CSE 635, Spring 2021, Homework 6
Sima Shafaei

1. **First Create your own variable for survive using ifelse command to convert y/n to 1/0 then select best subset of variables from pclass, sex, age, fare,embarked that have highest R-Square in lm**

Code:

```
d=read.csv("TitanicPassengers1.csv",header = TRUE)
head(d)
dt=na.omit(d)
head(dt)
survived=ifelse(dt$Survived=="Yes",1,0) #convert Survived to numerical variable
sex=ifelse(dt$Sex=="male",1,0) #convert Sex to numerical variable
#================================================
m=lm(survived~dt$Pclass)
summary(m)  #R-squared:  0.1294

m=lm(survived~sex)
summary(m)  #R-squared:  0.2903

m=lm(survived~dt$Age)
summary(m)  #R-squared:  0.005963

m=lm(survived~dt$Fare)
summary(m)  #R-squared:  0.07193

m=lm(survived~dt$Embarked)
summary(m)  #R-squared:  0.03846

m=lm(survived~dt$Pclass+sex)
summary(m)  #R-squared:0.3683

m=lm(survived~dt$Pclass+dt$Age)
summary(m)  #R-squared:0.1804

m=lm(survived~dt$Pclass+dt$Fare)
summary(m)  #R-squared:0.1362

m=lm(survived~dt$Pclass+dt$Embarked)
summary(m)  #R-squared: 0.139

m=lm(survived~sex+dt$Age)
summary(m)  #R-squared: 0.2911

m=lm(survived~sex+dt$Fare)
summary(m)  #R-squared:0.3197
```

```
m=lm(survived~sex+dt$Embarked)
summary(m)  #R-squared:0.3124

m=lm(survived~dt$Age+dt$Fare)
summary(m)  #R-squared:0.08263

m=lm(survived~dt$Age+dt$Embarked)
summary(m)  #R-squared:0.04567

m=lm(survived~dt$Fare+dt$Embarked)
summary(m)  #R-squared: 0.08655

m=lm(survived~dt$Pclass+sex+dt$Age)
summary(m)  #R-squared:0.3902

m=lm(survived~dt$Pclass+sex+dt$Fare)
summary(m)  #R-squared:0.3689

m=lm(survived~dt$Pclass+sex+dt$Embarked)
summary(m)  #R-squared:0.3734

m=lm(survived~dt$Pclass+dt$Age+dt$Fare)
summary(m)  #R-squared:0.1831

m=lm(survived~dt$Pclass+dt$Age+dt$Embarked)
summary(m)  #R-squared:0.1872

m=lm(survived~dt$Pclass+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.1433

m=lm(survived~sex+dt$Age+dt$Fare)
summary(m)  #R-squared:0.322

m=lm(survived~sex+dt$Age+dt$Embarked)
summary(m)  #R-squared:0.3135

m=lm(survived~sex+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.3306

m=lm(survived~dt$Age+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.09753

m=lm(survived~dt$Pclass+sex+dt$Age+dt$Fare)
summary(m)  #R-squared:0.3902

m=lm(survived~dt$Pclass+dt$Age+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.1886
```

```
m=lm(survived~dt$Pclass+sex+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.3736

m=lm(survived~dt$Pclass+sex+dt$Age+dt$Embarked)
summary(m)  #R-squared:0.3939

m=lm(survived~sex+dt$Age+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.333

m=lm(survived~dt$Pclass+sex+dt$Age+dt$Fare+dt$Embarked)
summary(m)  #R-squared:0.3939
```

Results:

	Selected Variable	Residual Error
1	Pclass	0.1294
2	Sex	0.2903
3	Age	0.005963
4	Fare	0.07193
5	Embarked	0.03846
6	Pclass+Sex	0.3683
7	Pclass+Age	0.1804
8	Pclass+Fare	0.1362
9	Pclass+Embarked	0.139
10	Sex+Age	0.2911
11	Sex+Fare	0.3197
12	Sex+Embarked	0.3124
13	Age+Fare	0.08263
14	Age+Embarked	0.04567
15	Fare+Embarked	0.08655
16	Pclass+Sex+Age	0.3902
17	Pclass+Sex+Fare	0.3689
18	Pclass+Sex+Embarked	0.3734
19	Pclass+Age+Fare	0.1831
20	Pclass+Age+Embarked	0.1872
21	Pclass+Fare+Embarked	0.1433
22	Sex+Age+Fare	0.322
23	Sex+Age+Embarked	0.3135
24	Sex+Fare+Embarked	0.3306
25	Age+Fare+Embarked	0.09753
26	Pclass+Sex+Age+Fare	0.3902
27	Pclass+Age+Fare+Embarked	0.1886
28	Pclass+Sex+Fare+Embarked	0.3736
29	Pclass+Sex+Age+Embarked	0.3939
30	Sex+Age+Fare+Embarked	0.333
31	Pclass+Sex+Age+Fare+Embarked	0.3939

Both (Pclass+Sex+Age+Embarked) and (Pclass+Sex+Age+Fare+Embarked) have the same R-Square. So we selected the combination with fewer number of variables: Pclass+Sex+Age+Embarked

2. Run lm and predict survivability and calculate accuracy for selected variables

Code:

```
EmbarkedQ=ifelse(dt$Embarked=="Q",1,0)
EmbarkedS=ifelse(dt$Embarked=="S",1,0)
m1=lm(survived~dt$Pclass+sex+dt$Age+dt$Embarked)
summary(m1)
c=coef(m1)
c
y1=c[1]+c[2]*dt$Pclass+c[3]*sex+c[4]*dt$Age+c[5]*EmbarkedQ+c[6]*EmbarkedS
#survival prediction of linear model

summary(y1)
z=ifelse(y1>0.5,1,0)
t=table(survived,z)
t
sum(diag(t))/sum(t)  #Accuracy of LM=0.7955182
```

Results:

Summary of prediction:

Min	1st Qu.	Median	Mean	3rd Qu.	Max
-0.1842	0.1249	0.3379	0.4062	0.6603	1.0819

Table of prediction

	0	1
0	360	64
1	82	208

Accuracy: 0.7955182

3. Run logit and predict survivability and calculate accuracy for selected variables

Code:

```
m2=glm(survived~dt$Pclass+sex+dt$Age+dt$Embarked,family="binomial"("logit"))
summary(m2)
c=coef(m2)
c
y2=1/(1+exp(-c[1]-c[2]*dt$Pclass-c[3]*sex-c[4]*dt$Age-c[5]*EmbarkedQ-
c[6]*EmbarkedS))  #predictor of logit model
```

```
z=ifelse(y2>0.5,1,0)
t=table(survived,z)
t
sum(diag(t))/sum(t)  #Accuracy of Logit=0.7955182
```

Results:

Summary of prediction:

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.01404	0.10959	0.30921	0.40616	0.69634	0.97113

Table of prediction

	0	1
0	357	67
1	79	211

Accuracy: 0.7955182

4. Run probit and predict survivability and calculate accuracy for selected variables

Code:

```
m3=glm(survived~dt$Pclass+sex+dt$Age+dt$Embarked,family="binomial"("probit"))
summary(m3)
c=coef(m3)
c
y3=pnorm(c[1]+c[2]*dt$Pclass+c[3]*sex+c[4]*dt$Age+c[5]*EmbarkedQ+c[6]*EmbarkedS)
z=ifelse(y3>0.5,1,0)
t=table(survived,z)
t
sum(diag(t))/sum(t)  #Accuracy of Probit=0.7983193
```

Results:

Summary of prediction:

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.008119	0.114718	0.325170	0.409959	0.692514	0.978019

Table of prediction

	0	1
0	360	64
1	80	210

Accuracy: 0.7983193

We can see that logit and linear model have the same accuracy for this dataset and probit is slightly better than them