**Data Mining Assignment**

Name :- Magan Uttam Jadhav Roll N0. :- 2117

The data is given for mortality rate for the male and female for the different year (excel file), the excel file is have two data set of same kind, as mortality rate for male and mortality rate for female seperatelly, I was merge those both sheet.

Firstly I have made one excel sheet for those and take age, gender and year as the regressor and mortality rate as the response.

In the given data I use representation for male is 1 and for female is 2. And after that I divide this data into two part as the separate distribution for male (mortality) and for female (mortality).

After that I have done EDA on this new sheet, and fit the different model on this data.

For this analysis I have use correlation analysis, scatter plot, 5-point summary, decision tree model, random forest model, Linear Regression Model, anova, t-SNE method, KL divergence

**##Data**

Data=read.csv("clipboard",sep="\t",header=T)

head(Data)

> > head(Data)

age Gender year Rate

1 0-1 1 1986-1990 0.0928

2 1-5 1 1986-1990 0.0110

3 5-10 1 1986-1990 0.0026

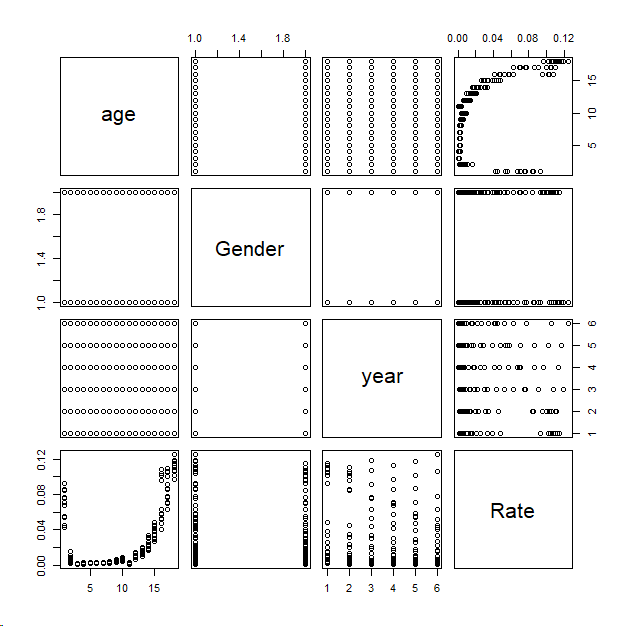
4 10-15 1 1986-1990 0.0014

5 15-20 1 1986-1990 0.0018

6 20-25 1 1986-1990 0.0023

**matrix plot for the given data, this plot shown the correlation between variables.**

plot(Data)



dim(Data)

[1] 216 4 **## dimention of the given data**

library(tidyr)

numeric\_data = Data[, sapply(Data, is.numeric)]

any(is.na(numeric\_data))

[1] FALSE

correlation\_matrix = cor(numeric\_data)

correlation\_matrix **### This is the correlation matrix for numerica variables**

Gender Rate

Gender 1.00000000 -0.04805744

Rate -0.04805744 1.00000000

D1 = subset(Data, Gender == "1");D1

summary(D1) **##@ summary**

age Gender year Rate

Length:108 Min. :1 Length:108 Min. :0.000700

Class :character 1st Qu.:1 Class :character 1st Qu.:0.002475

Mode :character Median :1 Mode :character Median :0.007500

Mean :1 Mean :0.026976

3rd Qu.:1 3rd Qu.:0.041275

Max. :1 Max. :0.124800

> mean(D1[,4]) **## mean of the mortality columns**

[1] 0.02697583

> var(D1[,4]) **## variance of the mortality columns**

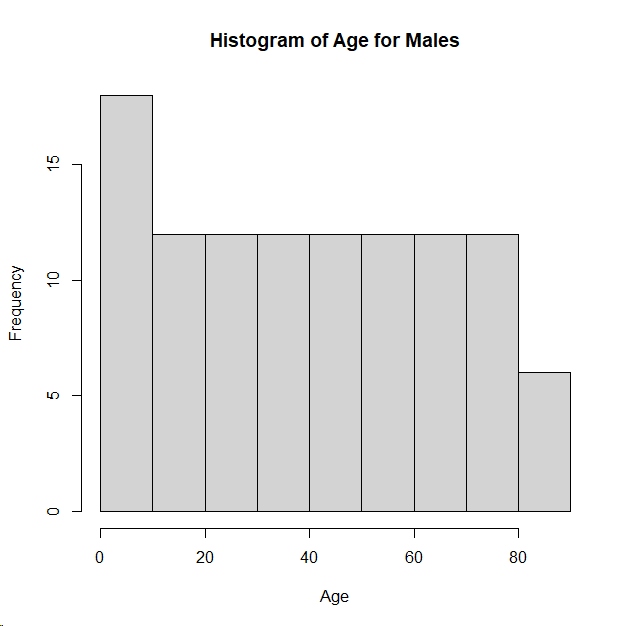
[1] 0.001311513

> sd(D1[,4]) **## standard deviation of the mortality columns**

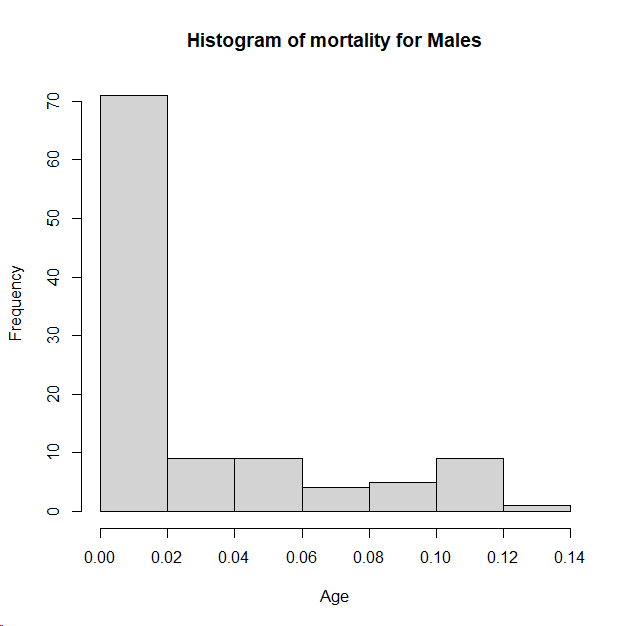
[1] 0.03621482

D1$age\_midpoint = sapply(strsplit(as.character(D1$age), "-"), function(x) mean(as.numeric(x)))

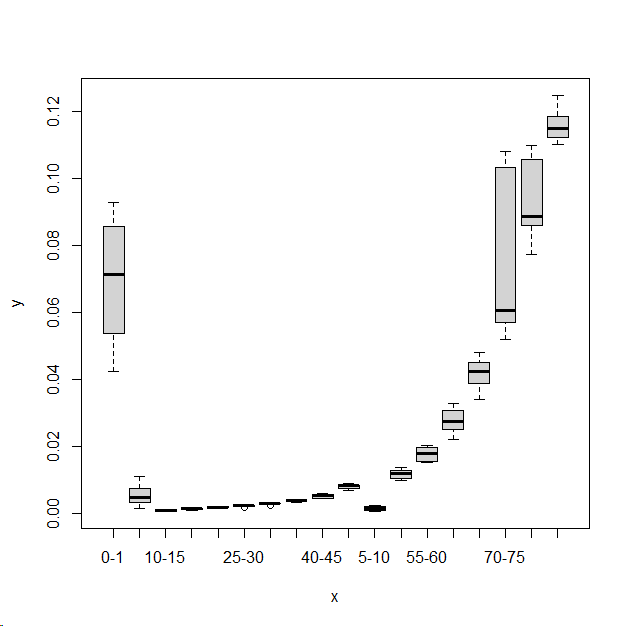
hist(D1$age\_midpoint, main = "Histogram of Age for Males", xlab = "Age")



hist(D1$Rate, main = "Histogram of mortality for Males", xlab = "Age")

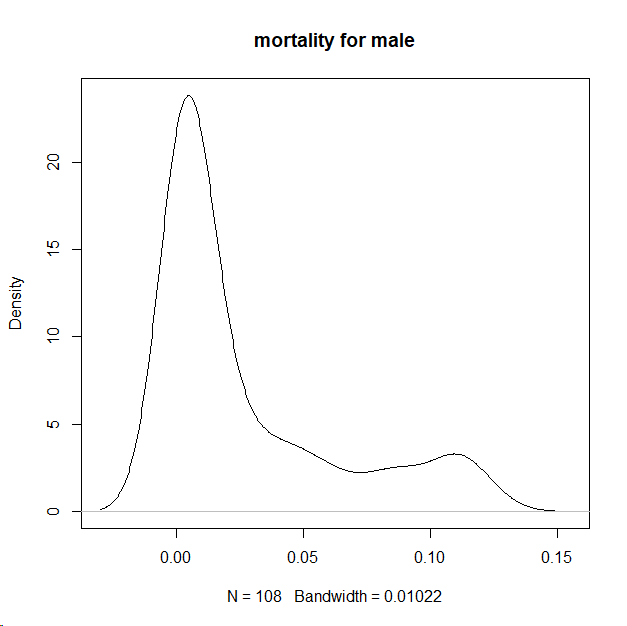


plot(as.factor(D1$age),D1[,4])



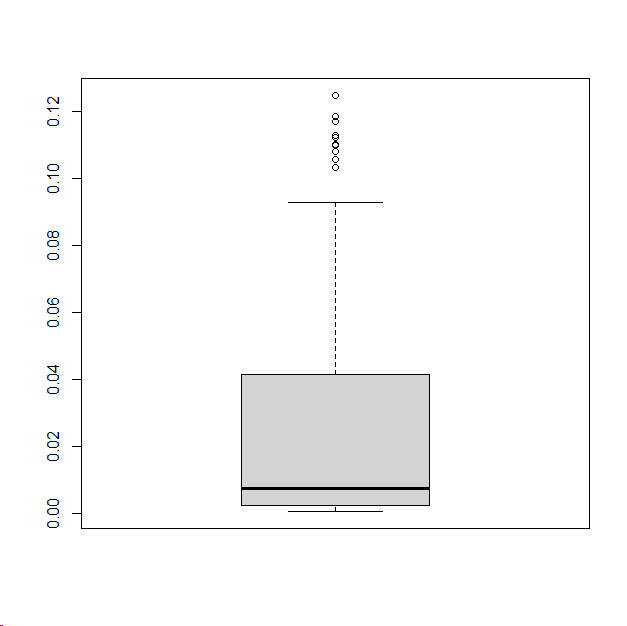
**## from the above graph we conclude that the mortality is converge toward zero (aproximataly)**

plot(density(D1[,4]),main="mortality for male")



**## From this plot we say that the distribution of mortality for the male is positively skewed.**

boxplot(D1[,4]) **## Boxplot for the male mortality**



D2 = subset(Data, Gender == "2");D2

summary(D2) **### Summary foe the male mortality**

age Gender year Rate

Length:108 Min. :2 Length:108 Min. :0.00060

Class :character 1st Qu.:2 Class :character 1st Qu.:0.00255

Mode :character Median :2 Mode :character Median :0.00540

Mean :2 Mean :0.02364

3rd Qu.:2 3rd Qu.:0.03110

Max. :2 Max. :0.11500

> mean(D1[,4]) **## mean for male mortality**

[1] 0.02697583

> var(D1[,4]) **## variance for male mortality**

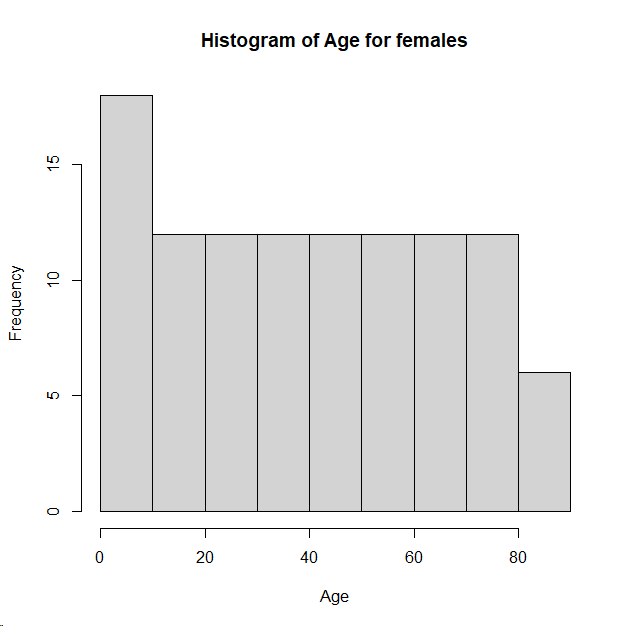
[1] 0.001311513

> sd(D1[,4]) **## standard deviation for the male mortality**

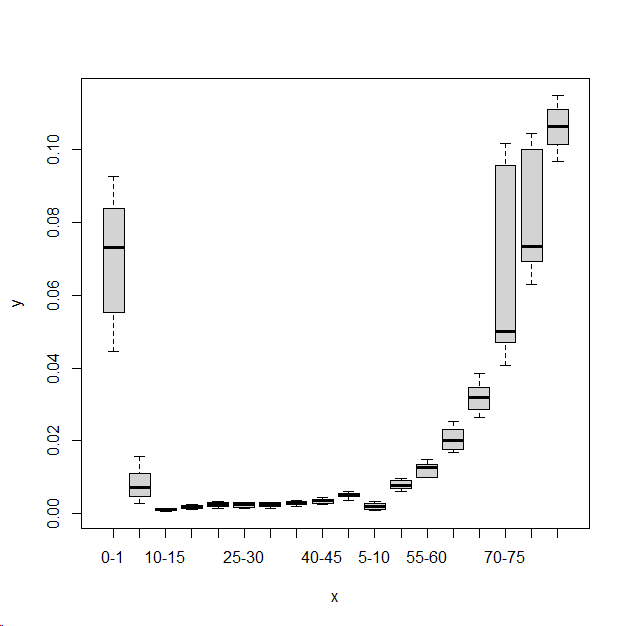
[1] 0.03621482

D2$age\_midpoint = sapply(strsplit(as.character(D2$age), "-"), function(x) mean(as.numeric(x)))

hist(D2$age\_midpoint, main = "Histogram of Age for females", xlab = "Age")

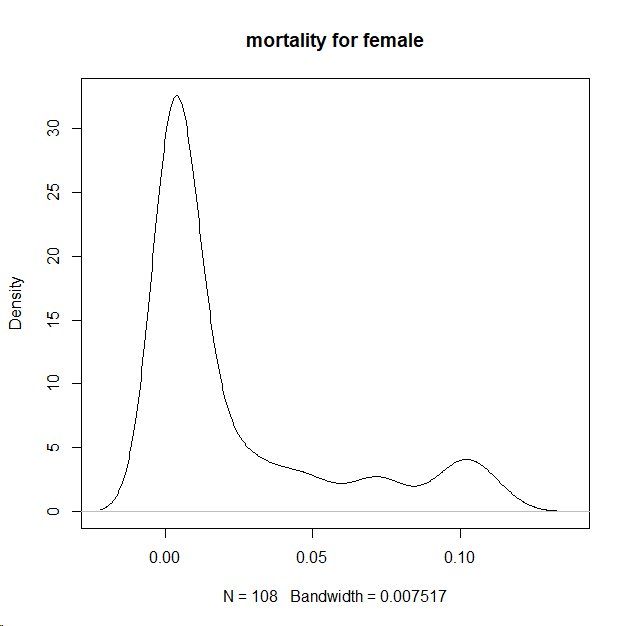


plot(as.factor(D2$age),D2[,4])



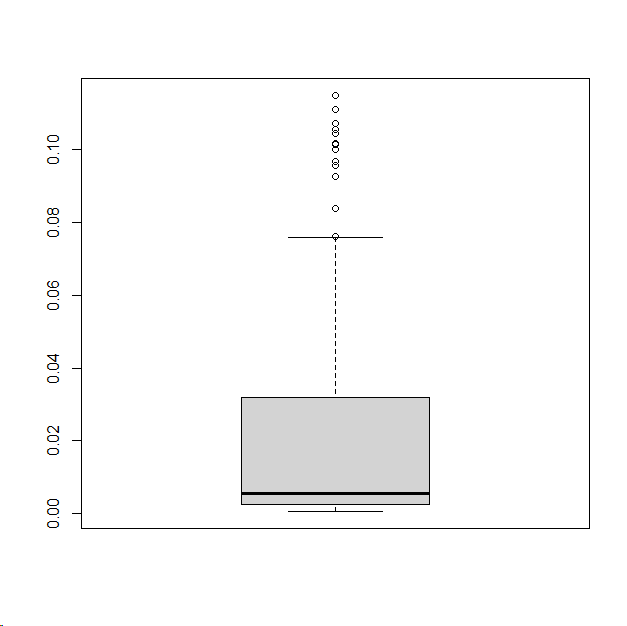
**## from the above graph we conclude that the mortality is converge toward zero (aproximataly)**

plot(density(D2[,4]),main="mortality for female")



**## From this plot we say that the distribution of mortality for the male is positively skewed.**

boxplot(D2[,4])



**# Calculate density for male and female mortalities**

density\_male = density(D1[, 4])

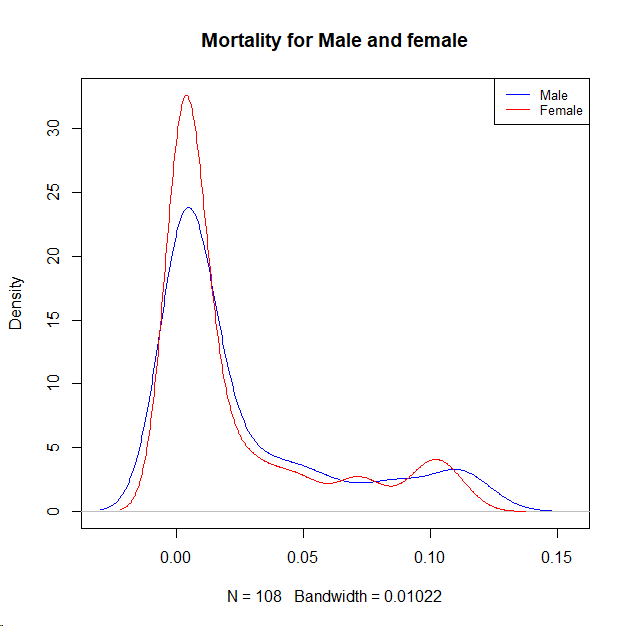
density\_female = density(D2[, 4])

**# Plot both densities on the same graph with different colors**

plot(density\_male, main = "Mortality for Male and female", col = "blue", xlim = range(density\_male$x, density\_female$x), ylim = range(density\_male$y, density\_female$y))

lines(density\_female, col = "red")

legend("topright", legend = c("Male", "Female"), col = c("blue", "red"), lty = 1, cex = 0.8)



**## This is the density plot of the male and female together. From this graph we say that the male mortality is follow heavy tell distribution than the female mortality rate. And the peakedness of the female mortality is sliedly hight than the male mortality.**

#######

X = Data %>% select(age\_midpoint, Gender, year\_midpoint)

y = Data$Rate

**# Splitting the data into training and testing sets**

split = sample(2, nrow(Data), replace = TRUE, prob = c(0.8, 0.2))

X\_train = X[split == 1,]

X\_test = X[split == 2,]

y\_train = y[split == 1]

y\_test = y[split == 2]

**# Linear Regression Model**

linear\_model = lm(y\_train ~ ., data = data.frame(cbind(y\_train, X\_train)))

> summary(linear\_model)

Call:

lm(formula = y\_train ~ ., data = data.frame(cbind(y\_train, X\_train)))

Residuals:

Min 1Q Median 3Q Max

-0.030800 -0.020339 -0.008893 0.005635 0.092252

Coefficients:

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

(Intercept) 1.087e+00 5.221e-01 2.082 0.0388 \*

age\_midpoint 7.587e-04 8.552e-05 8.872 8.77e-16 \*\*\*

GenderFemale -4.589e-03 4.398e-03 -1.043 0.2982

year\_midpoint -5.446e-04 2.608e-04 -2.088 0.0383 \*

---

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

Residual standard error: 0.02922 on 173 degrees of freedom

Multiple R-squared: 0.3318, Adjusted R-squared: 0.3202

F-statistic: 28.64 on 3 and 173 DF, p-value: 4.375e-15

**# Predictions**

linear\_predictions = predict(linear\_model, newdata = data.frame(X\_test))

mse\_linear = mean((linear\_predictions - y\_test)^2); mse\_linear

[1] 0.0007351338

r2\_linear = summary(linear\_model)$r.squared; r2\_linear

[1] 0.3318299

**#### ANOVA**

model = aov(Rate ~ age+year+Gender, data = Data)

model

> model

Call:

aov(formula = Rate ~ age + year + Gender, data = Data)

Terms:

age year Gender Residuals

Sum of Squares 0.24477006 0.00338567 0.00060033 0.01118276

Deg. of Freedom 17 5 1 192

Residual standard error: 0.007631746

Estimated effects may be unbalanced

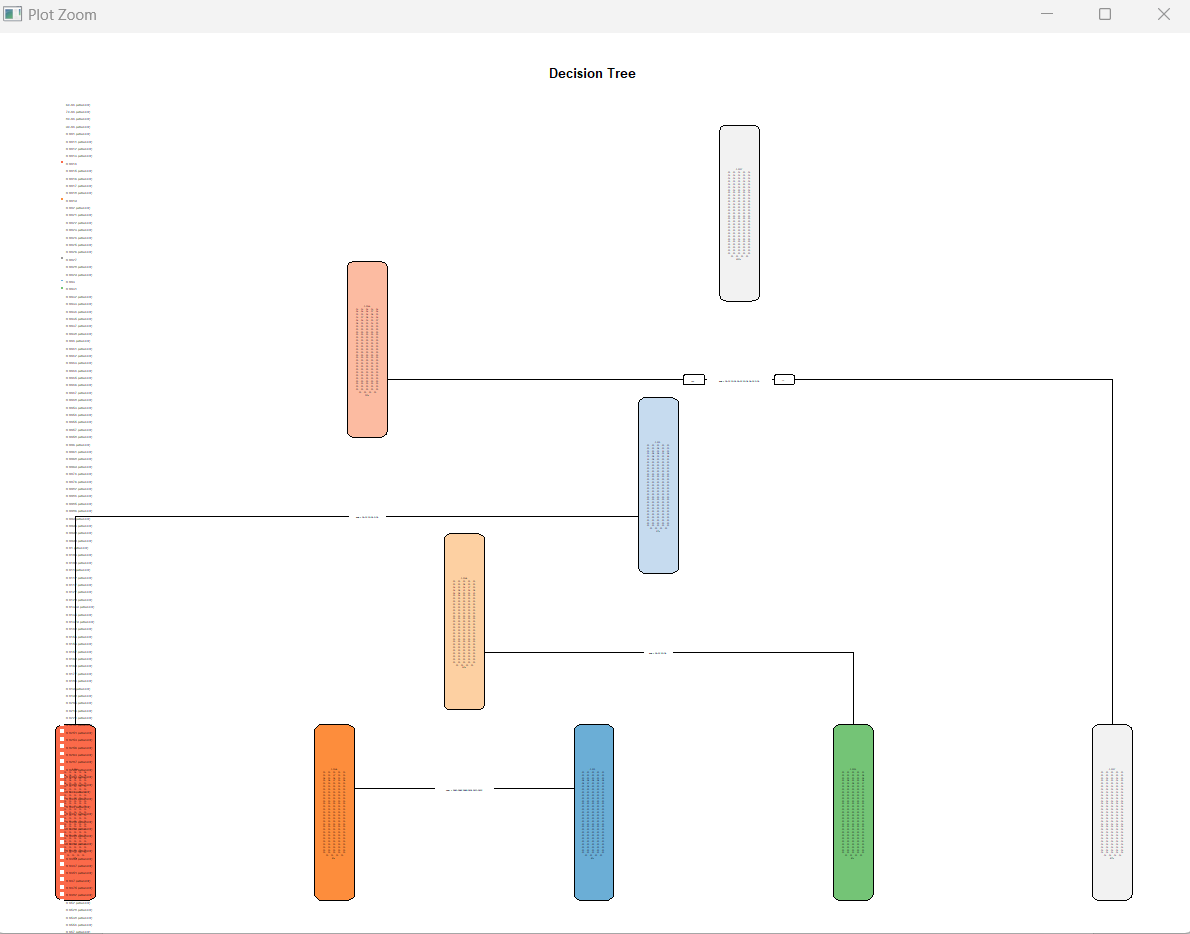
**### Decision Tree**

library(rpart)

library(rpart.plot)

tree\_model = rpart(Rate ~ age + year + Gender , data = Data, method = "class")

rpart.plot(tree\_model, main = "Decision Tree")



predictions = predict(tree\_model, Data, type = "class")

accuracy = mean(predictions == Data$Rate)

accuracy

[1] 0.08796296

rmse = sqrt(mean((predictions - Data$Rate)^2))

rmse

[1] 0.04257183 **### mean square erroe for decision tree**

**# Linear Regression Model by using 4 fold cross validation**

library(caret)

train\_control = trainControl(method = "cv", number = 4)

lm\_model\_cv = train(Rate ~ .,

data = Data[, c("Rate", "age", "year", "Gender")],

method = "lm",

trControl = train\_control)

lm\_model\_cv

> lm\_model\_cv

Linear Regression

216 samples

3 predictor

No pre-processing

Resampling: Cross-Validated (4 fold)

Summary of sample sizes: 162, 162, 161, 163

Resampling results:

RMSE Rsquared MAE

0.008269518 0.9502695 0.005607388

Tuning parameter 'intercept' was held constant at a value of TRUE

**### random forest model with cross-validation (k=4)**

train\_control = trainControl(method = "cv", number = 4)

rf\_model\_cv = train(Rate ~ age + year + Gender,

data = Data,

method = "rf",

trControl = train\_control)

rf\_model\_cv

**Random Forest**

216 samples

3 predictor

No pre-processing

Resampling: Cross-Validated (4 fold)

Summary of sample sizes: 162, 160, 162, 164

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

2 0.02082497 0.8160104 0.016218473

12 0.01409106 0.8258691 0.009940718

23 0.01358445 0.8381929 0.009723926

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 23.

**## Comparison which model is perform better**

lm\_rmse = min(lm\_model\_cv$results$RMSE)

rf\_rmse = min(rf\_model\_cv$results$RMSE)

**# Compare the RMSE values**

if (lm\_rmse < rf\_rmse) {

cat("Linear Regression performs better with RMSE:", lm\_rmse, "\n")

} else if (rf\_rmse < lm\_rmse) {

cat("Random Forest performs better with RMSE:", rf\_rmse, "\n")

} else {

cat("Both models have similar performance with RMSE:", lm\_rmse, "\n")

}

**Linear Regression performs better with RMSE: 0.008269518**

**## Distance between the two distribution (male and female distribution for mortality)**

dist1 = dist(D1[,4])

dist2 = dist(D2[,4])

**# Measure distance between dist1 (male) and dist2 (female) using Euclidean distance**

distance\_euclidean = sqrt(sum((as.matrix(dist1) - as.matrix(dist2))^2))

distance\_euclidean

[1] 0.7573659 **## This is the distance between two distribution**

**The eucludean distance between the distributions of male and femalemortality rates is [1] 0.7573659 . this distance a moderate separation or dissimilarity between the two distribution. Since the distance is not vry large, it suggest that, in terms of mortality rates, there exists some similarity or overlap between male and female distribution.**

**#####**

install.packages("Rtsne")

library(Rtsne)

combined\_data = rbind(D1[,4], D2[,4])

**# Perform t-SNE with 2 dimensions and perplexity with 10**

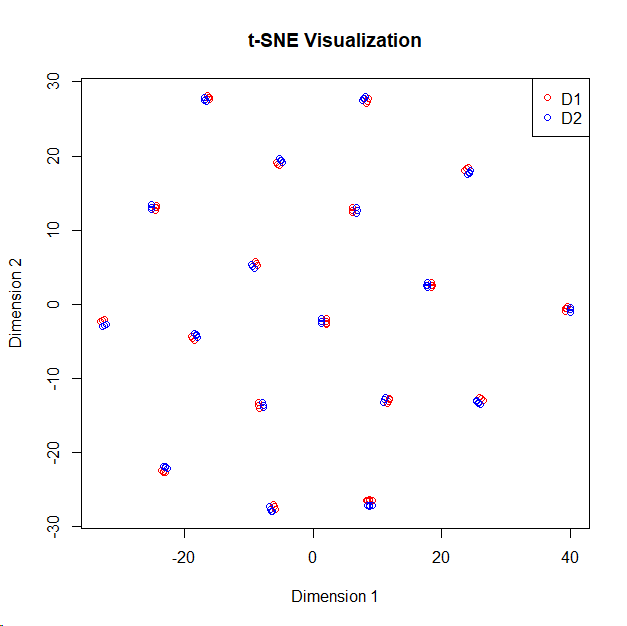
tsne\_result = Rtsne(combined\_data, dims = 2, perplexity = 10)

# Create a scatter plot of the t-SNE result

plot(tsne\_result$Y, col = c(rep("red", nrow(D1)), rep("blue", nrow(D2))),

main = "t-SNE Visualization", xlab = "Dimension 1", ylab = "Dimension 2")

legend("topright", legend = c("D1", "D2"), col = c("red", "blue"), pch = 1)



**## t-SNE coordinates**

tsne\_coordinates = tsne\_result$Y

> tsne\_coordinates

[,1] [,2]

[1,] -8.842991 5.4903610

[2,] 8.762227 -26.4757736

[3,] 8.469656 27.4040868

**# Function to estimate KL divergence with smoothing**

KL\_divergence = function(P, Q, bins = 50, epsilon = 1e-10) {

min\_val = min(min(P), min(Q))

max\_val = max(max(P), max(Q))

hist\_P = hist(P, breaks = seq(min\_val, max\_val, length.out = bins), plot = FALSE)

hist\_Q = hist(Q, breaks = seq(min\_val, max\_val, length.out = bins), plot = FALSE)

**# Normalize histograms to obtain probability distributions**

prob\_P = hist\_P$density / sum(hist\_P$density)

prob\_Q = hist\_Q$density / sum(hist\_Q$density)

**# smoothing to avoid zero probability values**

prob\_P = prob\_P + epsilon

prob\_Q = prob\_Q + epsilon

**# Compute KL divergence**

KL = sum(prob\_P \* log(prob\_P / prob\_Q), na.rm = TRUE)

return(KL)

}

**# Estimate KL divergence between 'D1[,4]' and 'D2[,4]' with smoothing**

estimated\_KL = KL\_divergence(D1[,4], D2[,4])

estimated\_KL

[1] 2.475463

The kullback leibler divergence is a measure of how one probalility distribution is diverges from the other, expected probability distribution. In this case estimate of kl is divergence between the distributions for the male mortality and for the female mortality is [1] 2.475463

By the above analysis we conclude that there is no big difference between mortality of male and mortality of female, means the gender is not affected so much on the mortality of the person.

**## relation between age and mortality**

install.packages("ggplot2")

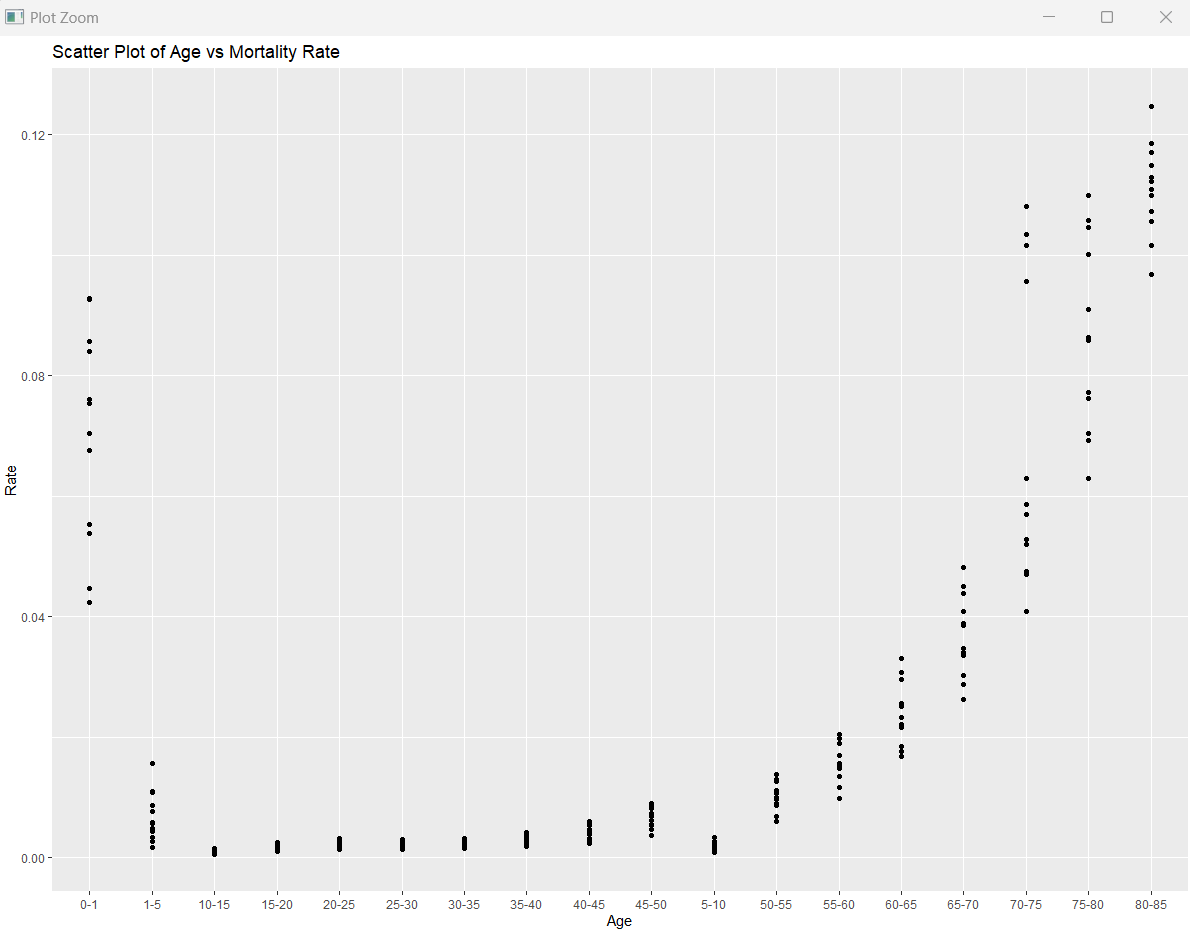
library(ggplot2)

ggplot(Data, aes(x = age, y = Rate)) +

geom\_point() +

labs(x = "Age", y = "Rate") +

ggtitle("Scatter Plot of Age vs Mortality Rate")



**the mortality rate is high in the age group 0-5 and above 60, and for other age group(5-60) is the mortality rate is less as compare to other. But the positive correlation between the mortality and age, means the increses age of the people then increasing the mortality rate as well.**

install.packages("ggplot2")

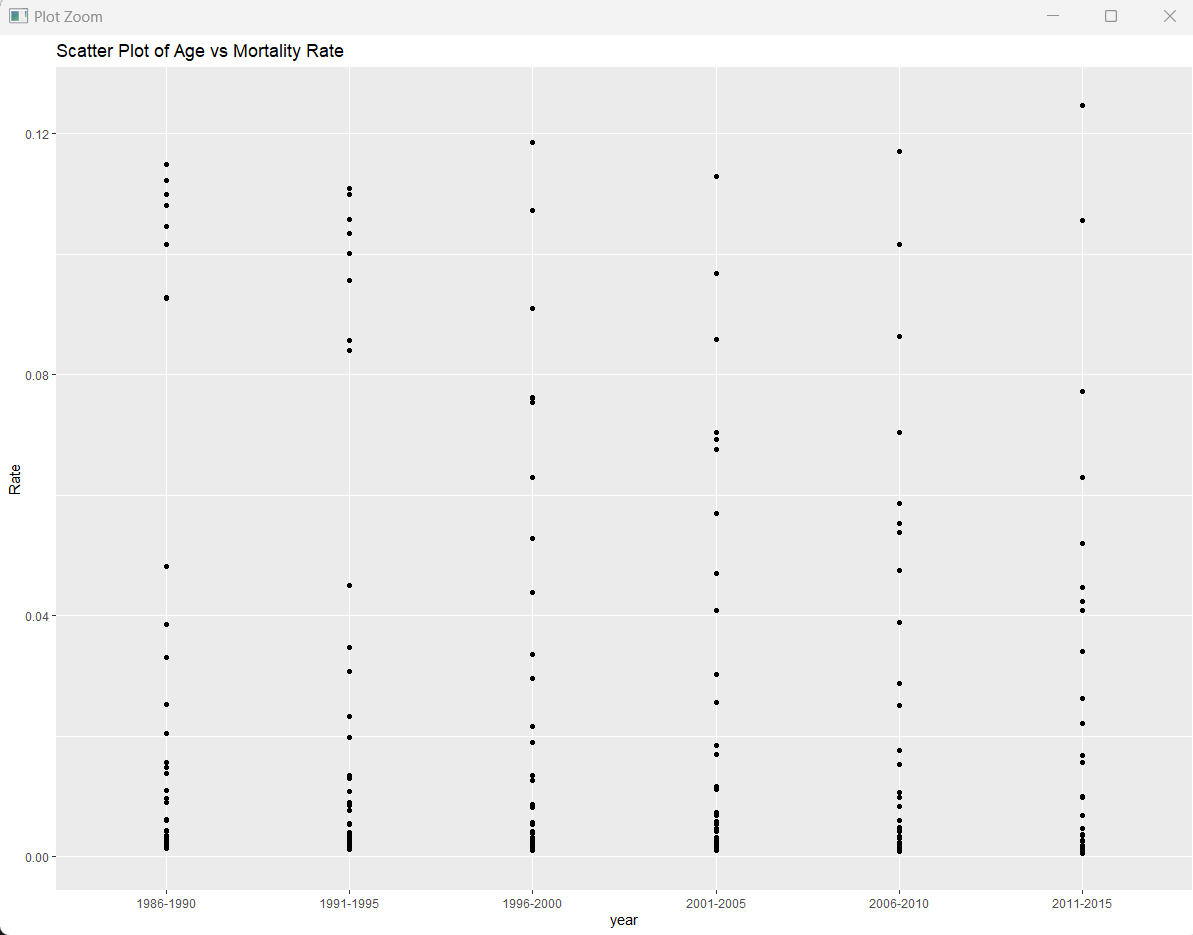
library(ggplot2)

ggplot(Data, aes(x = year, y = Rate)) +

geom\_point() +

labs(x = "year", y = "Rate") +

ggtitle("Scatter Plot of year vs Mortality Rate")



**## there is some inpact of the time on the mortality rate between the age group 0-10 in male and female as well the rate is decreases over the time passes. And for the other age group rather than the age group 0-10 there is not significant inpact of time, means for age group 10-85 mortality rate not change so much**.

**## gender of the people is also not so much related to the mortality rate.**