Special Module Assignment RMarkdown

Alka Santosh Naik

2023-09-09

# Preprocessing

1. Loading the file “6304 Module 1 Assignment Data.xlsx”
2. Taking the numerical portion of the U number as a random number seed, to a random sample of 30 cars.

rm(list=ls())  
library(rio)  
library(moments)  
data=import("6304 Module 1 Assignment Data.xlsx")  
colnames(data)=tolower(make.names(colnames(data)))  
set.seed(16999752)  
sample.data=data[sample(1:nrow(data),30),]  
sample.data

## price mileage age make  
## 94 6454 61417 11 Nissan  
## 19 6991 154711 13 Lexus  
## 50 12389 6778 2 Nissan  
## 47 12991 44220 5 Honda  
## 55 7498 13702 2 Nissan  
## 49 8271 86054 10 Honda  
## 84 16000 29530 4 Nissan  
## 44 21867 22543 3 Honda  
## 13 30000 10705 1 Lexus  
## 48 21990 33835 3 Honda  
## 40 16999 28326 4 Honda  
## 74 11177 34758 2 Nissan  
## 90 3000 161953 18 Nissan  
## 66 5983 74398 9 Nissan  
## 98 9881 73979 7 Nissan  
## 34 26995 403 0 Honda  
## 58 8690 44738 3 Nissan  
## 3 29000 12886 1 Lexus  
## 7 13940 72424 7 Lexus  
## 92 7997 93116 6 Nissan  
## 24 9995 112034 10 Lexus  
## 20 8689 126278 12 Lexus  
## 30 24891 1880 1 Honda  
## 65 22995 7638 1 Nissan  
## 54 18499 3669 1 Nissan  
## 16 27990 30218 3 Lexus  
## 76 13599 40450 3 Nissan  
## 96 20796 13461 3 Nissan  
## 26 6980 94537 6 Honda  
## 85 2995 186000 12 Nissan

# Analysis

1. The structure of the data object using the str() command.

str(sample.data)

## 'data.frame': 30 obs. of 4 variables:  
## $ price : num 6454 6991 12389 12991 7498 ...  
## $ mileage: num 61417 154711 6778 44220 13702 ...  
## $ age : num 11 13 2 5 2 10 4 3 1 3 ...  
## $ make : chr "Nissan" "Lexus" "Nissan" "Honda" ...

1. Mean, Median, Standard Deviation, Skewness, and Kurtosis of the price variable.

mean(sample.data$price)

## [1] 14518.07

median(sample.data$price)

## [1] 12690

sd(sample.data$price)

## [1] 8128.776

skewness(sample.data$price)

## [1] 0.4856733

kurtosis(sample.data$price)

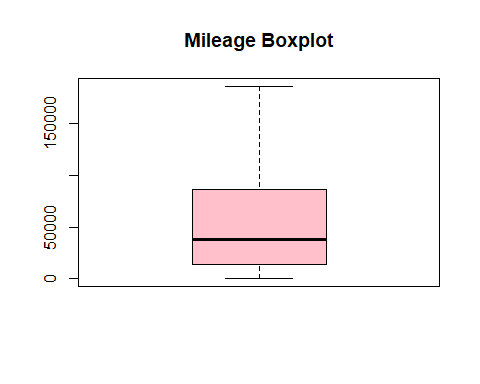
## [1] 1.989719

**Observation :**

* According to the normal distribution mean and median must be equal to each other, but here we see mean of sample data is greater then the median.
* In normal distribution skewness value is zero, skewness value for sample data is 0.4856733 which is greater than zero. Therefore, we can say the distribution is right or positively skewed.

1. A boxplot of the mileage variable.

boxplot(sample.data$mileage,col="pink",main="Mileage Boxplot",pch=19)



**Observation :**

* From the above boxplot we can say that the distribution is right or positively skewed because the median is closer to the bottom and the whisker is shorter at the lower end.

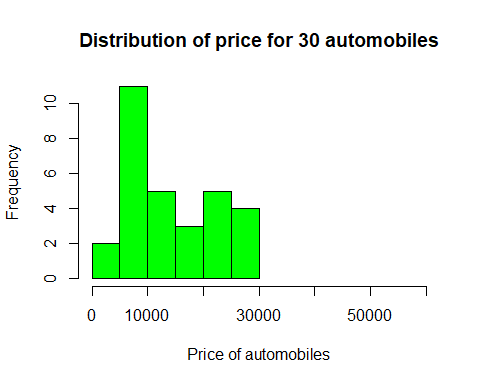
1. Quartiles of the mileage variable. Show your quartiles running from the minimum to maximum values for the variable, incrementing by .20.

quantile(sample.data$mileage,probs=seq(0,1,0.20))

## 0% 20% 40% 60% 80% 100%   
## 403.0 12449.8 29942.8 51409.6 93400.2 186000.0

1. A simple histogram of the price variable. Color your histogram green and give it an appropriate main title. Make sure the bottom axis of your histogram covers a range from 0 to $60,000. Based on this graphical tool would you say from this histogram the distribution of price follows a symmetric distribution, or a skewed distribution?

hist(sample.data$price,col="green",xlim=c(0,60000),xlab=("Price of automobiles"),main="Distribution of price for 30 automobiles")

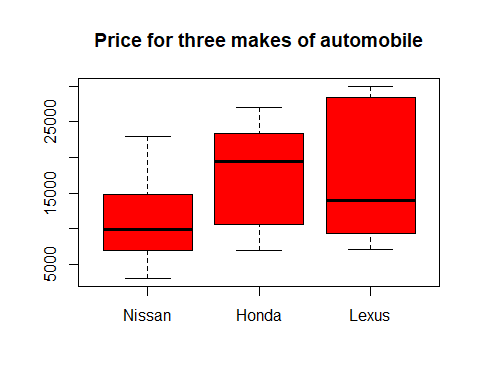


**Observations :**

* As we see most of the data is on the left side of the histogram and few data on the right side, therefore we can say the distribution is right or positively skewed.

1. Three comparative boxplots for the price variable, one for each of the three makes of automobile included in. Your boxplots should be colored red and shown side by side with an appropriate main title and labels for the makes on the bottom axis. Based on these boxplots what can you say about the similarity in price between the three makes based on your sampled data?

nissan=subset(sample.data,make=="Nissan")  
honda=subset(sample.data,make=="Honda")  
lexus=subset(sample.data,make=="Lexus")  
boxplot(nissan$price,honda$price,lexus$price,main="Price for three makes of automobile",col="red",pch=19,names=c("Nissan","Honda","Lexus"))

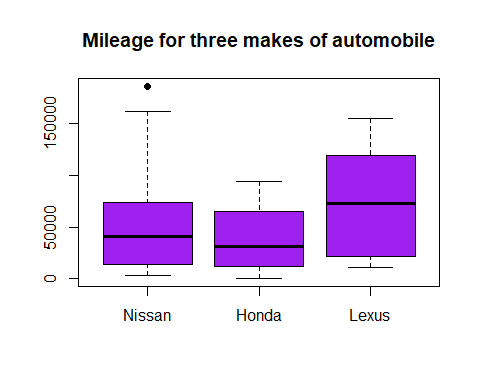


**Observations :**

* On comparing the minimum and maximum values of the 3 makes from the boxplot we can say Nissan automobile is quoted the least price less than 5000 and Lexus is quoted the heights price approaching 30000.
* Most of the automobiles from Nissan falls on the lower price range, whereas Honda automobiles falls on the middle price range.
* Automobiles from Lexus have wide distribution of price range most of them being priced higher.

1. Following Part 6 above, create three comparative boxplots for the mileage variable, one for each of the three makes of automobile included in. Your boxplots should be colored purple and shown side by side with an appropriate main title and labels for the makes on the bottom axis. Based on your boxplots for Parts 6 and 7, does there appear to be any difference in price or mileage across the three makes of automobile?

boxplot(nissan$mileage,honda$mileage,lexus$mileage,main="Mileage for three makes of automobile",col="purple",pch=19,names=c("Nissan","Honda","Lexus"))



**Observations :**

* Between Nissan and Honda we see Honda is priced higher than Nissan but Honda gives lesser mileage than Nissan.
* Comparing the minimum and maximum values we observe Nissan gives the highest mileage(including or excluding the outlier) and Honda gives the least mileage.
* The mileage distribution of Lexus is better compared to Nissan and Honda for its price distribution.