**DATA SCIENCE CERTIFICATION**

**Assignment** – Extrapolate Data Analysis of House Data Price



**Step 1:** From the following commands given sample data converts into Data Frame in R, this can be used for processing.

houseData <- read.csv("D:/DataScientist/Data Science Learner/Datascience/Datsets/houseprices.csv")

**Step 2**: Need to verify the loaded data, by basic retrievals

> head(houseData,10)

Price Living.Area Bathrooms Bedrooms Lot.Size Age Fireplace

1 142212 1982 1.0 3 2.00 133 0

2 134865 1676 1.5 3 0.38 14 1

3 118007 1694 2.0 3 0.96 15 1

4 138297 1800 1.0 2 0.48 49 1

5 129470 2088 1.0 3 1.84 29 1

6 206512 1456 2.0 3 0.98 10 0

7 50709 960 1.5 2 0.01 12 0

8 108794 1464 1.0 2 0.11 87 0

9 68353 1216 1.0 2 0.61 101 0

10 123266 1632 1.5 3 0.23 14 0

> colMeans(houseData)

Price Living.Area Bathrooms Bedrooms Lot.Size Age Fireplace

1.638621e+05 1.807303e+03 1.918338e+00 3.183381e+00 5.695798e-01 2.806113e+01 5.931232e-01

Step 3 : We can understand the data by the preliminary analysis.

**Analysis 1**

Nature of data

> summary(houseData)

Price Living.Area Bathrooms Bedrooms

Min. : 16858 Min. : 672 Min. :1.000 Min. :1.000

1st Qu.:112014 1st Qu.:1336 1st Qu.:1.500 1st Qu.:3.000

Median :151917 Median :1672 Median :2.000 Median :3.000

Mean :163862 Mean :1807 Mean :1.918 Mean :3.183

3rd Qu.:205235 3rd Qu.:2206 3rd Qu.:2.500 3rd Qu.:4.000

Max. :446436 Max. :4534 Max. :4.500 Max. :6.000

Lot.Size Age Fireplace

Min. :0.0000 Min. : 0.00 Min. :0.0000

1st Qu.:0.2100 1st Qu.: 6.00 1st Qu.:0.0000

Median :0.3900 Median : 18.00 Median :1.0000

Mean :0.5696 Mean : 28.06 Mean :0.5931

3rd Qu.:0.6000 3rd Qu.: 34.00 3rd Qu.:1.0000

Max. :9.0000 Max. :247.00 Max. :1.0000

From the above statics we can understand:

**Price:**

1. There are gap between minimum value and maximum value data from rest of the data
2. There is no major gap between mean and median – means data value are in continues series

**Living.Area:**

1. There are gap between minimum value and maximum value data from rest of the data
2. There is no major gap between mean and median – means data value are in continues series

**Bathrooms:**

1. There are no major gap between minimum value and maximum value data value
2. There is no major gap between mean and median – means data value are in continues series

**Bedrooms:**

1. There are no major gap between minimum value and maximum value data value
2. There is no major gap between mean and median – means data value are in continues series

**Lot.Size:**

1. There are gap between minimum value and maximum value data value from rest of the data
2. There are major gap between mean and median – means data value are not in continues series

**Age:**

1. There are large gap between maximum value data value from rest of the data
2. There are major gap between mean and median – means data value are not in continues series

**Fireplace:**

1. The range of data minimal (It contain only boolean data value)

**Analysis 2**

We can determine the outlier data

par(mfrow=c(6,1))

boxplot(houseData$Price, main="Price", horizontal = TRUE)

boxplot(houseData$Living.Area, main="Living.Area", horizontal = TRUE)

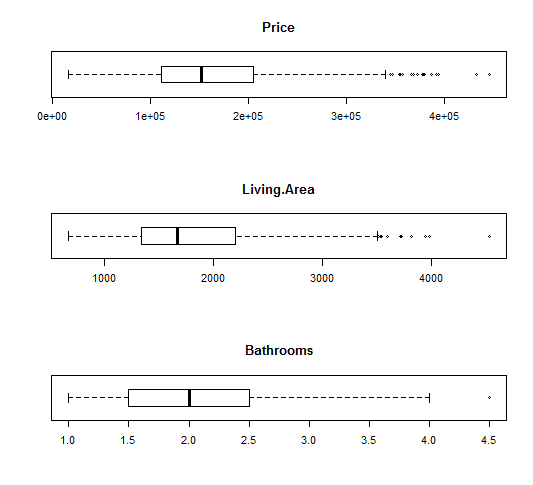
boxplot(houseData$Bathrooms, main="Bathrooms", horizontal = TRUE)

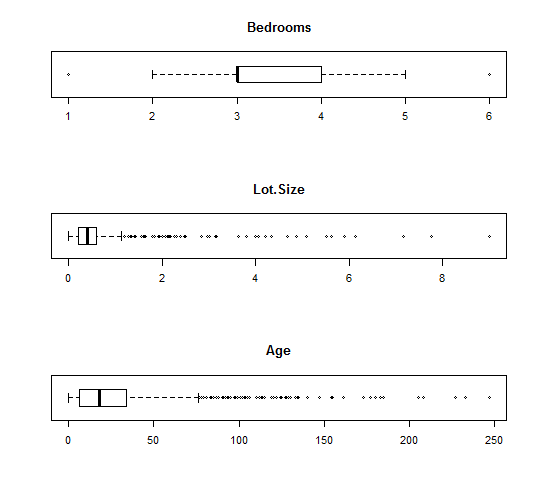
boxplot(houseData$Bedrooms, main="Bedrooms", horizontal = TRUE)

boxplot(houseData$Lot.Size, main="Lot.Size", horizontal = TRUE)

boxplot(houseData$Age, main="Age", horizontal = TRUE)

par(mfrow=c(1,1))





**Price:**

1. There are few data which are outliers in the maximum value
2. The price serious is not so continues for the given sample

**Living.Area:**

1. There are few data which are outliers in the maximum value
2. Most of the sample data are continues

**Bathrooms:**

1. There are very few data which are outliers in the maximum value
2. Most of the data point in align with Mean

**Bedrooms:**

1. There are very few data which are outliers in both minimum and maximum value
2. Most of the data point in align with Mean and this sample data serious is continues

**Lot.Size:**

1. There are more data points which are outliers in maximum value
2. The sample data serious is not continues or this data set may need validation / correction

**Age:**

1. There are more data points which are outliers in maximum value
2. The sample data serious is not continues or this data set may need validation / correction

**Analysis 3**

We can determine correlation relationship within the data points

> cor(houseData)

Price Living.Area Bathrooms Bedrooms Lot.Size

Price 1.0000000 0.7763965 0.6701894 0.4710743 0.15528413

Living.Area 0.7763965 1.0000000 0.7226488 0.6643470 0.20018004

Bathrooms 0.6701894 0.7226488 1.0000000 0.4917984 0.10099313

Bedrooms 0.4710743 0.6643470 0.4917984 1.0000000 0.14068157

Lot.Size 0.1552841 0.2001800 0.1009931 0.1406816 1.00000000

Age -0.3633536 -0.2631685 -0.4438296 -0.0605977 0.01513502

Fireplace 0.4602370 0.4814361 0.4446997 0.2958730 0.05276519

Age Fireplace

Price -0.36335361 0.46023700

Living.Area -0.26316849 0.48143613

Bathrooms -0.44382957 0.44469967

Bedrooms -0.06059770 0.29587297

Lot.Size 0.01513502 0.05276519

Age 1.00000000 -0.24879438

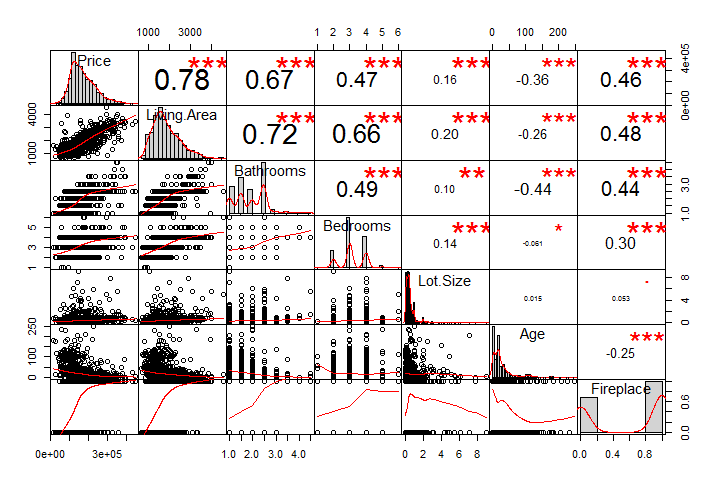
Fireplace -0.24879438 1.00000000

1. Except the Age all other data are positively correlated with price. It show most the data are valid. ie. If the age of the building increases then price decrease.
2. Data for Living area and Bathrooms are highly correlated, means Price is directly proportional to this parameter
3. Lot size is least correlated. So, price is not related with the Lot.size

**Analysis 4**

We can determine more from correlation analysis chart

> chart.Correlation(houseData, histogram=TRUE, pch=19)

Pr

1. Price, Living Area are normally distributed and lean towards the left side.
2. Price and Living Area are tightly correlated with Fireplace
3. Price increase based upon Living area, Bathrooms and Bedrooms
4. There are positive relationship between Bedroom count and Bathroom count
5. Price is negatively correlated with Age

**Analysis 5**

We can determine relationship with Price and Living Area with respect to Bedroom count

par(mfrow=c(3,2))

for( i in 1:length(unique(datafile1$Bedrooms))){

plot(datafile1$Living.Area[datafile1$Bedrooms==i],

datafile1$Price[datafile1$Bedrooms==i],

col=datafile1$Bedrooms[i],

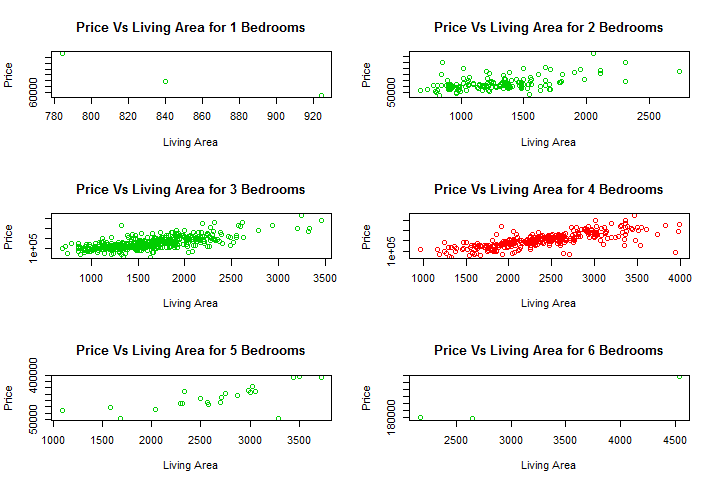
main=paste0("Price Vs Living Area for ",i," Bedrooms"),

xlab="Living Area",

ylab="Price")

}

par(mfrow=c(1,1))



1. More data points available for 2,3,4 and 5 Bedrooms
2. Price is higher end for 1 Bedroom and 6 Bed rooms living area when compare with other data
3. From the given sample 3 bedroom price (with respect to Living area)seems cheaper when compare with others.