Project #1: Predicting Housing Prices with Linear Regression
Majdal Hindi, 300991890

Introduction

The source of dataset is California Housing Prices | Kaggle . This dataset contains collected information about the houses and their prices in California. This data was used to determine which factors affect the price of the property, such as location, number of rooms, season of the year, area, etc. In this project, I will focus on studying the relationship between the area of the house and its price.

Location and Distribution Metrics

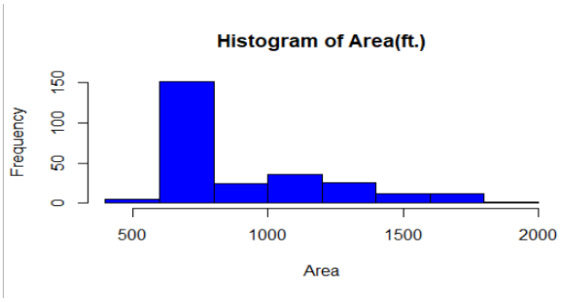
The dataset contains 267 observations. Area in feet is the **predictor variable** (X), Price in $ is the **response variable** (Y). We can see from the tables below, the first observations of the dataset, and some descriptive statistics of the studied variables (Area vs. Price). The mean of the houses area is 936.2 feet, median =50% quantile =789.3 feet, mode 794.52 feet, and the (25%,75%) quantile (756.213 feet, 1121.9 feet). On the other hand, the mean of the houses price 281k $, median =50% quantile =249k $, mode 460k $ and the (25%,75%) quantile = (217.55k $,327k $). We notice that the value of the metrics is not equal, because of the asymmetric distribution as we will see in next section.

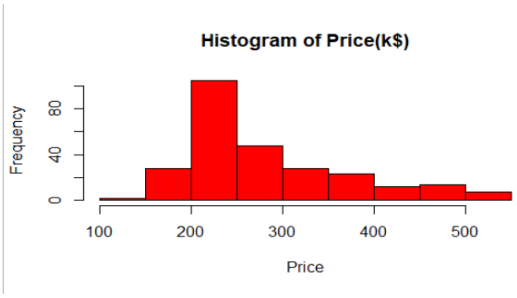
```
> head(data)
       Area    Price
1  743.0856 246172.7
2  756.2128 246331.9
3  587.2808 209280.9
4 1604.7464 452667.0
5 1375.4508 467083.3
6  675.1900 203491.8
```

```
> summary(data)
      Area            Price
 Min.   : 410.7   Min.   :117564
 1st Qu.: 756.2   1st Qu.:217553
 Median : 798.3   Median :249076
 Mean   : 936.2   Mean   :281172
 3rd Qu.:1121.9   3rd Qu.:326965
 Max.   :1942.5   Max.   :538272
```

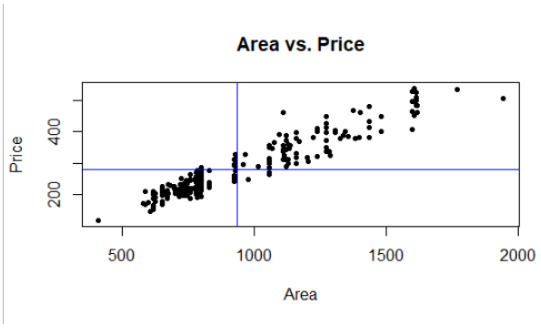
Data Visualization

From the histograms below, we can see that most of the houses have area between 600-800 feet, and the prices between 200-250 thousand $. Both of the histograms are skewed right. For a skewed distribution, however, there is no "center" in the usual sense of the word, so the mean, median and the mode (less recommended in this case) together reflect the different aspect of "centerness". The Standard deviation of x = 284.895 feet and y= 89.11912k $ (far from the mean).





The scatter graph shows the relationship between the area of the house and its price. The correlation between the area and the price is ***0.9510874***. That means strong positive relationship between the two variables, they move together in the same direction. The ablines (vertical and horizontal) cut the scatter graph on the mean, so we can see the observations above/under the mean.

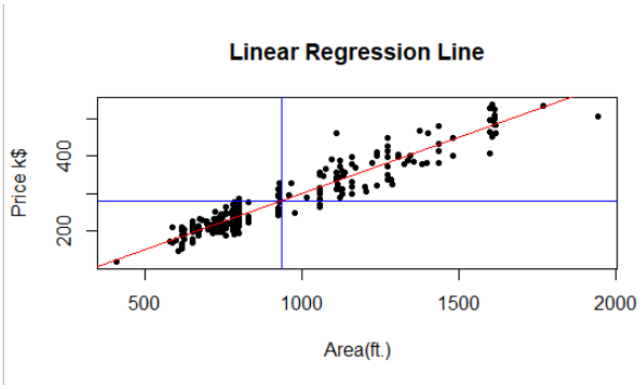


Linear Regression

Linear regression deals with the relationship between two variables and how one variable can explain or predict its value of another variable using the following equation: y *=ax+ b+ ε*

a: slope = 297.513
b: intercept = 2633.620
R-squared = 0.9046
e: Residual standard error = 27580

From the graph, we can see the linear regression line in red:
Expected Price = 297.513 Area + 2633.62



Interpretation

The regression line means that the best linear prediction in terms of SSE described by the equation above. The slope indicates that there is a positive relationship between the area of the house and the price of the house. An increase in apartment size of 1 foot is associated with a ***$297.513*** increase in price. The intercept of the regression tells us the average expected value for the response variable when the predictor variable is equal to zero, but in our case, its meaningless; there is no apartment with area = zero, therefore the meaning of the intercept: regardless to the area of the apartment, the average price of the apartment is ***$2633.62.***

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-76764 -19693    224  17481 127351

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2633.620   5808.341   0.453    0.651
x            297.513      5.936  50.118   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27580 on 265 degrees of freedom
Multiple R-squared:  0.9046,    Adjusted R-squared:  0.9042
F-statistic:  2512 on 1 and 265 DF,  p-value: < 2.2e-16
```

Residuals identify the deviation of observed values from the expected values. It gives an insight into how good our model is against the actual value. In this case the observed values fall on average ***$27,580*** from the regression line. From the graph of linear regression, we can see that the observation with the largest standardized residual is far from the regression line is the apartment with area ***1109.25 feet (calculated by R).***

R-squared is the proportion of the variance in the response variable that can be explained by the predictor variable. In this case, ***90.46%*** of the variance in the apartment price can be explained by the foot- area of the apartment. When we have only one explanatory variable, R^2 equals the square of the correlation = (***0.9510874) ^2 = 0.9045672.***

Predictions

Let’s do some predictions using the linear regression line **Expected Price = 297.513 Area + 2633.62**

Apartment Area (ft.)	Expected Price	Actual price	Error
720.81	$217,083.966	$198,591.85	$18,492.116
1109.25	$322,649.915	$460,001.26	$-127,351.345
1942.5	$580,552.623	503,790.23 $	$76,762.393
697.89	$210,264.968	$219,865.76	$-9600.792
927.83	$278,675.107	$293,876.27	$-15,201.163
579.75	$175,116.782	$171,262.65	$3,854.132