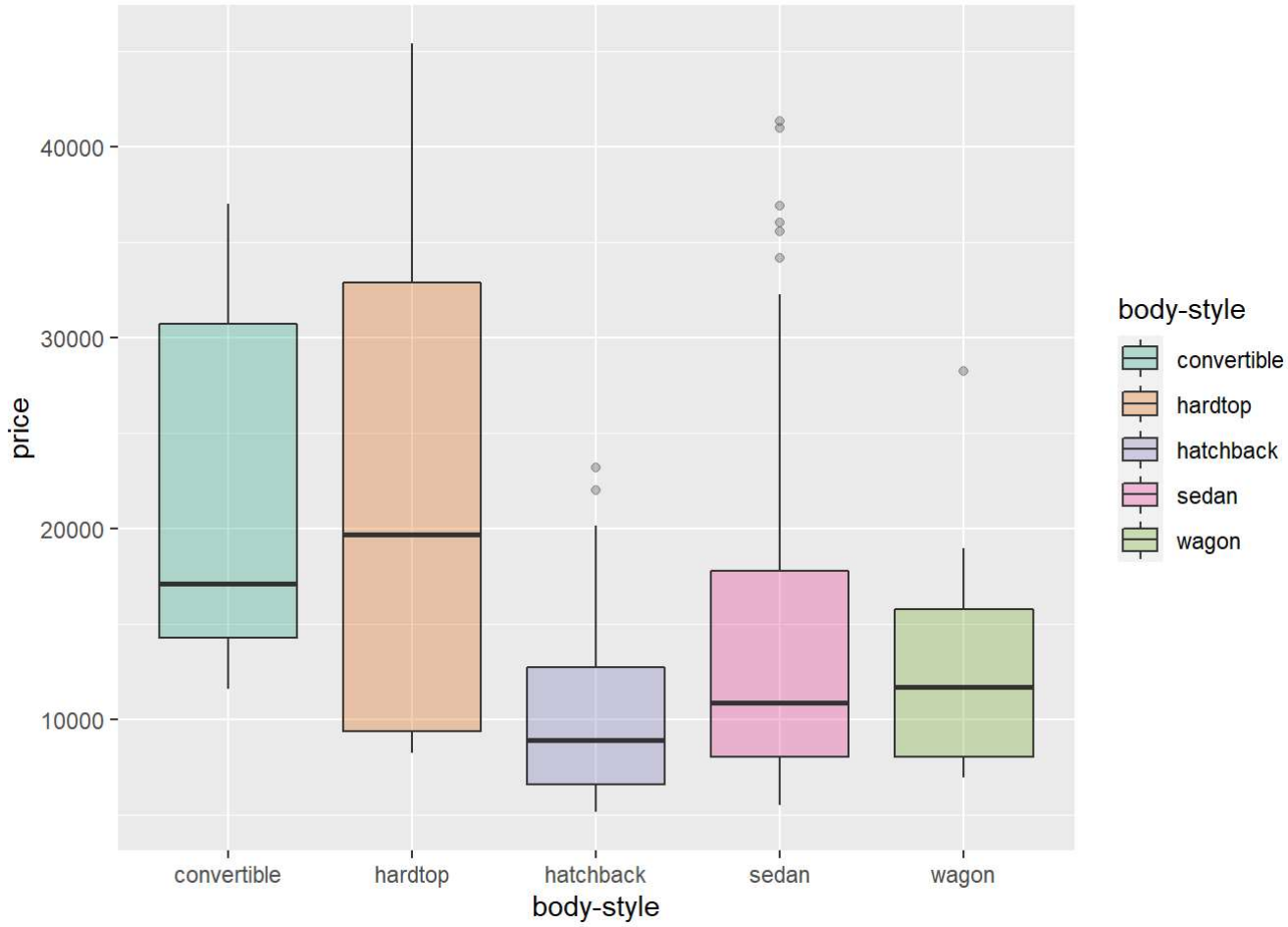



```
# Mean price of std is approximately 10000 whereas the mean price of turbo is approxmately $1500
0 i.e. std is cheaper than turbo. The lowest price of std is less than $10000 and highest value
 is approximately $25000. There are many outliers for values greater than 30000. The lowest pric
e of turbo is less than $10000 and the highest value of turbo is less than $30000

ggplot(autoData, aes(x=`num-of-doors`, y=price, fill=`num-of-doors`)) + geom_boxplot(alpha=0.3)
 + scale_fill_brewer(palette="Set3")
```

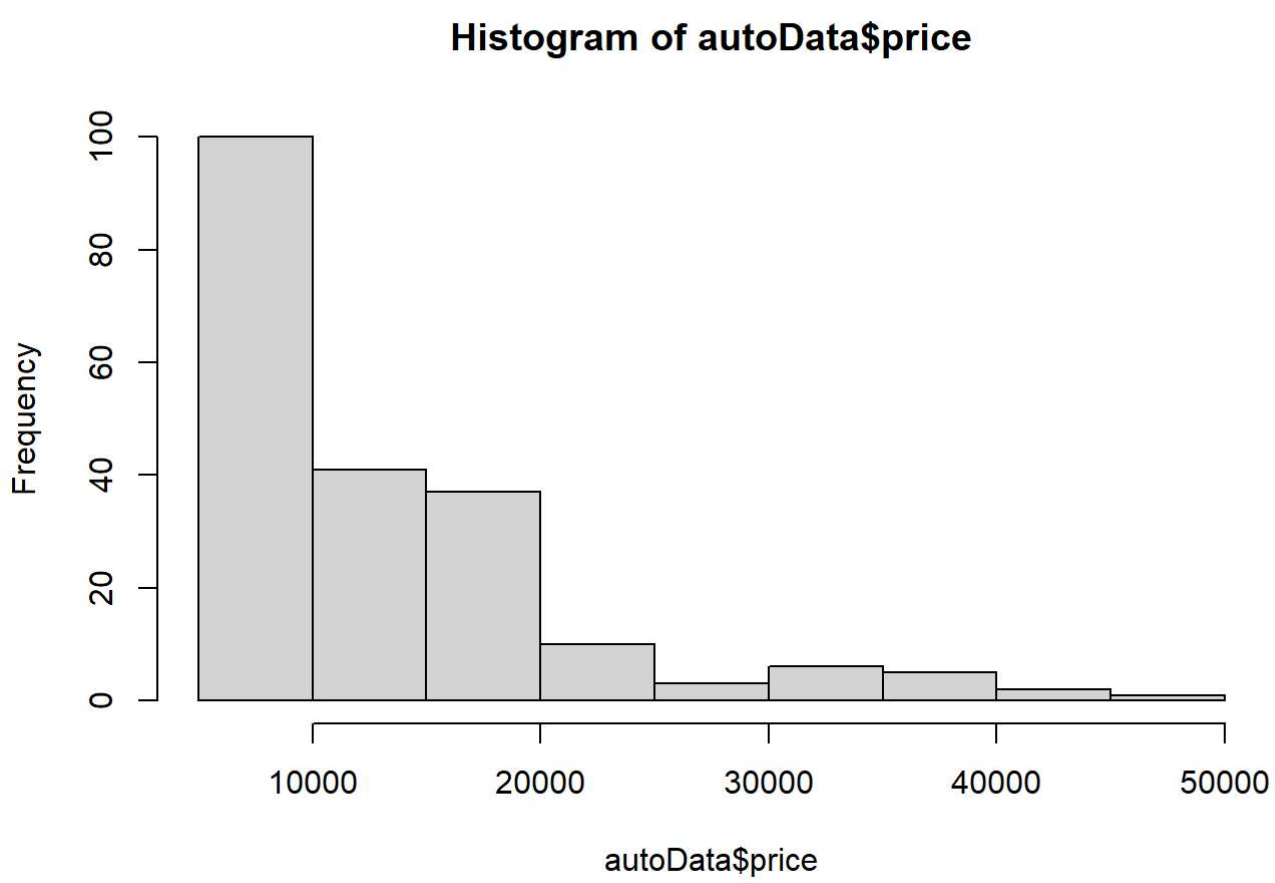


 # Mean price of 4 no. of doors is slightly greater than 2 doors. The lowest value of 4 no. of doors is slightly less than $10000 and the highest value of 4 no. of doors is slightly less than $30000. There are outliers present after the range of $30000. The lowest range of 2 no. of doors is very less than $10000 and the highest range of 2 no. of doors is approximately the same value as 4 no. of doors. The range of the outliers for the doors are almost similar like the 4. no of doors

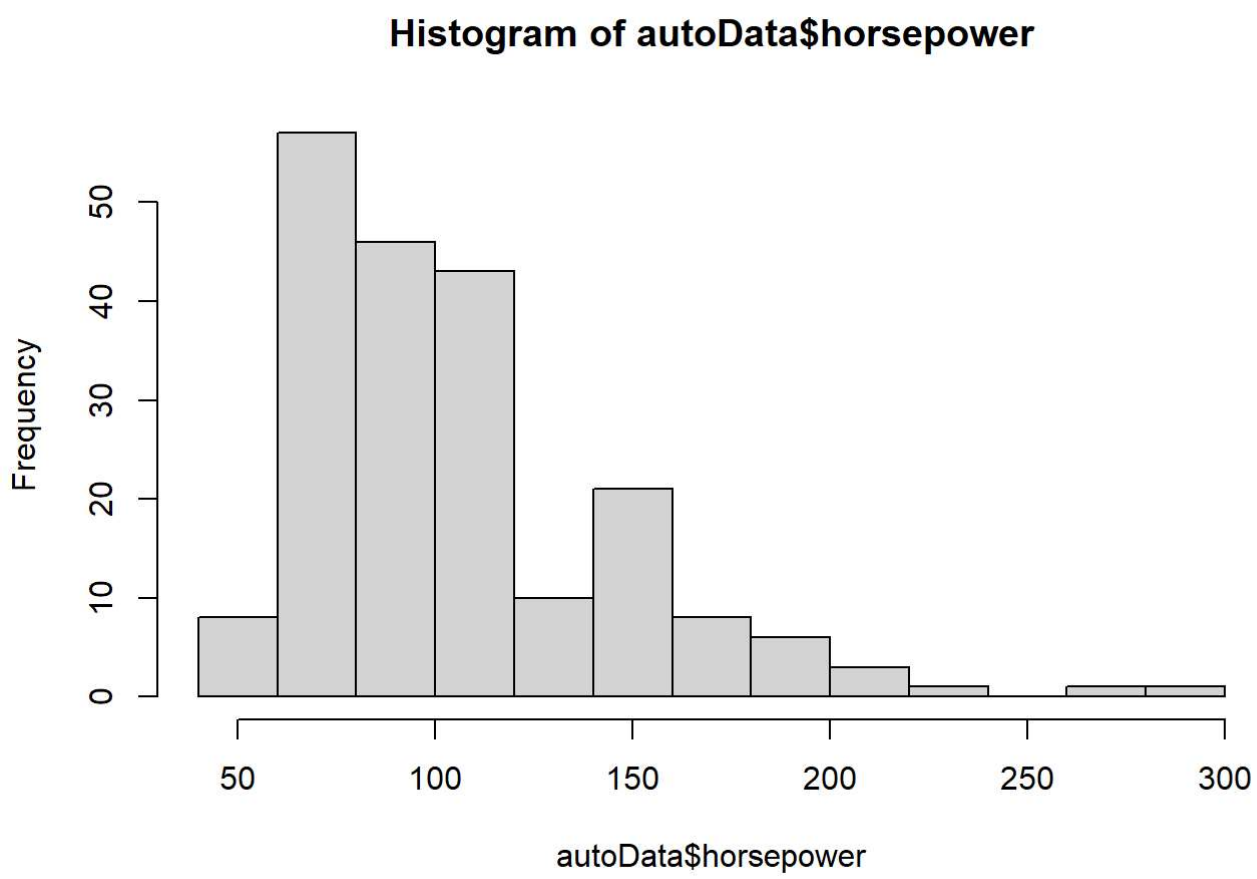
```
ggplot(autoData, aes(x=`body-style`, y=price, fill=`body-style`)) + geom_boxplot(alpha=0.3) + scale_fill_brewer(palette="Dark2")
```



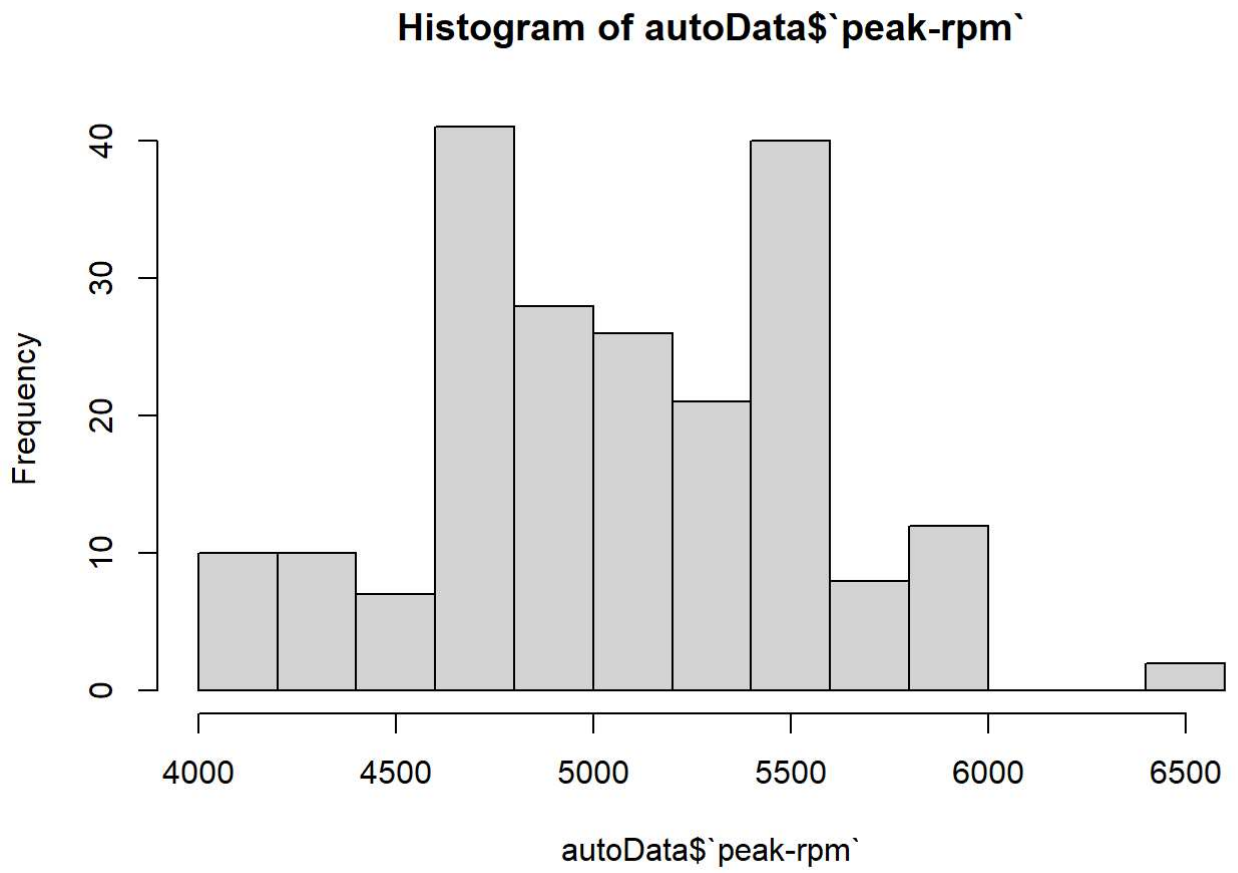

```
#mean price of hatchback is found to be the least whereas the mean price of hardtop is found to
 be the highest. The lowest value is less than $10000(hatchback) whereas the highest value is gr
eater than $40000(hardtop).


hist(autoData$price) # The distribution of the price of the cars is right-skewed.
```



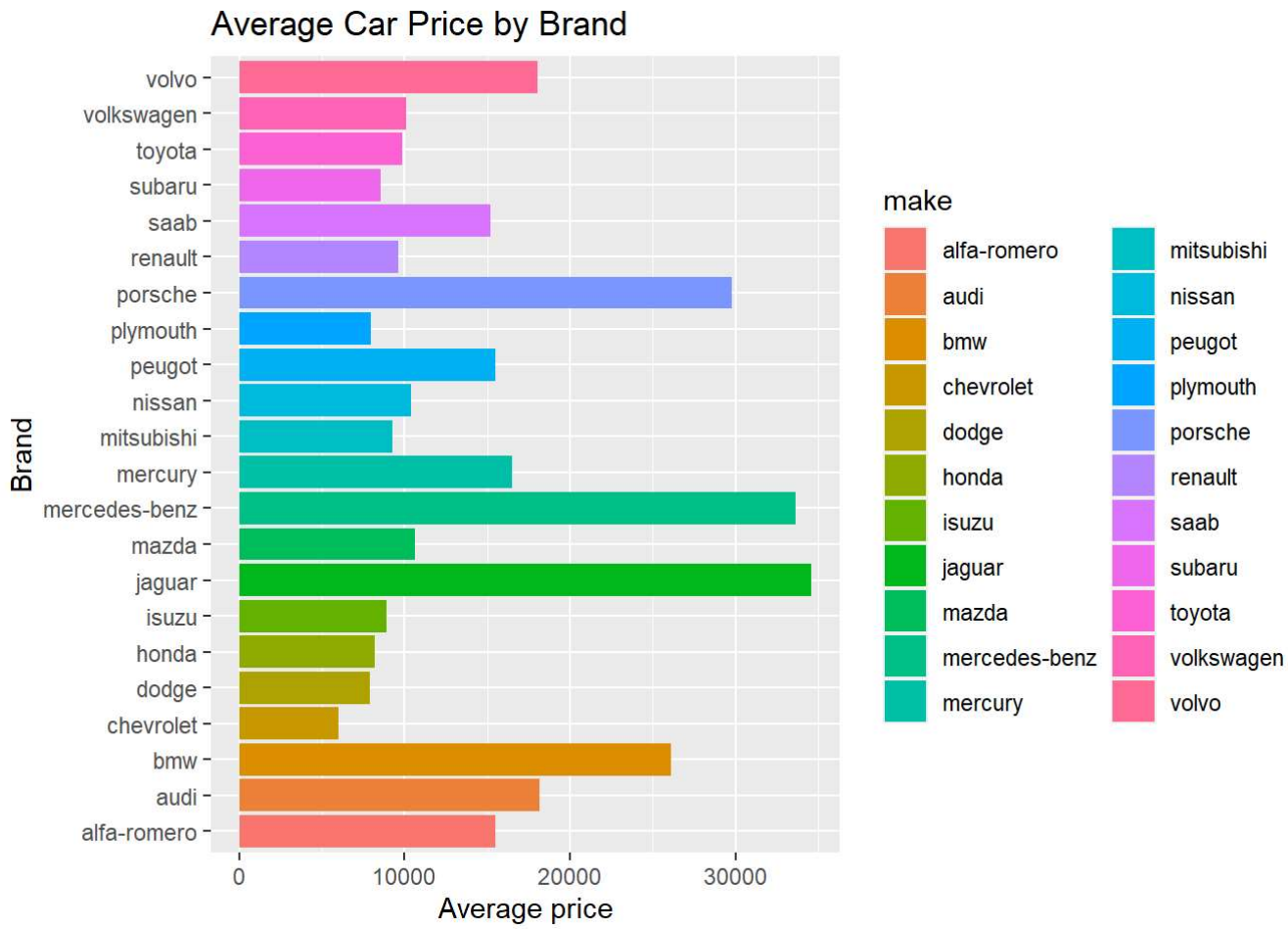

```
hist(autoData$horsepower) # distribution of the horsepower of cars is right-skewed.
```



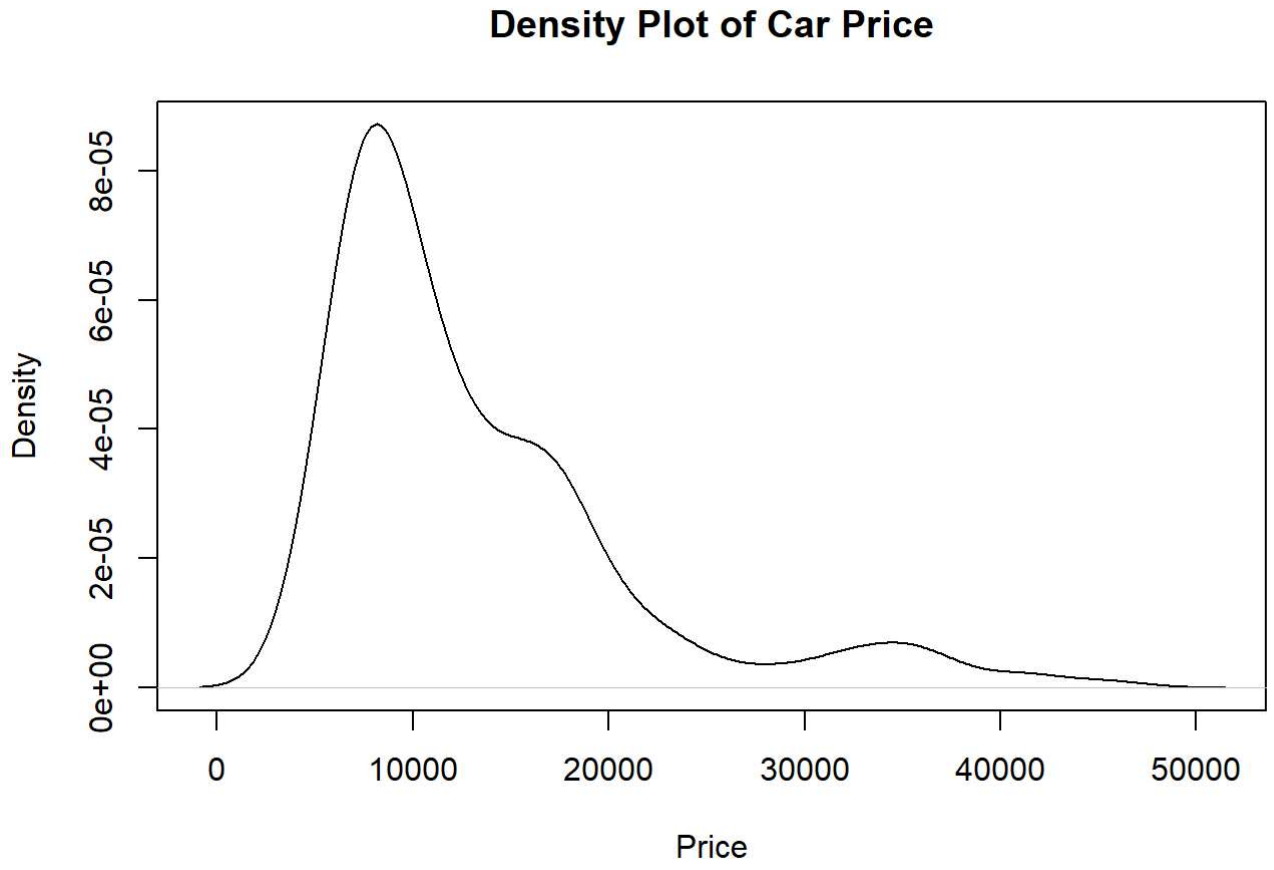

```
hist(autoData$`peak-rpm`) #distribution of the peak-rpm of the cars is normal distribution
```



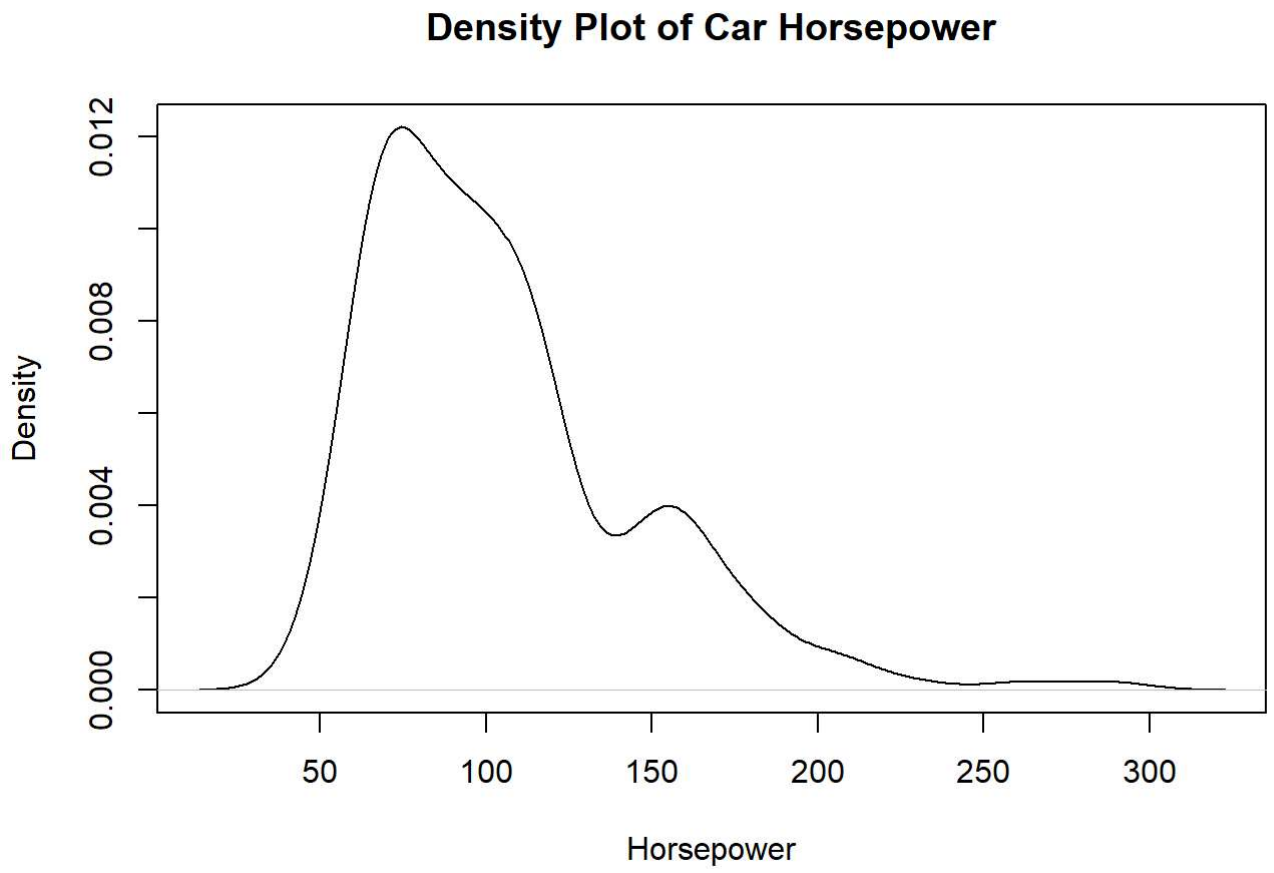

```
# Average sale prices of cars by brand
avgbymake <- autoData%>% group_by(make) %>% summarise(mean = mean(price))
ggplot(avgbymake, aes(x = make, y = mean, fill = make)) + geom_bar(stat = "identity", position =
"dodge") + coord_flip() + scale_colour_gradientn(colours=rainbow(4)) + labs(title = "Average Car
Price by Brand", y= "Average price", x="Brand")
```




```
# Mercedes-Benz, Jaguar, Porsche and BMW are the only four brands with average car prices over
 $25,000.

# Density plot of car price
plot(density(autoData$price), xlab='Price', main ="Density Plot of Car Price")
```




```
#Density plot of car horsepower
plot(density(autoData$horsepower), xlab='Horsepower', main ="Density Plot of Car Horsepower")
```



Correlation between variables

```
library(corrplot)
```

```
## corrplot 0.88 loaded
```

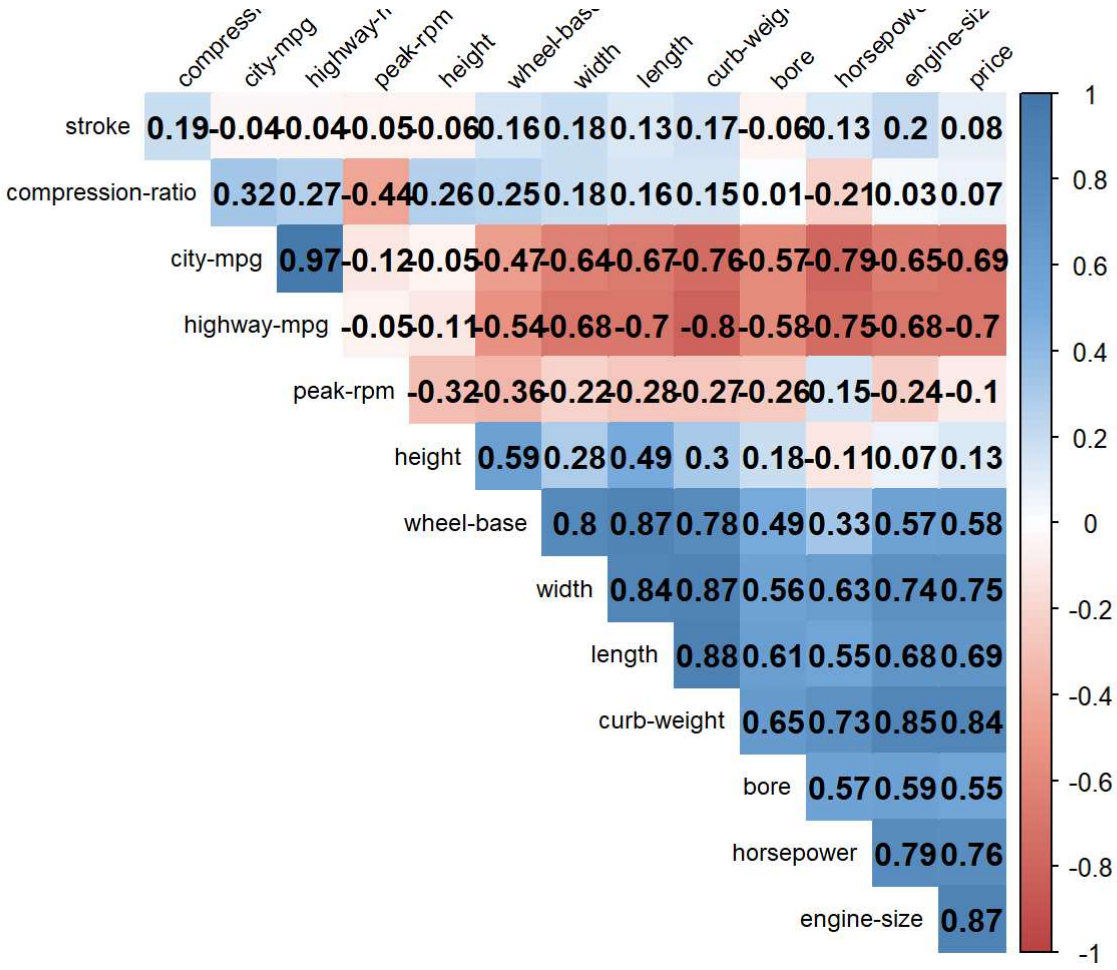
```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 4.0.5
```

```
numData <- select_if(autoData, is.numeric)
M <- round(cor(numData),2)
M
```

```
##                   wheel-base length width height curb-weight engine-size  bore
## wheel-base              1.00   0.87  0.80   0.59        0.78        0.57  0.49
## length                  0.87   1.00  0.84   0.49        0.88        0.68  0.61
## width                   0.80   0.84  1.00   0.28        0.87        0.74  0.56
## height                  0.59   0.49  0.28   1.00        0.30        0.07  0.18
## curb-weight             0.78   0.88  0.87   0.30        1.00        0.85  0.65
## engine-size             0.57   0.68  0.74   0.07        0.85        1.00  0.59
## bore                    0.49   0.61  0.56   0.18        0.65        0.59  1.00
## stroke                  0.16   0.13  0.18  -0.06        0.17        0.20 -0.06
## compression-ratio       0.25   0.16  0.18   0.26        0.15        0.03  0.01
## horsepower              0.33   0.55  0.63  -0.11        0.73        0.79  0.57
## peak-rpm               -0.36  -0.28 -0.22  -0.32       -0.27       -0.24 -0.26
## city-mpg               -0.47  -0.67 -0.64  -0.05       -0.76       -0.65 -0.57
## highway-mpg            -0.54  -0.70 -0.68  -0.11       -0.80       -0.68 -0.58
## price                   0.58   0.69  0.75   0.13        0.84        0.87  0.55
##                   stroke compression-ratio horsepower peak-rpm city-mpg
## wheel-base          0.16              0.25       0.33    -0.36    -0.47
## length              0.13              0.16       0.55    -0.28    -0.67
## width               0.18              0.18       0.63    -0.22    -0.64
## height             -0.06              0.26      -0.11    -0.32    -0.05
## curb-weight         0.17              0.15       0.73    -0.27    -0.76
## engine-size         0.20              0.03       0.79    -0.24    -0.65
## bore               -0.06              0.01       0.57    -0.26    -0.57
## stroke              1.00              0.19       0.13    -0.05    -0.04
## compression-ratio   0.19              1.00      -0.21    -0.44     0.32
## horsepower          0.13             -0.21       1.00     0.15    -0.79
## peak-rpm           -0.05             -0.44       0.15     1.00    -0.12
## city-mpg           -0.04              0.32      -0.79    -0.12     1.00
## highway-mpg        -0.04              0.27      -0.75    -0.05     0.97
## price               0.08              0.07       0.76    -0.10    -0.69
##                   highway-mpg price
## wheel-base              -0.54  0.58
## length                  -0.70  0.69
## width                   -0.68  0.75
## height                  -0.11  0.13
## curb-weight             -0.80  0.84
## engine-size             -0.68  0.87
## bore                    -0.58  0.55
## stroke                  -0.04  0.08
## compression-ratio        0.27  0.07
## horsepower              -0.75  0.76
## peak-rpm                -0.05 -0.10
## city-mpg                 0.97 -0.69
## highway-mpg              1.00 -0.70
## price                   -0.70  1.00
```

```
par(mfrow = c(1,1))
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(M, method="color", col=col(100),
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, tl.cex = 0.75, #Text label color and rotation
         diag=FALSE
         )
```




```
# Please use the show in new window option to view the correlation matrix values correctly.

# The outcome varaible price has the highest correlation with curb weight and engine size. The c
orrelation coeeficient is 0.84 and 0.87 for the same respectively. There is strong negative corr
elation between price and cit-mpg and highway-mpg. This indicates that those variables might hav
e to be inverse transformed for use in the models ahead.
```

t-test and Hypothesis testing

```
# Is there a significant difference between the mean values of the city mileage of a car and the
highway mileage of a car?

t.test(autoData$`city-mpg`, autoData$`highway-mpg`)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$`city-mpg` and autoData$`highway-mpg`
## t = -8.3383, df = 406.93, p-value = 1.176e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.835843 -4.227572
## sample estimates:
## mean of x mean of y
##  25.21951  30.75122
```

```
# We conduct a t-test to compare the mean values of the mileage of a car in the city vs on the h
ighway. The t-test gives us a result proving that there is a difference in the average mileage o
f car in the city vs the highway. The difference has a 95% confidence interval from -6.84 to -4.
22 which indicates that highway mileage is better than city mileage on an average by 4 to 7 mile
s per gallon.
# The p-value indicates that the test is significant and we reject the Null hypothesis that the
 difference in means is equal to 0.

# Does average price of cars change because of the type of fuel or the aspiration of the engine?

t.test(autoData$price[autoData$`fuel-type`=='gas'], autoData$price[autoData$`fuel-type`=='diese
l'])
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$price[autoData$`fuel-type` == "gas"] and autoData$price[autoData$`fuel-type`
== "diesel"]
## t = -1.5685, df = 23.482, p-value = 0.1301
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6652.0510   911.0699
## sample estimates:
## mean of x mean of y
##  12967.66  15838.15
```

```
# The result of the t-test comparing the price of cars which use gas vs diesel shows that there
 might be a credible difference. From the data, we know that we have very few observations with
 diesel fuel type. The results of this test are not significant based on the p-value. We fail to
reject the Null hypothesis.

t.test(autoData$price[autoData$aspiration =='std'], autoData$price[autoData$aspiration =='turbo'
])
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$price[autoData$aspiration == "std"] and autoData$price[autoData$aspiration ==
"turbo"]
## t = -3.2012, df = 67.126, p-value = 0.002092
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6165.925 -1429.914
## sample estimates:
## mean of x mean of y
##  12562.23  16360.15
```

```
# The result of the t-test comparing the price of cars which have standard aspiration vs turbo a
spiration shows that there might be a credible difference. The p-value of 0.0020 is significant
 at a standard alpha value of 0.05 and it means there is a significant difference between the av
erage price of cars of the two groups.
```

Chi-Squared Test

```
ch1 <- table(autoData$`num-of-doors`, autoData$`engine-location`)
ch2 <- table(autoData$`body-style`, autoData$aspiration)
ch3 <- table(autoData$`num-of-cylinders`, autoData$aspiration)
chOut <- chisq.test(ch1, correct=FALSE)
```

```
## Warning in chisq.test(ch1, correct = FALSE): Chi-squared approximation may be
## incorrect
```

```
chOut1 <- chisq.test(ch2, correct=FALSE)
```

```
## Warning in chisq.test(ch2, correct = FALSE): Chi-squared approximation may be
## incorrect
```

```
chOut2 <- chisq.test(ch3, correct=FALSE)
```

```
## Warning in chisq.test(ch3, correct = FALSE): Chi-squared approximation may be
## incorrect
```

```
chOut
```

```
##
##  Pearson's Chi-squared test
##
## data:  ch1
## X-squared = 3.9682, df = 1, p-value = 0.04637
```

```
#The reported value of chi-square, 3.90 on one degree of freedom, has a low corresponding p-valu
e 0.04, which is   just below the standard alpha level of p < 0.05. Thus we reject the null hypo
thesis of independence between number of doors and location of the engine.

chOut1
```

```
##
##  Pearson's Chi-squared test
##
## data:  ch2
## X-squared = 1.5971, df = 4, p-value = 0.8093
```

```
# The p-value of 0.8093 is not significant and we fail to reject the null hypothesis of independ
ence between body style and aspiration.

chOut2
```

```
##
##  Pearson's Chi-squared test
##
## data:  ch3
## X-squared = 13.864, df = 6, p-value = 0.03119
```

```
#The reported value of chi-square, 13.86 on 6 degrees of freedom, has a low corresponding p-valu
e 0.03, which is  below the standard alpha level of p < 0.05. Thus we reject the null hypothesis
of independence between number of cylinders and aspiration of the engine.

chOut$residuals # There are no particularly large residuals for this Chi-squared test. Largest t
wo values are two and four doors with rear engine location.
```

```
##
##             front       rear
##   four  0.1587807 -1.3029048
##   two  -0.1812724  1.4874648
```

```
chOut1$residuals # No large residuals here too. Largest here is turbo and convertible.
```

```
##
##                       std       turbo
##   convertible  0.48836640 -1.04063770
##   hardtop      0.17336636 -0.36941847
##   hatchback   -0.04830373  0.10292822
##   sedan       -0.07589478  0.16172072
##   wagon       -0.10777013  0.22964247
```

```
chOut2$residuals # Largest residual is turbo aspiration with five cylinder engine.
```

```
##
##                  std       turbo
##   eight   0.44581549 -0.94996791
##   five   -1.33712473  2.84921814
##   four   -0.02649486  0.05645668
##   six     0.52576383 -1.12032618
##   three   0.19937475 -0.42483856
##   twelve  0.19937475 -0.42483856
##   two     0.39874950 -0.84967713
```

Multiple Linear Regression Models

```
# 3 highest correlated variables
lm1 <- lm(price~horsepower+`curb-weight`+`engine-size`, data = autoData)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ horsepower + `curb-weight` + `engine-size`,
##     data = autoData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8742.4 -1757.0    34.4  1310.2 14315.5
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.371e+04  1.372e+03  -9.999  < 2e-16 ***
## horsepower     2.669e+01  1.020e+01   2.618  0.00952 **
## `curb-weight`  4.735e+00  9.302e-01   5.090 8.19e-07 ***
## `engine-size`  9.498e+01  1.299e+01   7.309 6.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3581 on 201 degrees of freedom
## Multiple R-squared:  0.7988, Adjusted R-squared:  0.7958
## F-statistic: 266.1 on 3 and 201 DF,  p-value: < 2.2e-16
```

```
# Adjusted R-squared value is 0.7958

lm2 <- lm(price~horsepower+`curb-weight`+`engine-size`+`city-mpg`+`highway-mpg`, data = autoDat
a)
summary(lm2)
```

```
##
## Call:
## lm(formula = price ~ horsepower + `curb-weight` + `engine-size` +
##     `city-mpg` + `highway-mpg`, data = autoData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8857.6 -1791.0   151.3  1348.4 14295.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -10596.824   4109.078  -2.579 0.010634 *
## horsepower        18.323     13.155   1.393 0.165228
## `curb-weight`      4.265      1.133   3.763 0.000221 ***
## `engine-size`     99.387     13.697   7.256 8.71e-12 ***
## `city-mpg`      -115.300    186.781  -0.617 0.537744
## `highway-mpg`     42.684    175.468   0.243 0.808058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3590 on 199 degrees of freedom
## Multiple R-squared:  0.7999, Adjusted R-squared:  0.7949
## F-statistic: 159.1 on 5 and 199 DF,  p-value: < 2.2e-16
```

```
# Adjusted R-squared value is 0.7949

lm3 <- lm(price~`city-mpg`+`highway-mpg`, data = autoData)
summary(lm3)
```

```
##
## Call:
## lm(formula = price ~ `city-mpg` + `highway-mpg`, data = autoData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8666  -3449  -1273   1156  20889
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37852.44    1985.07  19.069  < 2e-16 ***
## `city-mpg`      -86.53     255.01  -0.339  0.73473
## `highway-mpg`  -729.16     242.26  -3.010  0.00295 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5664 on 202 degrees of freedom
## Multiple R-squared:  0.4943, Adjusted R-squared:  0.4893
## F-statistic: 98.74 on 2 and 202 DF,  p-value: < 2.2e-16
```

```
# Adjusted R-squared value is 0.4893. This shows that a model with just these variables is not s
uch a good model.

# 5 highest correlated variables
lm4 <- lm(price~horsepower+`curb-weight`+`engine-size`+width+length, data = autoData)
summary(lm4)
```

```
##
## Call:
## lm(formula = price ~ horsepower + `curb-weight` + `engine-size` +
##     width + length, data = autoData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8118.7 -1615.3   -49.5  1386.6 14957.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -37243.323  13338.391  -2.792  0.00575 **
## horsepower        25.156     10.348   2.431  0.01595 *
## `curb-weight`      3.920      1.553   2.524  0.01239 *
## `engine-size`     93.640     13.083   7.157 1.55e-11 ***
## width            489.383    249.338   1.963  0.05107 .
## length           -36.255     47.370  -0.765  0.44497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3565 on 199 degrees of freedom
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7977
## F-statistic: 161.9 on 5 and 199 DF,  p-value: < 2.2e-16
```

```
# Adjusted R-squared value is 0.7977

lm1BF <- lmBF(price~horsepower+`curb-weight`+`engine-size`, data = autoData)
```

```
## Warning: data coerced from tibble to data frame
```

```
summary(lm1BF)
```

```
## Bayes factor analysis
## --------------
## [1] horsepower + `curb-weight` + `engine-size` : 1.306696e+66 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# The Bayes factor linear model summary gives an odds ratio that is highly in favor of the alter
nate hypothesis that having horsepower, curb weight and engine size as predictors of the price o
f the car is significant. We fail to reject the bull hypothesis. The linear regression model out
put gives an adjusted R-squared of 0.7958. This means that the model accounts for almost 80% of
 variability in the price of the car.

# The addition of the variables city-mpg and highway-mpg causes a decrease in the adjusted R-squ
ared score.  Hence, these variables should not be included for the best possible model.

lm4BF <- lmBF(price~horsepower+`curb-weight`+`engine-size`+width+length, data = autoData)
```

```
## Warning: data coerced from tibble to data frame
```

```
summary(lm4BF)
```

```
## Bayes factor analysis
## --------------
## [1] horsepower + `curb-weight` + `engine-size` + width + length : 4.268319e+64 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# The Bayes factor linear model summary gives an odds ratio that is highly in favor of the alter
nate hypothesis that having horsepower, curb weight and engine size as predictors of the price o
f the car is significant. We fail to reject the bull hypothesis. The linear regression model out
put gives an adjusted R-squared of 0.7977. This means that the model accounts for almost 80% of
 variability in the price of the car.
# This is slightly higher than the model using only the top 3 highest correlated variables to pr
edict the price. As we can see from the model, the width and lenght variables are mot significan
t.

lm3BF <- lmBF(price~`city-mpg`+`highway-mpg`, data = autoData)
```

```
## Warning: data coerced from tibble to data frame
```

```
summary(lm3BF)
```

```
## Bayes factor analysis
## --------------
## [1] `city-mpg` + `highway-mpg` : 2.48648e+27 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
# The Bayes Factor linear model shows an odds ratio highly in favor of the alternaitve hypotheis
that having city-mpg and highway-mpg variables to help predict the price of a car is signigican
t. This means that we fail to reject the null hypothesis. Based on the linear regression model o
utput, we can conclude that there is not sufficient evidence to prove that the inclusion of city
-mpg and highway-mpg in our model accounts for more variability of the price variable.
```

AOV

```
# Are the factors of groups like body style i.e. convertible, sedan, hatchback, etc., aspiratio
n, engine location, or type of wheel drive sampled from the same population?

aov1 <- aov(price~`body-style`, data = autoData)
summary(aov1)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## `body-style`   4 1.837e+09 459203739   8.365 2.93e-06 ***
## Residuals    200 1.098e+10  54895882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov2 <- aov(price~`drive-wheels`+aspiration, data = autoData)
summary(aov2)
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## `drive-wheels`   2 5.226e+09 2.613e+09  70.610 <2e-16 ***
## aspiration       1 1.518e+08 1.518e+08   4.101 0.0442 *
## Residuals      201 7.438e+09 3.701e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov3 <- aov(price~aspiration+`body-style`+`engine-location`, data = autoData)
summary(aov3)
```

```
##                    Df    Sum Sq   Mean Sq F value   Pr(>F)
## aspiration          1 4.374e+08 437369658   8.843 0.003308 **
## `body-style`        4 1.955e+09 488716478   9.881 2.61e-07 ***
## `engine-location`   1 6.309e+08 630893183  12.756 0.000446 ***
## Residuals         198 9.793e+09  49458902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AOVBayesOut <- anovaBF(price~`body-style`, data = autoData)
```

```
## Warning: data coerced from tibble to data frame
```

```
summary(AOVBayesOut)
```

```
## Bayes factor analysis
## --------------
## [1] body-style : 3416.592 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
 # Show the range of values for the grand mean

# The Pr(>F) values for all three ANOVA tests were significant which means the variable body-sty
le i.e., convertible,hatchback,sedan and the variable drive-wheels i.e., whether the car is rea
r/front wheel drive or 4 wheel drive, makes a significant difference on the determning the price
of the car.
# The F-value were significantly greater than 1, which means that the tests were significant and
the means of the car price for the groups differ from each other.

# The Bayes analysis using ANOVA gave us an odds ratio of 3416:1 in favor of the alternate hypot
hesis, which is a very strong result according to the rule of thumb.  This result confirms our p
revious evidence suggesting support for an alternative hypothesis of credible differences among
 these means of the body style groups.
```

GLM Output - Transformed Predictor

```
glmOut <- glm(formula = aspiration ~ horsepower,
              family = binomial(link="logit"),
              data = autoData)
summary(glmOut)
```

```
##
## Call:
## glm(formula = aspiration ~ horsepower, family = binomial(link = "logit"),
##     data = autoData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4818  -0.6258  -0.5327  -0.4794   2.0862
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.904869   0.519162  -5.595  2.2e-08 ***
## horsepower   0.012489   0.004146   3.012  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 193.57  on 204  degrees of freedom
## Residual deviance: 184.45  on 203  degrees of freedom
## AIC: 188.45
##
## Number of Fisher Scoring iterations: 4
```

```
confint(glmOut)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.970731305 -1.92202767
## horsepower   0.004428591  0.02084332
```

```
exp(confint(glmOut))
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %   97.5 %
## (Intercept) 0.01885964 0.146310
## horsepower  1.00443841 1.021062
```

```
# The results are not intuitive.
# Logistic regression might not be the ideal test for these variables and the dataset.
```

```
glmOut2 <- glm(formula = `engine-location` ~ horsepower,
               family = binomial(link="logit"),
               data = autoData)
summary(glmOut2)
```

```
##
## Call:
## glm(formula = `engine-location` ~ horsepower, family = binomial(link = "logit"),
##     data = autoData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45571  -0.10251  -0.07164  -0.04675   2.11984
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.21161    2.21854  -4.152 3.29e-05 ***
## horsepower   0.03419    0.01130   3.025  0.00249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31.302  on 204  degrees of freedom
## Residual deviance: 20.017  on 203  degrees of freedom
## AIC: 24.017
##
## Number of Fisher Scoring iterations: 8
```

```
confint(glmOut2)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) -14.93130370 -5.71916184
## horsepower    0.01433319  0.06168339
```

```
exp(confint(glmOut2))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) 3.276553e-07 0.003282461
## horsepower  1.014436e+00 1.063625538
```

```
# The secong model is a bit more intuitive. The 95% credible interval is still extremely small.
# The AIC of the model is 23.275. The variables are significant for the logistic regression mode
l.
```

Pseudo-Rsquared values

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.0.5
```

```
PseudoR2(glmOut, which="Nagelkerke")
```

```
## Nagelkerke
## 0.07122233
```

```
PseudoR2(glmOut2, which="Nagelkerke")
```

```
## Nagelkerke
##  0.3782546
```

```
# The Pseudo R squared value of 0.0712 is pretty low. It indicates that the horsepower can accou
nt for only 7% variation in the aspiration of a car.
# The Pseudo R squared value of 0.3782 is a decent value. It indicates 37% variability in engine
location because of the horsepower value of a car.
```

Conclusion

1. There was a significant difference in the average mileage of a car in the city versus on the highway. The average price of a car with standard aspirated engine versus turbo aspirated engine also showed a significant difference.
2. The number of doors in a car and the location of the engine were dependent on each other. The number of cylinders and aspiration of the engine were also variables that were dependent on each other.
3. The top three highest correlated variables i.e., curb weight, engine size and horsepower were the best to predict the price of a car. Adding the two largest negatively correlated variables did not improve the model but rather decreased the R-squared value of the model by a bit. If the top five highest correlated variables are considered in the model, the adjusted R-squared value increased by a negligible amount and those additional varaibles did not show a significant p-value anyway.

4. The ANOVA tests answered our questions regarding the difference in the mean price of the car for different groups. Factors like body-style, type of wheel drive, enine aspiration and engine location showed a significant difference in the average car prices.
5. The logistic regression models did not yield intuitive results for this dataset. There was no significant result for prediciting aspiration type or engine location using the variable horsepower. At most, the model could account for 37% variability of the dependent variable.