```
#Q1

rm(list=ls(all=TRUE))

library(data.table)
library(sandwich)
library(lmtest)
library(ggplot2)

context1 <- fread("htv.csv")

context1$abilsq <- context1$abil^2
context1$educsq <- context1$educ^2
context1$expersq <- context1$exper^2

model1 <- lm(log(wage)~abil+educ+exper,data=context1) #1961.569
c(AIC(model1), BIC(model1))
summary(model1)

model2 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*educ)+(abil*exper)+(educ*exper),data=context1)
c(AIC(model2),BIC(model2))
summary(model2)

model90<-lm(log(wage)~abil+educ+exper+educ*exper,data=context1)
AIC(model90)
BIC(model90)

model3 <- lm(log(wage)~abil+educ+exper+abilsq,data=context1)
model4 <- lm(log(wage)~abil+educ+exper+educsq,data=context1)
model5 <- lm(log(wage)~abil+educ+exper+expersq,data=context1)
c(AIC(model3),AIC(model4),AIC(model5))
c(BIC(model3),BIC(model4),BIC(model5))

#All of the following should be abandoned, since "sq" is nonsense:
summary(model3)
summary(model4)
summary(model5)

#
model6 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq,data=context1)
c(BIC(model6)) #1971.098
summary(model6)

model7 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*educ),data=context1)
model8 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*exper),data=context1)
model9 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(educ*exper),data=context1)
c(BIC(model7),BIC(model8),BIC(model9)) #1972.140 1974.875 1976.231


#
model10 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*educ)+(abil*exper),data=context1)
model11 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*educ)+(educ*exper),data=context1)
model12 <- lm(log(wage)~abil+educ+exper+abilsq+educsq+expersq+(abil*exper)+(educ*exper),data=context1)
c(BIC(model10),BIC(model11),BIC(model12)) #1978.680 1976.902 1978.143

#
model13 <- lm(log(wage)~abilsq+(abil*educ)+(educ*exper),data=context1)
c(BIC(model13)) #1964.304
summary(model13)

#
model14 <- lm(log(wage)~abil+educ+exper+(abil*educ)+(educ*exper),data=context1)
c(BIC(model14)) #1964.784
summary(model14)

#
model15 <- lm(log(wage)~abil+educ+exper+(abil*educ),data=context1)
model16 <- lm(log(wage)~abil+educ+exper+(abil*exper),data=context1)
model17 <- lm(log(wage)~abil+educ+exper+(educ*exper),data=context1)
c(BIC(model15),BIC(model16),BIC(model17))
summary(model17)
# (Intercept)  1.331373   0.284573   4.678 3.21e-06 ***
#   abil         0.052907   0.008667   6.105 1.38e-09 ***
#   educ         0.048998   0.019102   2.565  0.01044 *
#   exper       -0.037966   0.023596  -1.609  0.10787
# educ:exper   0.005602   0.001742   3.215  0.00134 **

c(AIC(model17)) # 1927.66

model18 <- lm(log(wage)~abil+educ+exper+(abil*educ)+(abil*exper),data=context1)
model19 <- lm(log(wage)~abil+educ+exper+(abil*educ)+(educ*exper),data=context1)
model20 <- lm(log(wage)~abil+educ+exper+(abil*exper)+(educ*exper),data=context1)

c(BIC(model18),BIC(model19),BIC(model20)) # 1973.427 1964.784 1964.244 Not good.

model21 <- lm(log(wage)~abil+educ+exper+(abil*educ)+(abil*exper)+(educ*exper),data=context1)
c(BIC(model21)) #1971.249


# model17 is the greatest so far. So, remove some insignificant variables now.


# Final version of model 2
model2 <- lm(log(wage)~abil+educ+exper+(educ*exper),data=context1)
BIC(model2)

model13<-lm(log(wage)~abil+educ*exper,data=context1)
BIC(model13)

Interpretations:
  #a. only abil,educ,exper and educ*exper are the variables that best fit the model and have least BIC in model2
  #b. Interaction variable can have the combined effect of educ and exper.

  ###########################################################
#Q2
rm(list=ls(all=TRUE))

## Import packages
library(data.table)
library(ggplot2)
library(mfx)
library(pscl)
context2 <- fread('loanapp.csv')

model3 <- glm(approve~white,family=binomial(),data=context2)
coeftest(model3, vcov.=vcovHC)
```

```
summary(model3)

model4 <- glm(approve~white+hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+chist+pubrec+mortlat1+mortlat2+vr,family=binomial(link='logit'),data=context2)
coeftest(model4, vcov.=vcovHC)
summary(model4)

model5 <-
glm(approve~white+hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+chist+pubrec+mortlat1+mortlat2+vr+I(white*obrat),family=binomial(link='logit'),data=context2)

coeftest(model5, vcov.=vcovHC)
summary(model5)


Interpretations:
#a. it's equal to 1.4094, indicating if the applicant was white, the probability of approving the loan
#140.94% higher than the person is not white.
#b. It decreased to 0.93776, but still positive and significant in the approval rate.
#c. It decreased to 0.29688,and became insignificant.
#d. White *obrat is an interaction variable. Something like bad records to repay.

###############################################################
#Q3
rm(list=ls(all=TRUE))

library(data.table)
library(ggplot2)
library(lmtest)
library(sandwich)

context3 <- fread('smoke.csv')

context3$agesq <- context3$age^2
model6 <- glm(cigs~educ+age+agesq+log(income)+restaurn,family=poisson(),data=context3)
coeftest(model6, vcov.=vcovHC)
coeftest(model6)
summary(model6)


# #look into the difference between lm model and glm model.
model7 <- lm(cigs~educ+age+agesq+log(income)+restaurn,data=context3)
summary(model7)
coeftest(model7)
coeftest(model7, vcov.=vcovHC)

Interpretations:
#a. One year of schooling increase is associated with 0.05952 decrease
#in cigs. smoked per day controlling for other variables.

#b.
#when a person is 20, 1.140e-01+2*(-1.368e-03)*20 =  0.05928
# a person is 60, 1.140e-01+2*(-1.368e-03)*60 = -0.05016

###############################################################
#Q4

rm(list=ls())

library(data.table)

context4 <- fread('hdisease.csv')
context4$exang          <-      ifelse(context4$exang=="Yes",1,0)  # forget this step, resulting in no-outcome.
frmla <-hdisease~age+cp+trestbps+thalach+exang

library(evtree)
model7 <-evtree(frmla,data=context4)
plot(model7)


library(party)
model8<-        ctree(frmla,data=context4)
plot(model8,main="Conditional Inference Tree (context4)")
model8 <-ctree(hdisease~age+cp+trestbps+thalach+exang,data=context4)
plot(model8)

model8

context5 <- fread('hdisease-new.csv')
context5$exang          <-      ifelse(context5$exang=="Yes",1,0)
hdisease_pred <- predict(model8, context5)
hdisease_pred
plot(hdisease_pred)

summary(model7)
summary(model8)

Interpretation
##
## 1. Model8 is over-fitting, Model7 is under-fitting
##
## 2. Since dset has so many values, it will make the tree with too many branches.
##    Too many classifications may lead non-representative



#Q5
rm(list=ls(all=TRUE))

library(data.table)
install.packages("expm", dependencies = TRUE)

context5<-fread("WAGE1.csv")

seed<-2
maxClusters<-10

## Use within-group variation to choose k
wss     <- rep(-1,maxClusters)
for (i in 1:maxClusters) {
  set.seed(seed)
  model <- kmeans(context5,centers=i,nstart=10)
  wss[i] <- model$tot.withinss
}
plot(1:maxClusters,     wss, type="b",
     xlab="Number of Clusters",
     ylab="Aggregate Within Group SS")
?kmeans
```

```
## Run the model
set.seed(seed)
model9<-kmeans(context5,centers=3,nstart=10)
model9
model9$centers
groups1<-model9$cluster
groups1
context5$cluster <-groups1
model10 <- lm(wage~educ+exper+tenure,data=context5[cluster==1])
model11 <- lm(wage~educ+exper+tenure,data=context5[cluster==2])
model12 <- lm(wage~educ+exper+tenure,data=context5[cluster==3])
summary(model10)
summary(model11)
summary(model12)
?kmeans

Interpretations
#a.Using k-means cluster, the optimal number of clusters for this data set is 2.

#b. Cluster 1 has lowest education and highest exper and tenure values for centers.
#Cluster 2 has highest education and lowest exper and tenure.
#Cluster 3 has 2nd highest education and low tenure center values. Cluster 1 can be thought of as workers who have comparatively
#low level of education but a lot of experience as well as highest number of years of experience. Cluster 2 is for workers with comparatively
#higher level of education but very low experience and tenure, they can be classified as workers in initial phase of their careers. Cluster 3
#has workers with medium level of education and a lot of experience but lesser tenure as compared to cluster 1

#c. In model 2, the intercept is positive and the intercept for model 3 and 4 are negative. So, for cluster 1, if educ, exper and tenure are all zero
#then the model predicts a positive number for wage. However, it is the opposite for model3 and model4.Education has a similar effect
#on all three models and it has the highest effect in model4. Experience has a negative coefficient in model2 and positive coefficients in
#model3 and 4. Tenure has the similar effect in all the three models. The differences that we observed in these three models in terms of a positve
#or negative effect are for intercept and exper only.

#Q6
rm(list=ls(all=TRUE))

library(ggplot2)

context6<-fread("murder.csv")

## Run model
model13 <-prcomp(context6[1:50,2:52])
## Generating screeplot
screeplot(model13,type="lines")
model13$rotation[,1]*100

## get the principal components
context6$factor<-model13$x[,1]
#model13$ rotation:method to rotate the axis, x:values(components) on axis
head(context6)
summary(context6$factor)
ts.plot(context6$factor)

Interpretations
# a. 1 principal components because the elbow appears to be at n=2.
# b. component?

library(ggplot2)
context6<-fread("murder.csv")
xdata<-context6[1:50,2:52]
model13<-prcomp(xdata)
eig<-model13$sdev
variance<-sum(eig)-cumsum(eig)
plot(0:10,variance[1:11])
lines(0:10,variance[1:11])
screeplot(model13,type='lines')
factor<-model13$x[,1]
mean(model13$rotation[,1])
ts.plot(factor)
context6[27,1]
screeplot(model13)
```