Automotive Insights: Exploratory Data Analysis of Car Sales datasets using R

Hrithvik Jeevankumar Shetty

2023-04-05

Loading of Library

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1  
## ✔ tibble 3.1.8 ✔ dplyr 1.1.0  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 1.0.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(matlib)

## Warning: package 'matlib' was built under R version 4.2.3

## *Introduction*

The dataset Car\_sales.csv consists of information regarding total sales of different car brand and their model. I chose this dataset out of all is because I am an automobile enthusiast since my childhood and loved everything about it. Here cars are grouped under its brands, price, engine capacity, power, etc. It is differentiated in the most efficient way according to buyer’s point of view so that he can get to know about all the information of different cars in one dataset. Buyer can compare according to their requirement like price of the car, engine capacity and make right decision about his purchase. It can answer some of the question such as which automaker has achieved the highest overall sales till date, when was the newest automobile model introduced, what is the estimated year-to-year resale value, comparison of the car's brand, price, and engine power.

#Data 1. Data source: Include the citation for your data & provide the link to the source. <https://www.kaggle.com/datasets/gagandeep16/car-sales>

1. Data collection: This data set was obtained from Analytixlabs in to predict future car sales and assist customers with their purchases.
2. Cases: The rows represent various attributes of the cars & its sales.
3. Variables
4. Manufacturer- Brand who manufactures the car.
5. Model- model from certain manufacturer of the car.
6. Sales\_in\_thousands- Number of cars sold.
7. \_\_year\_resale\_value- Resale value of the car after a year.
8. Vehicle\_type- Type of vehicle depending on its seating.
9. Price\_in\_thousands- Price of the car.
10. Engine\_size- Capacity of the engine in the car.
11. Horsepower- Horsepower of the car engine.
12. Wheelbase- Length from front wheel to rare wheel.
13. Width- Width of the car.
14. Type of study: It is an observational dataset with data collected purely for the intention of prediction by Analytixhub.

# Data Quality

Adding Dataset

TheDataset <- read.csv("C:\\Users\\user\\Downloads\\Car-sales.csv")  
head(TheDataset)

## Manufacturer Model Sales\_in\_thousands X\_\_year\_resale\_value Vehicle\_type  
## 1 Acura Integra 16.919 16.360 Passenger  
## 2 Acura TL 39.384 19.875 Passenger  
## 3 Acura CL 14.114 18.225 Passenger  
## 4 Acura RL 8.588 29.725 Passenger  
## 5 Audi A4 20.397 22.255 Passenger  
## 6 Audi A6 18.780 23.555 Passenger  
## Price\_in\_thousands Engine\_size Horsepower Wheelbase Width Length Curb\_weight  
## 1 21.50 1.8 140 101.2 67.3 172.4 2.639  
## 2 28.40 3.2 225 108.1 70.3 192.9 3.517  
## 3 NA 3.2 225 106.9 70.6 192.0 3.470  
## 4 42.00 3.5 210 114.6 71.4 196.6 3.850  
## 5 23.99 1.8 150 102.6 68.2 178.0 2.998  
## 6 33.95 2.8 200 108.7 76.1 192.0 3.561  
## Fuel\_capacity Fuel\_efficiency Latest\_Launch Power\_perf\_factor  
## 1 13.2 28 2/2/2012 58.28015  
## 2 17.2 25 6/3/2011 91.37078  
## 3 17.2 26 1/4/2012 NA  
## 4 18.0 22 3/10/2011 91.38978  
## 5 16.4 27 10/8/2011 62.77764  
## 6 18.5 22 8/9/2011 84.56511

(dim(TheDataset))

## [1] 157 16

The dataset has 16 columns & 157 observations

For finding missing value

sum(is.na(TheDataset))

## [1] 51

There are 51 missing values in this Dataset. Imputing the missing values in the data set with the mean of that variable

colSums(is.na(TheDataset))

## Manufacturer Model Sales\_in\_thousands   
## 0 0 0   
## X\_\_year\_resale\_value Vehicle\_type Price\_in\_thousands   
## 36 0 2   
## Engine\_size Horsepower Wheelbase   
## 1 1 1   
## Width Length Curb\_weight   
## 1 1 2   
## Fuel\_capacity Fuel\_efficiency Latest\_Launch   
## 1 3 0   
## Power\_perf\_factor   
## 2

As we can see, the data is now free from all missing values & the dimension of the dataset is still same

For finding duplicates

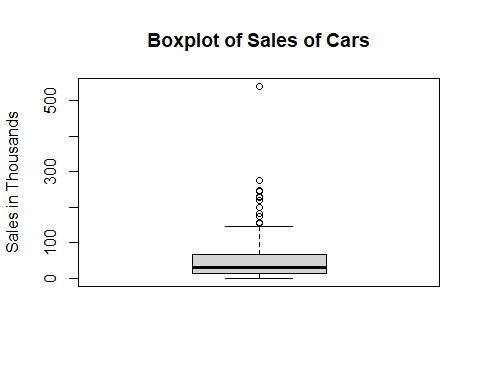
sum(duplicated(TheDataset))

## [1] 0

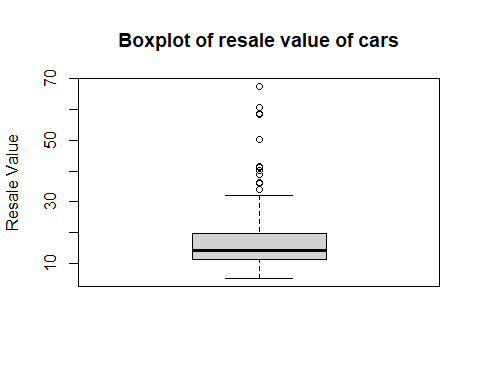
There are zero duplicate values in this Dataset

# Data Cleaning

Sales\_Outliers <- boxplot(TheDataset$Sales\_in\_thousands, ylab = "Sales in Thousands",main = "Boxplot of Sales of Cars")$out



resale\_outliers <- boxplot(TheDataset$X\_\_year\_resale\_value, ylab = "Resale Value",main = "Boxplot of resale value of cars")$out

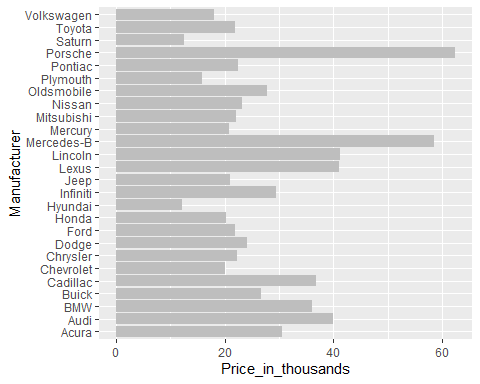


Data <- TheDataset  
Data <- Data[-which(Data$Sales\_in\_thousands %in% Sales\_Outliers),]  
Data <- Data[-which(Data$X\_\_year\_resale\_vallue %in% resale\_outliers),]

# Data Visualization

Comparing average price for different automobile manufacturers.

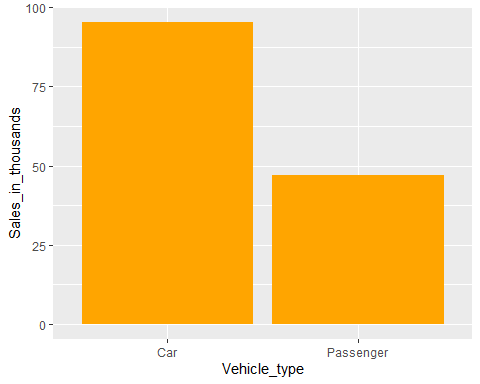
TheData <- na.omit(TheDataset)  
ggplot(TheData,aes(Manufacturer,Price\_in\_thousands))+  
geom\_bar(stat = "summary", fun = "mean", fill = "Grey")+coord\_flip()



We can say from the above graph Porsche and Mercedes Benz are the most expensive while Hyundai and Saturn being the cheapest

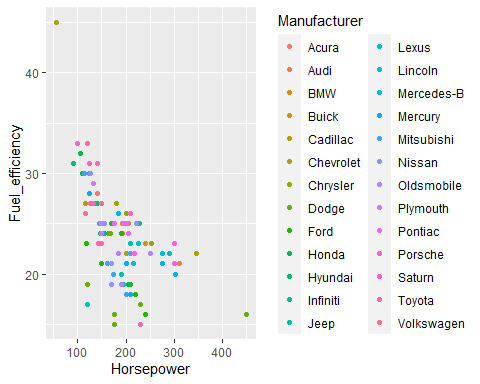
Comparing sales of cars and passengers.

ggplot(TheData,aes(Sales\_in\_thousands, Vehicle\_type)) +  
geom\_bar(stat = "summary",fun = "mean", fill = "ORANGE")+ coord\_flip()

 We can say from the graph that car was getting sold more compared to passenger by margin of almost 45,000 units.

Comparing horsepower with fuel efficiency for all Automobile manufacturers.

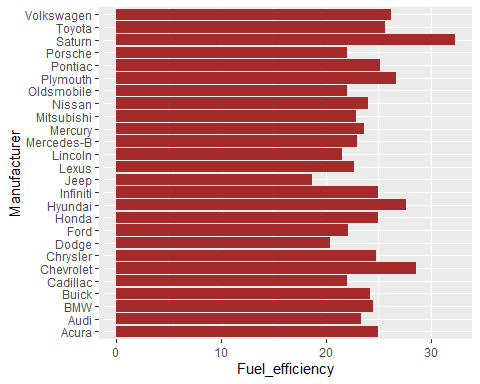
ggplot(TheData,aes(Horsepower,Fuel\_efficiency, color= Manufacturer)) +  
geom\_point()



We can observe from the graph that as the horsepower of car increases the fuel efficiency decreases.

Determining fuel efficiency for different Automobile manufacturers.

ggplot(TheData,aes(Manufacturer,Fuel\_efficiency)) +  
geom\_bar(position = "dodge", stat = "summary",fun = "mean", fill = "BROWN") +  
coord\_flip()



We can observe from the graph that Saturn and Chevrolet has highest fuel efficiency compared to other car manufacturer while Jeep being the lowest efficiency.

### **Questions for next stage**

1. Is the sales of car more of high fuel efficient cars?
2. Is cost associated with the engine size
3. Is fuel capacity associated with the curb weight of the car?

# HYPOTHESIS 1

In EDA, I evaluated average fuel efficiency of all the car manufacturers. In this, I want to test if the average fuel efficiency of all the vehicles is 25 miles per gallon or less.

H0 = The mean of fuel efficiency was proven to be 25 H1 = Mean is less than 25

res <- t.test(TheData$Fuel\_efficiency, mu = 25)  
res

##   
## One Sample t-test  
##   
## data: TheData$Fuel\_efficiency  
## t = -2.162, df = 116, p-value = 0.03267  
## alternative hypothesis: true mean is not equal to 25  
## 95 percent confidence interval:  
## 23.31316 24.92616  
## sample estimates:  
## mean of x   
## 24.11966

The p-value t-test is less than the significance level of 0.05. In conclusion, there is a difference between the mean. Hence we reject the NULL Hypothesis Ho, i.e µ = 25.

# HYPOTHESIS 2

As the horsepower of the car engine & its fuel efficiency are correlated, I’m trying to figure out whether their variance matches up for the vast variety of car models.

H0: σ1 = σ2 (the variances are equal)

H1: σ1 ≠ σ2 (the variances are not equal)

var.test(TheData$Horsepower,TheData$Fuel\_efficiency)

##   
## F test to compare two variances  
##   
## data: TheData$Horsepower and TheData$Fuel\_efficiency  
## F = 176.96, num df = 116, denom df = 116, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 122.7560 255.1114  
## sample estimates:  
## ratio of variances   
## 176.9645

It is obvious that because the P value is significantly below 0.05, the null hypothesis can be rejected. As a result, the alternative hypothesis is correct. Additionally, it is noticed that difference between 2 samples for 95% confidence interval is very large. The two variances therefore cannot be equivalent to one another.

## Principal Component Analysis

pca\_data <- TheData[c(1:17),c(3, 4, 6, 7, 8, 9, 10, 11, 12, 14, 16)]  
head(pca\_data)

## Sales\_in\_thousands X\_\_year\_resale\_value Price\_in\_thousands Engine\_size  
## 1 16.919 16.360 21.50 1.8  
## 2 39.384 19.875 28.40 3.2  
## 4 8.588 29.725 42.00 3.5  
## 5 20.397 22.255 23.99 1.8  
## 6 18.780 23.555 33.95 2.8  
## 7 1.380 39.000 62.00 4.2  
## Horsepower Wheelbase Width Length Curb\_weight Fuel\_efficiency  
## 1 140 101.2 67.3 172.4 2.639 28  
## 2 225 108.1 70.3 192.9 3.517 25  
## 4 210 114.6 71.4 196.6 3.850 22  
## 5 150 102.6 68.2 178.0 2.998 27  
## 6 200 108.7 76.1 192.0 3.561 22  
## 7 310 113.0 74.0 198.2 3.902 21  
## Power\_perf\_factor  
## 1 58.28015  
## 2 91.37078  
## 4 91.38978  
## 5 62.77764  
## 6 84.56511  
## 7 134.65686

Eliminated columns that are not of interest.

scaled\_pca <- scale(pca\_data)  
pca\_cor <- cor(scaled\_pca)  
is\_orthogonal\_matrix(pca\_cor)

## [1] FALSE

Generated correlation matrix for PCA.

eigen\_pca = eigen(pca\_cor)  
eigen\_value <- eigen\_pca$values  
eigen\_vectors <- eigen\_pca$vectors

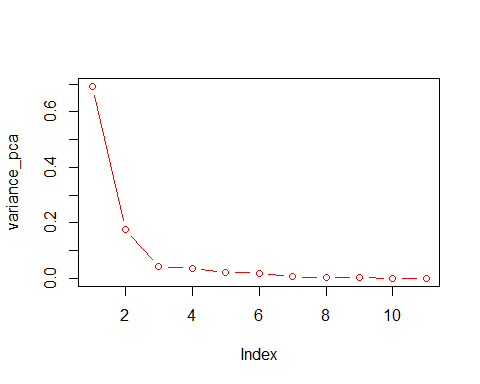
The eigen values and eigen vectors of the correlation matrix are obtained.

variance\_pca <- eigen\_value / sum(eigen\_value)  
cumsum(variance\_pca)

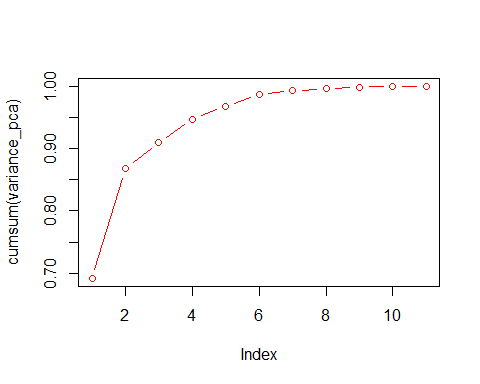
## [1] 0.6916401 0.8675291 0.9095907 0.9469400 0.9681197 0.9873364 0.9926799  
## [8] 0.9965007 0.9990179 1.0000000 1.0000000

The first two principal components are sufficient for analysis as it explains around 86% variability in the data.

plot(variance\_pca,type = "b",col="red")



plot(cumsum(variance\_pca),type = "b",col="red")



computed = eigen\_vectors[,1:2]  
colnames(computed) = c("pc1", "pc2")  
row.names(computed) = colnames(pca\_data)  
computed

## pc1 pc2  
## Sales\_in\_thousands 0.1687070 0.546833857  
## X\_\_year\_resale\_value -0.2149825 -0.527012309  
## Price\_in\_thousands -0.3109547 -0.324140646  
## Engine\_size -0.3205674 0.262738093  
## Horsepower -0.3405950 -0.026825922  
## Wheelbase -0.3014049 0.177505565  
## Width -0.2959573 0.231253348  
## Length -0.2931460 0.388335812  
## Curb\_weight -0.3473876 0.073031241  
## Fuel\_efficiency 0.3250451 0.001317556  
## Power\_perf\_factor -0.3451906 -0.085359986

The first principal component (PC1) is inversely linked with price and residual value, indicating that as a car’s size, weight, horsepower, and other characteristics decline, so too do its price and residual value.

The second principal component (PC2) is mostly linked to elements that reflect a vehicle’s fuel economy, such as fuel economy and fuel capacity. Better fuel efficiency in an automobile is indicated by higher PC2 values.

total, these principle components indicate that there are two key elements that determine a car’s total “size” and value: its weight and physical dimensions (as determined by PC1), and its fuel economy (as determined by PC2).

## Conclusion

We have analysed the data set to give insight for car buyers on important considerations. It is important to take into account variables like pricing, horsepower, and fuel efficiency.

Hyundai, Saturn, and Chevrolet vehicles are more affordable and have superior fuel efficiency than those from other companies.Given the exorbitant cost of premium brands, there aren’t many people buying expensive cars. Additionally, we see an inverse relationship between horsepower and fuel economy, meaning that as horsepower rises, fuel efficiency falls.