University of Moratuwa
Department of Computer Science & Engineering
MBA in Information Technology - 2018

Name & Index	: D .N.K. Medawatta - 189115D

Title of Assignment	: Exercise 6.3 – House Price Prediction		
Assignment No	:	Group □	Individual ■

Subject Code	: CS5122
Subject	: Descriptive Predictive Analytics

Lecturer	: Dr. Uthayashanker Thayasivam

Student's Statement :

We certify that we have not plagiarized the work of others or participated in unauthorized collusion when preparing this assignment

Office use only :

On/ before deadline Extension Given Late Submission

Marks Given :

Question 01

List 4 question that you may want to explore from the dataset

- How many distinct records are there in the dataset?
- What columns contain outliers?
- What is the average square feet of a house?
- What is the average market value of a house?
- What is the average age of a house?
- What is the relationship between the area and the market value of a house?
- What is the relationship between the age and the market of a house?
- What is the most frequent age of houses?

Question 02

By analyzing statistical properties of data (e.g mean, std, min, correlation , etc.) and Visualization what you can claim about the data set?

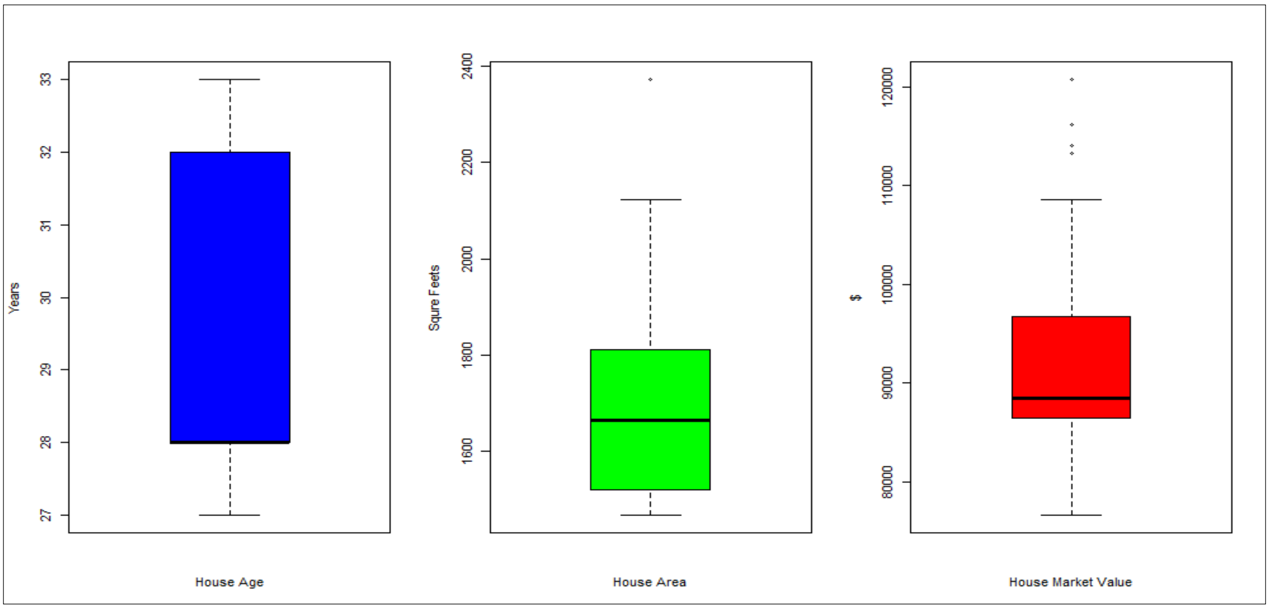
Mean Median and Mode or Original Data Set

```
# Read house price data
houseData = read.csv("Home Market Value.csv",header = TRUE)
# Remove $ marks
houseData$Market.Value = as.numeric(gsub("[\\$,]","",
houseData$Market.Value))
# Remove commas
houseData$Square.Feet = as.numeric(gsub("[\\,]","",
houseData$Square.Feet))
# View Summary
summary(houseData)
```

```
  House.Age      Square.Feet    Market.Value
 Min.   :27.00   Min.   :1468   Min.   : 76600
 1st Qu.:28.00   1st Qu.:1520   1st Qu.: 86575
 Median :28.00   Median :1666   Median : 88500
 Mean   :29.83   Mean   :1695   Mean   : 92069
 3rd Qu.:32.00   3rd Qu.:1807   3rd Qu.: 96525
 Max.   :33.00   Max.   :2372   Max.   :120700
```

Finding Outliers of Original Data Set

```
# Finding Outliers
par(mfrow=c(1,3))
boxplot(dataSet$House.Age, xlab="House Age", ylab="Years", col =
"Blue")
boxplot(dataSet$Square.Feet, xlab="House Area", ylab="Squre
Feets", col = "Green")
boxplot(dataSet$Market.Value, xlab="House Market Value",
ylab="$", col = "Red")
par(mfrow=c(1,1))
```



- Market values of houses are left skewed
- Area of houses also has outliers

However it’s clear that the original data set contains some outliers. Above box plot on House Market Value and House Area shows some outliers. Therfore, for further analysis it is required to eliminate outliers.

Removing Outliers

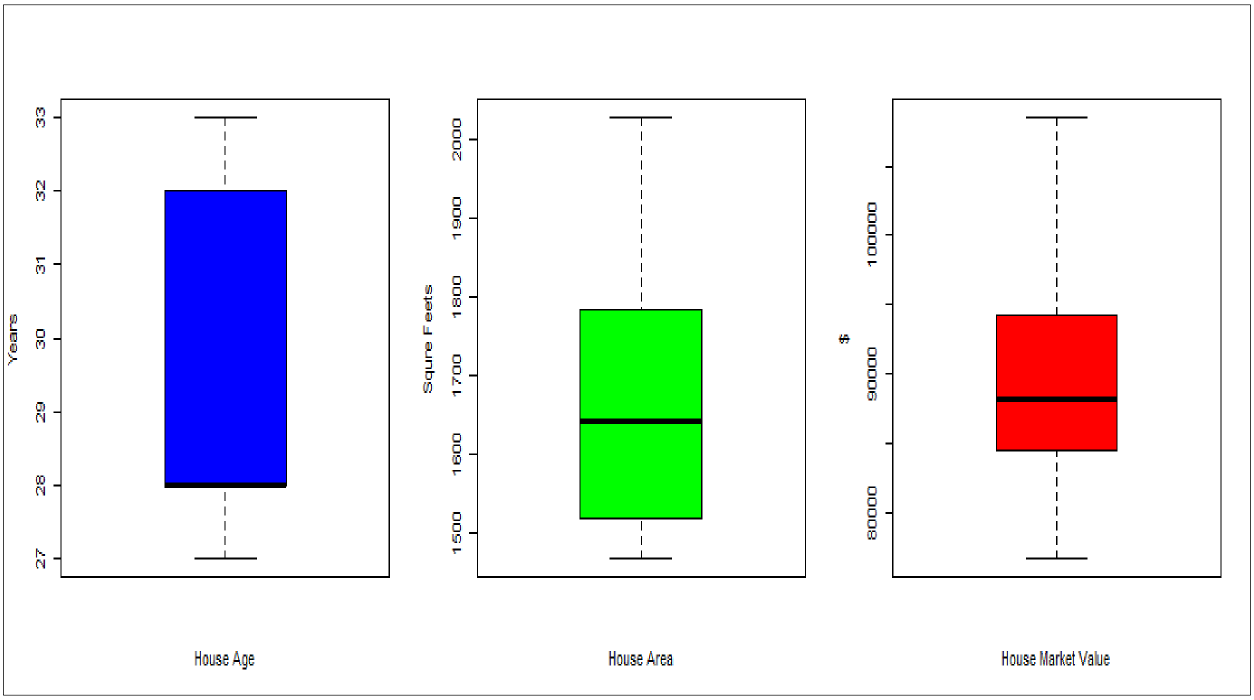
```
nrow(dataSet)
dataSet %>% distinct(dataSet$Square.Feet, dataSet$Market.Value, 
dataSet$House.Age, .keep_all = TRUE)
nrow(dataSet)
houseData <- subset(dataSet, dataSet$Market.Value < 110000)
nrow(houseData)
houseData <- subset(houseData, houseData$Square.Feet < 2200)
nrow(houseData)
```

```
> nrow(dataSet)
[1] 42
> houseData <- subset(dataSet, dataSet$Market.Value < 110000)
> nrow(houseData)
[1] 38
> houseData <- subset(houseData, houseData$Square.Feet < 2200)
> nrow(houseData)
[1] 38
> nrow(dataSet)
[1] 42
> dataSet %>% distinct(dataSet$Square.Feet, 
dataSet$Market.Value, dataSet$House.Age, .keep_all = TRUE)
   House.Age Square.Feet Market.Value dataSet$Square.Feet 
dataSet$Market.Value dataSet$House.Age
…

> nrow(dataSet)
```

```
[1] 42
> houseData <- subset(dataSet, dataSet$Market.Value < 110000)
> nrow(houseData)
[1] 38
> houseData <- subset(houseData, houseData$Square.Feet < 2200)
> nrow(houseData)
[1] 38
```

Check If Outliers Exist

```
# Check if outliers exits
par(mfrow=c(1,3))
boxplot(houseData$House.Age, xlab="House Age", ylab="Years", 
col = "Blue")
boxplot(houseData$Square.Feet, xlab="House Area", ylab="Squre 
Feets", col = "Green")
boxplot(houseData$Market.Value, xlab="House Market Value", 
ylab="$", col = "Red")
par(mfrow=c(1,1))
```



- Skewedness of market values of houses and house areas are reduced due to the elimination of outliers

Find the Mode

```
findMode <- function(x){
  uniqx <- unique(x)
  uniqx[which.max(tabulate(match(x,uniqx)))]
}
```

```
> findMode(houseData$House.Age)
```

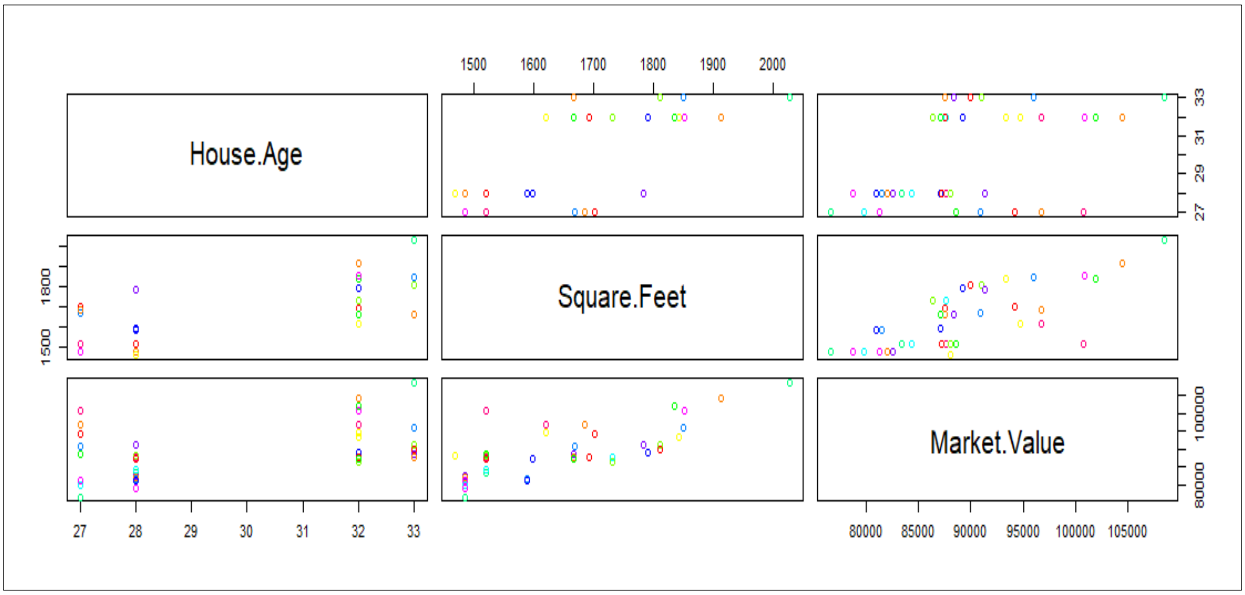
```
[1] 28
> findMode(houseData$Square.Feet)
[1] 1520
> findMode(houseData$Market.Value)
[1] 87600
```

```
# Check corelations
cor(houseData)
```

```
> cor(houseData)
             House.Age Square.Feet Market.Value
House.Age    1.0000000   0.7392243    0.4624017
Square.Feet  0.7392243   1.0000000    0.7316613
Market.Value 0.4624017   0.7316613    1.0000000
```

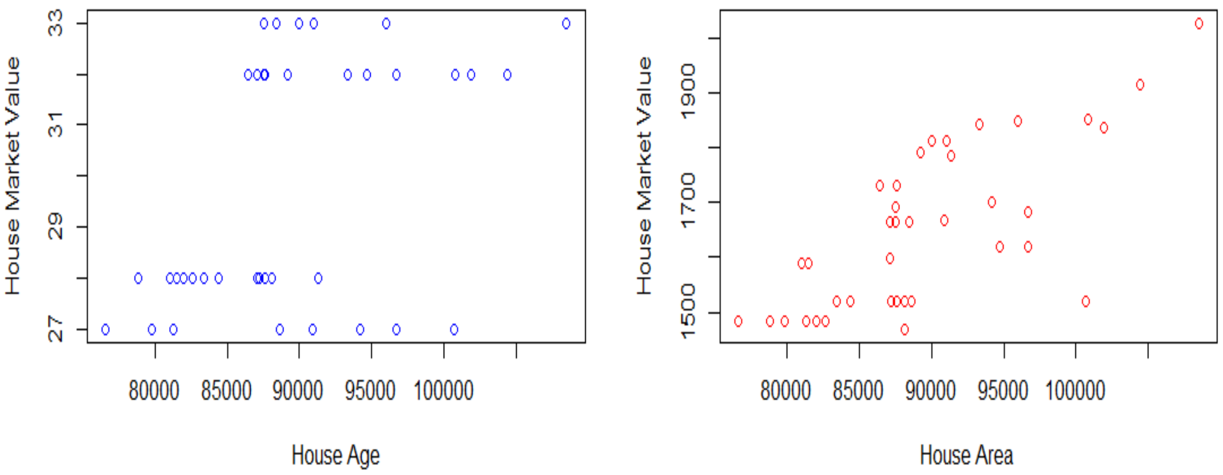
- Market value of a house has a strong positive correlation with the house area

```
# Plot the house data
plot(houseData)
```



- Above resulted plot of the house dataset reveals that there is a strong correlation between the House Area and the Market Value

```
par(mfrow=c(1,2))

plot(houseData$Market.Value,houseData$House.Age, ylab = "House
Market Value", xlab="House Age", col = "Blue")

plot(houseData$Market.Value,houseData$Square.Feet, ylab = "House
Market Value", xlab="House Area", col = "Red")

par(mfrow=c(1,1))
```



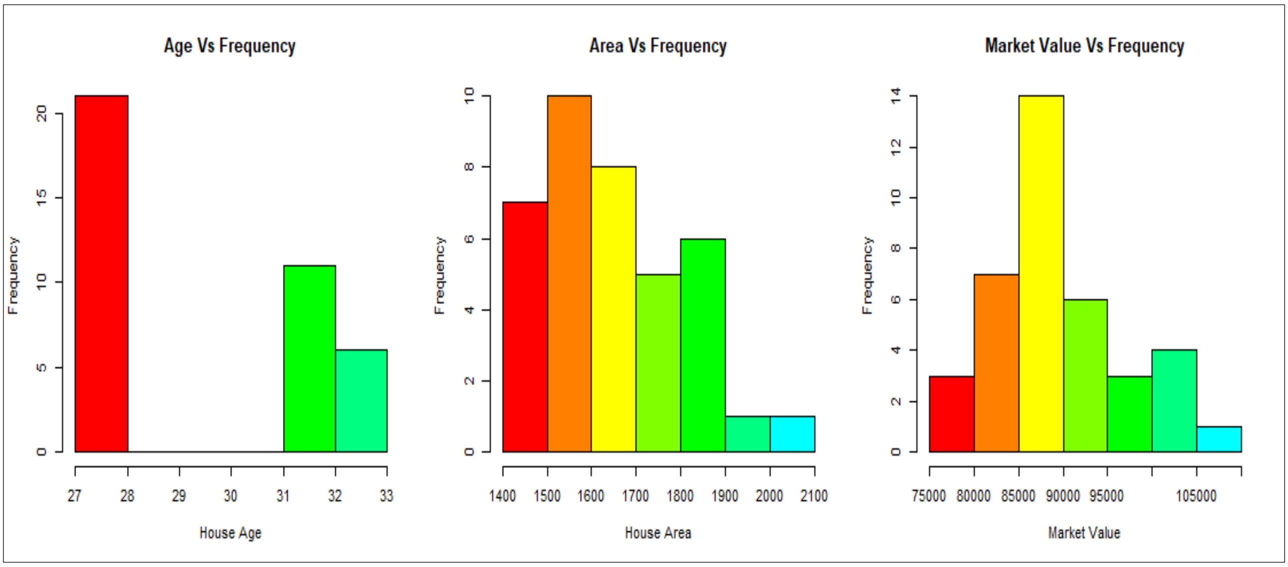
Standard Deviation

```
# Find standard deviation
sd(houseData$House.Age)
sd(houseData$Square.Feet)
sd(houseData$Market.Value)
```

```
> sd(houseData$House.Age)
[1] 2.434928
> sd(houseData$Square.Feet)
[1] 147.405
> sd(houseData$Market.Value)
[1] 7352.578
```

Distribution of Age, Area and Market Value

```
par(mfrow=c(1,3))

hist(houseData$House.Age, ylab="Frequency", xlab = "House Age",
main = "Age Vs Frequency", col=rainbow(12))

hist(houseData$Square.Feet, ylab="Frequency", xlab = "House
Area", main = "Area Vs Frequency", col=rainbow(12))

hist(houseData$Market.Value, ylab="Frequency", xlab = "Market
Value", main = "Market Value Vs Frequency", col=rainbow(12))

par(mfrow=c(1,1))
```



- Houses are either below 28 years or above 31 years
- Majority of the houses have comparatively less amount of square feet.
- There is very limited number of houses which has large amount of square feet of area.
- Prices of houses are normally distributed, however prices are left skewed.
- Highest number of houses’ market value is in between $85000 and $90000

Coefficient of Variance of age and area of the house

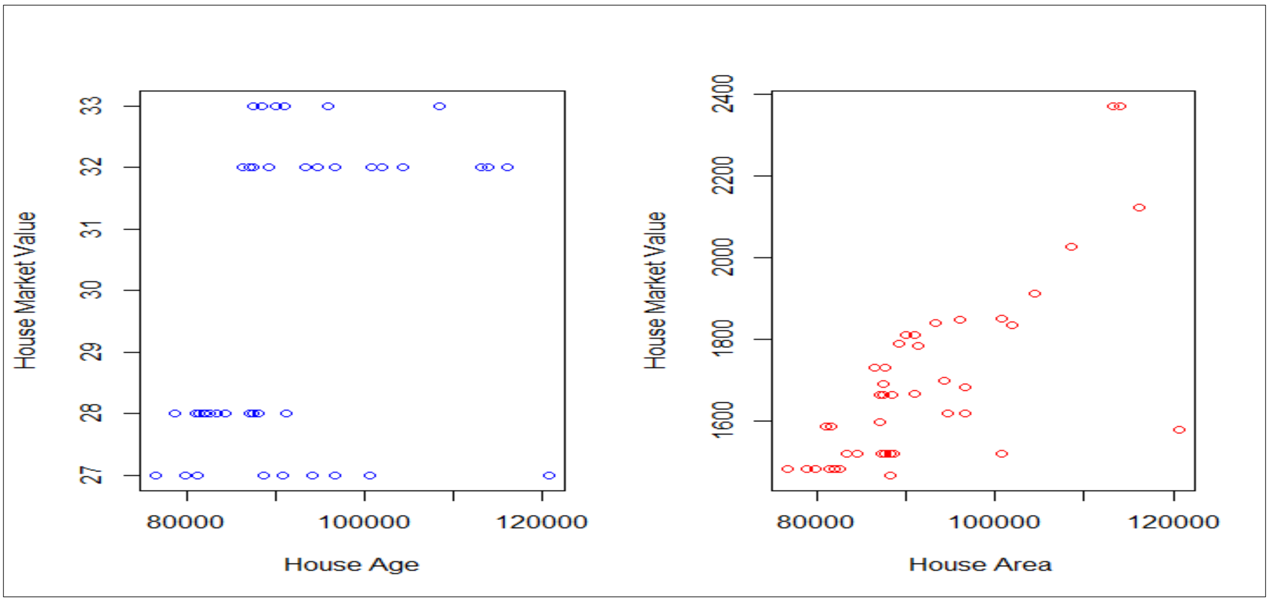
```
mean_age <- mean(houseData$House.Age)
sd_age <- sd(houseData$House.Age)
mean_area <- mean(houseData$Square.Feet)
sd_area <- sd(houseData$Square.Feet)

cov_age = sd_age/mean_age
cov_area = sd_area/mean_area
```

```
> cov(houseData$Square.Feet,houseData$Market.Value)
[1] 792979.7
> cov(houseData$House.Age,houseData$Market.Value)
[1] 8278.378
```

- It is clear that the Coefficient of Variation of age of the house is greater than to the Coefficient of Variation of the area of the house. That means compared to the area of the house, can make a higher impact on ditermining the market value of the house

Measures of Association

```
# Covariance between each Age and Feet
cov(houseData$Square.Feet,houseData$Market.Value)
cov(houseData$House.Age,houseData$Market.Value)

par(mfrow=c(1,2))

plot(houseData$Market.Value,houseData$House.Age, ylab = "House
Market Value", xlab="House Age", col = "Blue")

plot(houseData$Market.Value,houseData$Square.Feet, ylab = "House
Market Value", xlab="House Area", col = "Red")

par(mfrow=c(1,1))
```



Question 03

What regression analysis technique, that is suitable to predict the market value, given the age of a house and square feet? Justify.

This can be predicted using Linear Regression.
Both Simple and Multiple Linear Regression is suitable for predicting the market value of the given houses.

House market values can be predicted based on either area or the age of the house. In such cases simple linear regression can be used.

However in the case of predicting the house market value based on both area and the age of the house, then the multiple linear regressions can be used.

Therefore, here both of them will be described as below.

Question 04

Approach 1: Simple Linear Regression – Area Vs Market Value

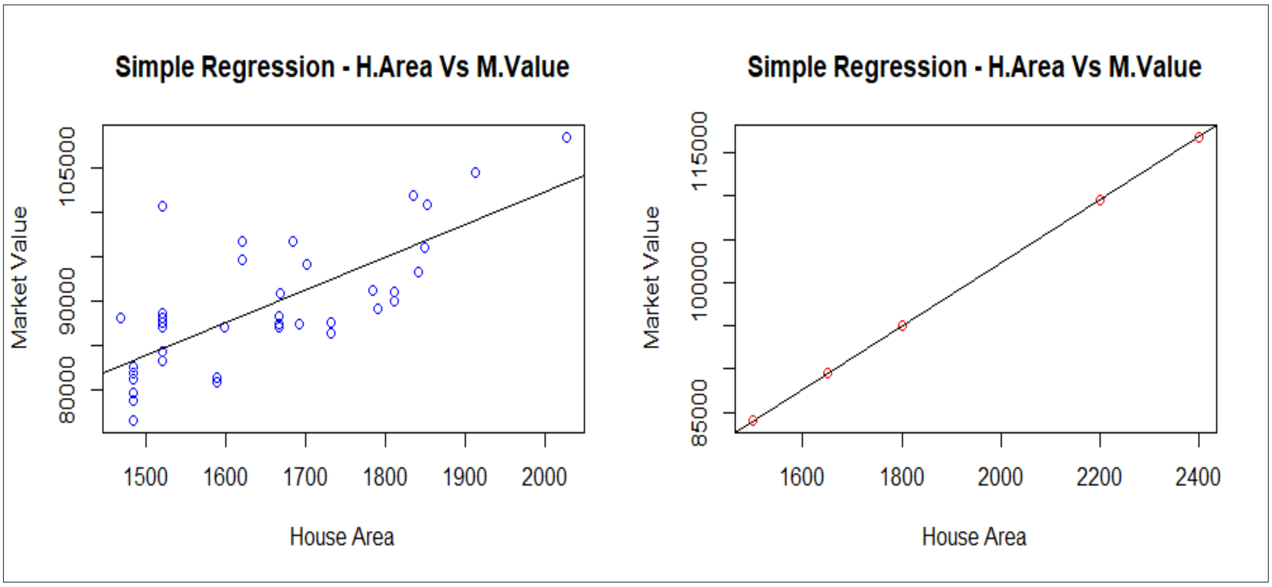
```
# Linear Regression Analysis
#=======================================================
price= houseData$Market.Value
age= houseData$House.Age
area = houseData$Square.Feet

sd(area, na.rm = FALSE)
sd(age, na.rm = FALSE)

newHoseAreas = c(1650,1500,1800,2200,2400)
newHouseAge = c(26,28,29,30,31)
```

```
# Simple Linear Regression Analysis | x=area, y=market value
#=======================================================

slm_1.houseData = lm(price~area)
summary(slm_1.houseData)

newHoseData = data.frame(area=newHoseAreas, age = newHouseAge)
predictedHousePrices = predict(slm_1.houseData, newHoseData,
level = 0.95, interval = "confidence")

predictedHousePrices
```

```
> summary(slm_1.houseData)

Call:
lm(formula = price ~ area)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
 -6841  -3614  -1098   3720  15945 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29281.786   9394.597   3.117  0.00358
area           36.495      5.667   6.440  1.8e-07

(Intercept) **
area        ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5081 on 36 degrees of freedom
Multiple R-squared:  0.5353,  Adjusted R-squared:  0.5224 
F-statistic: 41.47 on 1 and 36 DF,  p-value: 1.802e-07

> predictedHousePrices
       fit       lwr       upr
1  89499.1  87827.32  91170.88
2  84024.8  81611.88  86437.71
3  94973.4  92583.50  97363.30
4 109571.5 103048.51 116094.56
5 116870.6 108105.91 125635.30
```

```
par(mfrow=c(1,2))

plot(area,price,col="Blue", xlab = "House Area", ylab = "Market
Value", main = "Simple Regression - H.Area Vs M.Value")

abline(slm_1.houseData)

plot(newHoseAreas,predictedHousePrices[,1],col="Red", xlab =
"House Area", ylab = "Market Value", main = "Simple Regression -
H.Area Vs M.Value")

abline(slm_1.houseData)

par(mfrow=c(1,1))
```



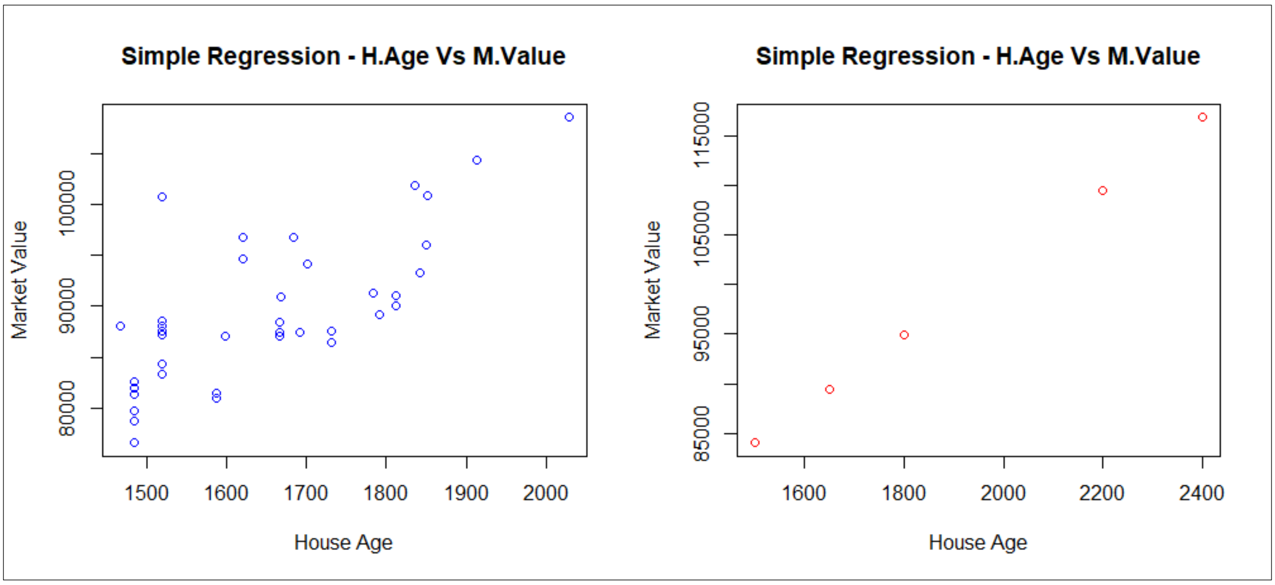
Approach 2: Simple Linear Regression – Age Vs Market Value

```
# Simple Linear Regression Analysis | x=age, y=market value
#==============================================================
slm_2.houseData = lm(price~age)
summary(slm_2.houseData)

newHoseData = data.frame(area=newHoseAreas, age = newHouseAge)
predictedHousePrices = predict(slm_1.houseData, newHoseData,
level = 0.95, interval = "confidence")

predictedHousePrices
```

```
> summary(slm_2.houseData)

Call:
lm(formula = price ~ age)

Residuals:
   Min      1Q Median      3Q     Max
 -9129  -5189  -1375    3710   14971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48029.0    13312.9   3.608  0.00093 ***
age           1396.3      446.2   3.129  0.00347 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6609 on 36 degrees of freedom
```

```
Multiple R-squared:  0.2138,   Adjusted R-squared:  0.192
F-statistic: 9.791 on 1 and 36 DF,  p-value: 0.003468

> predictedHousePrices
       fit        lwr       upr
1  89499.1  87827.32  91170.88
2  84024.8  81611.88  86437.71
3  94973.4  92583.50  97363.30
4 109571.5 103048.51 116094.56
5 116870.6 108105.91 125635.30
```

```
par(mfrow=c(1,2))

plot(area,price,col="Blue", xlab = "House Age", ylab = "Market
Value", main = "Simple Regression - H.Age Vs M.Value")

abline(slm_2.houseData)

plot(newHoseAreas,predictedHousePrices[,1],col="Red", xlab =
"House Age", ylab = "Market Value", main = "Simple Regression -
H.Age Vs M.Value")

abline(slm_2.houseData)

par(mfrow=c(1,1))
```

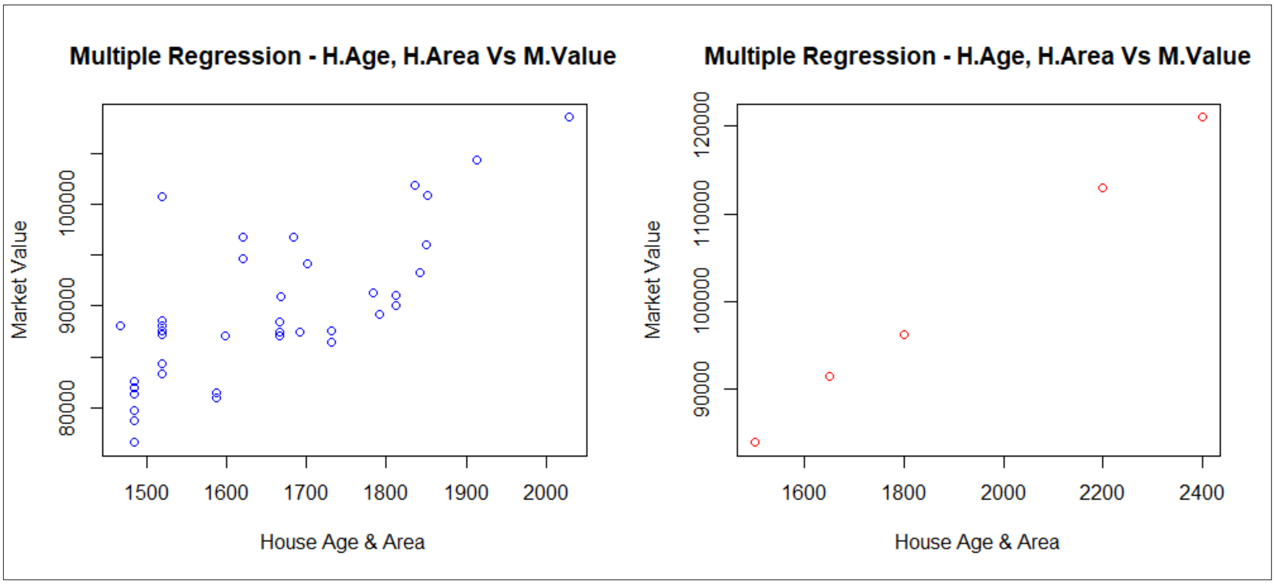


Approach 3: Multiple Linear Regression – Age, Area Vs Market Value

```
# Multiple Linear Regression Analysis
#                | x1=age, x2=area, y=market value
#=============================================================
lm_3.houseData = lm(price~age+area)
summary(slm_2.houseData)

newHoseData = data.frame(area=newHoseAreas, age = newHouseAge)
predictedHousePrices = predict(lm_3.houseData, newHoseData, level
= 0.95, interval = "confidence")

predictedHousePrices
```

```
> summary(slm_2.houseData)

Call:
lm(formula = price ~ age)

Residuals:
   Min     1Q Median     3Q    Max 
 -9129  -5189  -1375   3710  14971 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  48029.0    13312.9   3.608  0.00093 ***
age           1396.3      446.2   3.129  0.00347 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6609 on 36 degrees of freedom
Multiple R-squared:  0.2138,  Adjusted R-squared:  0.192 
F-statistic: 9.791 on 1 and 36 DF,  p-value: 0.003468

> predictedHousePrices
        fit       lwr       upr
1  91442.22  87250.25  95634.20
2  83966.38  81550.08  86382.68
3  96306.21  92747.22  99865.20
4 112933.44 103616.59 122250.28
5 120985.87 109021.92 132949.81
```

```
par(mfrow=c(1,2))

plot(area,price,col="Blue", xlab = "House Age & Area", ylab =
"Market Value", main = "Multiple Regression - H.Age, H.Area Vs
M.Value")

abline(lm_3.houseData)

plot(newHoseAreas,predictedHousePrices[,1],col="Red", xlab =
"House Age & Area", ylab = "Market Value", main = "Multiple
Regression - H.Age, H.Area Vs M.Value")

abline(lm_3.houseData)

par(mfrow=c(1,1))
```


Multiple Regression - H.Age, H.Area Vs M.Value
Market Value
House Age & Area
Multiple Regression - H.Age, H.Area Vs M.Value
Market Value
House Age & Area