Exploring and understanding data

We use Toyota Used car dataset selected from  [http://www.biz.uiowa.edu/faculty/jledolter/DataMining/datatext.html](http://www.biz.uiowa.edu/faculty/jledolter/DataMining/dataexercises.html). The characteristics describing these Toyota used cars include price and age (in months), accumulated kilometers on the odometer (in kilometer), fuel type (there are three: petrol, diesel, and compressed natural gas CNG), horsepower, color (whether metallic = 1, or not), transmission (whether automatic = 1, or not), cylinder volume (in cubic centimeters), doors (number of), and weight (in kilograms).

Save your code and answer the following questions:

1. Import the data and check the structure of the dataset. How many cars are in this dataset?

Ans 1: This dataset has 1436 cars

Code1:

data <- read.csv(file.choose(), stringsAsFactors = F)

str(data)

2. Use summary () to summarize variables Price, Age, KM and Weight. What is the maximum price of cars? What is the median value of Age? What is the mean of Weight and KM?

Ans 2:

Max Price - 32500

Median -61

Mean of KM -68533.26

Median of KM - 1072.46

Code 2:

col <- c('Price','Age','KM','Weight')

summary(data[col])

max <-max(data$Price)

max

median(data$Age)

mean(data$KM)

mean(data$Weight)

3. Use mean () and median () separately to check the mean and median of variable horsepower (HP).

Ans 3:

Mean =101.5021

Median =110

Code 3:

mean(data$HP)

median(data$HP)

4. Use the range () function to inspect the minimum and maximum value of Toyota car price. What is the difference between the maximum and minimum prices (use diff ())?

Ans 4:

Range = 4350 32500

Diff   = 28150

Code 4:

ran <- range(data$Price)

ran

diff(ran)

5. What is the IQR for Toyota car prices?

Ans 5:

3500

Code 5:

IQR(data$Price)

 6. Use quantile to calculate five number summaries for the variable price.

Ans 6:

   0%   25%   50%   70%  100%

 4350  8450  9900 11250 32500

Code 6:

quantile(data$Price, c(0,0.25,0.50,0.70,1))

1. Create boxplots for car age and kilometer. What does the horizontal line at the bottom stands for?

Chart, box and whisker chart

Description automatically generated

Ans 7 : The horizontal line at the bottom represents the lowest actual value within 25th percentile.

Code 7:

boxplot(data$Age, main="Boxplotof Age", ylab="Age")

boxplot(data$KM, main="Boxplot of KM", ylab="KM")

1. Create histograms for Toyota car price and car weight. What price range has the highest frequency? What weight range has the highest frequency?

Ans 8:

Price Range : 5000 - 10000

Weight Range : 1000 - 1100

Code 8:

hist(data$Price, main="Spread of Prices", xlab="Prices")

hist(data$Weight, main="Spread of Car Weight", xlab="Weight")

1. What are the variance and standard deviation of variable Price?

Ans 9:

13154872

3626.965

Code 9:

var(data$Price)

sd(data$Price)

1. Use one-way table to check the variables MetColor, Doors and FuelType. How many cars are with metallic color? How many cars use CNG as their fuel type? How many cars have 5 doors?

Ans 10:

Metallic color - 969

CNG                 - 17

5 doors            - 674

Code 10:

table(data$MetColor)

table(data$Doors)

table(data$FuelType)

1. Learn to use prop.table () to inspect the proportions of cars using diesel as fuel type. What value do you have?

Ans 11:

10.793872

Code 11:

x <- prop.table(table(data$FuelType))\*100

x

12. Round the proportion of cars using diesel as fuel type, leaving two-digit number. What number do you get now?

Ans 12:

10.79

Code 12:

round(x,2)

13. Plot the relationship between kilometers and Price. If we have a car with more kilometers, with other factors fixed, what direction does the car price usually move?

Ans 13: When the kilometers are more the car price usually tend to be decreased almost merging the x -axis

Code 13:

plot(data$KM , data$Price,,xlab="KM",ylab="Price", main = "relation btwn KM and Price")

14. Use the Cross-table to see how many cars with automatic transmission use diesel as fuel type? How many cars with 5 doors have automatic transmission?

Ans 14:

O

30

Code 14:

CrossTable(data$Automatic,data$FuelType)

CrossTable(data$Automatic,data$Doors)