```
#####################################
## Author:   Minke Li
## Date:     2018-04-14
## Title:    ps4.R
## Purpose:  Analyze the data in file htv.csv, loanapp.csv, smoke.csv, hdisease.csv, WAGE1.csv, murder.csv
#####################################
rm(list=ls()) #drop all variables

library(data.table)
library(sandwich)
library(lmtest)
library(tseries)
library(plm)
library(party)
library(evtree)
library(glmx)
library(ggplot2)

##Question1
context1 <- fread('htv.csv')

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

model2 <- lm(log(wage)~abil+educ+exper+I(abil^2)+I(educ^2)+I(exper^2)+abil:educ+abil:exper+educ:exper,data=context1)
AIC(model2)
BIC(model2)

model2   <- step(model2,k=log(nrow(context1)))
model2   <- step(model2)
model2 <- lm(log(wage)~abil+educ+exper+educ:exper,data=context1)
BIC(model2)

##Interpretations
#a. model2 considers the interaction between variables education and experience, while model 1 only considers their
individual effects
#b. the interaction variable makes the model more complex

rm(list=ls())

##Question2
context2 <- fread('loanapp.csv')

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

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

model5 <-
glm(approve~white+hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+chist+pubrec+mortlat1+mortlat2+vr+white:obrat,data
 = context2)
coeftest(model5,.vcov=vcovHC)


##Interpretations
#a. when other factors don't remain the same, if applicant was white, the approval rate for loan would increases.
#b. bata1 becomes smaller, but it is still significant
#c. bata1 becomes negative, which makes 'white' not significant
#d. the interaction between white and obrat. since white and obrat have strong relationship

rm(list=ls())

##Question3
context3 <- fread('smoke.csv')

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

##Interpretations
#a. Every one year increase in education is associated with a 5.95% decrease in daily cigarette consumption
#b. when a person is 20, the rate is 5.927%; when a person is 60, the rate is -5.016%

rm(list=ls())

##Question4

context4 <- fread('hdisease.csv')
context4$exang          <-      ifelse(context4$exang=="Yes",1,0)  # forget this step, resulting in no-outcome.

model8 <- ctree(hdisease~age+cp+trestbps+thalach+exang,data=context4)
model7 <- evtree(hdisease~age+cp+trestbps+thalach+exang,data=context4)

context45 <- fread('hdisease-new.csv')
```

```
context45$hdisease_pred <- predict(model8,context45)

plot(model7)
plot(model8)

##Interpretations
#a. model8 is over-fitting, model7 is under-fitting
#b. the contents of dset is text, and it doesn't have a good way to identify it as number

rm(list=ls())

##Question5
context5 <- fread('WAGE1.csv')

seed        <-  2
maxClusters     <-      10
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")

set.seed(seed)
model9 <- kmeans(context5,centers=3,nstart=10)
model9$centers

group1 <-model9$cluster
context5$cluster<-group1

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)

##Interpretations
#a. 2
#b. cluster1: low education,high experience and tenure
#   cluster2: high education, low experience and tenure
#   cluster3: high education, experience, and tenure
#c.


rm(list=ls())

##Question6
context6 <- fread('murder.csv')
xdata <- context6[1:50,2:52]
model13 <-prcomp(xdata)
eig <- model13$sdev
variance <- sum(eig)-cumsum(eig)
screeplot(model13,type="lines")

factor <- model13$x[,1]
ts.plot(factor)
context6[27,1]

##Interpretations
#a.one
#b. around 1991 the factor is the greatest
```