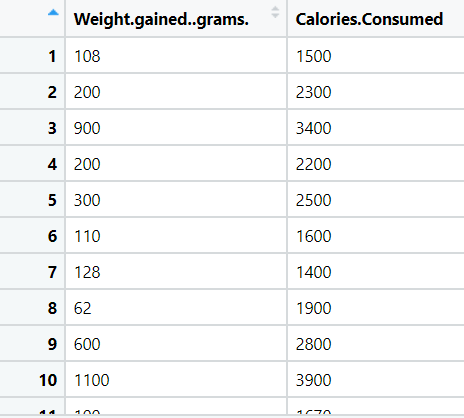
# Simple Linear Regression (Module - 6)

## Do the necessary transformations for input variables for getting better R^2 value for the model prepared. Build the model and predict for the output variables.

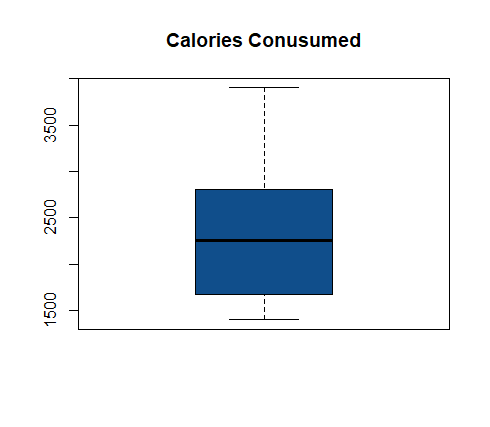
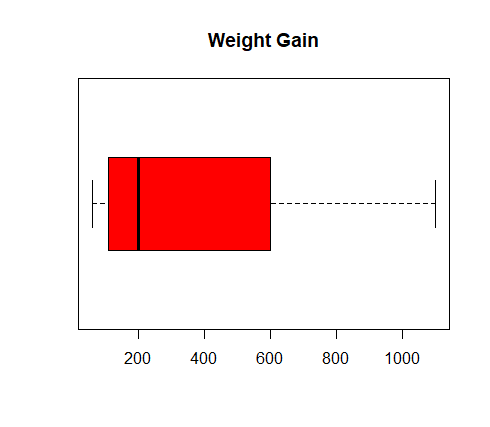
1.) Calories\_consumed -> predict weight gained using calories consumed



**Ans)**

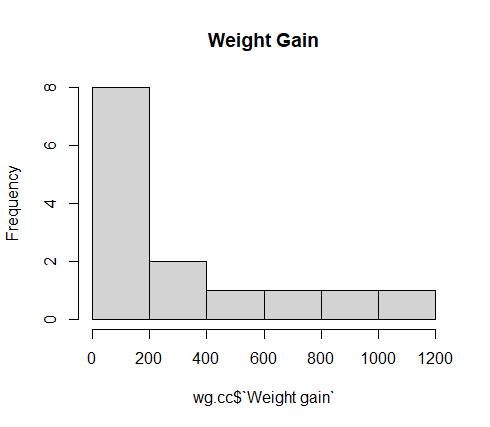
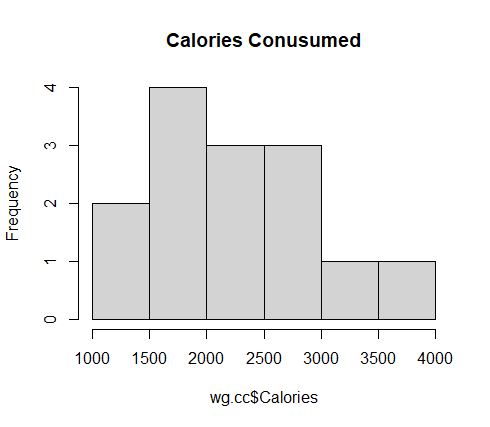
**Inferences from the below Graphical Representation:**

**Box Plot**

****

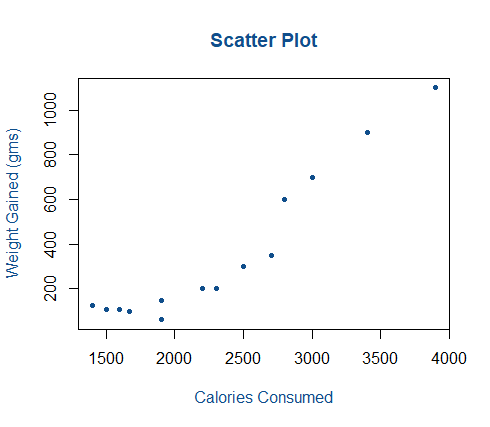
From the above box plots, we can say that there are no outliers exist.

**Histogram:**

****

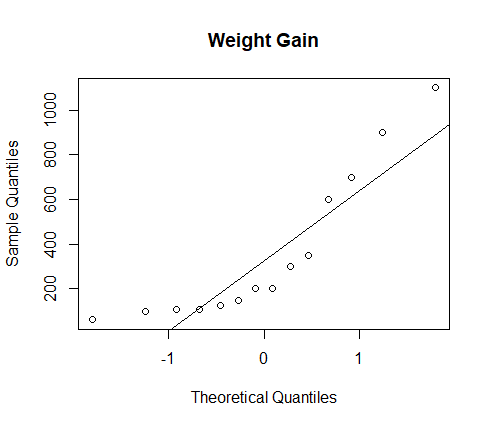
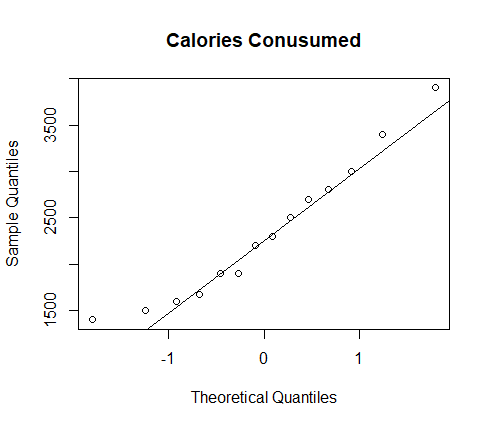
From the above histogram, the data is normally distributed for the calories consumed graph whereas in the weight gain graph, the data is right skewed in nature.

**Scatter Plot:** Coefficient correlation = 0.94



From the above scatter plot, we can say that a strong positive correlation between the variables exists.

**QQ Plot:**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R ^2** | **Adj R ^2** | **Coefficient** | **RMSE** |
| Simple Linear Reg | 0.89 | 0.88 | 0.95 | 103.3025 |
| Logarithmic Reg -Log(x) | 0.81 | 0.77 | 0.91 | 141.001 |
| Exponential Reg – Log(y) | 0.87 | 0.86 | 0.94 | 118.04 |
| Polynomial Reg  Log(y) & x+I(x\*x) | 0.87 | 0.85 | 0.94 | 117.41 |

**Inferences from the above table:**

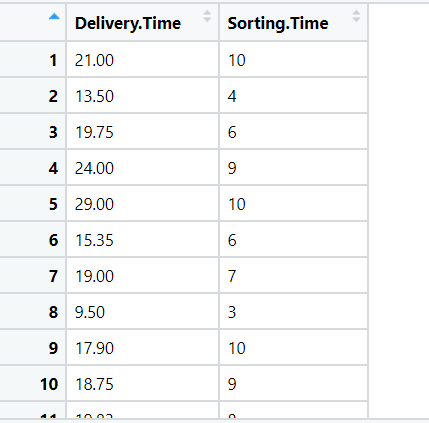
* From all the models the R^2 and Adjusted R^2 value is > 0.80 which indicates that all the above models are strong.
* All the model coefficient correlation represents strong positive relation
* But comparing the models with RMSE, the least RMSE need to be considered as the best fit model
* So, from the above observation, Simple linear regression is best fit model as it is with least RMSE and there is a strong correlation comparing with the other models.

**Results from the model:**

* Test RMSE = 108.55
* Train RMSE = 101.82

AS there is no much difference between train error and test error, the model is **Right Fit.**

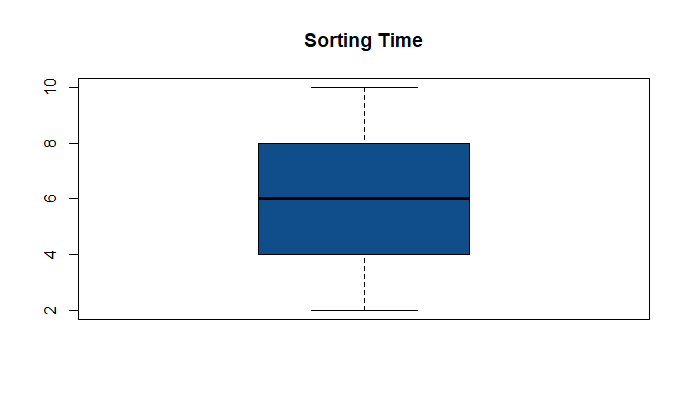
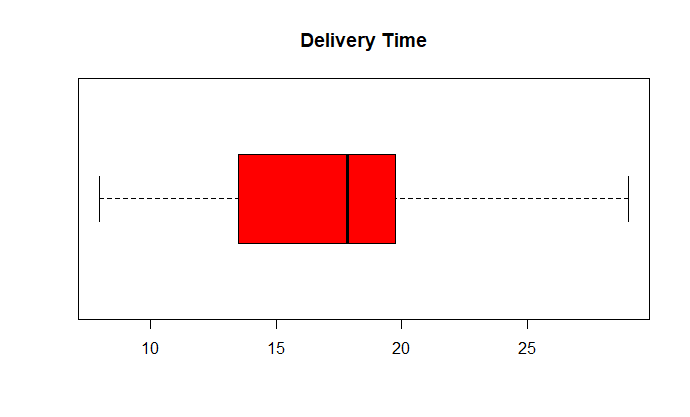
2) Delivery\_time -> Predict delivery time using sorting time



**Ans)**

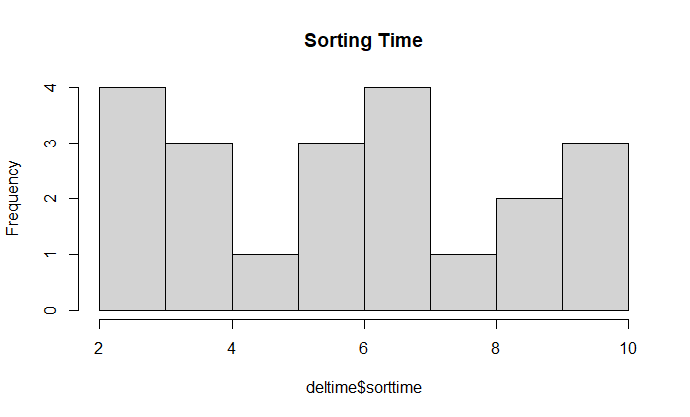
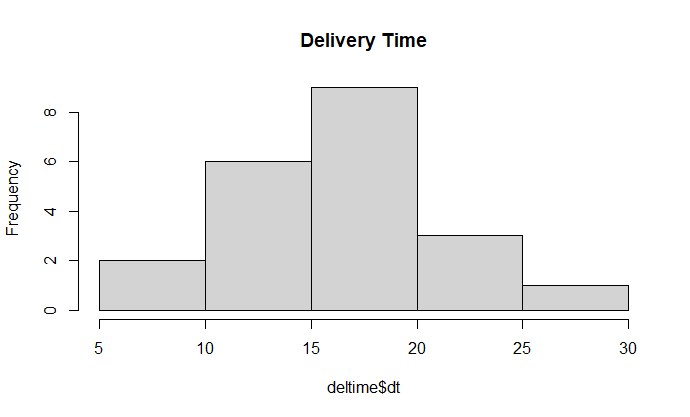
**Inferences from the below Graphical Representation:**

**Box Plot:**

****

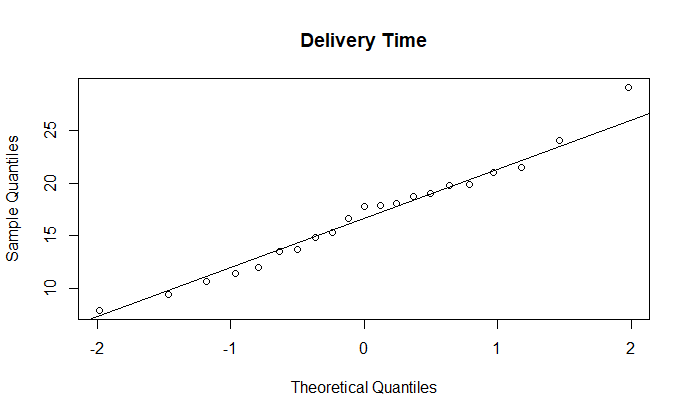
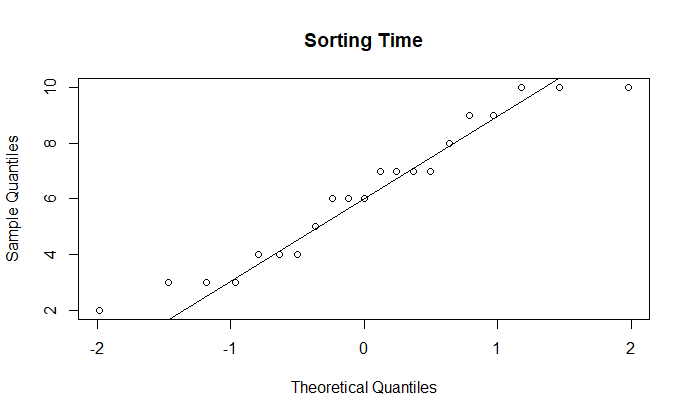
* From the above box plots, we can say that there are no outliers exist.

**Histogram:**

****

* From the histogram we can that the data is normally distributed

**QQ Norm:**

****

* Both graph represents that data is normally distributed.

**Coeffiencent Correlation = 0.83 indicates strong positive correlation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R ^2** | **Adj R ^2** | **Coefficient** | **RMSE** |
| Simple Linear Reg | 0.68 | 0.66 | 0.82 | 2.79 |
| Logarithmic Reg -Log(x) | 0.70 | 0.68 | 0.83 | 2.73 |
| Exponential Reg – Log(y) | 0.71 | 0.69 | 0.80 | 2.94 |
| Polynomial Reg  Log(y) & x+I(x\*x) | 0.76 | 0.73 | 0.83 | 2.79 |

**Inferences from the above table:**

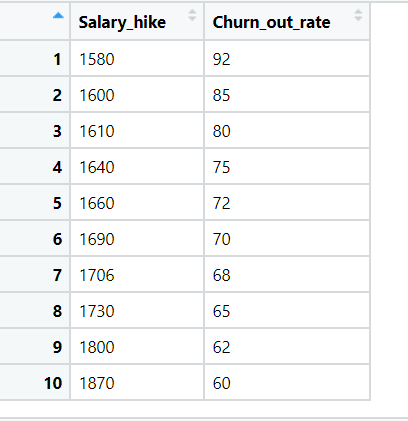
* From all the models the R^2 and Adjusted R^2 value is between 0.65 to 0.80 which indicates that all the above models are moderately strong.
* All the model coefficient correlation represents strong positive relation
* But comparing the models with RMSE, the least RMSE need to be considered as the best fit model
* So, from the above observation, Logarithmic regression is best fit model as it is with least RMSE and there is a strong correlation comparing with the other models.

**Results from the model:**

* Test RMSE = 3.89
* Train RMSE = 2.34

As train error is low and test error is high the model is over fit, we need to apply **regularization methods** to get the right fit model.

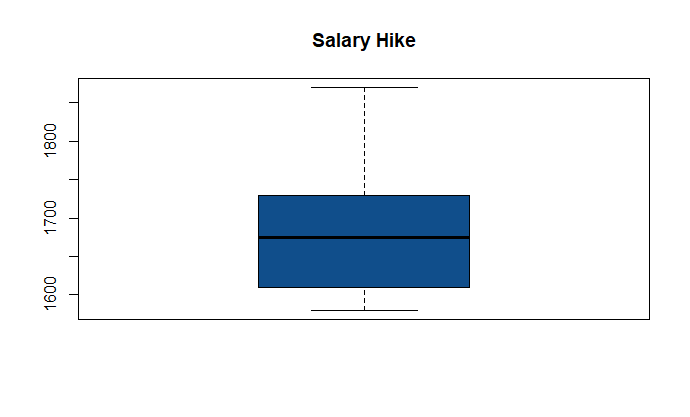
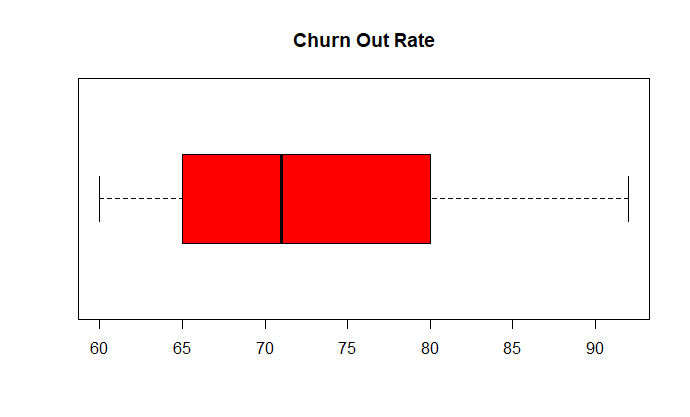
1. Emp\_data -> Build a prediction model for Churn\_out\_rate



**Ans)**

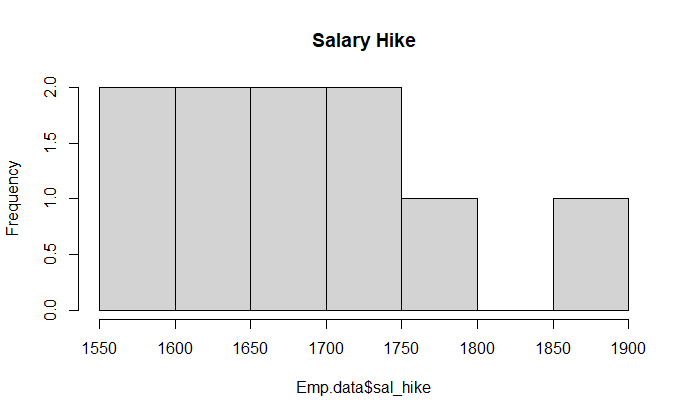
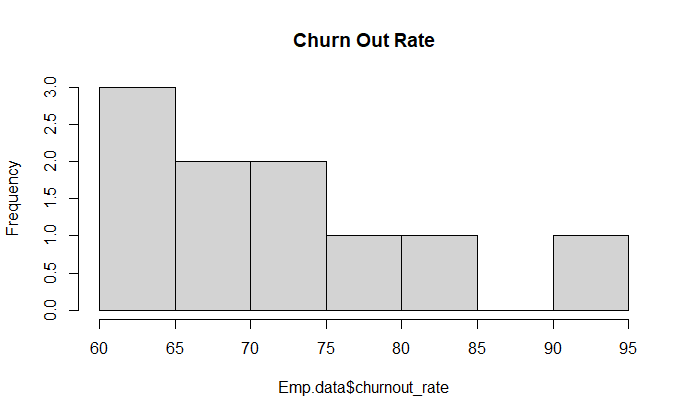
**Inferences from below Graphical Representation:**

**Box Plot:**

****

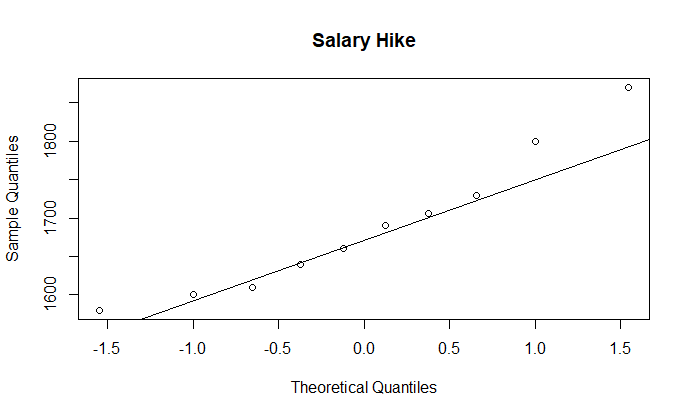
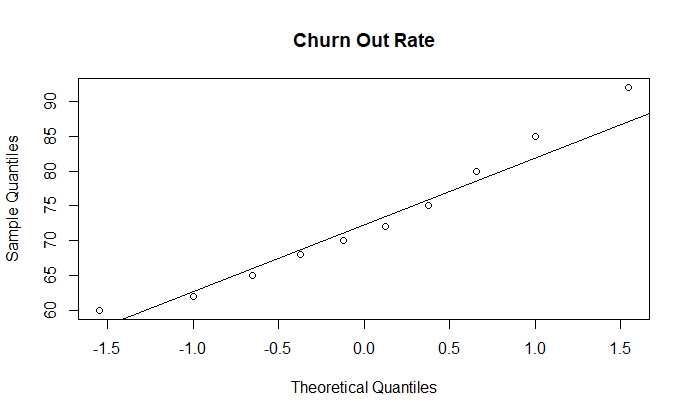
* From the above box plots, we can say that there are no outliers exist.

**Histogram:**

****

* From the histogram we can that the data is normally distributed

**QQ Plots:**

****

* Both graph represents that data is normally distributed
* **Coeffiencent Correlation = -0.911 (Strong Negative correlation occurs)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R ^2** | **Adj R ^2** | **Coefficient** | **RMSE** |
| Simple Linear Reg | 0.83 | 0.81 | 0.91 | 3.99 |
| Logarithmic Reg -Log(x) | 0.84 | 0.82 | 0.92 | 3.78 |
| Exponential Reg – Log(y) | 0.87 | 0.85 | 0.93 | 3.54 |
| Polynomial Reg  Log(y) & x+I(x\*x) | 0.98 | 0.97 | 0.99 | 1.32 |

**Inferences from the above table:**

* From all the models the R^2 and Adjusted R^2 value is > 0.80 which indicates that all the above models are strong.
* All the model coefficient correlation represents strong positive relation
* But comparing the models with RMSE, the least RMSE need to be considered as the best fit model
* So, from the above observation, Polynomial regression is best fit model as it is with least RMSE and there is a strong correlation comparing with the other models.

**Results from the model:**

* Test RMSE = 1.04
* Train RMSE = 1.38

As there is a slight variation in the test and train RMSE approximately both are closer so the model is **Right Fit.**

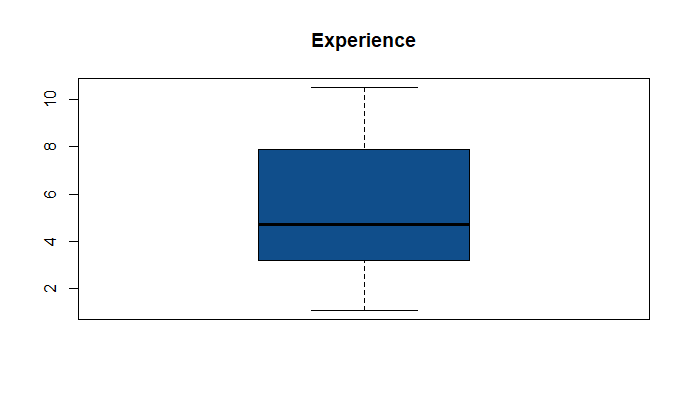
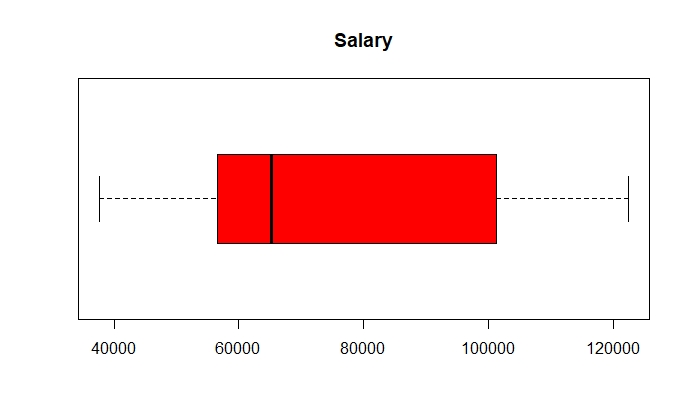
## 4)Salary\_hike -> Build a prediction model for Salary\_hike



**Ans)**

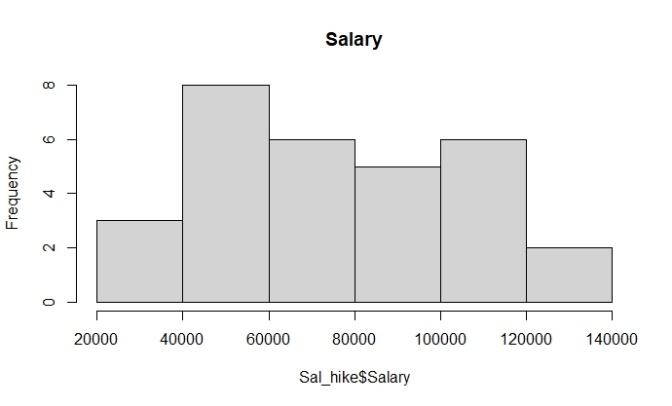
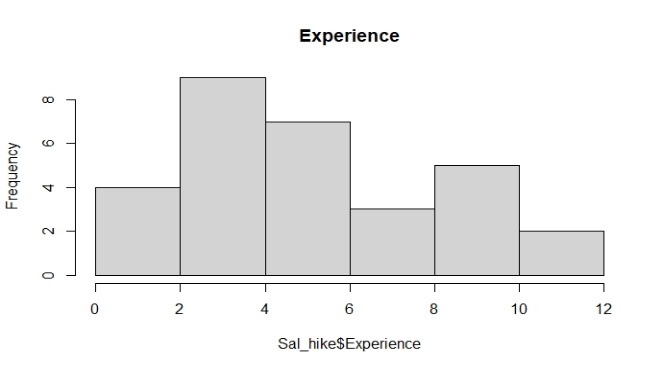
**Inferences from the below Graphical Representation:**

**Box Plot:**

****

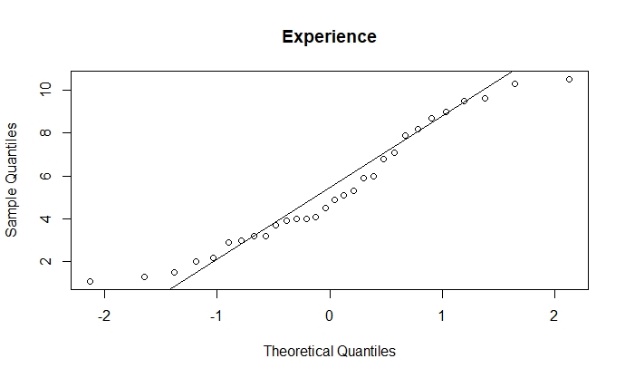
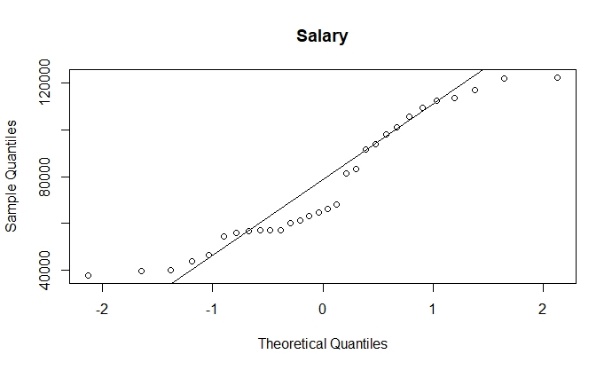
* From the above box plots, we can say that there are no outliers exist

**Histogram:**

****

* From the histogram we can that the data is normally distributed

**QQ Plots:**

****

* Each graph represents that data is normally distributed

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R ^2** | **Adj R ^2** | **Coefficient** | **RMSE** |
| Simple Linear Reg | 0.95 | 0.95 | 0.97 | 5592.04 |
| Logarithmic Reg -Log(x) | 0.85 | 0.84 | 0.92 | 10302.89 |
| Exponential Reg – Log(y) | 0.93 | 0.92 | 0.96 | 7213.235 |
| Polynomial Reg  Log(y) & x+I(x\*x) | 0.94 | 0.94 | 0.97 | 5391.082 |

**Inferences from the above table:**

* From all the models the R^2 and Adjusted R^2 value is > 0.80 which indicates that all the above models are strong.
* All the model coefficient correlation represents strong positive relation
* But comparing the models with RMSE, the least RMSE need to be considered as the best fit model
* So, from the above observation, Polynomial regression is best fit model as it is with least RMSE and there is a strong correlation comparing with the other models.

**Results from the model:**

* Test RMSE = 6780.185
* Train RMSE = 5151.434

As train error is low and test error is high the model is over fit, we need to apply **regularization methods** to get the right fit model

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without errors
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.