#Assignment 3 lab1  
library(readxl)  
library(dplyr)  
library(tidyr)  
#Part 1  
diabetes = read.csv('pima-indians-diabetes.csv')  
diabetes1 = as.data.frame(diabetes)  
x = diabetes1[[8]] #Vector of plasma glucose concentration levels  
y = diabetes1[[2]] #Vector of ages  
plot(x,y, col=as.factor(diabetes1$X1), pch=19,   
 main="Plasma glucose concentration on Age",   
 xlab="Age", ylab="Plasma glucose concentration")

#Part 2  
set.seed(12345)  
train=diabetes%>%select(X1, X148, X50)  
m1=glm(as.factor(diabetes1$X1)~., train, family="binomial")  
coef(m1)

## (Intercept) X148 X50   
## -5.89785793 0.03558250 0.02450157

Prob=predict(m1, type="response")  
Pred=ifelse(Prob>0.5, "1", "0")  
table(train$X1, Pred)

summary(m1)

missclass=function(X,X1) {  
 n=length(X)  
 return(1-sum(diag(table(X,X1)))/n)  
}  
  
missclass(as.factor(diabetes1$X1), Pred)

## [1] 0.2659713

#Part 3  
  
plot(x, y, col=as.factor(Pred), pch=19,   
 main="Plasma glucose concentration on Age",   
 xlab="Age", ylab="Plasma glucose concentration")  
#The values in abline below are calculated by hand. Can be seen in the report  
abline(165.7539767, -0.6886066)

#Part 4  
# r = 0.2  
Pred=ifelse(Prob>0.2, "1", "0")  
table(train$X1, Pred)

summary(m1)

plot(x, y, col=as.factor(Pred), pch=19,   
 main="Plasma glucose concentration on Age",   
 xlab="Age", ylab="Plasma glucose concentration")

# r = 0.8  
Pred=ifelse(Prob>0.8, "1", "0")  
table(train$X1, Pred)

summary(m1)

plot(x, y, col=as.factor(Pred), pch=19,   
 main="Plasma glucose concentration on Age",   
 xlab="Age", ylab="Plasma glucose concentration")

#Part 5  
head(df)

z1 = c((diabetes1$X50)^4)  
z2 = c(((diabetes1$X50)^3)\*diabetes1$X148)  
z3 = c(((diabetes1$X50)^2)\*((diabetes1$X148)^2))  
z4 = c(diabetes1$X50\*((diabetes1$X148)^3))  
z5 = c(diabetes$X148^4)  
X1 = diabetes1$X50  
X2 = diabetes1$X148  
df = data.frame(z1, z2, z3, z4, z5, X1, X2)  
head(df)

train=df%>%select(X1, X2, z1, z2, z3, z4, z5)  
m1=glm(as.factor(diabetes1$X1)~., train, family="binomial")  
coef(m1)

Prob=predict(m1, type="response")  
Pred=ifelse(Prob>0.5, "1", "0")  
table(train$X1, Pred)

summary(m1)

missclass(as.factor(diabetes1$X1), Pred)

## [1] 0.2451108

plot(x, y, col=as.factor(Pred), pch=19,   
 main="Plasma glucose concentration on Age",   
 xlab="Age", ylab="Plasma glucose concentration")