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

**Ans :**

**R Code :**

## Simple Linear Regression

########## Emp Data Set #########

emp <- read.csv('D:\\Data Science\\Excelr\\Assignments\\Assignment\\Simple Linear Regression\\emp\_data.csv')

CR <- emp$Churn\_out\_rate

SH <- emp$Salary\_hike

boxplot(CR, col="dodgerblue4")

boxplot(SH,col="dodgerblue4")

plot(SH,CR,main="Scatter Plot",

col="Dodgerblue4",

col.main="Dodgerblue4",

col.lab="Dodgerblue4",

xlab="Churn Out Rate",

ylab="Salary Hike", pch=20)

reg.model<-lm(CR~SH, data=emp)

summary(reg.model)

plot(SH,CR,main="Line of Best Fit",col="Dodgerblue4", col.main="Dodgerblue4")

abline(reg.model, col="red")

res <- signif(residuals(reg.model))

pre <- predict(reg.model)

segments(SH, CR, SH, pre)

predict(reg.model,newdata= data.frame(SH=c(2000,2200)))

###### Squared Data Transformation ###########

SHsqr <- SH \* SH

empsqr <- cbind(emp,SHsqr)

reg.modelsqr<-lm(CR~SHsqr, data=empsqr)

summary(reg.modelsqr)

plot(SHsqr,CR,main="Line of Best Fit",col="Dodgerblue4", col.main="Dodgerblue4")

abline(reg.modelsqr, col="red")

res <- signif(residuals(reg.modelsqr))

pre <- predict(reg.modelsqr)

segments(SHsqr, CR, SHsqr, pre)

###### Square-Root Data Transformation ###########

SHsqrt <- sqrt(SH)

empsqrt <- cbind(emp,SHsqrt)

reg.modelsqrt<-lm(CR~SHsqrt, data=empsqrt)

summary(reg.modelsqrt)

plot(SHsqrt,CR,main="Line of Best Fit",col="Dodgerblue4", col.main="Dodgerblue4")

abline(reg.modelsqrt, col="red")

res <- signif(residuals(reg.modelsqrt))

pre <- predict(reg.modelsqrt)

segments(SHsqrt, CR, SHsqrt, pre)

###### Log Data Transformation ###########

SHlog <- log(SH)

emplog <- cbind(emp,SHlog)

reg.model.log<-lm(CR~SHlog, data=emplog)

summary(reg.model.log)

plot(SHlog,CR,main="Line of Best Fit",col="Dodgerblue4", col.main="Dodgerblue4")

abline(reg.model.log, col="red")

res <- signif(residuals(reg.model.log))

pre <- predict(reg.model.log)

segments(SHlog, CR, SHlog, pre)

**Results :**

> reg.model<-lm(CR~SH, data=emp)

> summary(reg.model)

Call:

lm(formula = CR ~ SH, data = emp)

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 \*\*\*

SH -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

> predict(reg.model,newdata= data.frame(SH=c(2000,2200)))

1 2

41.27962 20.97109

> reg.modelsqr<-lm(CR~SHsqr, data=empsqr)

> summary(reg.modelsqr)

Call:

lm(formula = CR ~ SHsqr, data = empsqr)

Residuals:

Min 1Q Median 3Q Max

-3.993 -3.296 -1.766 2.586 8.519

Coefficients:

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

(Intercept) 1.563e+02 1.422e+01 10.99 4.17e-06 \*\*\*

SHsqr -2.918e-05 4.946e-06 -5.90 0.000362 \*\*\*

---

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

Residual standard error: 4.703 on 8 degrees of freedom

Multiple R-squared: 0.8131, Adjusted R-squared: 0.7897

F-statistic: 34.8 on 1 and 8 DF, p-value: 0.0003619

> reg.modelsqrt<-lm(CR~SHsqrt, data=empsqrt)

> summary(reg.modelsqrt)

Call:

lm(formula = CR ~ SHsqrt, data = empsqrt)

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 \*\*\*

SHsqrt -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

> reg.model.log<-lm(CR~SHlog, data=emplog)

> summary(reg.model.log)

Call:

lm(formula = CR ~ SHlog, data = emplog)

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 \*\*\*

SHlog -176.1 26.3 -6.697 0.000153 \*\*\*

---

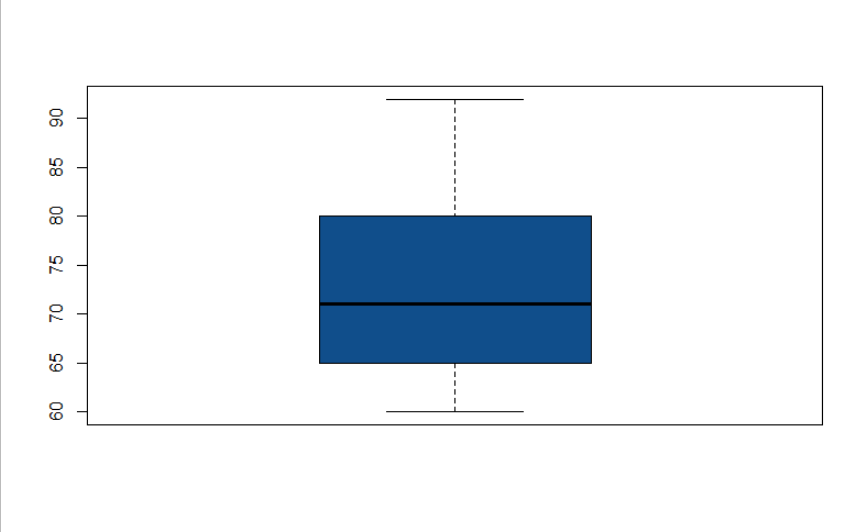
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

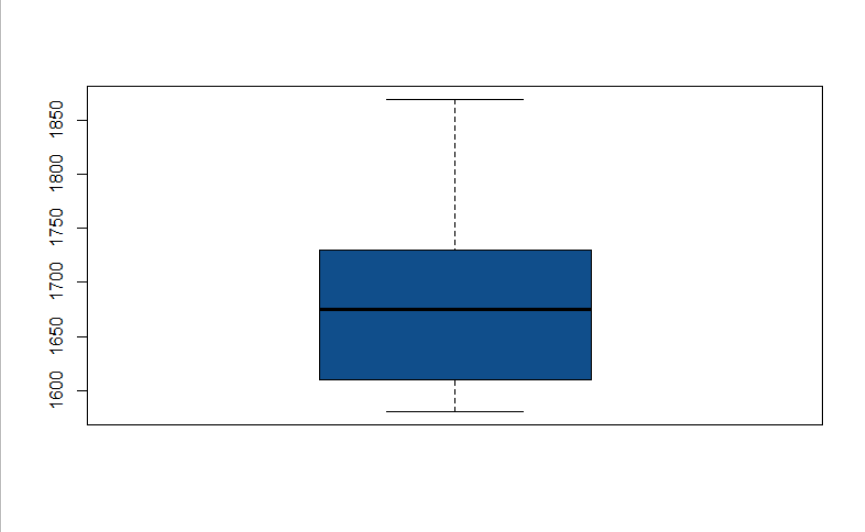
Residual standard error: 4.233 on 8 degrees of freedom

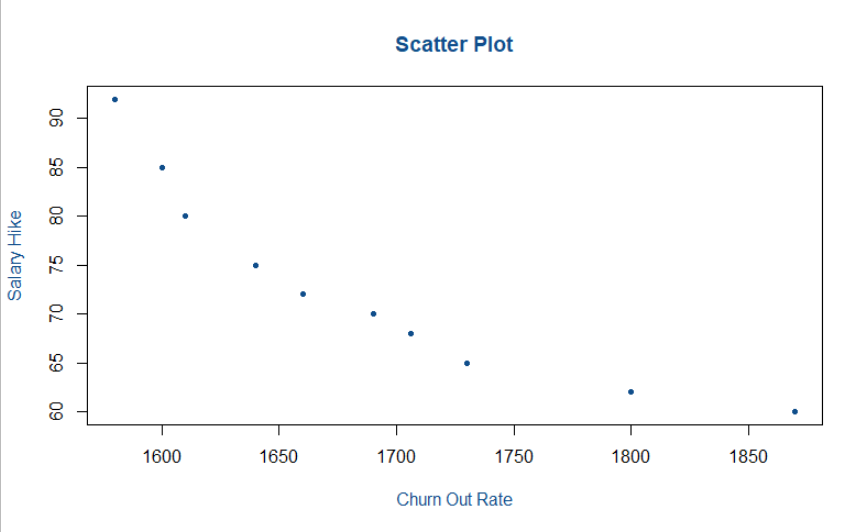
Multiple R-squared: 0.8486, Adjusted R-squared: 0.8297

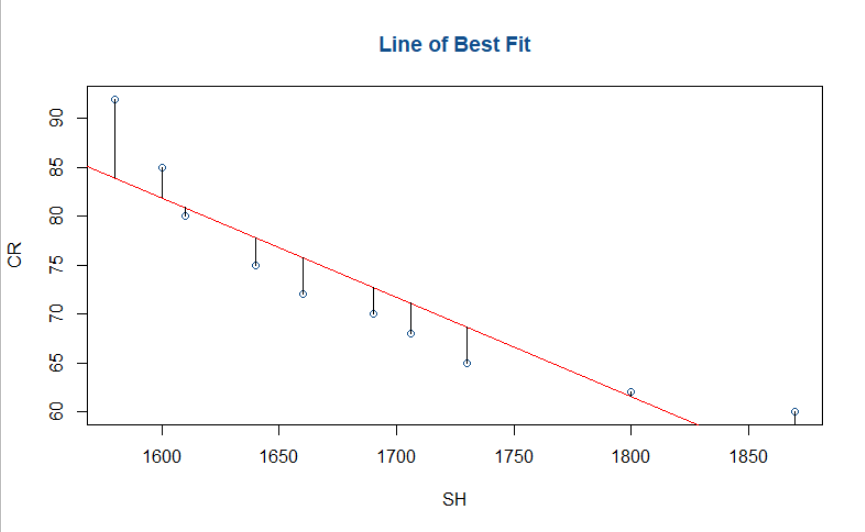
F-statistic: 44.85 on 1 and 8 DF, p-value: 0.0001532

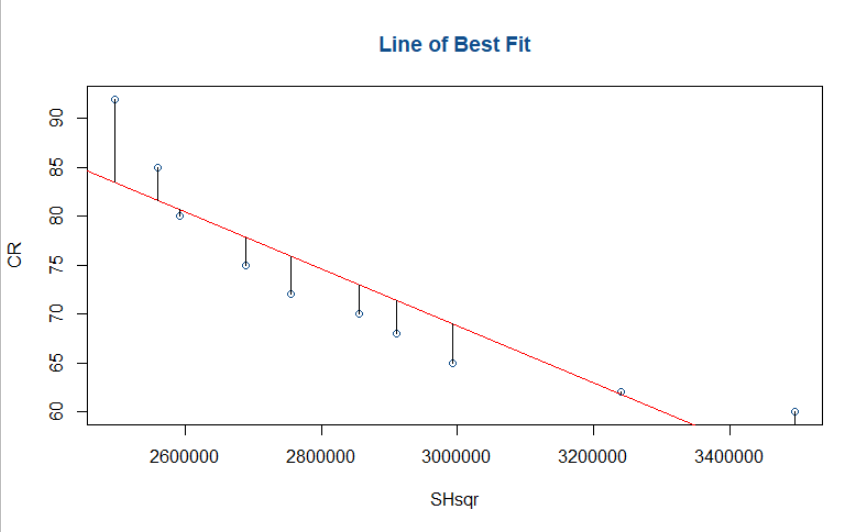
**Plots :**

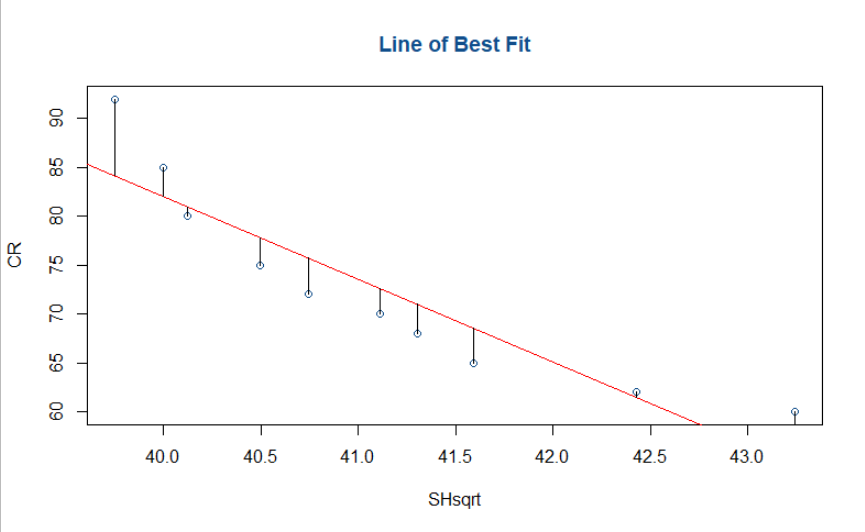


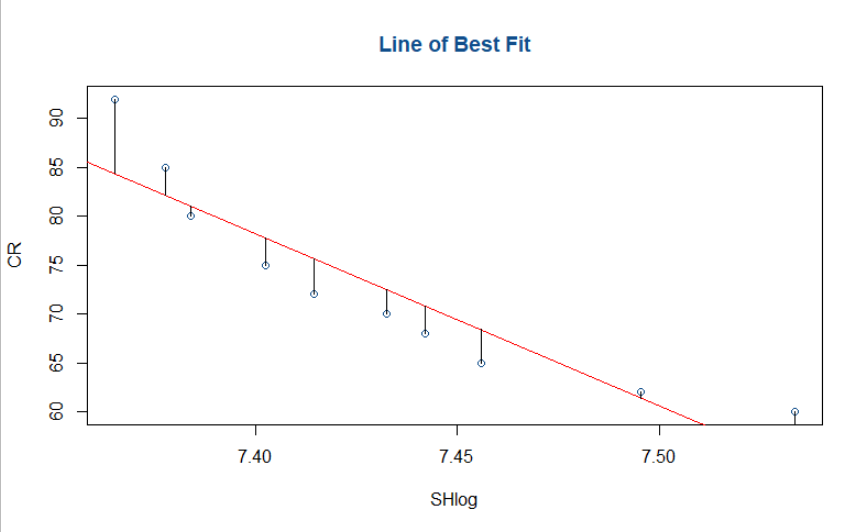












**Inference :**

Performed regression models using various data transformations.

The best results we are getting in Log Data Transformation.

|  |  |  |
| --- | --- | --- |
| **Data Transformation** | **Multiple R-squared** | **Adjusted R-squared** |
| Normal | 0.8312 | 0.8101 |
| Squared | 0.8131 | 0.7897 |
| Square Root | 0.84 | 0.82 |
| Log | 0.8486 | 0.8297 |