 Exploring and understanding data

We use House Price dataset selected from <http://www.biz.uiowa.edu/faculty/jledolter/DataMining/dataexercises.html>. The HousePrices.csv data set includes prices and characteristics of houses in a major US metropolitan area. The variables include Price (sale price in dollars), SqFt (size in square feet), Bedrooms (number of), Bathrooms (number of), Offers (number of offers the house has received while on the market), Brick (whether it is brick construction; Yes/No), and Neighborhood (East/North/West).

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

Ans 1: The dataset has 128 observations.

Code 1:

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

str(data)

2. Use summary () to summarize variables Price, SqFt and Offers. What is the maximum price of houses? What is the median value of offers? What is the mean of SqFt?

Ans 2:

Max Price              --   211200

Median of offers  --   3

Mean of SqFt        --   2000.938

Code 2:

col <- c('Price','SqFt','Offers')

summary(data[col])

max <-max(data$Price)

max

median(data$Offers)

mean(data$SqFt)

1. Use mean () and median () separately to check the mean and median of bedrooms.

Ans 3:

Mean - 3.023438

Median – 3

Code 3:

mean(data$Bedrooms)

median(data$Bedrooms)

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

Ans 4:

Range = 69100 211200

Difference = 142100

Code 4:

ran <- range(data$Price)

ran

diff(ran)

1. What is the IQR for house prices?

Ans 5:

36925

Code 5:

IQR(data$Price)

1. Use quantile function to find percentiles at 0, 20, 40, 60, 80, 100 for house prices.

Ans 6:

    0%    20%    40%    60%    80%   100%

 69100 107780 117800 134200 151780 211200

Code 6:

quantile(data$Price, c(0,0.2,0.4,0.6,0.8,1))

7. Create boxplots for house price and SqFt. What does the horizontal line in the middle stands for?

Chart, box and whisker chart

Description automatically generated

Ans 7:  The horizontal line represents the median

Code 7:

boxplot(data$Price, main="Plot of House Price", ylab="house Price")

boxplot(data$SqFt, main="Plot of SqFt", ylab="house SqFt")

8. Create histograms for house price and house offers. What price range has the highest frequency? What offer range has the highest frequency?

Ans 8:

Price Range   : 100000 - 150000

Offers Range : 2-3

Code 8:  
hist(data$Price, main="Spread of Prices", xlab="Prices")  
hist(data$Offers, main="Spread of offers", xlab="Offers")

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

Ans 9:

[1] 721930821

[1] 26868.77

Code 9:

var(data$Price)

sd(data$Price)

 10. Use one-way table to check the variables Brick and Neighborhood. How many houses are built with bricks? How many houses are in the West side of the neighborhood?

Ans 10 :

42

39

Code 10 :

table(data$Brick)

table(data$Neighborhood)

11. Learn to use prop.table () to inspect the proportions of houses in the East side of the neighborhood. What is the proportion of houses in the East side?

Ans 11:

35.15625

Code 11:

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

x

12. Round the proportion of houses with bricks, leaving two-digit number. What number do you get?

Ans 12:

32.81

Code 12:

round(prop.table(table(data$Brick))\*100,2)

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

Ans 13:

When we have a larger house with more SqFt , the house price tend to move upwards which in-turn causes increase in the prices of the houses.

Code 13:

plot(data$SqFt , data$Price,xlab="Sqft",ylab="housePrice", main = "relation btwn house and SqFt")

1. Use the Cross table to see how many houses built with bricks are in the North neighborhood?

Ans 14:

7

Code 14:

CrossTable(data$Neighborhood,data$Brick)