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

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

3) Emp\_data -> Build a prediction model for Churn\_out\_rate

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

Do the necessary transformations for input variables for getting better R^2 value for the model prepared.

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

Ans:-

**Business Problem:-** To predict the weight gained using calories consumed

**Datasets:-**

Independent Variable (x): calories\_consumed

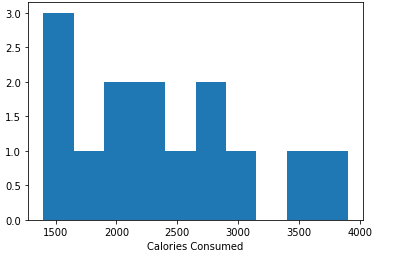
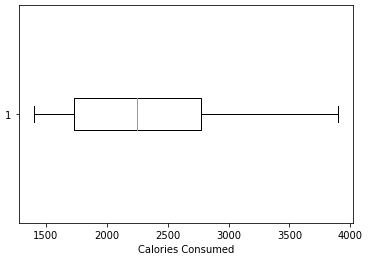
Dependent variable(y): weight\_gained

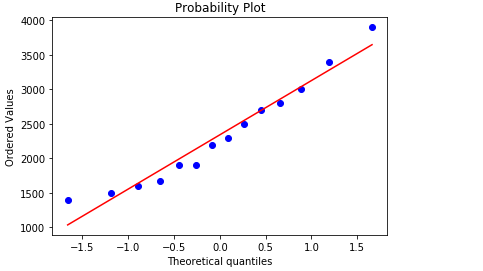
Both x and y are continuous variable

EDA

Graphical Representation:-

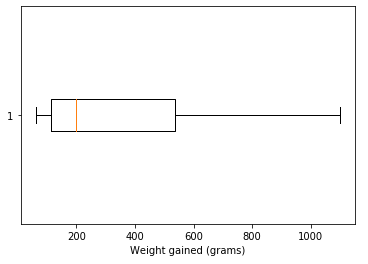
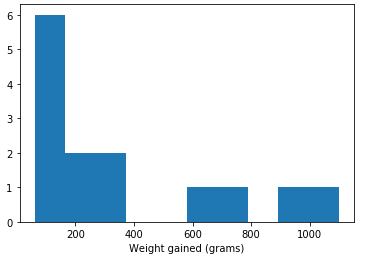
**For Calories\_consumed variable**

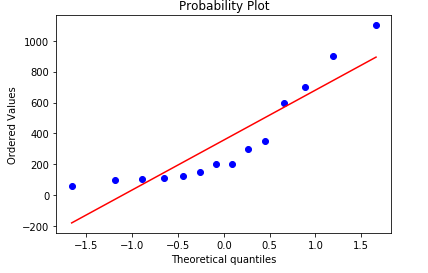
 



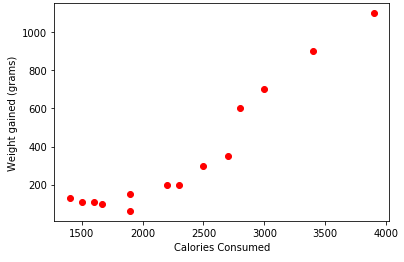
The boxplot and histogram represents that it is not normally distributed and the data is positively skewed. From the above qq plot it is found that most of the data points are near to the line but not on the line.

**For Weight Gained Variable**





The box plot and histogram represents that the data is not normally distributed and it is positively skewed. From the above qq plot it is found that most of the data points are away from the line.



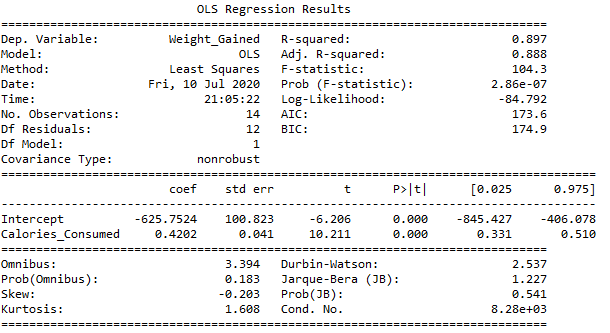
**Co-relation**

Cor-coeff = 0.9469910088554458

The correlation value is high which indicates there is good or strong co relation calories consumed and weight gained and sign is positive (which means if calories consumed increases, weight gained also increases). From the scatter plot it can be easily identified that there is linear relationship between calories consumed and weight gained.

**Model Building**

**1. Linear Model**



**Analysis**

From the summary,

p-value is < 0.05, have to reject null hypothesis and accept the alternative hypothesis, so the data is significant( variables are related to each other)

R square value is 0.897which indicates there is good fit of the data points in the regression line

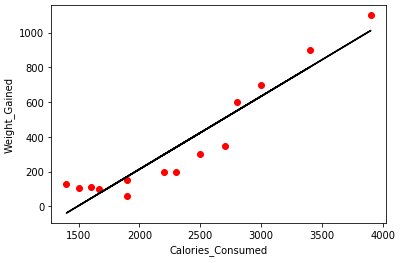
R-squared = 0.897

**RMSE =** 103.3025

Need to be reduced. Lesser the RMSE value better the model is so we have to do necessary

transformation for reducing the RMSE value.

Regression Line:



**weight\_gained = -625.752356 + 0.** **0.420157\* calories\_consumed**

**Transformation**

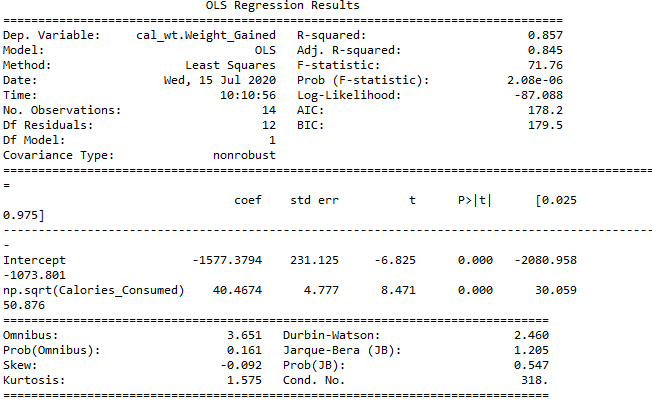
2. Square Root Model

X : square root(calories\_consumed)

Y : weight\_gained

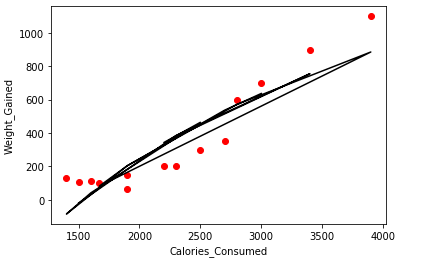
Cor-coef = 0.926

**Model Summary**



R-squared= 0.857

RMSE=121.71



**Regession Line**

Weight\_gained = -1577.379417+ 40.46736 √calories\_consumed

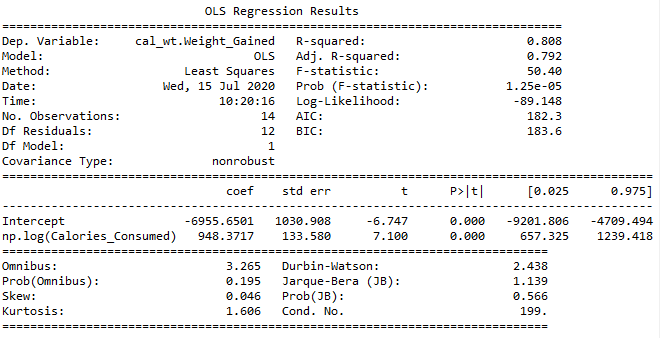
**3. Logarithmic Model**

X: log(calories\_consumed)

Y: weight\_gained

Cor-coef= 0.8987

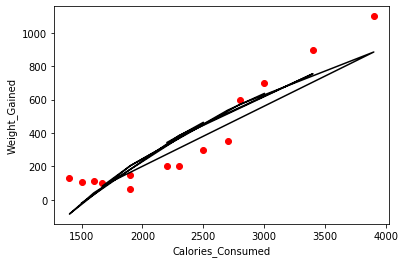
Model Summary



R-squared = 0.8077

RMSE= 141.005

Regression Line



Weight\_gained = -6955.65+ 948.37 log(calories\_consumed)

**4. Quadratic Model**

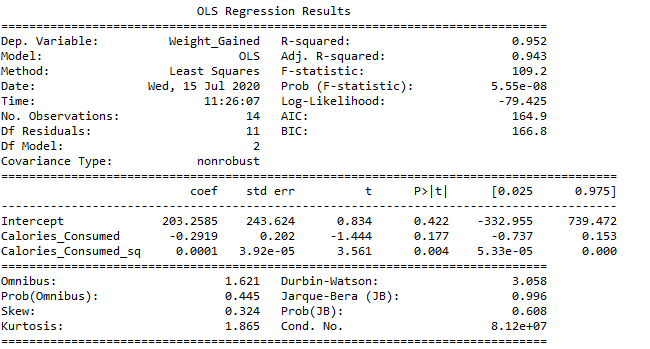
X: calories \_consumed \* calories\_consumed

Y: weight\_gained

Model Building

Cor-coef = 0.9710636

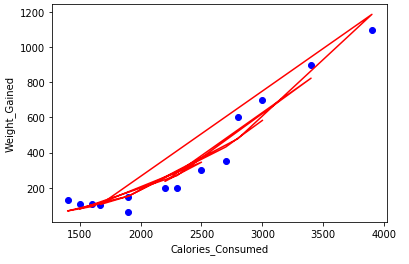
**Model Summary**



**R-Squared=** 0.952

**RMSE=** 70.40

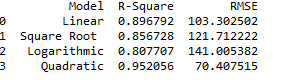
Regression Line



**Model**

Weight\_gained = 203.258458-0.291894calories\_consumed + 1. 0.000140-09 (calories\_consumed^2)

|  |
| --- |
|  |
|  |
| |  | | --- | |  | |



From the above model quadratic model is the best model as R-squared value is the highest

and RMSE is the lowest.

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

1.) **Business Objective**: - To predict delivery time using sorting time

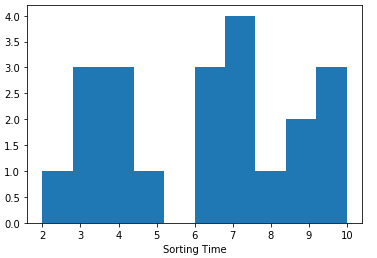
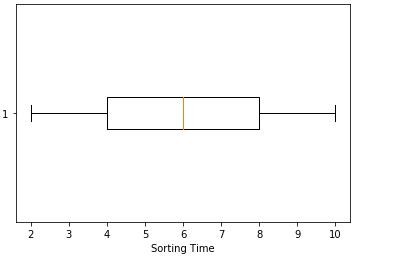
2.) **Datasets: -** Independent Variable (x): sorting\_time

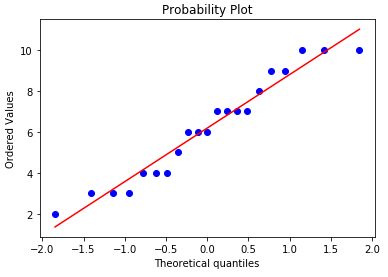
Dependent Variable (y): delivery\_time

Both x and y are continuous variable

3.) **EDA**

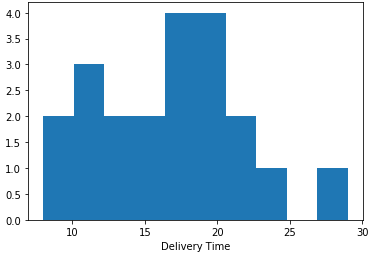
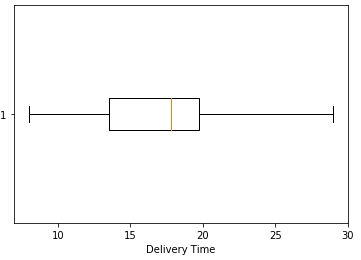
For delivery time

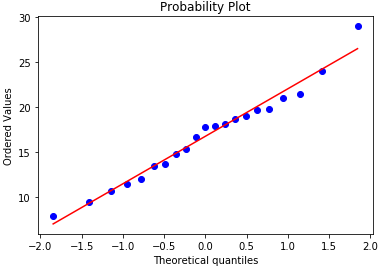
 



From the above plots, delivery time is not normally distributed, but tends to normal

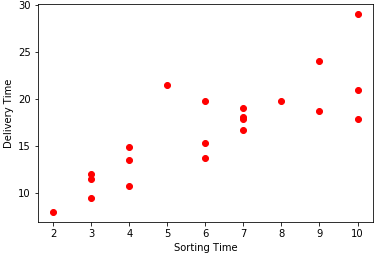
For sorting time



From the above plots, sorting time is not normally distributed

Scatter Plot:



Cor-coeff: 0.8259

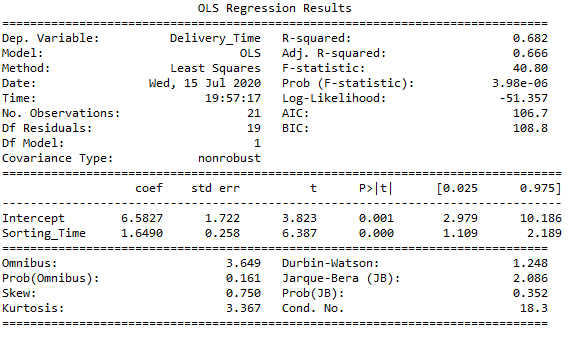
There is linear positive relationship with moderate co-relation between sorting\_time and

delivery\_time.

**Model Building**

**1. Linear Model**

**Model Summary**

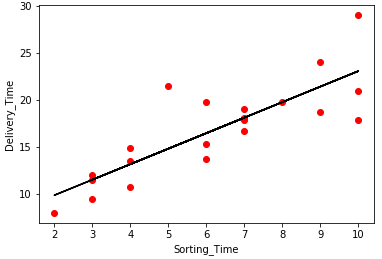


R-Squared = 0.682. This value is very less, thus need to be improved.

p-value < 0.05, Therefore variables are significant

**RMSE = 2.79**

**Regression Line:**



**delivery\_time = 6.582734 + 1.649020 (sorting\_time)**

Transformation

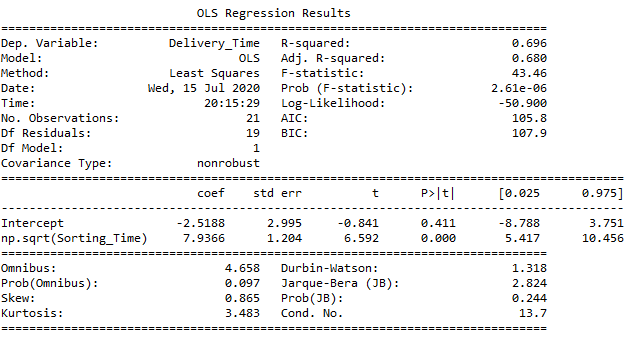
2.) Square root transformation

X: squareroot(sorting\_time)

Y: delivery\_time

Cor-coeff = 0.834

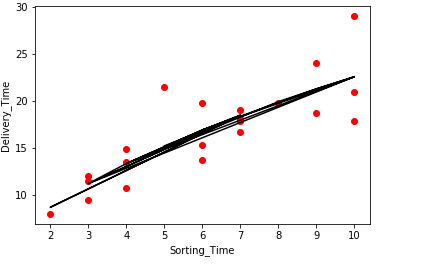
Model Summary:



**R-squared= 0.696**

**RMSE = 2.73**

**Regression Line**



**delivery\_time= -2.518837 + 7.936591√sorting\_time**

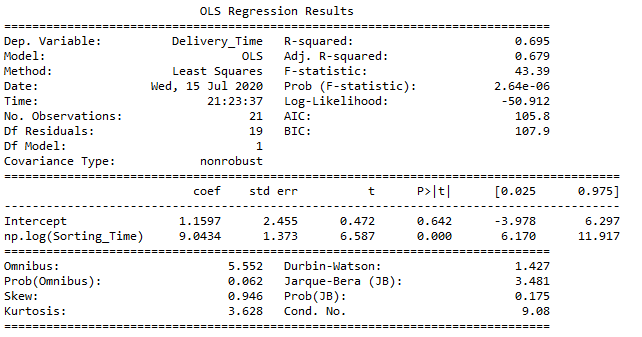
3.) Logarithmic model

X: log(sorting\_time)

Y: delivery\_time

Cor-coef = 0.83

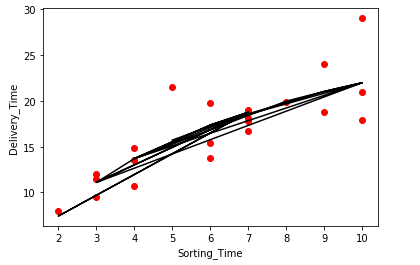
Model Summary:



R-squared: 0.695

RMSE = 2.733

Regression Line:



delivery\_time = 1.159684+ 9.043413 log(sorting\_time)

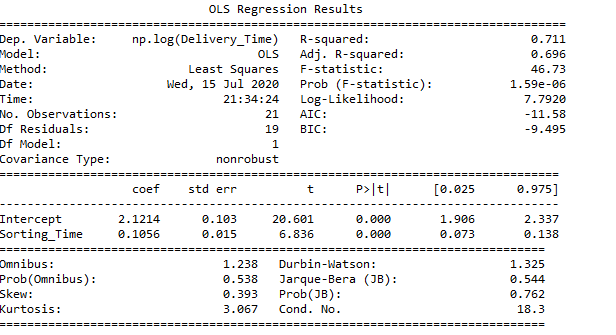
4.) Exponential Model

X: sorting\_time

Y: log(delivery\_time)

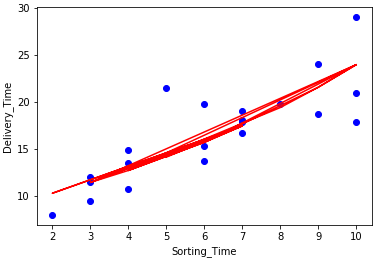
Cor-coef = 0.84

Model Summary

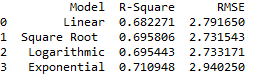


R-Squared=0.711

RMSE = 2.94



Log(delivery\_time) = 2.1213719 + 0.1055516(sorting\_time)



Here the highest R-squared value is in exponential model and lowest RMSE is in Square root model. Now as RMSE value is lowest in square root model, so that model is considered as

best.

3) Emp\_data -> Build a prediction model for Churn\_out\_rate

**Business Problem:-** To predict churn out rate using salary\_hike

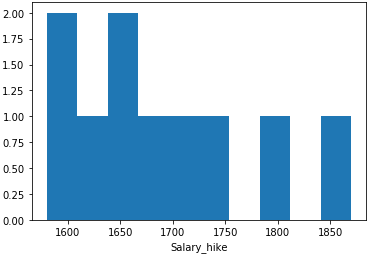
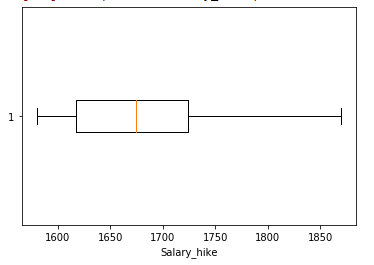
**Datasets:-**

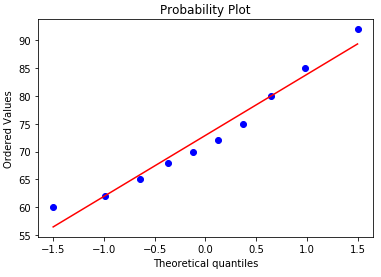
Independent variable(x): salary\_hike

Depenedent variable(y): churn\_out\_rate

**EDA:-**

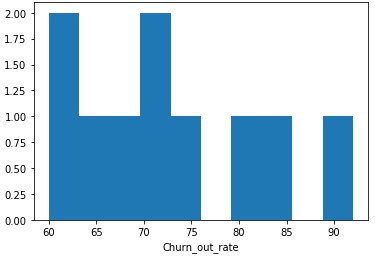
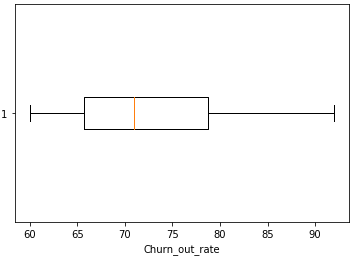
For Salary Hike

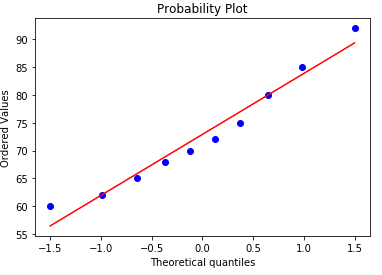
 



From the histogram and box plot, it is found that the data set is having positive skewness and the tail is extended towards right, Moreover the qq plot shows that the most of the datapoints are not on the line. Thus we can conclude that data is not normal.

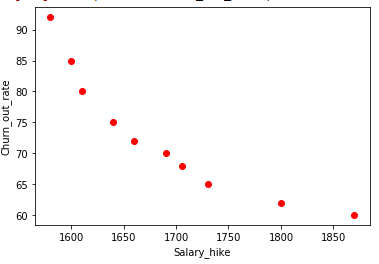
For Churn out rate



From the histogram and box plot, it is found that the data set is having positive skewness and the tail is extended towards right, Moreover the qq plot shows that the most of the data points are not on the line. Thus we can conclude that data is not normal.

Scatter Plot



The above scatter plot shows there is linear relationship between salary hike and churn out rate.

The relationship is negative, as it shows if salary hike increases churn out rate decreases and vice versa.

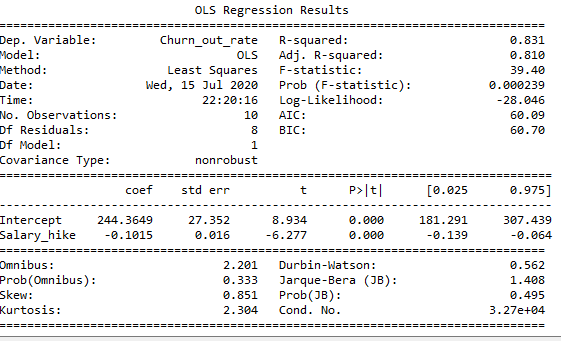
Co-relation = -0.9117216

This value indicates that that there is strong co-relation between salary hike and churn out rate.

**Model Building**

1. Linear Model

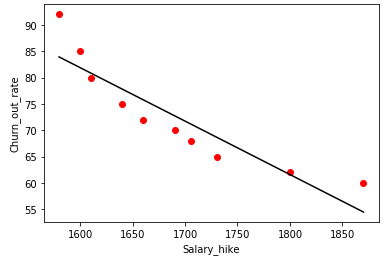
Model Summary:



The Multiple R-square value is 0.831 is quite high, which indicates there is good fit of the data points in the regression line.

Moreover the p- value is less than 0.05 which states that the variables are significant

Root Mean Square Error: 3.99



Regression line:- **Churn out rate = 244.3649111 + (-0.1015426) Salary hike**

Now Transformation:

2. Square Root Model

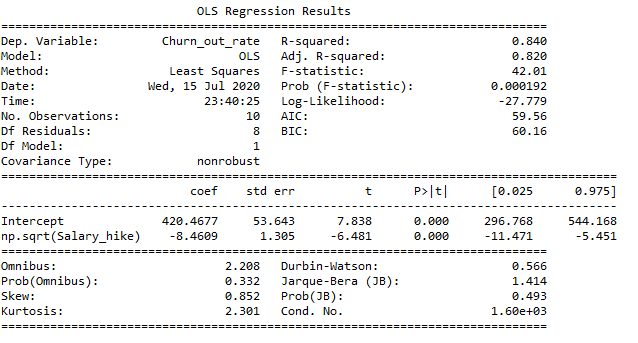
X= square root (salary\_hike)

Y= churn\_out\_rate

Cor-coeff = -0.916

This is showing there is strong co-relation with negative relationship between two models.

Model Summary



R-Square= 0.840

RMSE = 3.891995

Regression Line : -



**Churn\_out\_rate = 420.467749 -8.460948 √salary hike**

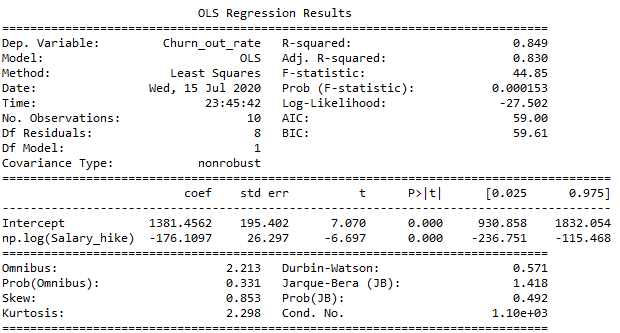
3. Logarithmic Model

Independent variable = log(salary\_hike)

Dependent variable = churn\_out\_rate

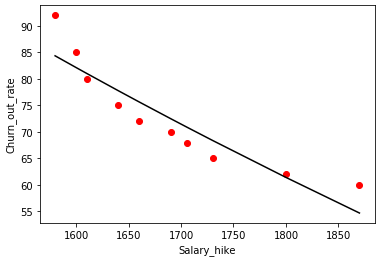
Cor = -0. 92

Model Summary:



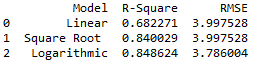
R-squared = 0.849

RMSE = 3.786



Regression Line :-

Churn out rate = 1381.4562 + ( -176.1097 ) log(Salary hike)



In this case,s R-squared value is highest and RMSE is lowest in logarithmic model. So it is the best model.

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

**Business Problem:-** Predict salary using years of experience

**Datasets:-**

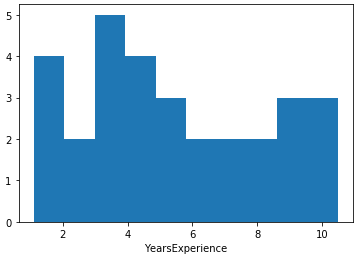
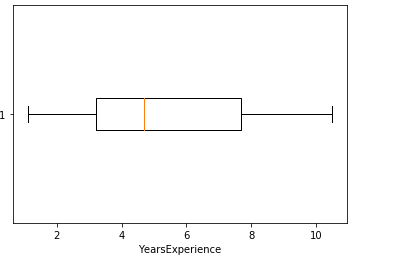
Independent Variable (x): Yearsexperience

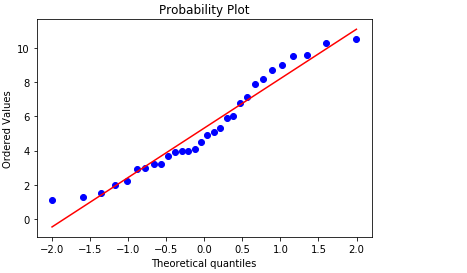
Dependent Variable(y): Salary

**EDA**

Graphical Representation

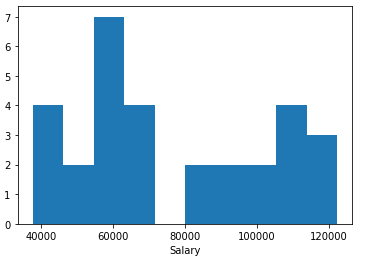
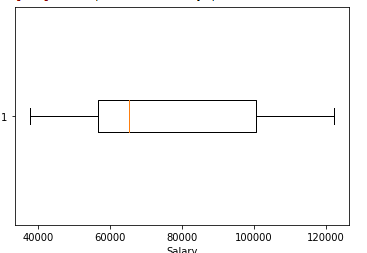
For YearsExperience

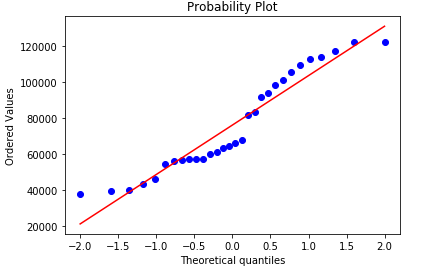
 



From the above plot, it can be said that the data is not normal and having slightly positive skewness. From the qq plot, it can be represented that the most of the data points are away from the optimal line.

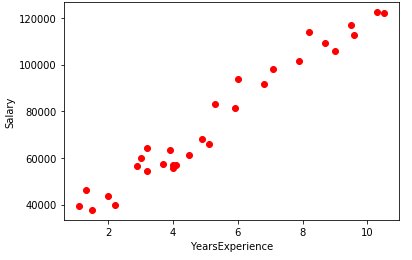
For Salary



From the above plot, it can be said that the data is not normal and having slightly positive skewness. From the qq plot, it can be represented that the most of the data points are away from the optimal line.

Scatter Plot



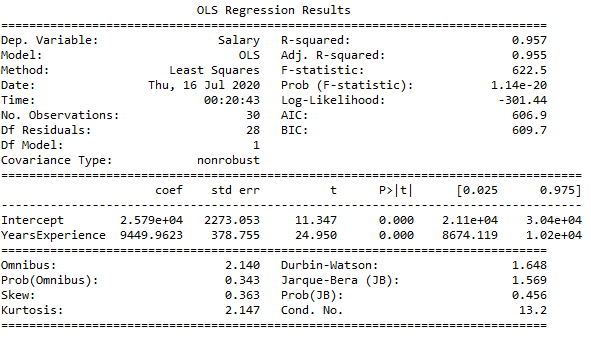
Cor-coeff = 0.978

The scatter plot and cor-coeffcient value indicates that there is strong positive linear co-relation between years of Experience and salary. That means if years of experience increases than salary increases.

**Model Building**

**1. Linear Model**

Model Summary



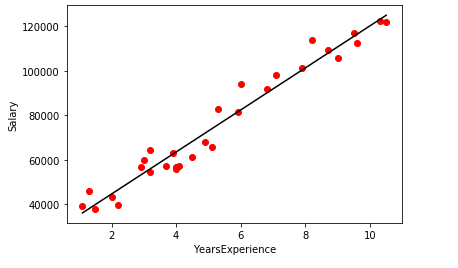
R- Squared = 0.957

This is quite high and indicates that there is good fit of data points in the regression line

p-value <0.05 which indicates that the variables are significant

RMSE = 5592.044

Regression Line



Salary = 25792.20+ 9449.96Years of experience

Transformation

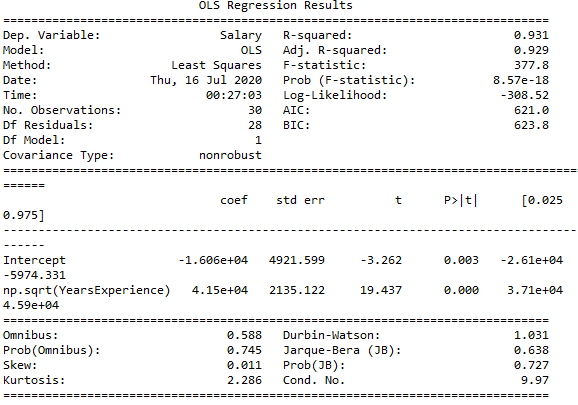
2. Square Root Model

x: square root(YearsExperience)

y: Salary

Cor- coeff = 0.964 (It is reduced)

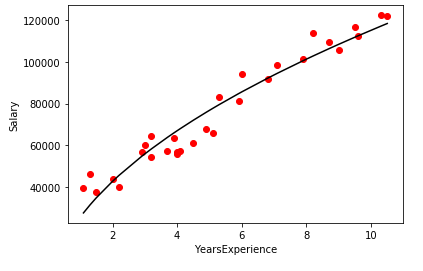
Model Summary



R- squared = 0.931 .There is no improvement in this value rather it decreased

RMSE = 7080.096 Moreover RMSE value also increased

So, that is not a good model.



Regression Line:- Salary = -16055.77 + 41500.68√Years of experience

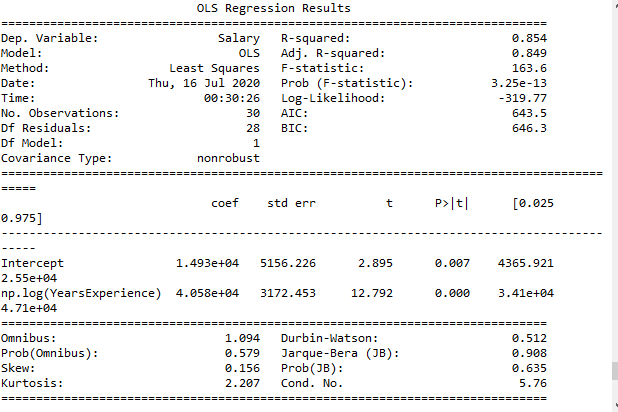
3. Logarithmic Model

X: log(YearsExperience)

Y: Salary

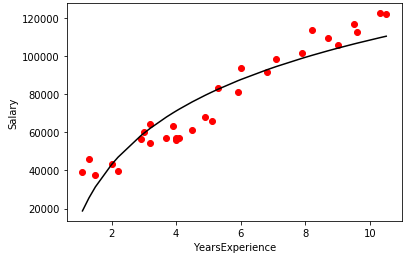
Cor-coef = 0.92

Model Summary:



R-Squared = 0.854 (Further reduced from previous model)

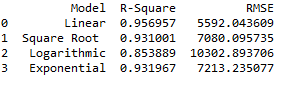
RMSE = 10302.89 (It increased further)



|  |
| --- |
| Regression Line : - Salary = 14927.97 + 40581.99 log(YearsExperience)  3.) Exponential model  X: YearsExperience  Y: log(Salary)  Cor-coef = 0.9653844  Model Summary    R-Squared = 0.932  RMSE = 7213.235 |
|  |
| |  | | --- | |  | |

Regression-Line :-

log (salary) = 10.5074019 + 0.1254529(YearsExperience)



In this case, linear model is having highest R- Squared value and lowest RMSE value, So this is considered as best model.