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

Ans:

> cor(cal$Weight.gained..grams.,cal$Calories.Consumed)

[1] 0.946991

Greater than 0.85

We have strong correlation

Thus, we can use calories value to find weight gained

Any missing values?

> sum(is.na(cal))

[1] 0

No missing values

Trying Different plots ->

# both are normally distributed

y<-cal$Weight.gained..grams.

x<-cal$Calories.Consumed

plot(x,y)

cor(x,y) #0.946991

plot(sqrt(x),y)

cor(sqrt(x),y)

plot(x,sqrt(y))

cor(x,sqrt(y)) #0.9559736

cor(sqrt(x),sqrt(y))

plot(log(x),y)

cor(log(x),y)

plot(exp(x),y)

cor(exp(x),y)

cor(y,exp(x))

plot(1/x,y)

cor(1/x,y)

cor(x,1/y)

cor(1/x,1/y)

plot(x^2,y)

cor(x^2,y) #0.9710636

cor(x^3,y) #0.971167

cor(x^4,y) #0.9534202

cor(x^3,sqrt(y))

plot(x^1/3,y)

cor(x^1/3,y)

cor(x,sqrt(y))

cor(sqrt(x),y^2)

* **#Base Model**

Modelb <- lm(y~x)

summary(Modelb)

Call:

lm(formula = y ~ x)

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 \*\*\***

x 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**

Modelb\_pred <- (Modelb$fitted.values)

Modelb\_err <- y - Modelb\_pred

RMSE\_Modelb <- sqrt(mean(Modelb\_err^2))

**RMSE\_Modelb**

[1] **103.3025**

* **#Improved model test 1**

Model1<- lm(y~x^3)

summary(Model1)

Call:

lm(formula = y ~ x^3)

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 \*\*\***

x 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**

Model1\_pred <- (Model1$fitted.values)

Model1\_err <- y - Model1\_pred

RMSE\_Model1 <- sqrt(mean(Model1\_err^2))

**RMSE\_Model1**

**[1] 103.3025**

* **#improved model test 2**

Model2<- lm(sqrt(y)~x)

summary(Model2)

Call:

lm(formula = sqrt(y) ~ x)

Residuals:

Min 1Q Median 3Q Max

-4.7448 -1.5770 -0.2277 1.8965 3.8881

Coefficients:

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

(Intercept) -7.1154342 2.2552542 -3.155 **0.0083 \*\***

x 0.0103864 0.0009204 11.285 **9.56e-08 \*\*\***

---

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

Residual standard error: 2.496 on 12 degrees of freedom

Multiple R-squared: **0.9139**, Adjusted R-squared: 0.9067

F-statistic: 127.3 on 1 and 12 DF, **p-value: 9.56e-08**

Model2\_pred <- (Model2$fitted.values^2)

Model2\_err <- y - Model2\_pred

RMSE\_Model2 <- sqrt(mean(Model2\_err^2))

**RMSE\_Model2**

**73.74643**

1. Delivery\_time -> Predict delivery time using sorting time

Answer ->#problem 2

dt <- read.csv(file.choose())

cor(dt$Delivery.Time,dt$Sorting.Time)

y<- dt$Delivery.Time

x<-dt$Sorting.Time

**sum(is.na(dt))**

**[1] 0**

#**No Missing values**

hist(x)

hist(y)

qqnorm(x)

qqline(x)

qqnorm(y)

qqline(y)

plot(x,y)

**cor(x,y) #0.8259973**

plot(sqrt(x),y)

cor(sqrt(x),y) #0.83415

plot(x,sqrt(y))

cor(x,sqrt(y)) #0.8390768

**cor(sqrt(x),sqrt(y)) # 0.8539328**

**> cor(x,y)**

**[1] 0.8259973**

>

> cor(sqrt(x),y)

[1] 0.83415

> cor(x,sqrt(y))

[1] 0.8390768

**> cor(sqrt(x),sqrt(y))**

**[1] 0.8539328**

>

> cor(log(x),y)

[1] 0.8339325

> cor(x,log(y))

[1] 0.8431773

**> cor(log(x),log(y))**

**[1] 0.8787271**

>

> cor(exp(x),y)

[1] 0.6009307

> cor(y,exp(x))

[1] 0.6009307

> cor(exp(x),exp(y))

[1] 0.5230692

>

> cor(1/x,y)

[1] -0.8020811

> cor(x,1/y)

[1] -0.8256215

**> cor(1/x,1/y)**

**[1] 0.9140341**

>

>

> cor(x^2,y)

[1] 0.7939063

> cor(x,y^2)

[1] 0.7763201

> cor(x^2,y^2)

[1] 0.7669344

>

> cor(x^3,y)

[1] 0.7540763

> cor(x^4,y)

[1] 0.7161257

> cor(x^3,sqrt(y))

[1] 0.7468065

* **#Base Model**

> Modelb <- lm(y~x)

> summary(Modelb)

Call:

lm(formula = y ~ x)

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 \*\***

x 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**

> Modelb\_pred <- (Modelb$fitted.values)

> Modelb\_err <- y - Modelb\_pred

> RMSE\_Modelb <- sqrt(mean(Modelb\_err^2))

**> RMSE\_Modelb**

**[1] 2.79165**

* **-#Improved model test 1**

> Model1<- lm(log(y)~log(x))

> summary(Model1)

Call:

lm(formula = log(y) ~ log(x))

Residuals:

Min 1Q Median 3Q Max

-0.23303 -0.09050 -0.00825 0.08897 0.36439

Coefficients:

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

(Intercept) 1.74199 0.13312 13.086 **5.92e-11 \*\*\***

log(x) 0.59752 0.07446 8.024 **1.60e-07 \*\*\***

---

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

Residual standard error: 0.1558 on 19 degrees of freedom

Multiple **R-squared: 0.7722**, Adjusted R-squared: 0.7602

F-statistic: 64.39 on 1 and 19 DF, p-value: 1.602e-07

>

> Model1\_pred <- (exp(Model1$fitted.values))

> Model1\_err <- y - Model1\_pred

>

> RMSE\_Model1 <- sqrt(mean(Model1\_err^2))

> RMSE\_Model1

**[1] 2.745829**

>

* **#improved model test 2**

>

> Model2<- lm(1/y ~ 1/x)

> summary(Model2)

Call:

lm(formula = 1/y ~ 1/x)

Residuals:

Min 1Q Median 3Q Max

-0.031086 -0.014936 -0.009484 0.008505 0.059431

Coefficients:

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

(Intercept) 0.065569 0.004896 13.39 1.91e-11 \*\*\*

---

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

Residual standard error: 0.02244 on 20 degrees of freedom

> plot(1/x,1/y)

> plot(x,y)

> Model2\_pred <- (1/Model2$fitted.values)

> Model2\_err <- y - Model2\_pred

>

> RMSE\_Model2 <- sqrt(mean(Model2\_err^2))

**> RMSE\_Model2**

**[1] 5.186454**

>

* **#improved model test 2**

>

> Model3<- lm(sqrt(y) ~ sqrt(x))

> summary(Model3)

Call:

lm(formula = sqrt(y) ~ sqrt(x))

Residuals:

Min 1Q Median 3Q Max

-0.55193 -0.20240 -0.00951 0.11906 0.78231

Coefficients:

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

(Intercept) 1.6135 0.3486 4.628 **0.000183 \*\*\***

sqrt(x) 1.0022 0.1401 7.153 **8.48e-07 \*\*\***

---

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

Residual standard error: 0.3342 on 19 degrees of freedom

Multiple **R-squared: 0.7292**, Adjusted R-squared: 0.7149

F-statistic: 51.16 on 1 and 19 DF, **p-value: 8.476e-07**

>

> Model3\_pred <- (Model3$fitted.values^2)

> Model3\_err <- y - Model3\_pred

>

> RMSE\_Model3 <- sqrt(mean(Model3\_err^2))

**> RMSE\_Model3**

**[1] 2.762294**

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

> x<- cr$Salary\_hike

> y<-cr$Churn\_out\_rate

>

**> sum(is.na(dt))**

**[1] 0**

#No missing values

> hist(x)

> hist(y)

>

> qqnorm(x)

> qqline(x)

>

> qqnorm(y)

> qqline(y)

>

> plot(x,y)

>

**> cor(x,y)**

**[1] -0.9117216**

>

> cor(sqrt(x),y)

[1] -0.9165311

> cor(x,sqrt(y))

[1] -0.9235755

> cor(sqrt(x),sqrt(y))

[1] -0.9280927

>

> cor(log(x),y)

[1] -0.9212077

> cor(x,log(y))

[1] -0.9346361

**> cor(log(x),log(y))**

**[1] -0.942917**

>

> cor(exp(x),y)

[1] NaN

> cor(y,exp(x))

[1] NaN

> cor(exp(x),exp(y))

[1] NaN

>

> cor(1/x,y)

[1] 0.9301463

**> cor(x,1/y)**

**[1] 0.9541312**

> cor(1/x,1/y)

[1] -0.9674839

>

>

> cor(x^2,y)

[1] -0.9017223

> cor(x,y^2)

[1] -0.8859551

> cor(x^2,y^2)

[1] -0.8749097

>

> cor(x^3,y)

[1] -0.8912506

> cor(x^4,y)

[1] -0.8803527

> cor(x^3,sqrt(y))

[1] -0.9042215

>

>

> cor(x^1/3,y)

[1] -0.9117216

> cor(x,y^1/3)

[1] -0.9117216

> cor(x^1/3,y^1/3)

[1] -0.9117216

>

> cor(x,sqrt(y))

[1] -0.9235755

> cor(sqrt(x),y^2)

[1] -0.8913034

>

* **#Base Model**

>

> Modelb <- lm(y~x)

> summary(Modelb)

Call:

lm(formula = y ~ x)

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 \*\*\***

x -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**

>

> Modelb\_pred <- (Modelb$fitted.values)

> Modelb\_err <- y - Modelb\_pred

>

> RMSE\_Modelb <- sqrt(mean(Modelb\_err^2))

**> RMSE\_Modelb**

**[1] 3.997528**

* **#Improved model test 1**

>

> Model1<- lm(log(y)~log(x))

> summary(Model1)

Call:

lm(formula = log(y) ~ log(x))

Residuals:

Min 1Q Median 3Q Max

-0.04433 -0.03234 -0.01865 0.02737 0.08377

Coefficients:

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

(Intercept) 22.2472 2.2436 9.916 **9.04e-06 \*\*\***

log(x) -2.4180 0.3019 -8.008 **4.33e-05 \*\*\***

---

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

Residual standard error: 0.0486 on 8 degrees of freedom

Multiple **R-squared: 0.8891**, Adjusted R-squared: 0.8752

F-statistic: 64.13 on 1 and 8 DF, **p-value: 4.335e-05**

>

> Model1\_pred <- (exp(Model1$fitted.values))

> Model1\_err <- y - Model1\_pred

>

> RMSE\_Model1 <- sqrt(mean(Model1\_err^2))

**> RMSE\_Model1**

**[1] 3.318328**

* **#improved model test 2**

>

> Model2<- lm(1/y ~ x)

> summary(Model2)

Call:

lm(formula = 1/y ~ x)

Residuals:

Min 1Q Median 3Q Max

-0.0009743 -0.0003475 0.0001896 0.0003924 0.0006278

Coefficients:

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

(Intercept) -1.884e-02 3.643e-03 **-5.172 0.000851 \*\*\***

x 1.942e-05 2.154e-06 **9.014 1.83e-05 \*\*\***

---

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

Residual standard error: 0.0005952 on 8 degrees of freedom

Multiple **R-squared: 0.9104**, Adjusted R-squared: 0.8992

F-statistic: 81.25 on 1 and 8 DF, **p-value: 1.832e-05**

>

> Model2\_pred <- (1/Model2$fitted.values)

> Model2\_err <- y - Model2\_pred

>

> RMSE\_Model2 <- sqrt(mean(Model2\_err^2))

**> RMSE\_Model2**

**[1] 3.164457**

1. Salary\_hike -> Build a prediction model for Salary\_hike

> sh<- read.csv(file.choose())

> x<-sh$YearsExperience

> y<-sh$Salary

> sum(is.na(dt))

[1] 0

#No missing values

> hist(x)

> hist(y)

> qqnorm(x)

> qqline(x)

> qqnorm(y)

> qqline(y)

**> cor(x,y)**

**[1] 0.9782416**

>

> cor(sqrt(x),y)

[1] 0.9648839

**> cor(x,sqrt(y))**

**[1] 0.974595**

> cor(sqrt(x),sqrt(y))

[1] 0.9705406

>

> cor(log(x),y)

[1] 0.9240611

> cor(x,log(y))

[1] 0.9653844

> cor(log(x),log(y))

[1] 0.9514279

>

> cor(exp(x),y)

[1] 0.6873023

> cor(y,exp(x))

[1] 0.6873023

> cor(exp(x),exp(y))

[1] NaN

>

> cor(1/x,y)

[1] -0.7671995

> cor(x,1/y)

[1] -0.9278293

> cor(1/x,1/y)

[1] 0.8777211

>

>

> cor(x^2,y)

[1] 0.9567235

> cor(x,y^2)

[1] 0.9731053

> cor(x^2,y^2)

[1] 0.9767743

>

> cor(x^3,y)

[1] 0.9133658

> cor(x^4,y)

[1] 0.8683542

> cor(x^3,sqrt(y))

[1] 0.8877667

>

>

**> cor(x^1/3,y)**

**[1] 0.9782416**

> cor(x,y^1/3)

[1] 0.9782416

> cor(x^1/3,y^1/3)

[1] 0.9782416

>

> cor(x,sqrt(y))

[1] 0.974595

> cor(sqrt(x),y^2)

[1] 0.9437863

>

* **#Base Model**

>

> Modelb <- lm(y~x)

> summary(Modelb)

Call:

lm(formula = y ~ x)

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 \*\*\***

x 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**

>

> Modelb\_pred <- (Modelb$fitted.values)

> Modelb\_err <- y - Modelb\_pred

>

> RMSE\_Modelb <- sqrt(mean(Modelb\_err^2))

**> RMSE\_Modelb**

**[1] 5592.044**

* **#Improved model test 1**

>

> Model1<- lm(sqrt(y)~x)

> summary(Model1)

Call:

lm(formula = sqrt(y) ~ x)

Residuals:

Min 1Q Median 3Q Max

-18.497 -10.005 -1.529 7.423 23.457

Coefficients:

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

(Intercept) 180.7002 4.4457 40.65 **<2e-16 \*\*\***

x 17.0565 0.7408 23.02 **<2e-16 \*\*\***

---

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

Residual standard error: 11.32 on 28 degrees of freedom

Multiple R-squared: 0.9498, Adjusted R-squared: 0.948

F-statistic: 530.2 on 1 and 28 DF, p-value: < 2.2e-16

>

> Model1\_pred <- (Model1$fitted.values)

> Model1\_err <- y - Model1\_pred

>

> RMSE\_Model1 <- sqrt(mean(Model1\_err^2))

**> RMSE\_Model1**

**[1] 80369.63**

Here, The base model is Strong itself.

Other transformation doesn’t show much promise

So we will finalize the base models itself.