SOWK3136 Individual Lab Assignment

Student Name: Fung Kong Hoi

UID: 3035857517

Q1

For “weekly”, There are 7987 rows and 10 variables. In terms of country, there are 170 unique countries. The period covered was not mentioned specifically but the dataset recorded from week 1 to week 56. It is around a year. From the boxplot of all new cases for all countries, those of them are relatively concentrated from the range of 0 to 30000. However, there are some outliers for the US. which exceeds 1000000 cases weekly. For “daily”, There are 56477 rows and 15 variables. The period covered form 2020-01-03 to 2021-01-31. From the boxplot of all new cases for all countries, there is just an outlier. However, the outliers are greater than 3000 death cases daily. The cases are relatively stable and concentrated from the range of 0 to 2000. For other variables such as the school closing, cancel public events and vaccination policy, the value ranges from 0 to 5 and no outliers exist in those datasets. After investigating the basic data structure of the dataset, some research questions are initiated that highlight the relationship between new cases and new deaths, and how different health measures affect those two outcomes. More importantly, the questions hope to observe some pattern or correlation with specific countries with similar characteristics. To begin, some assumptions such as the relationship between new cases and new deaths will be made. After that, the project will try to find out some countries with familiar or special features. Visualization can help in this process. Finally, the scatter plot can define the outliers and the time series graphs can illustrate how government response affect the COVID-19 situation in different countries in a different time frame.

Research question 1: what is the relationship between new cases and new deaths? Is the relationship the same as what you expected, and can you find some abnormal cases in some countries/from daily or weekly data? Chart, scatter chart

Description automatically generated

To compare the relationship between the new cases and new deaths on a global scale, it is shown that they are positively correlated especially when the new cases range from 0 to 4000. However, when the new cases reach 5000 or 6000, the new death number is near the level to that of 0-1000. Those two dots are treated as outliers. In general, most of the weekly data of new cases are lower than 1500 cases per week for most of the countries. When the new cases increase, the new deaths will increase as well.Chart, line chart

Description automatically generated

The graph shows the cumulative COVID-19 cases and death numbers in the world. The steeper the curve, the higher the death rate. Most countries, such as India, Brazil, and The United Kingdom have a normal steepness curve. Even though The U.S. has the most cumulative cases, the steepness of the curve is similar to other countries. The blue curve which represents Mexico has the steepest curve while Turkey has the flattest curve among other countries. The death rate for Mexico and Turkey are 8.5% and 1.05% respectively.

Research question 2: Try to group some countries with familiar features (can be a similar trend/ health measure in a specific period/population) and analyze the pattern of those countries. Can you find some interesting cases?Chart, histogram

Description automatically generated

India, The U.S. Brazil were the countries that recorded the most confirmed COVID-19 cases throughout the whole-time timeframe. The new cases in those countries started increasing in April 2020. In October, those countries faced the second phase of COVID-19 confirmed cases. For example, the new cases in Brazil were 32058 while the cases in the U.S. was 38476. After two months, the new cases skyrocketed to 69826 and 402270 at the highest rate. One interesting finding is that from July to October, India, The U.S. Brazil had a very similar trend and confirmed cases number.Chart, line chart, histogram

Description automatically generated

United States of America, India, and Brazil had serious daily deaths number from April to July. They are the countries with more than 200 million of population. In terms of development level, The US is the top country with the highest GDP contribution to the world. India and Brazil are the “Bricks’ countries that have rapid economic development. Peru, Ecuador and Mexico, Argentina suffered the most from July to October. They are developing countries located on the American continent. After October 2020, the new daily death number of the four countries dropped significantly and became stable to around 1000 death cases. Chart, scatter chart

Description automatically generated

According to the scatter plot, For Brazil, there is a positive correlation between the new cases and new deaths. The plot may show that Brazil may have a higher death rate than the two countries. To be more specific, when the new cases rise from 1000 to 5000, the new deaths increase from 100 to more than 300. In contrast, the new deaths do not vary with the number of new cases in India and the US. In general, the number of new deaths is lower than 250. All of the outliers are in Brazil.

Chart, scatter chart

Description automatically generated

According to the scatter plot, Ecuador has relatively lower new cases than the three countries. However, it can be observed that Mexico has more new cases and deaths. In general, the death cases in four countries except Mexico are mostly lower than 250 cases. All of the outliers have come from Mexico. There is no strong correlation between new cases and death numbers for Peru, Ecuador, and Argentina

Research question3: As different variables can be grouped into specific types of health measures, try to use the grouped variable to find out the effects of those measures in terms of the numeric variables (new cases, new deaths, accumulate new cases, accumulate deaths)

Graphical user interface, application, Word

Description automatically generated

For Brazil, the new deaths skyrocketed from 200 to 1500 in April. The time frame was a coincidence with the abolishment of the testing policy. Before April, the new deaths were around 0 and Brazil started the testing policy in late March. Also, the contract tracing started in April. It can be explained by worsening conditions in Brazil. The Brazilian government started to be aware of the virus. When Brazil recorded the highest number of new deaths of 1700, the number declined gradually since the international travel restriction was imposed.

Chart, line chart

Description automatically generated

As concluded from the previous question, Mexico has different observation results from the other three countries. From this graph, Mexico had the highest daily new cases of 30000 cases in October. Before October, the new cases were around 4000 to 10000 in the previous three months. It began to rise significantly after the international travel policy was abolished while the other policies were not. Therefore, from these cases, the above-mentioned policy had a huge impact on the new daily cases in Mexico. Graphical user interface, application, Word

Description automatically generated

Ecuador started to adopt policies such as wearing masks, COVID-19 testing, and contact tracing in April 2020. Even though the cases reached an all-time high of 12000 cases in a day, the cases started to become stable from June. After that, most of the day recorded around 1500 per day given that most of the government responses are imposed. Therefore, it can conclude that adopting multiple measures can help to control the spread of the virus and new daily confirmed cases.

Research question 4: In addition, try to add a new insight/finding in your exploratory data analysis.Graphical user interface, application

Description automatically generated

It is assumed that the number of new deaths will drop when a tight government response is adopted such as adopting a wide variety of health measures at the same time. However, For Ecuador, a strict government response was imposed starting in April 2020 and the number of new deaths number risen dramatically to around 5000 in September. The outcome cannot be explained by the level of the government responses as it was constant. It could be caused by factors outside the measured variables. Therefore, more variables such as hospital numbers, and government medical expenses should be added to provide a more all-around analysis.

Q2

(c) y is 100. β0is -1, β1 is 0.5.

(d) X and y are positively correlated

(e)

The linear regression fits a model which is closer to the true value of the coefficients. The model has a large F-statistic will have a near-zero p-value

(g)

When the model fits more on the training data, it results in a slight increase in R^2 and MSE. Also, the p-value of the t-statistic point out there is no strong relationship between y and X^2

Q3

(iv) #2.54% test error rate in the validation set

(c) #first result: 0.0254

#second result: 0.0244

#third result: 0.0242

(d) #0.0278

*The “student” dummy variable does not reduce test error rate.*

*in the validation set estimate.*

Q4

(b) #The test MSE is 4.191075.

© Pruning the tree makes the test MSE to be 4.13135

(d) #the MSE becomes 2.58. The Price, ShelveLoc and Age are the three most influencing predictors of 𝚂𝚊𝚕𝚎Sale.

(e,f)

#2.67

#The random forest worsens the MSE to be 2.67. it is lower than the result in Bagging.  The m will make the MSE value to lie between 2.6 to 3. The Price, ShelveLoc and Age are the three most influencing predictors of 𝚂𝚊𝚕𝚎Sale.

Appendix

Q1

# Charge libraries:

library(ggplot2)

library(naniar)

library (readr)

library(plotly)

library(forcats)

library(cowplot)

library(ggstance)

library(RColorBrewer)

library(gridExtra)

#setwd("C:/Users/hongheikenny/Downloads/covid.RData")

load("/Users/hongheikenny/Downloads/covid.RData")

summary("covid.RData")

weekly

daily

unique(daily$Country)

#table(daily$New\_cases)

#ggplot(daily, aes(x = New\_cases)) +

+     geom\_bar()

tab\_cnt <- table(daily$New\_cases, daily$New\_deaths)

tab\_cnt

prop.table(tab\_cnt)

sum(prop.table(tab\_cnt))

#weekly

gg\_miss\_upset(weekly)

myweekly <- weekly[complete.cases(weekly$New\_cases), ]

#myweekly$Day <- as.Date(myweekly$Week, "%b")

weekly$New\_cases[weekly$New\_cases<0] <- 0

end <- myweekly[myweekly$Week=="56",]

myweekly <- myweekly %>%

  dplyr::rename(

    C1\_school = "C1\_School closing", C2\_workplace = "C2\_Workplace closing", C3\_public\_events = "C3\_Cancel public events",

    C4\_gathering = "C4\_Restrictions on gatherings", C5\_transport = "C5\_Close public transport",

    C6\_stayathome = "C6\_Stay at home requirements", C7\_internal\_move = "C7\_Restrictions on internal movement",

    C8\_inter\_travel = "C8\_International travel controls", E1\_income = "E1\_Income support",

    E2\_relief= "E2\_Debt/contract relief", H1\_campaigns = "H1\_Public information campaigns",

    H2\_testing = "H2\_Testing policy", H3\_contact\_tracing = "H3\_Contact tracing",

    H6\_mask = "H6\_Facial Coverings", H7\_vaccination = "H7\_Vaccination policy"

  )

#Boxplot of new cases for all countries

p <- ggplot(daily, aes(x= Country, y=New\_cases))

p <- p + geom\_boxplot()

p <- p + theme\_classic()

p <- p + labs(title = "Box plot of new cases for all countries")

p <- p + labs(subtitle = "daily")

p

#Boxplot of new cases for all countries

p\_death <- ggplot(daily, aes(x= Country, y=New\_deaths))

p\_death <- p\_death + geom\_boxplot()

p\_death <- p\_death + theme\_classic()

p\_death <- p\_death+ labs(title = "Box plot of new death cases for all countries")

p\_death <- p\_death+ labs(subtitle = "daily")

p\_death

#Boxplot of new cases for all countries

p\_cases <- ggplot(weekly, aes(x= CountryCode, y=New\_cases, group = CountryCode ))

p\_cases <- p\_cases + geom\_boxplot()

p\_cases <- p\_cases + theme\_classic()

p\_cases <- p\_cases+ labs(title = "Box plot of new death cases for all countries")

p\_cases <- p\_cases+ labs(subtitle = "weekly")

p\_cases

############research question#############################

#Q1. multi\_variable scatter plot between new cases and new deaths

plot(New\_cases ~ New\_deaths,

     data = daily,

     main = "graph of new cases and new deaths",

     xlab ="New cases",

     ylab = "New deaths",

     cex = 0.1)

#Cumulative cases in the world

cases\_1 <- ggplot(mydaily, aes(x=Cumulative\_cases, y=Cumulative\_deaths, colour = Country)) +

  geom\_line() +

  theme(legend.position="none")

fig.cases <- ggplotly(cases\_1)

fig.cases

#Q2

#world cases

cases <- ggplot(mydaily, aes(x=Day, y=New\_cases, colour = Country)) +

  geom\_line() +

  theme(legend.position="none")

fig.cases <- ggplotly(cases)

fig.cases

#new deaths in the world

deaths <- ggplot(mydaily, aes(x=Day, y=New\_deaths, colour = Country)) +

  geom\_line() +

  theme(legend.position="none")

fig.deaths <- ggplotly(deaths)

fig.deaths

# gorup the 4 countries with similar features: "Peru" , "Ecuador" , "Mexico", "Argentina"

country <- c("Peru" , "Ecuador" , "Mexico", "Argentina")

mydata <- mydaily[mydaily$Country %in% country,]

mydata <- mydata %>%

  dplyr::rename(

    C1\_school = "C1\_School closing", C2\_workplace = "C2\_Workplace closing", C3\_public\_events = "C3\_Cancel public events",

    C4\_gathering = "C4\_Restrictions on gatherings", C5\_transport = "C5\_Close public transport",

    C6\_stayathome = "C6\_Stay at home requirements", C7\_internal\_move = "C7\_Restrictions on internal movement",

    C8\_inter\_travel = "C8\_International travel controls", E1\_income = "E1\_Income support",

    E2\_relief= "E2\_Debt/contract relief", H1\_campaigns = "H1\_Public information campaigns",

    H2\_testing = "H2\_Testing policy", H3\_contact\_tracing = "H3\_Contact tracing",

    H6\_mask = "H6\_Facial Coverings", H7\_vaccination = "H7\_Vaccination policy"

  )

Peru <- mydata[mydata$Country=="Peru",]

Ecuador <- mydata[mydata$Country=="Ecuador",]

Mexico <- mydata[mydata$Country=="Mexico",]

Argentina <- mydata[mydata$Country=="Argentina",]

myfour <- mydata[mydata$Country %in% c("Peru","Ecuador","Mexico","Argentina"),]

ggplot(myfour, aes(x=New\_cases, y=New\_deaths, color=Country)) + geom\_point(alpha=0.3)

ggplot(myfour, aes(x=New\_cases, y=New\_deaths, color=Country)) + xlim(0,5000) +  ylim(0,1000) + geom\_point(alpha=0.3)

#Mexico case

p.Mexico.policy <- ggplot(Mexico, aes(x=Day)) +

  geom\_jitter(aes(y=C8\_inter\_travel, col=mycol[1]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H2\_testing,col=mycol[2]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H3\_contact\_tracing, col=mycol[3]),  alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H6\_mask, col=mycol[4]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H7\_vaccination, col=mycol[5]), alpha = 0.5, width = 0, height = 0.1) +

  xlab("Date") + ylab("Policy measures") +

  scale\_color\_identity(guide = "legend",

                       name = "Government responses:",

                       breaks = mycol[1:5],

                       labels = c("International travel","Testing","Contact tracing","Mask","Vaccination")) +

  theme(legend.position = "bottom")

p.Mexico.cases <- ggplot(Mexico, aes(x=Day)) +

  geom\_line(aes(y=New\_cases)) + ggtitle("Mexico") +

  xlab("Date") + ylab("New Cases") +

  theme(legend.position = "none")

Mexico.combine <- grid.arrange(p.Mexico.cases,p.Mexico.policy,nrow=2)

##Ecuador case

Health <- ggplot(mydata, aes(x = Day, y = H1\_campaigns, col=mycol[1])) +

  geom\_point(na.rm=T) +

  geom\_point(aes(y=H2\_testing+0.1), col=mycol[2]) +

  geom\_point(aes(y=H3\_contact\_tracing+0.2), col=mycol[3]) +

  geom\_point(aes(y=H6\_mask+0.3), col=mycol[4]) +

  geom\_point(aes(y=H7\_vaccination+0.4), col=mycol[5]) +

  ylab("Health system") +

  facet\_wrap( ~ Country, ncol=4) +

  theme(legend.position = "none")

fig.Health <- ggplotly(Health)

fig.Health

p.Ecuador.policy <- ggplot(Ecuador, aes(x=Day)) +

  geom\_jitter(aes(y=C8\_inter\_travel, col=mycol[1]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H2\_testing,col=mycol[2]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H3\_contact\_tracing, col=mycol[3]),  alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H6\_mask, col=mycol[4]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H7\_vaccination, col=mycol[5]), alpha = 0.5, width = 0, height = 0.1) +

  xlab("Date") + ylab("Policy measures") +

  scale\_color\_identity(guide = "legend",

                       name = "Government responses:",

                       breaks = mycol[1:5],

                       labels = c("International travel","Testing","Contact tracing","Mask","Vaccination")) +

  theme(legend.position = "bottom")

p.Ecuador.cases <- ggplot(Ecuador, aes(x=Day)) +

  geom\_line(aes(y=New\_cases)) + ggtitle("Ecuador") +

  xlab("Date") + ylab("New Cases") +

  theme(legend.position = "none")

Ecuador.combine <- grid.arrange(p.Ecuador.cases,p.Ecuador.policy,nrow=2)

#####group US,India, Brazil

Us <- mydata[mydata$Country=="United States of America",]

India <- mydata[mydata$Country=="India",]

Brazil <- mydata[mydata$Country=="Brazil",]

big\_countries <- mydata[mydata$Country %in% c("United States of America","India","Brazil"),]

ggplot(big\_countries, aes(x=New\_cases, y=New\_deaths, color=Country)) + geom\_point(alpha=0.3)

ggplot(big\_countries, aes(x=New\_cases, y=New\_deaths, color=Country)) + xlim(0,5000) +  ylim(0,1000) + geom\_point(alpha=0.3)

#Health measures in Brazil

p.Brazil.policy <- ggplot(Brazil, aes(x=Day)) +

  geom\_jitter(aes(y=C8\_inter\_travel, col=mycol[1]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H2\_testing,col=mycol[2]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H3\_contact\_tracing, col=mycol[3]),  alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H6\_mask, col=mycol[4]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H7\_vaccination, col=mycol[5]), alpha = 0.5, width = 0, height = 0.1) +

  xlab("Date") + ylab("Policy measures") +

  scale\_color\_identity(guide = "legend",

                       name = "Government responses:",

                       breaks = mycol[1:5],

                       labels = c("International travel","Testing","Contact tracing","Mask","Vaccination")) +

  theme(legend.position = "bottom")

#Brazil case

p.Brazil.deaths <- ggplot(Brazil, aes(x=Day)) +

  geom\_line(aes(y=New\_deaths)) + ggtitle("Brazil") +

  xlab("Date") + ylab("New deaths") +

  theme(legend.position = "none")

Brazil.combine <- grid.arrange(p.Brazil.deaths,p.Brazil.policy,nrow=2)

#Q4

#Ecuador's new deaths

p.Ecuador.policy <- ggplot(Ecuador, aes(x=Day)) +

  geom\_jitter(aes(y=C8\_inter\_travel, col=mycol[1]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H2\_testing,col=mycol[2]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H3\_contact\_tracing, col=mycol[3]),  alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H6\_mask, col=mycol[4]), alpha = 0.5, width = 0, height = 0.1) +

  geom\_jitter(aes(y=H7\_vaccination, col=mycol[5]), alpha = 0.5, width = 0, height = 0.1) +

  xlab("Date") + ylab("Policy measures") +

  scale\_color\_identity(guide = "legend",

                       name = "Government responses:",

                       breaks = mycol[1:5],

                       labels = c("International travel","Testing","Contact tracing","Mask","Vaccination")) +

  theme(legend.position = "bottom")

p.Ecuador.cases <- ggplot(Ecuador, aes(x=Day)) +

  geom\_line(aes(y=New\_deaths)) + ggtitle("Ecuador") +

  xlab("Date") + ylab("New deaths") +

  theme(legend.position = "none")

Ecuador.combine <- grid.arrange(p.Ecuador.cases,p.Ecuador.policy,nrow=2)

Q2

(a)

set.seed(7517)

x = rnorm(100, mean=0, sd=1)

(b)

eps = rnorm(100, 0, sqrt(0.25))

(c )

y = -1 + 0.5\*x + eps

(d)

plot(x, y)

(e)

lm.fit = lm(y~x)

summary(lm.fit)

(f)

plot(x, y)

abline(lm.fit, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)

(g)

lm.fit\_sq = lm(y~x+I(x^2))

summary(lm.fit\_sq)

(h)

set.seed(7517)

eps1 = rnorm(100, 0, 0.125)

x1 = rnorm(100)

y1 = -1 + 0.5\*x1 + eps1

plot(x1, y1)

lm.fit1 = lm(y1~x1)

summary(lm.fit1)

abline(lm.fit1, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)

the error observed in R^2 and RSE decrease.  
(i)

set.seed(7517)

eps2 = rnorm(100, 0, 0.5)

x2 = rnorm(100)

y2 = -1 + 0.5\*x2 + eps2

plot(x2, y2)

lm.fit2 = lm(y2~x2)

summary(lm.fit2)

abline(lm.fit2, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)

 the error observed in R^2 and RSE increase.

(j)

confint(lm.fit)

confint(lm.fit)

confint(lm.fit)

It is the narrowest for the lowest variance model and widest for the highest variance model The second fit’s interval is becoming narrower than the first fit’s interval and the third fit’s interval is wider than the first fit’s interval.

Q3

#ch5 Q5

(a)

library(ISLR)

attach(Default)

set.seed(1)

fit.glm <- glm(default ~ income + balance, data = Default, family = "binomial")

summary(fit.glm)

(b)(i)

train <- sample(dim(Default)[1], dim(Default)[1] / 2)

(II)

fit.glm <- glm(default ~ income + balance, data = Default, family = "binomial", subset = train)

summary(fit.glm)

(IIi)

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm <- rep("No", length(probs))

pred.glm[probs > 0.5] <- "Yes"

(Iv)

(pred.glm != Default[-train, ]$default)

(c)

train <- sample(dim(Default)[1], dim(Default)[1] / 2)

fit.glm <- glm(default ~ income + balance, data = Default, family = "binomial", subset = train)

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm <- rep("No", length(probs))

pred.glm[probs > 0.5] <- "Yes"

mean(pred.glm != Default[-train, ]$default)

train <- sample(dim(Default)[1], dim(Default)[1] / 2)

fit.glm <- glm(default ~ income + balance, data = Default, family = "binomial", subset = train)

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm <- rep("No", length(probs))

pred.glm[probs > 0.5] <- "Yes"

mean(pred.glm != Default[-train, ]$default)

train <- sample(dim(Default)[1], dim(Default)[1] / 2)

fit.glm <- glm(default ~ income + balance + student, data = Default, family = "binomial", subset = train)

pred.glm <- rep("No", length(probs))

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm[probs > 0.5] <- "Yes"

mean(pred.glm != Default[-train, ]$default)

(d)

train <- sample(dim(Default)[1], dim(Default)[1] / 2)

fit.glm <- glm(default ~ income + balance + student, data = Default, family = "binomial", subset = train)

pred.glm <- rep("No", length(probs))

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm[probs > 0.5] <- "Yes"

mean(pred.glm != Default[-train, ]$default)

*Q4*

*#Ch8 Q8*

(a)

library(tree)

library(ISLR)

attach(Carseats)

set.seed(0)

n <- nrow(Carseats)

p <- ncol(Carseats) - 1

train <- sample(1:n, n/2)

test <- (1:n)[-train]

# Part (b):

rtree.carseats <- tree(Sales ~ ., data = Carseats[train, ])

summary(rtree.carseats)

plot(rtree.carseats)

text(rtree.carseats, pretty = 0)

print(rtree.carseats)

#(c)

pred.carseats = predict(rtree.carseats, Carseats.test)

mean((Carseats.test$Sales - pred.carseats)^2)

####

cv.carseats = cv.tree(rtree.carseats, FUN = prune.tree)

par(mfrow = c(1, 2))

plot(cv.carseats$size, cv.carseats$dev, type = "b")

plot(cv.carseats$k, cv.carseats$dev, type = "b")

pruned.carseats = prune.tree(rtree.carseats, best = 9)

par(mfrow = c(1, 1))

plot(pruned.carseats)

text(pruned.carseats, pretty = 0)

pred.pruned = predict(pruned.carseats, Carseats.test)

mean((Carseats.test$Sales - pred.pruned)^2)

#Part d

library(randomForest)

bag.carseats = randomForest(Sales ~ ., data = Carseats.train, mtry = 10, ntree = 500,

                            importance = T)

bag.pred = predict(bag.carseats, Carseats.test)

mean((Carseats.test$Sales - bag.pred)^2)

importance(bag.carseats)

#part e

rf.carseats = randomForest(Sales ~ ., data = Carseats.train, mtry = 5, ntree = 500,

                           importance = T)

rf.pred = predict(rf.carseats, Carseats.test)

mean((Carseats.test$Sales - rf.pred)^2)