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

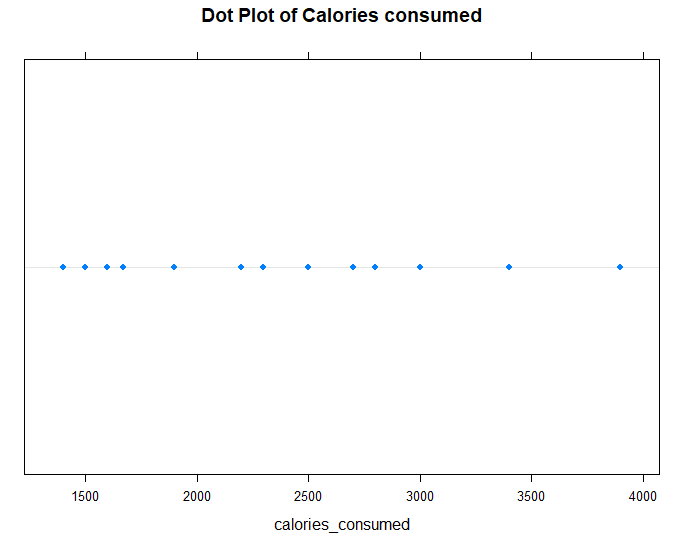
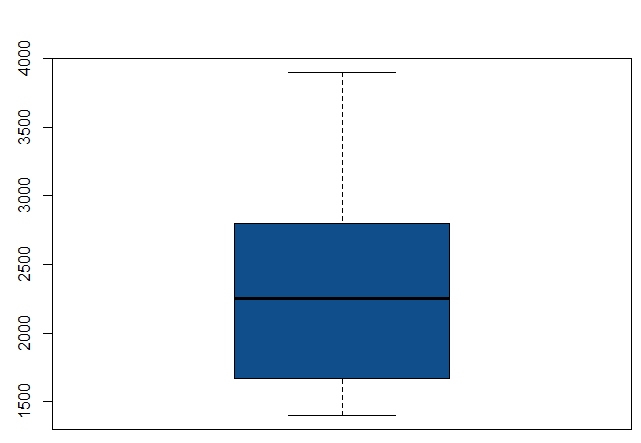
Dependent variable(y): weight\_gained

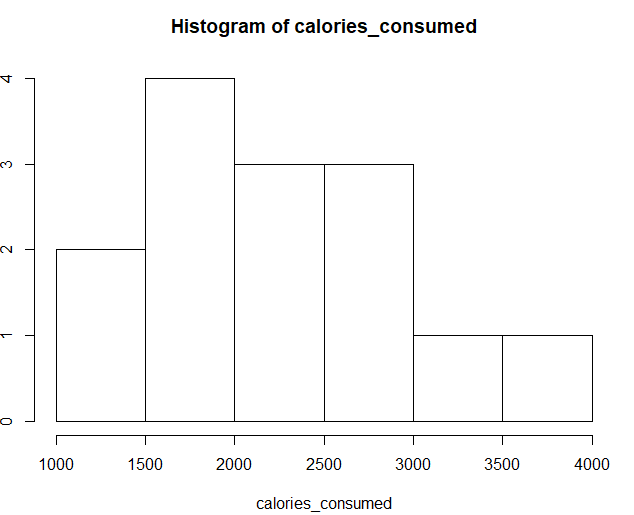
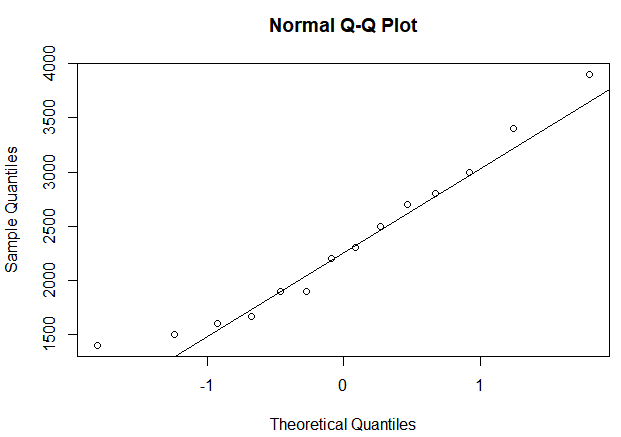
Both x and y are continuous variable

EDA

Graphical Representation:-

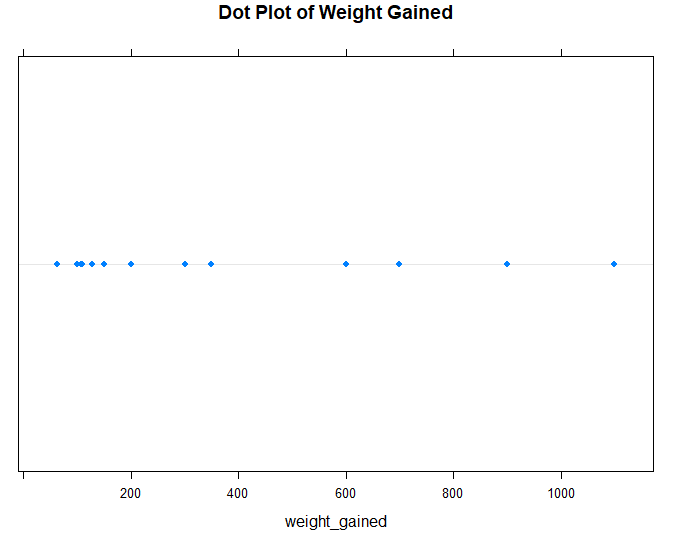
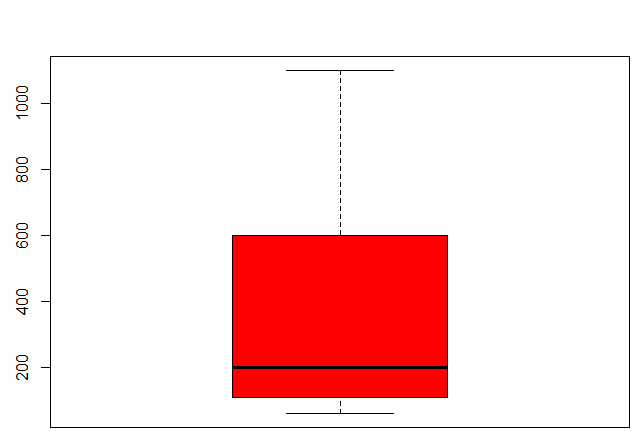
**For Calories\_consumed variable**

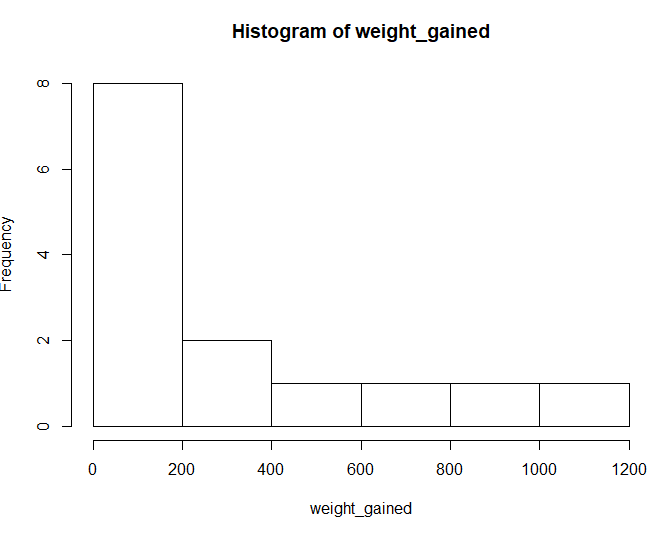
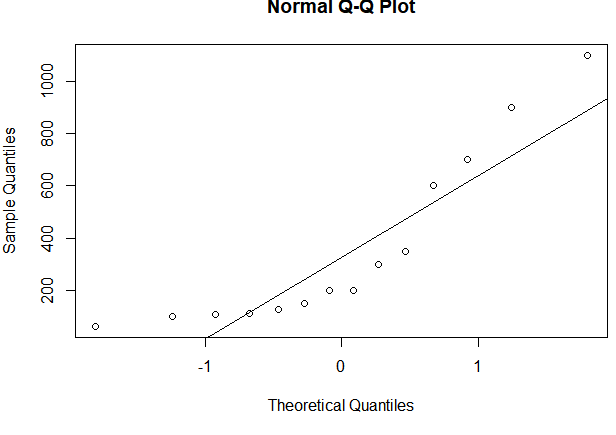
 

The boxplot and histogram represents that it is not normally distributed and the data is positively skewed . From the above qq plot it is found that most of the data points are near to the line but not on the line

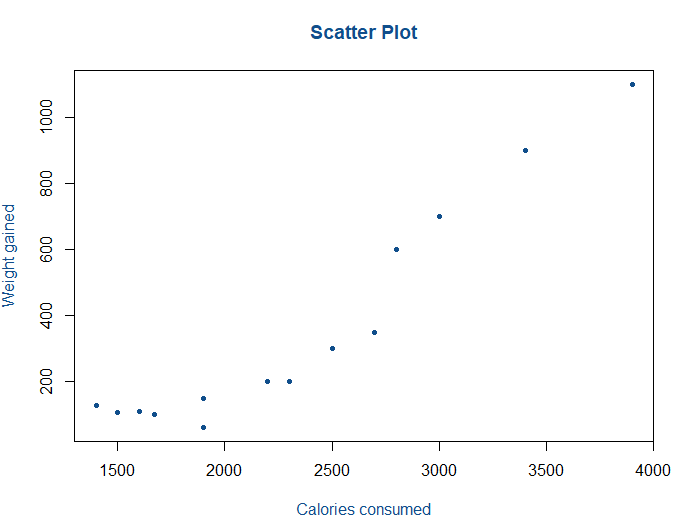
**For Weight gained variable**

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.

**Scatter plot**



**Co-relation**

Cor-coeff = 0.946911

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**

Linear Model

Model Summary:

Call:

lm(formula = weight\_gained ~ calories\_consumed, data = cal.wt)

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

calories\_consumed 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

**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)

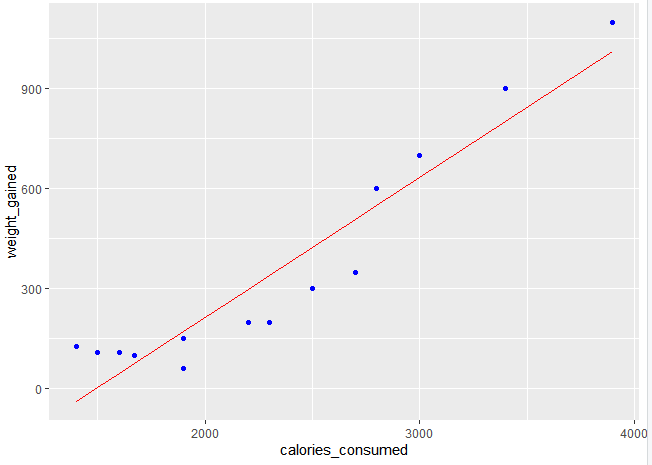
R square value is 0.8968 which indicates there is good fit of the data points in the regression line

R-squared = 0.8968

**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 = -625.7523557 + 0.4201566\* calories\_consumed**

**Evaluation**

Transformation

1. Square Root Model

X : square root(calories\_consumed)

Y : weight\_gained

Cor-coef = 0.92

Call:

lm(formula = weight\_gained ~ sqrt(calories\_consumed), data = cal.wt)

Residuals:

Min 1Q Median 3Q Max

-175.37 -123.59 29.85 105.48 191.23

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1577.379 231.125 -6.825 1.84e-05 \*\*\*

sqrt(calories\_consumed) 40.467 4.777 8.471 2.08e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

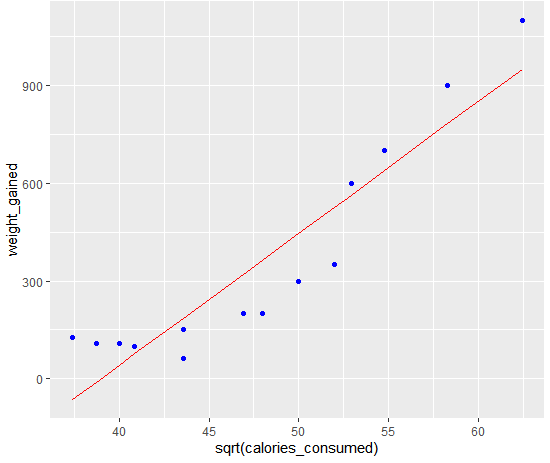
Residual standard error: 131.5 on 12 degrees of freedom

Multiple R-squared: 0.8567, Adjusted R-squared: 0.8448

F-statistic: 71.76 on 1 and 12 DF, p-value: 2.083e-06

R-squared= 0.8567

RMSE= 121.71



Regression Line

Weight\_gained = -1577.37942 + 40.46736 √calories\_consumed

2. Logarithmic Model

X: log(calories\_consumed)

Y: weight\_gained

Cor-coef= 0.89

Model Summary

Call:

lm(formula = weight\_gained ~ log(calories\_consumed), data = cal.wt)

Residuals:

Min 1Q Median 3Q Max

-187.44 -142.96 23.13 113.20 213.82

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6955.7 1030.9 -6.747 2.05e-05 \*\*\*

log(calories\_consumed) 948.4 133.6 7.100 1.25e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 152.3 on 12 degrees of freedom

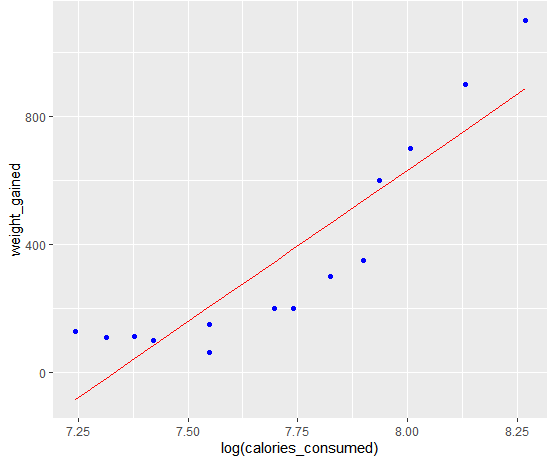
Multiple R-squared: 0.8077, Adjusted R-squared: 0.7917

F-statistic: 50.4 on 1 and 12 DF, p-value: 1.248e-05

R-squared = 0.8077

RMSE= 141.005

Regression Line



Weight\_gained = -6955.6501 + 948.3717 log(calories\_consumed)

3.) Quadratic Model

X: calories \_consumed \* calories\_consumed

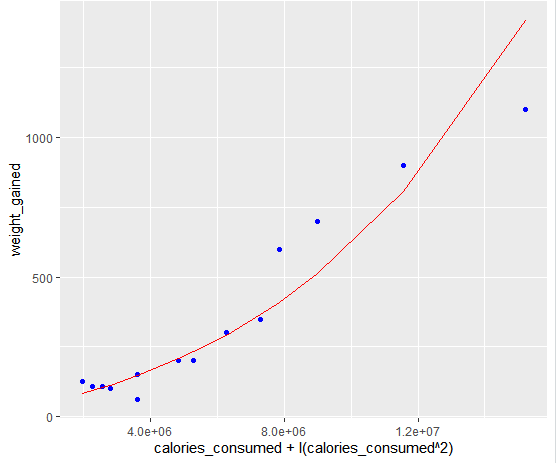
Y: weight\_gained

Cor-coef = 0.9710636

R-Suared= 0.8776

RMSE = 117.41

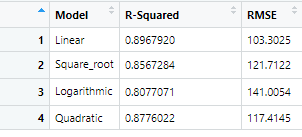
Regression Line



**Model**

Weight\_gained = 2.828719e+00 + 0.001142146e calories\_consumed + -1.675054e-09 (calories\_consumed^2)

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



From the above model Linear 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

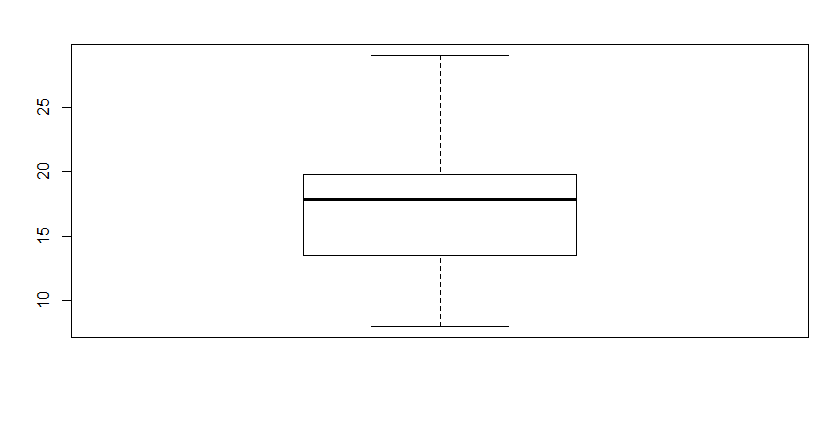
2.) **Datasets: -** Independent Variable (x): sorting\_time

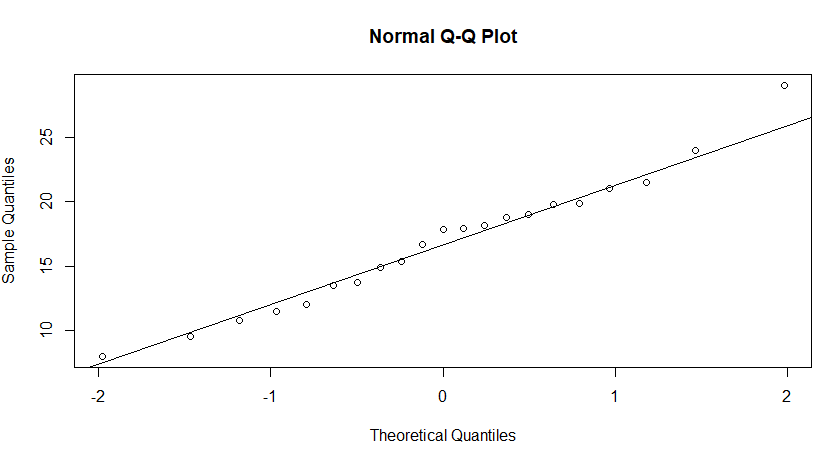
Dependent Variable (y): delivery\_time

Both x and y are continuous variable

3.) **EDA**

For delivery time

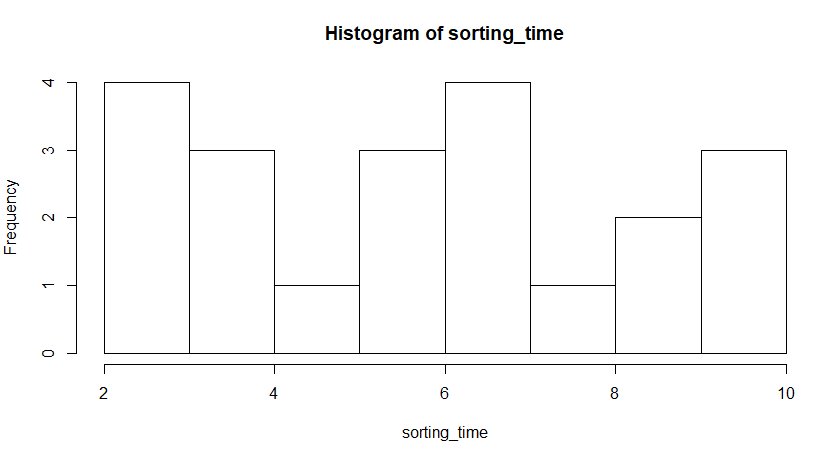
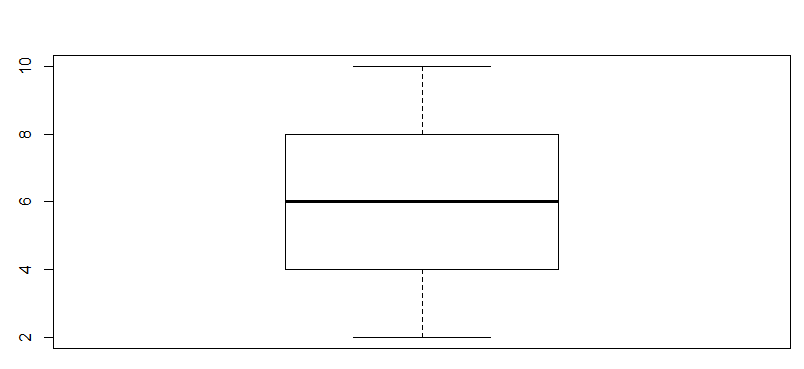
 

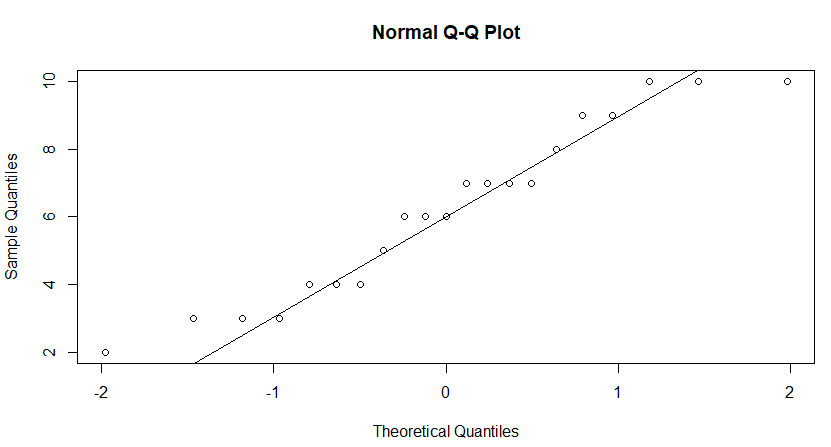
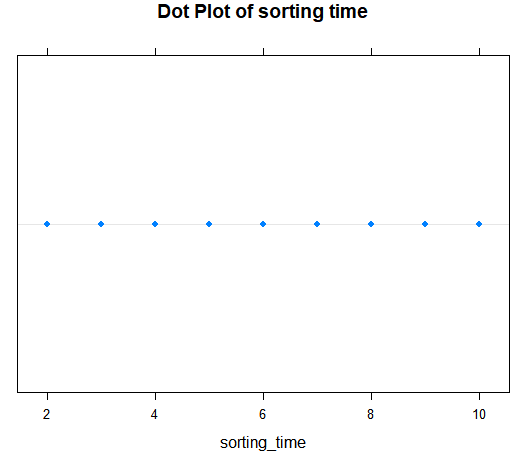




From the above plots, delivery time is not normally distributed, but tends to normal

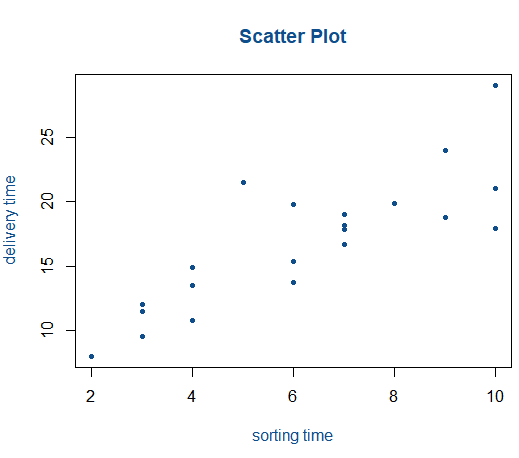
For sorting time

From the above plots, sorting time is not normally distributed

Scatter Plot:



Cor-coeff: 0.8259973

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

**Model Building**

Linear Model

Model Summary:

Call:

lm(formula = delivery\_time ~ sorting\_time, data = dvl.st)

Residuals:

Min 1Q Median 3Q Max

-5.1729 -2.0298 -0.0298 0.8741 6.6722

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.5827 1.7217 3.823 0.00115 \*\*

sorting\_time 1.6490 0.2582 6.387 3.98e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.935 on 19 degrees of freedom

Multiple R-squared: 0.6823, Adjusted R-squared: 0.6655

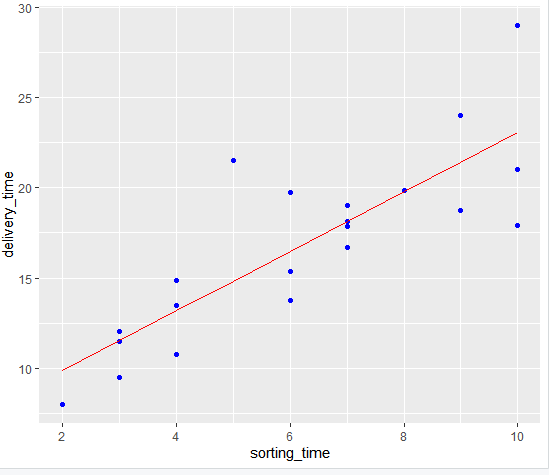
F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06

R-Squared = 0.6823. 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

1.) Square root Model

X: squareroot(sorting\_time)

Y: delivery\_time

Model Summary

lm(formula = delivery\_time ~ sqrt(sorting\_time), data = dvl.st)

Residuals:

Min 1Q Median 3Q Max

-4.6789 -1.7277 -0.3694 0.8023 6.4211

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.519 2.995 -0.841 0.411

sqrt(sorting\_time) 7.937 1.204 6.592 2.61e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.872 on 19 degrees of freedom

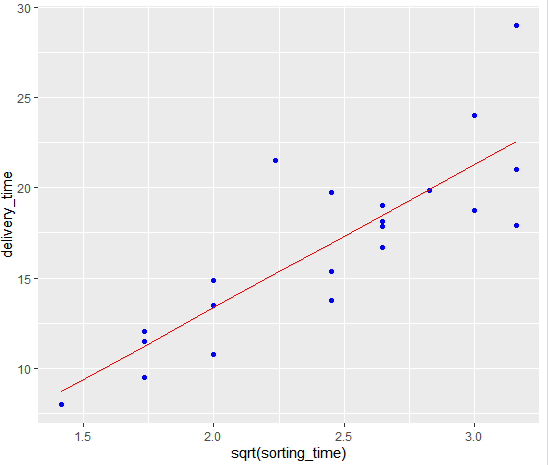
Multiple R-squared: 0.6958, Adjusted R-squared: 0.6798

F-statistic: 43.46 on 1 and 19 DF, p-value: 2.611e-06

**R-squared= 0.6958**

**RMSE = 2.73**

**Regression Line**



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

2.) Logarithmic model

X: log(sorting\_time)

Y: delivery\_time

Cor-coef = 0.83

Model Summary:

Call:

lm(formula = delivery\_time ~ log(sorting\_time), data = dvl.st)

Residuals:

Min 1Q Median 3Q Max

-4.0829 -2.0133 -0.1965 0.9351 7.0171

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.160 2.455 0.472 0.642

log(sorting\_time) 9.043 1.373 6.587 2.64e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.873 on 19 degrees of freedom

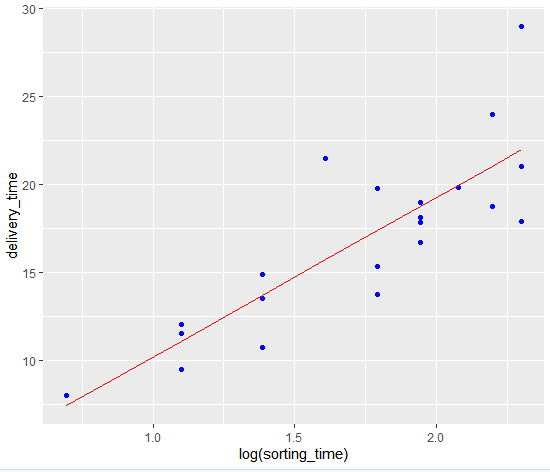
Multiple R-squared: 0.6954, Adjusted R-squared: 0.6794

F-statistic: 43.39 on 1 and 19 DF, p-value: 2.642e-06

R-squared: 0.69

RMSE = 2.73

Regression Line:



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

3.) Exponential Model

X: sorting\_time

Y: log(delivery\_time)

Cor-coef = 0.84

Model Summary

Call:

lm(formula = log(delivery\_time) ~ sorting\_time)

Residuals:

Min 1Q Median 3Q Max

-0.29209 -0.13364 0.02065 0.08421 0.41892

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.12137 0.10297 20.601 1.86e-14 \*\*\*

sorting\_time 0.10555 0.01544 6.836 1.59e-06 \*\*\*

---

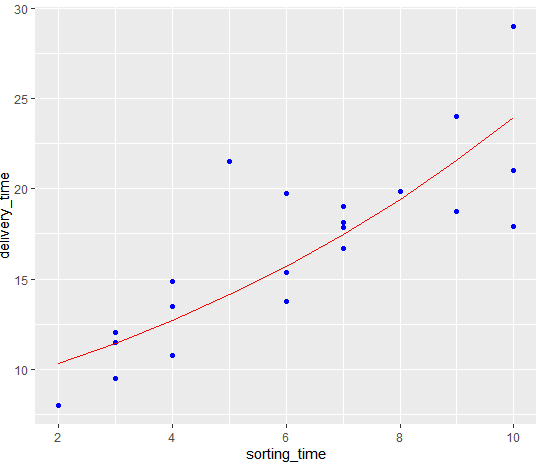
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1755 on 19 degrees of freedom

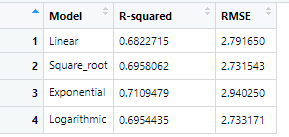
Multiple R-squared: 0.7109, Adjusted R-squared: 0.6957

F-statistic: 46.73 on 1 and 19 DF, p-value: 1.593e-06

R-Squared=0.7109

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

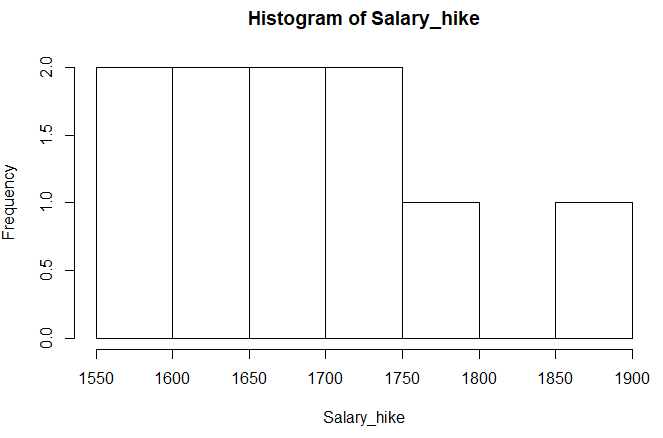
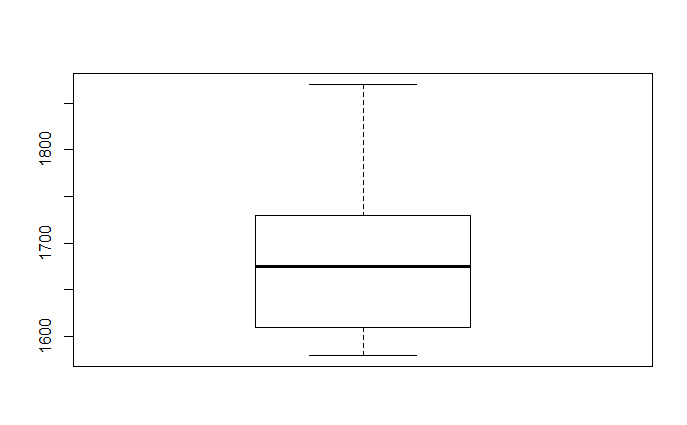
**Datasets:-**

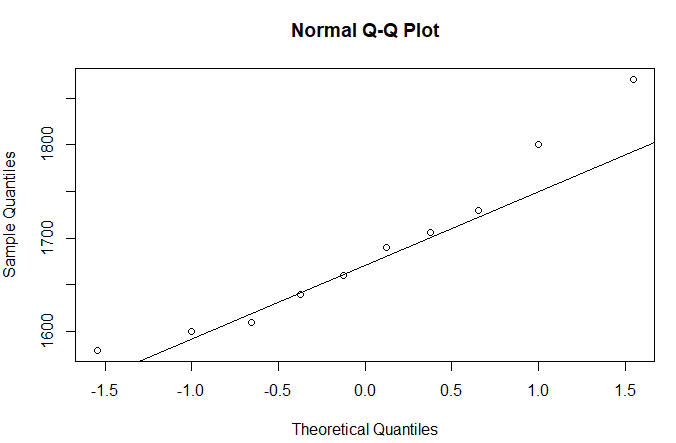
Independent variable(x): salary\_hike

Depenedent variable(y): churn\_out\_rate

**EDA:-**

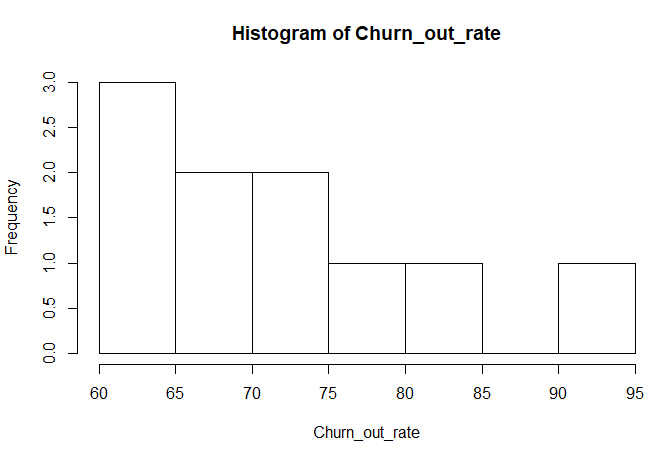
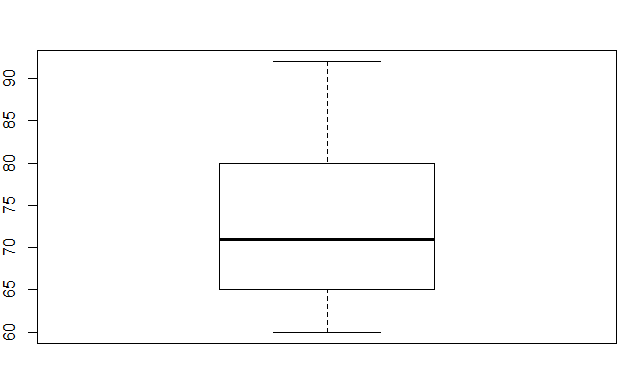
For Salary Hike

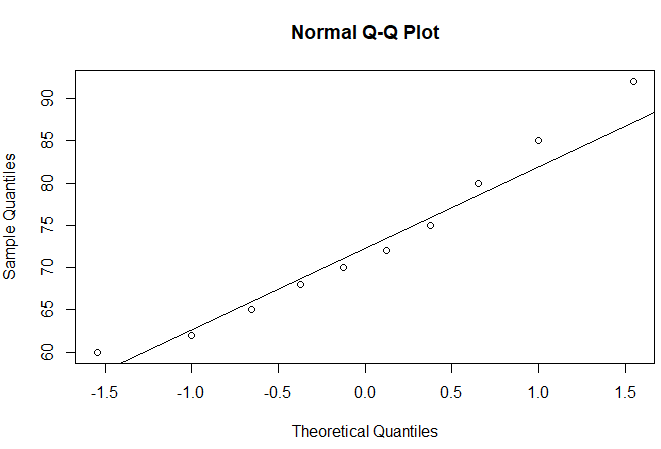
 



From the histogram and box plot, it is found that the data set is having positive skewness and the tail is extended towards right, Moreover the qq plot shows that the most of the datapoints are not on the line. Thus we can conclude that data is not normal.

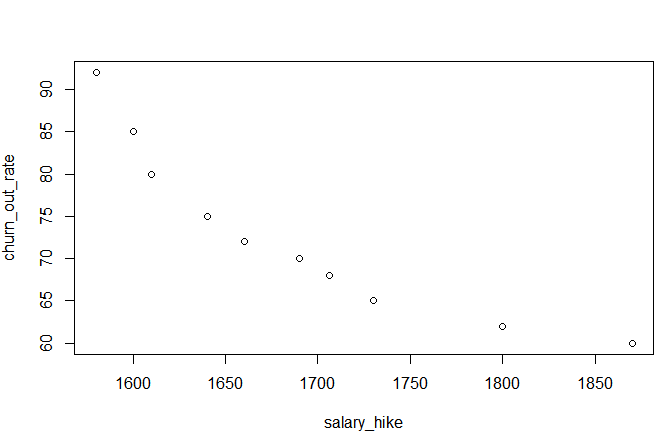
For Churn out rate



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**

Linear Model

Model Summary:

summary(reg\_sal\_churn)

Call:

lm(formula = Churn\_out\_rate ~ Salary\_hike, data = emp\_data)

Residuals:

Min 1Q Median 3Q Max

-3.804 -3.059 -1.819 2.430 8.072

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\*

Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.469 on 8 degrees of freedom

Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101

F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

The Multiple R-square value is 0.8312 is quite high, which indicates there is good fit of the data points in the regression line.

Moreover the 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**

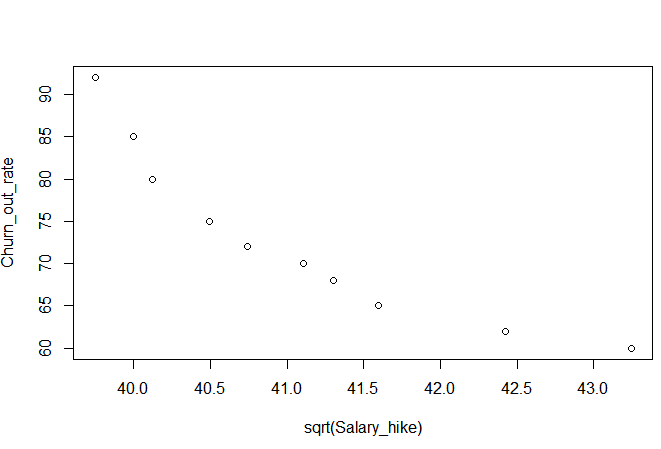
Now Transformation:

1. Square Root Model

X= square root (salary\_hike)

Y= churn\_out\_rate

Scatter plot



Cor = -0.91

This is showing there is strong corealation with negative relationship between two models.

Model Summary :

Summary: - Call:

lm(formula = Churn\_out\_rate ~ sqrt(Salary\_hike), data = emp\_data)

Residuals:

Min 1Q Median 3Q Max

-3.743 -2.955 -1.808 2.353 7.848

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 420.468 53.643 7.838 5.06e-05 \*\*\*

sqrt(Salary\_hike) -8.461 1.305 -6.481 0.000192 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.351 on 8 degrees of freedom

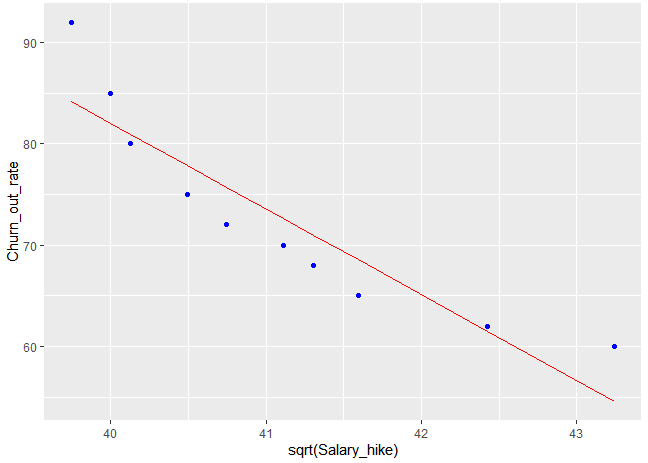
Multiple R-squared: 0.84, Adjusted R-squared: 0.82

F-statistic: 42.01 on 1 and 8 DF, p-value: 0.0001918

R 2 = 0.84 This is a little better model compared to previous model

p- value is less than 0.05

RMSE = 3.891995



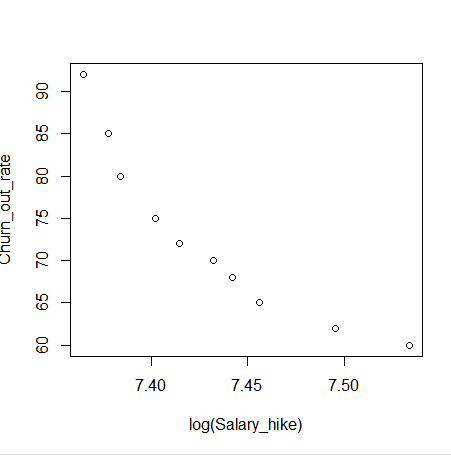
Regression Line : -

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

2. Logarithmic Model

Independent variable = log(salary\_hike)

Dependent variable = churn\_out\_rate



Cor = -0. 92

Model Building

Summary

lm(formula = Churn\_out\_rate ~ log(Salary\_hike), data = emp\_data)

Residuals:

Min 1Q Median 3Q Max

-3.678 -2.851 -1.794 2.275 7.624

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1381.5 195.4 7.070 0.000105 \*\*\*

log(Salary\_hike) -176.1 26.3 -6.697 0.000153 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

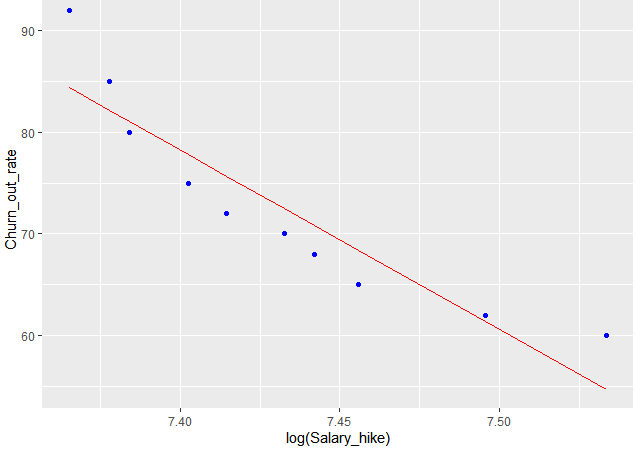
Residual standard error: 4.233 on 8 degrees of freedom

Multiple R-squared: 0.8486, Adjusted R-squared: 0.8297

F-statistic: 44.85 on 1 and 8 DF, p-value: 0.0001532

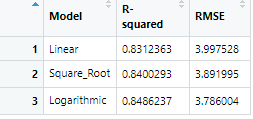
R-squared = 0.8486

RMSE = 3.786004



Regression Line :-

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



In this case, 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:-**

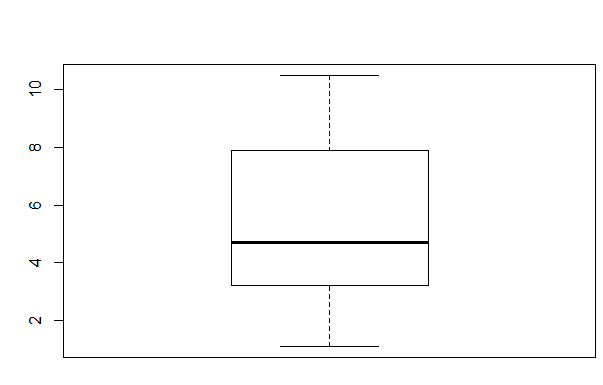
Independent Variable (x): Yearsexperience

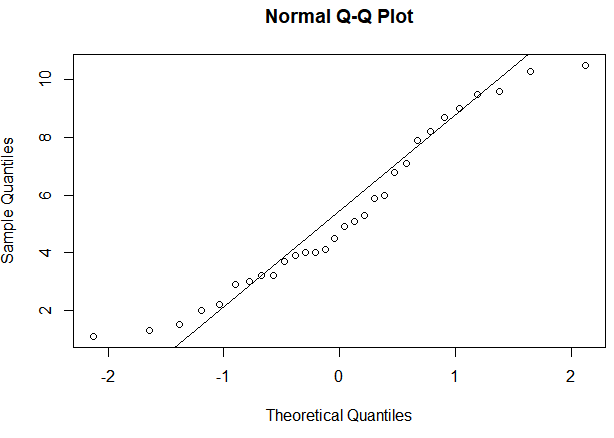
Dependent Variable(y): Salary

**EDA**

Graphical Representation

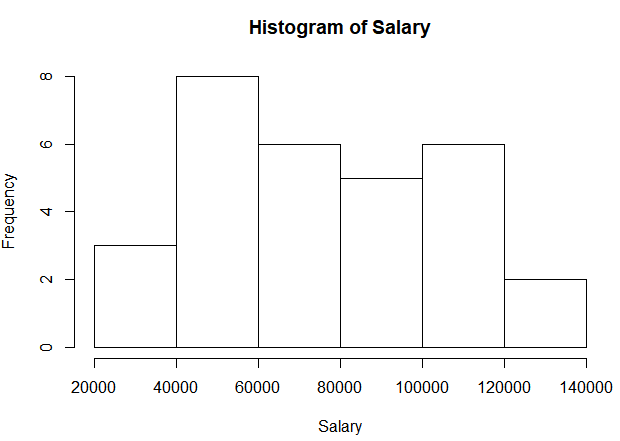
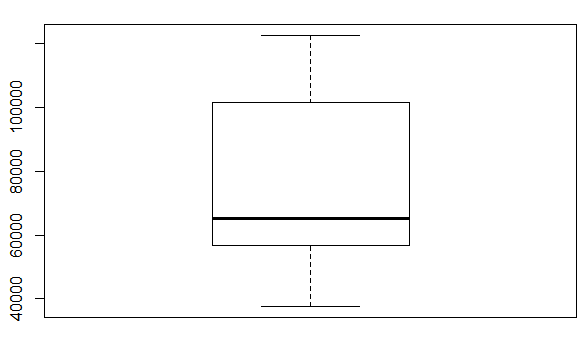
For YearsExperience

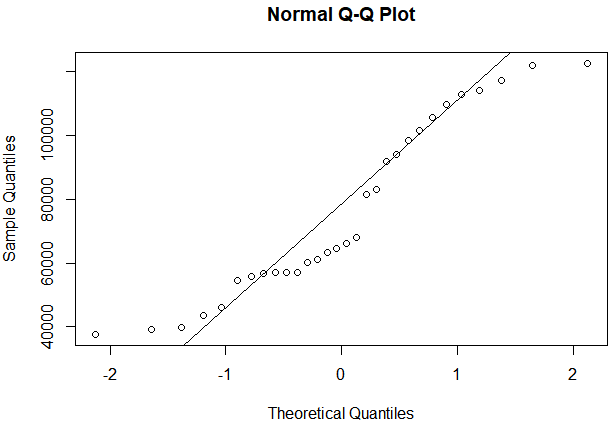
 



From the above plot, it can be said that the data is not normal and having slightly positive skewness. From the qq plot, it can be represented that the most of the data points are away from the optimal line.

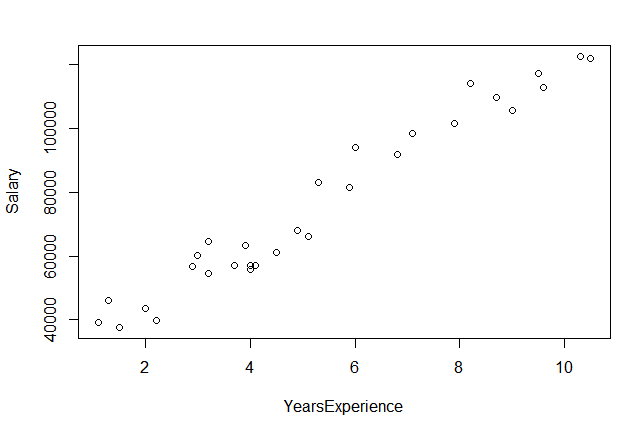
For Salary



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**

Linear Model

Model Summary:

Call:

lm(formula = Salary ~ YearsExperience, data = salary\_data)

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

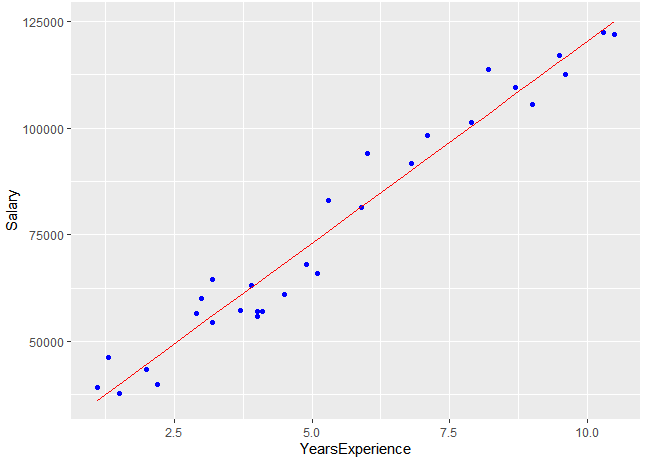
R- Squared = 0.957

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 = 25792.200 + 9449.962 Years of experience

**Evaluation**

Transformation

1.) Square Root Model

x: square root(YearsExperience)

y: Salary

Cor- coeff = 0.964 (It is reduced)

Model Summary

Call:

lm(formula = Salary ~ sqrt(YearsExperience), data = salary\_data)

Residuals:

Min 1Q Median 3Q Max

-11637 -5141 801 4180 14943

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -16056 4922 -3.262 0.00291 \*\*

sqrt(YearsExperience) 41501 2135 19.437 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7329 on 28 degrees of freedom

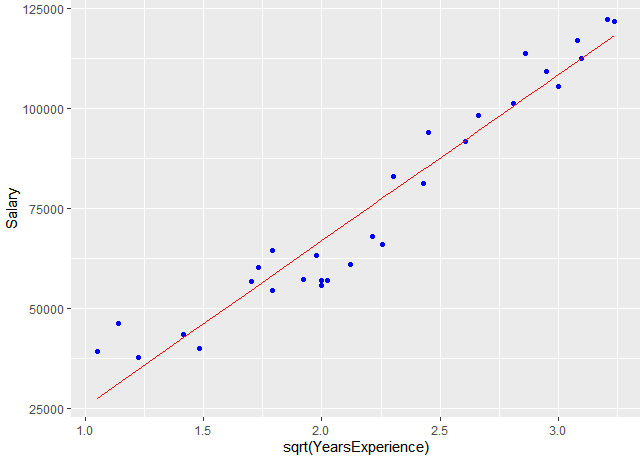
Multiple R-squared: 0.931, Adjusted R-squared: 0.9285

F-statistic: 377.8 on 1 and 28 DF, p-value: < 2.2e-16

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

2.) Logarithmic Model

X: log(YearsExperience)

Y: Salary

Cor-coef = 0.92

Model Summary:

Call:

lm(formula = Salary ~ log(YearsExperience), data = salary\_data)

Residuals:

Min 1Q Median 3Q Max

-15392.6 -7523.0 559.7 6336.1 20629.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 14928 5156 2.895 0.00727 \*\*

log(YearsExperience) 40582 3172 12.792 3.25e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

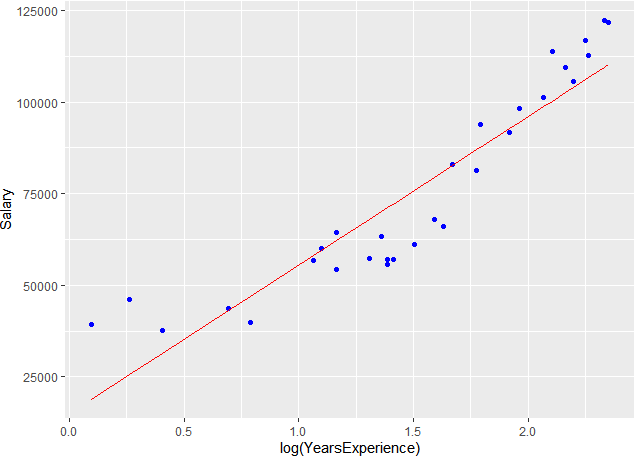
Residual standard error: 10660 on 28 degrees of freedom

Multiple R-squared: 0.8539, Adjusted R-squared: 0.8487

F-statistic: 163.6 on 1 and 28 DF, p-value: 3.25e-13

R-Squared = 0.85 (Further reduced from previous model)

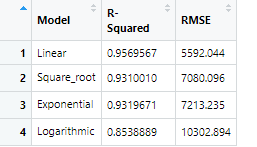
RMSE = 10302.89 (It increased further)



|  |
| --- |
| Regression Line : - Salary = 14927.97 + 40581.99 log(YearsExperience)  3.) Exponential model  X: YearsExperience  Y: log(Salary)  Cor-coef = 0.9653844  Model Summary  lm(formula = log(Salary) ~ YearsExperience)  Residuals:  Min 1Q Median 3Q Max  -0.18949 -0.06946 -0.01068 0.06932 0.19029  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 10.507402 0.038443 273.33 <2e-16 \*\*\*  YearsExperience 0.125453 0.006406 19.59 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.09789 on 28 degrees of freedom  Multiple R-squared: 0.932, Adjusted R-squared: 0.9295  F-statistic: 383.6 on 1 and 28 DF, p-value: < 2.2e-16  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.