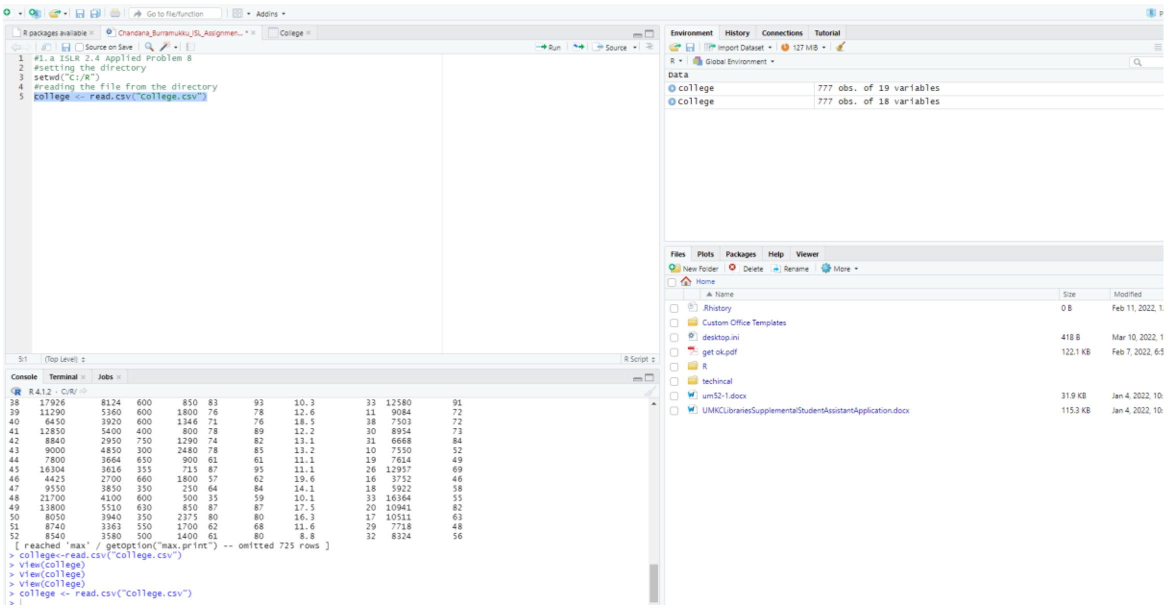
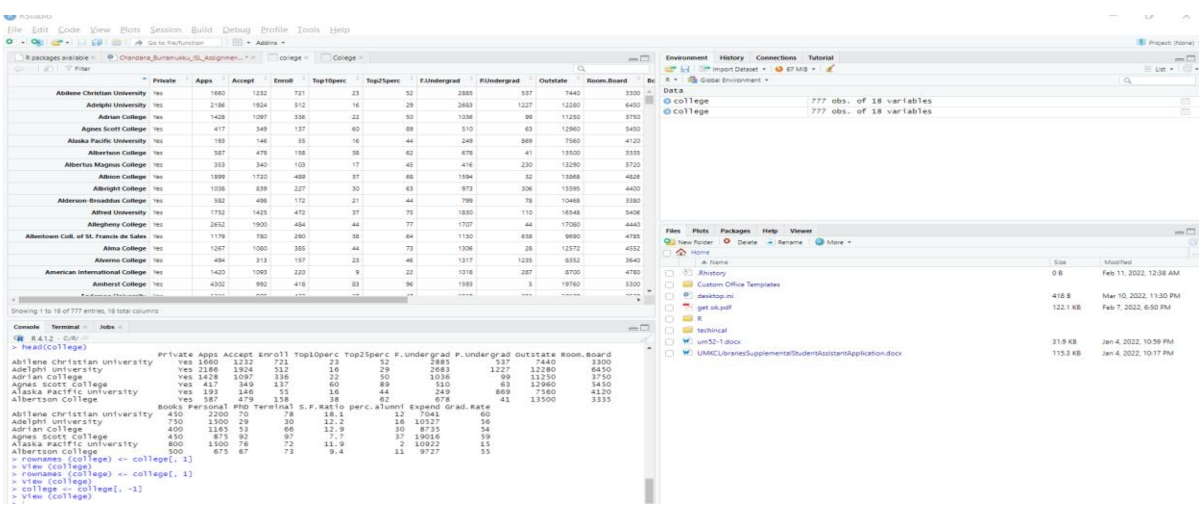
Name: Chandana Burramukku

ISL Assignment 2

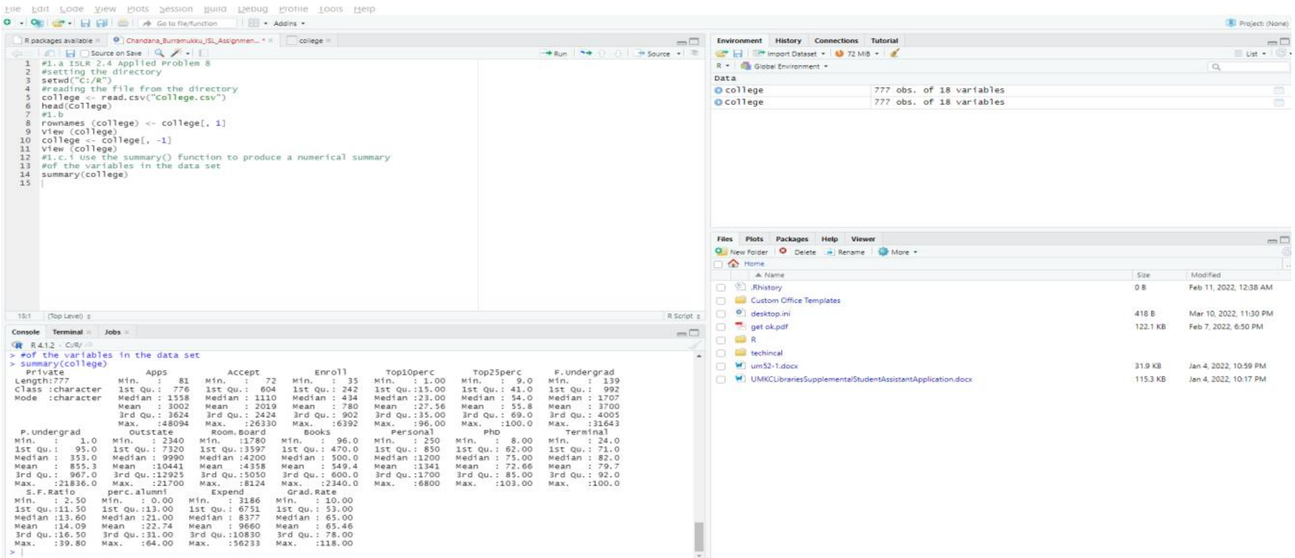
1.a



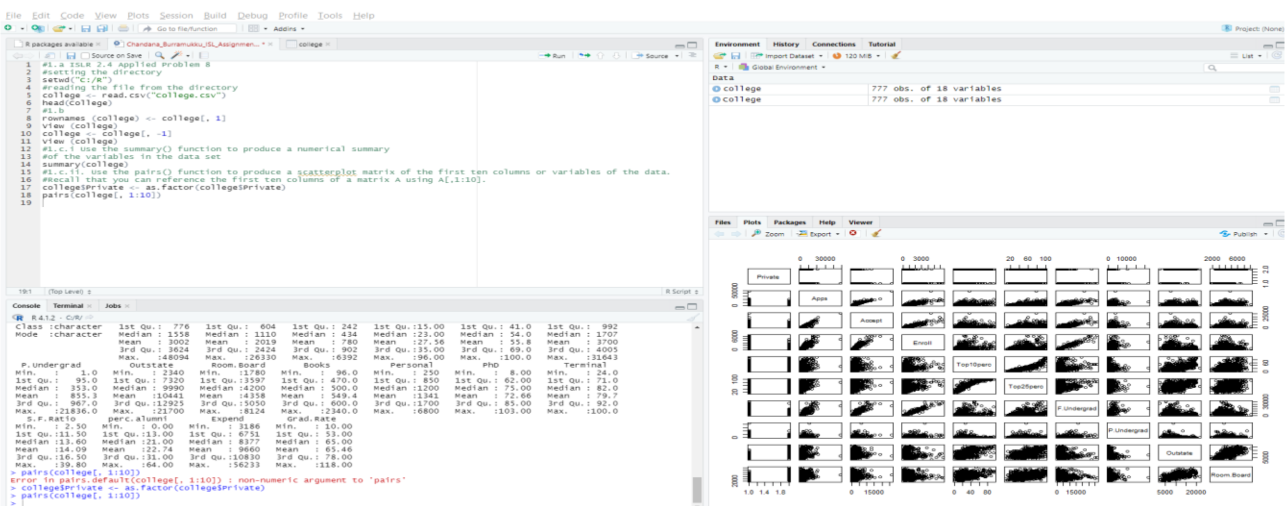
b.



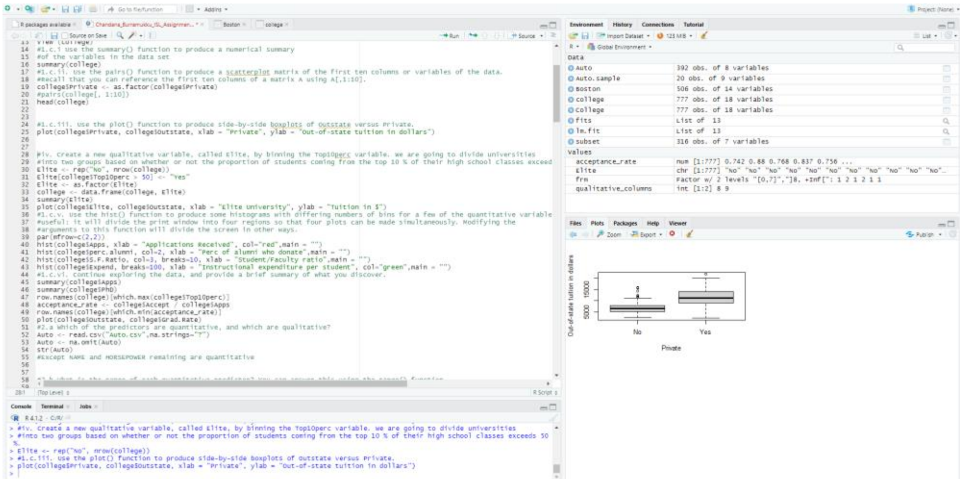
1.c.i



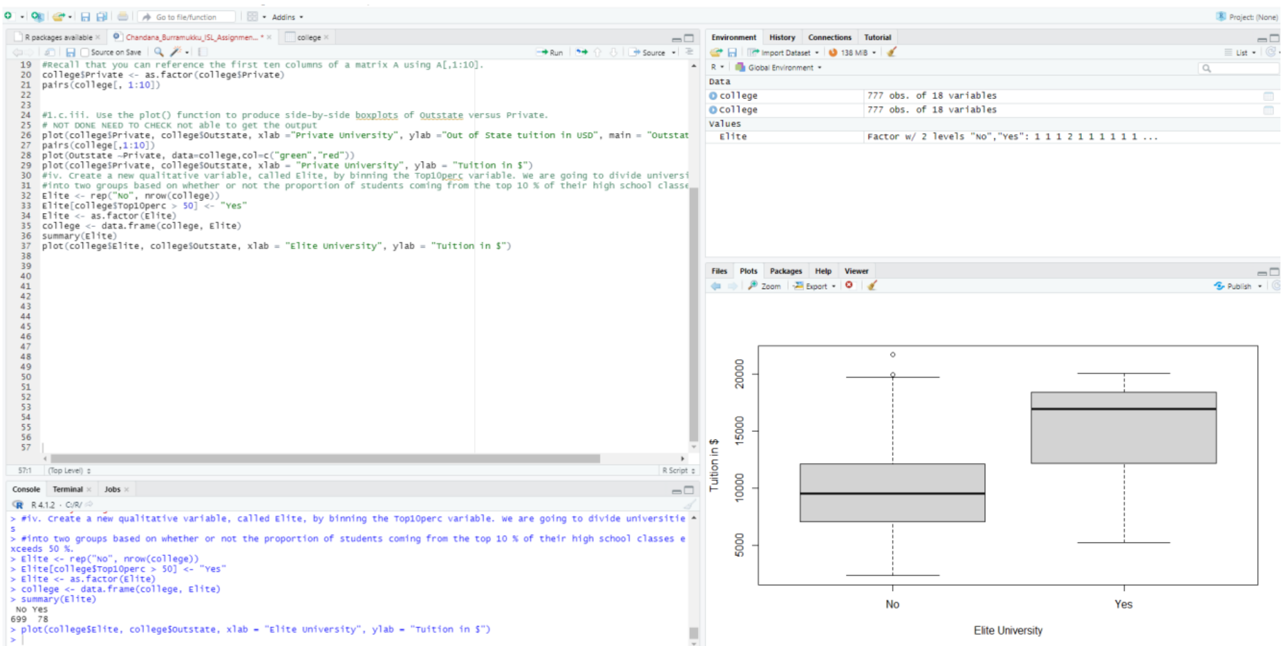
1.c.ii



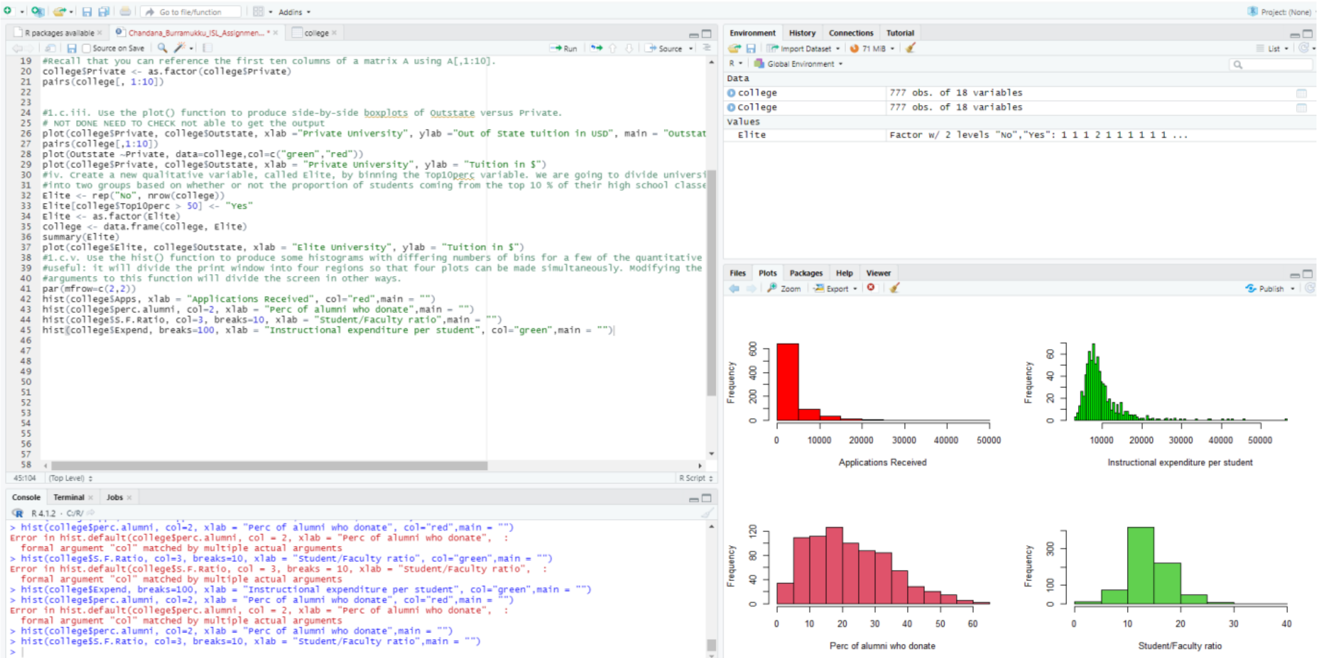
1.c.iii



1.c.iv

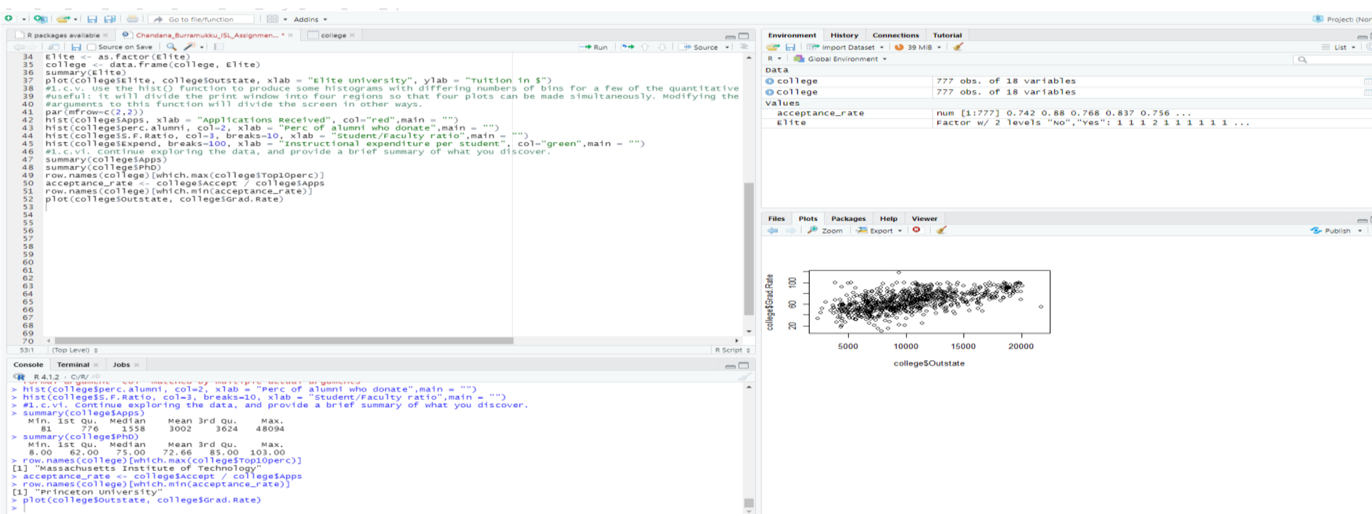

#Recall that you can reference the first ten columns of a matrix A using A[,1:10].
college$Private <- as.factor(college$Private)
pairs(college[, 1:10])
#1.c.iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.
NOT DONE NEED TO CHECK not able to get the output
plot(college$Private, college$Outstate, xlab ="Private University", ylab ="Out of State tuition in USD", main = "Outstat
pairs(college[,1:10])
plot(Outstate ~Private, data=college,col=c("green","red"))
plot(college$Private, college$Outstate, xlab = "Private University", ylab = "Tuition in $")
#iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universi
#into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classe
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
summary(Elite)
plot(college$Elite, college$Outstate, xlab = "Elite University", ylab = "Tuition in $")
college 777 obs. of 18 variables
college 777 obs. of 18 variables
Elite Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
No Yes
699 78
Tuition in $
Elite University
No
Yes

1.c.v



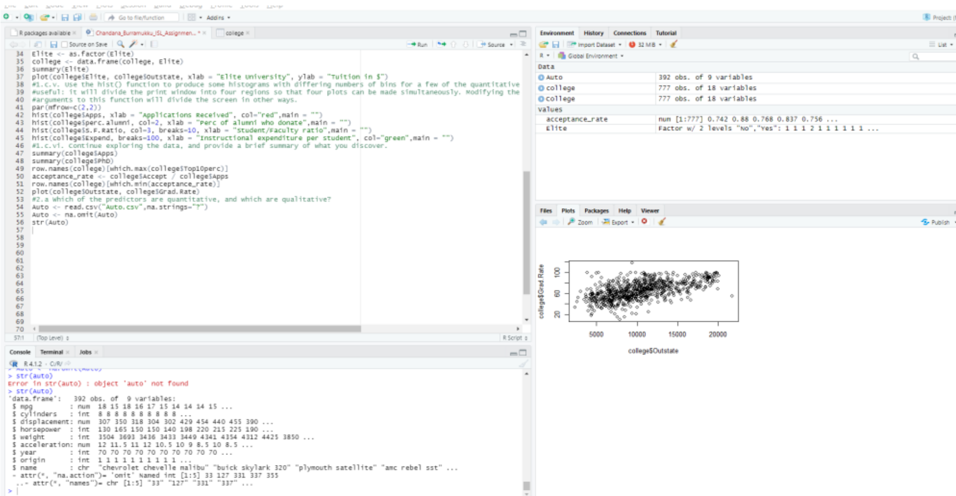
#Recall that you can reference the first ten columns of a matrix A using A[,1:10].
college$Private <- as.factor(college$Private)
pairs(college[, 1:10])
#1.c.iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.
NOT DONE NEED TO CHECK not able to get the output
plot(college$Private, college$Outstate, xlab ="Private University", ylab ="Out of State tuition in USD", main = "Outstat
pairs(college[,1:10])
plot(Outstate ~Private, data=college,col=c("green","red"))
plot(college$Private, college$Outstate, xlab = "Private University", ylab = "Tuition in $")
#iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universi
#into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classe
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
summary(Elite)
plot(college$Elite, college$Outstate, xlab = "Elite University", ylab = "Tuition in $")
#1.c.v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative
#useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the
#arguments to this function will divide the screen in other ways.
par(mfrow=c(2,2))
hist(college$Apps, xlab = "Applications Received", col="red",main = "")
hist(college$perc.alumni, col=2, xlab = "Perc of alumni who donate",main = "")
hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/Faculty ratio",main = "")
hist(college$Expend, breaks=100, xlab = "Instructional expenditure per student", col="green",main = "")
> hist(college$perc.alumni, col=2, xlab = "Perc of alumni who donate", col="red",main = "")
Error in hist.default(college$perc.alumni, col = 2, xlab = "Perc of alumni who donate", :
formal argument "col" matched by multiple actual arguments
> hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/Faculty ratio", col="green",main = "")
Error in hist.default(college$S.F.Ratio, col = 3, breaks = 10, xlab = "Student/Faculty ratio", :
formal argument "col" matched by multiple actual arguments
> hist(college$Expend, breaks=100, xlab = "Instructional expenditure per student", col="green",main = "")
> hist(college$perc.alumni, col=2, xlab = "Perc of alumni who donate", col="red",main = "")
Error in hist.default(college$perc.alumni, col = 2, xlab = "Perc of alumni who donate", :
formal argument "col" matched by multiple actual arguments
> hist(college$perc.alumni, col=2, xlab = "Perc of alumni who donate",main = "")
> hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/Faculty ratio",main = "")
college 777 obs. of 18 variables
college 777 obs. of 18 variables
Elite Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
Applications Received
Instructional expenditure per student
Perc of alumni who donate
Student/Faculty ratio
Frequency

1.c.vi

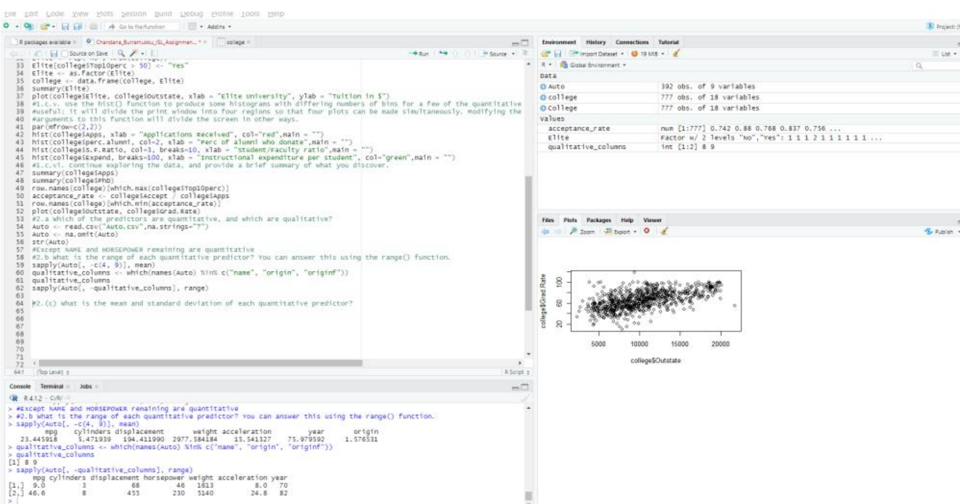


Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
summary(Elite)
plot(college$Elite, college$Outstate, xlab = "Elite University", ylab = "Tuition in $")
#1.c.v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative
#useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the
#arguments to this function will divide the screen in other ways.
par(mfrow=c(2,2))
hist(college$Apps, xlab = "Applications Received", col="red",main = "")
hist(college$perc.alumni, col=2, xlab = "Perc of alumni who donate",main = "")
hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/Faculty ratio",main = "")
hist(college$Expend, breaks=100, xlab = "Instructional expenditure per student", col="green",main = "")
#1.c.vi. Continue exploring the data, and provide a brief summary of what you discover.
summary(college$Apps)
summary(college$PhD)
row.names(college)[which.max(college$Top10perc)]
acceptance_rate <- college$Accept / college$Apps
row.names(college)[which.min(acceptance_rate)]
plot(college$Outstate, college$Grad.Rate)
> summary(college$Apps)
Min. 1st Qu. Median Mean 3rd Qu. Max.
81 776 1558 3002 3624 48094
> summary(college$PhD)
Min. 1st Qu. Median Mean 3rd Qu. Max.
8.00 62.00 75.00 72.66 85.00 103.00
> row.names(college)[which.max(college$Top10perc)]
[1] "Massachusetts Institute of Technology"
> acceptance_rate <- college$Accept / college$Apps
> row.names(college)[which.min(acceptance_rate)]
[1] "Princeton University"
> plot(college$Outstate, college$Grad.Rate)
acceptance_rate num [1:777] 0.742 0.88 0.768 0.837 0.756 ...
Elite Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
college$Outstate
college$Grad.Rate

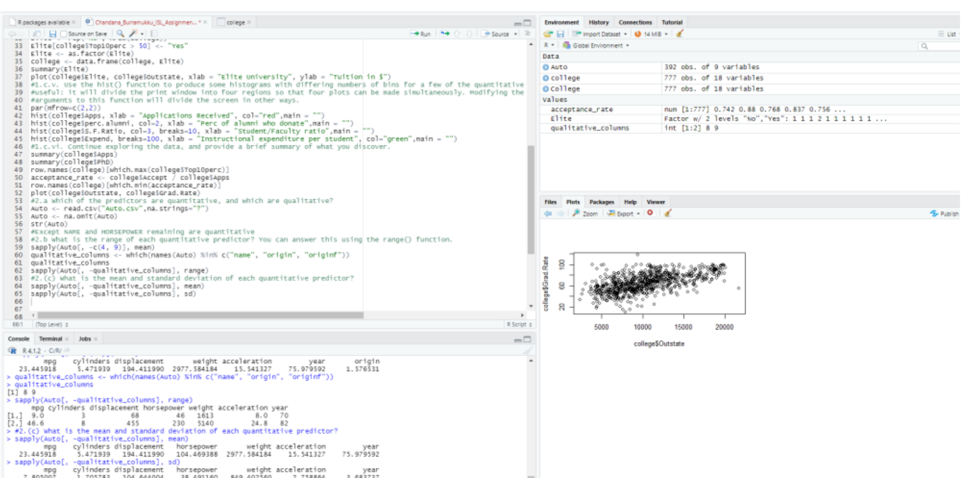
2.a



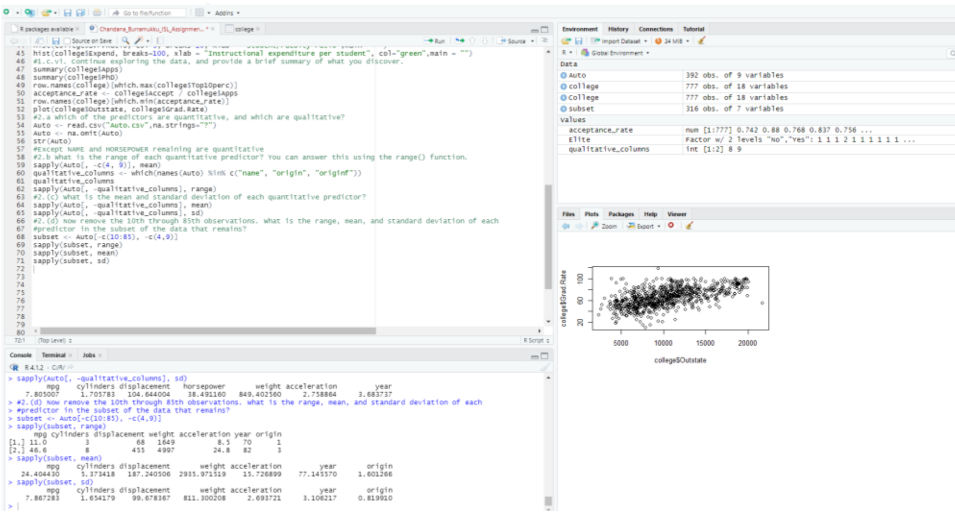
2.b



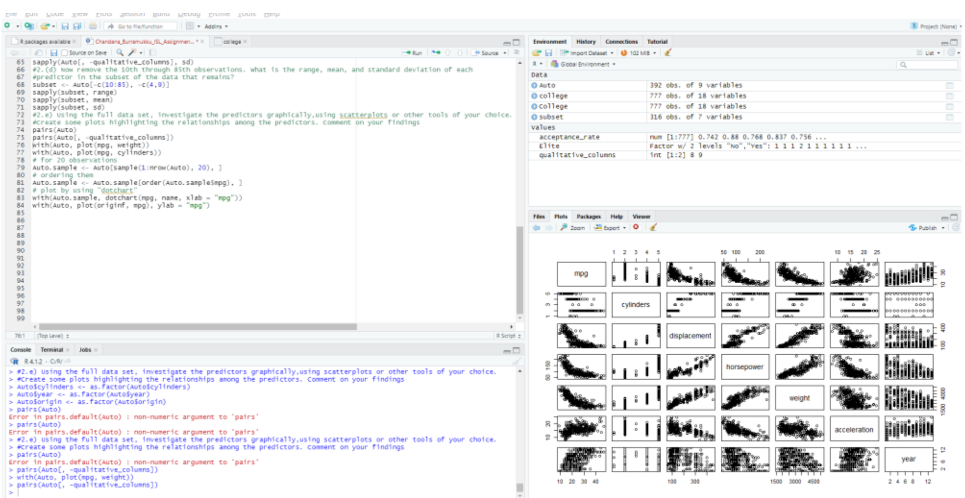
2.c

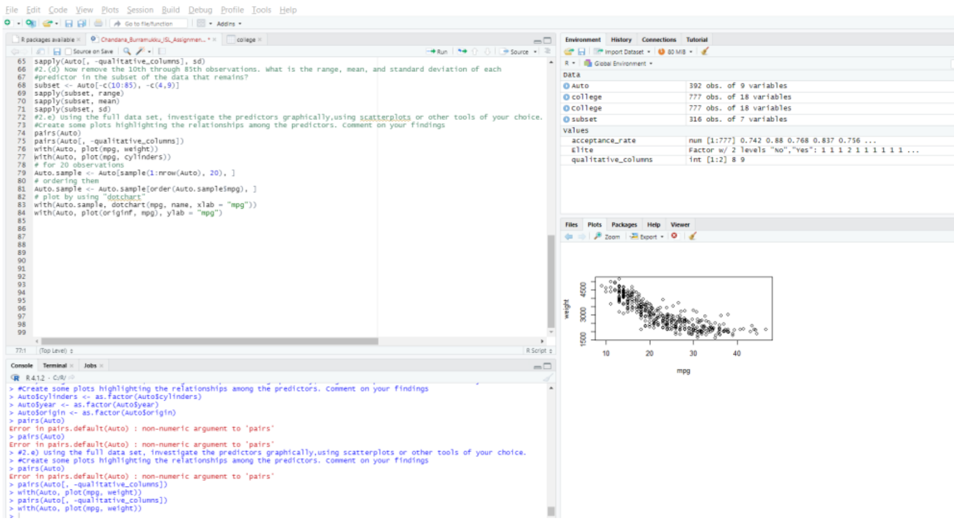


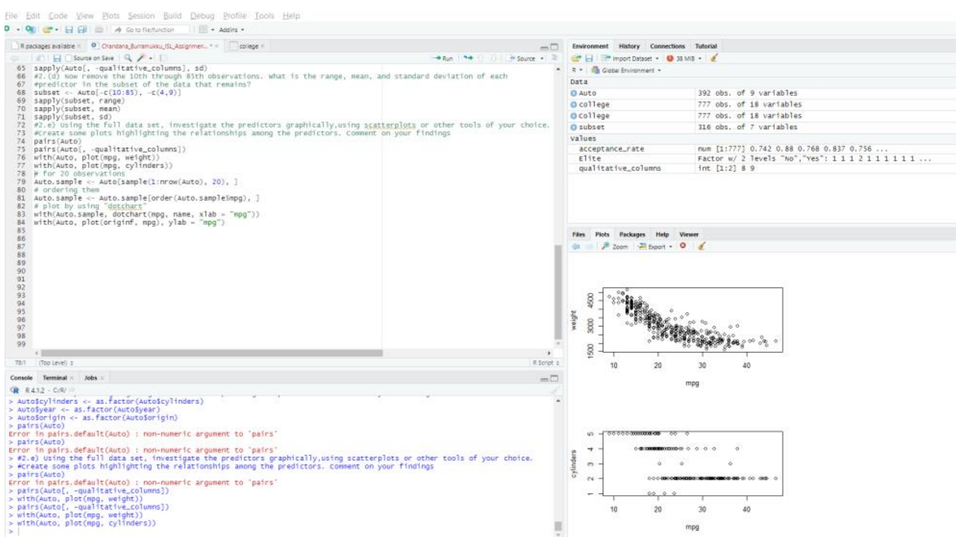
2.d

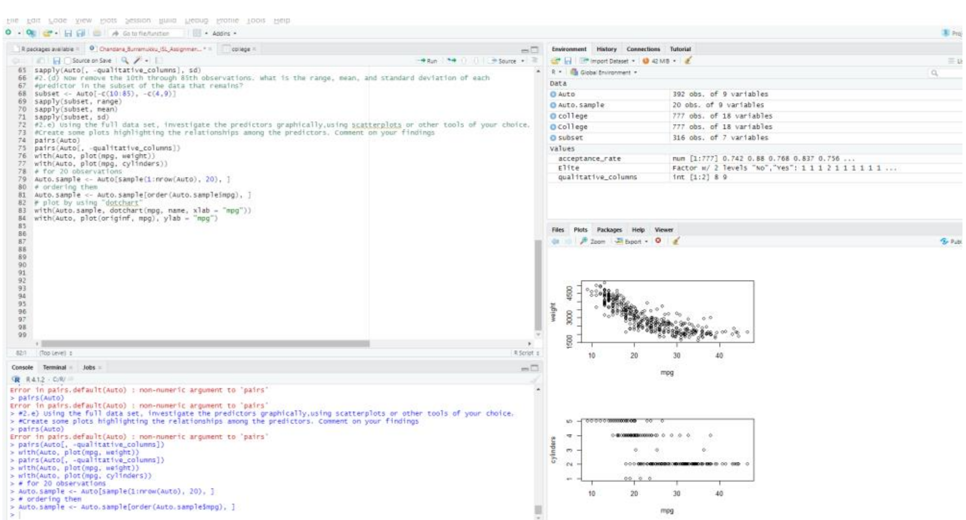


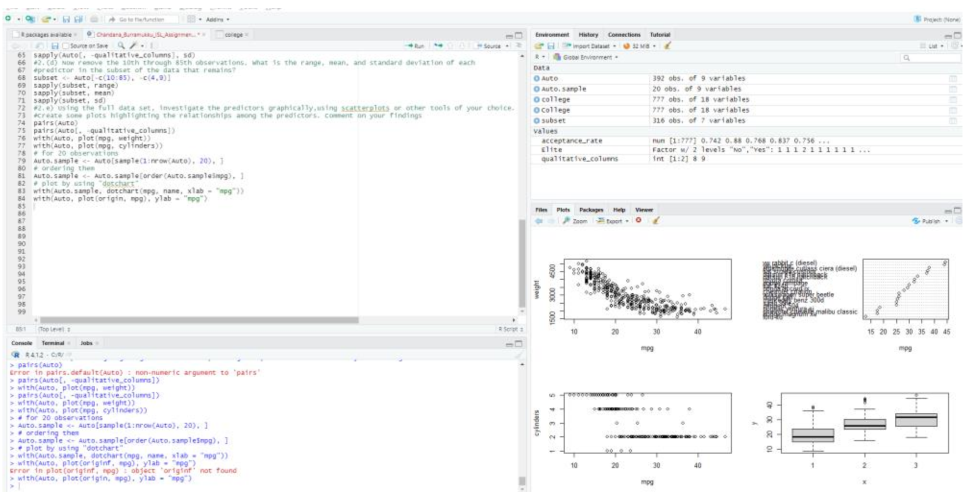
2.e



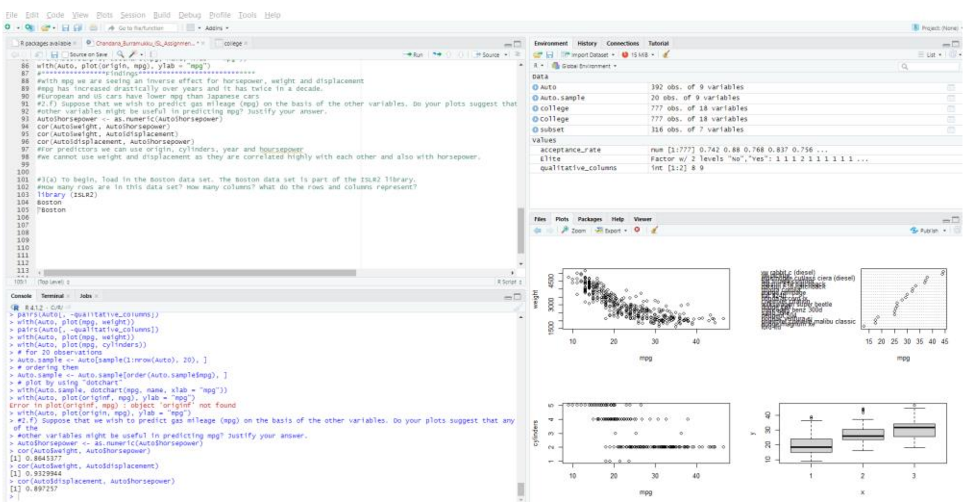

sapply(Auto[, -qualitative_columns], sd)
#2.(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each
#predictor in the subset of the data that remains?
subset <- Auto[-c(10:85), -c(4,9)]
sapply(subset, range)
sapply(subset, mean)
sapply(subset, sd)
#2.e) Using the full data set, investigate the predictors graphically,using scatterplots or other tools of your choice.
#Create some plots highlighting the relationships among the predictors. Comment on your findings
pairs(Auto)
pairs(Auto[, -qualitative_columns])
with(Auto, plot(mpg, weight))
with(Auto, plot(mpg, cylinders))
# for 20 observations
Auto.sample <- Auto[sample(1:nrow(Auto), 20),]
# ordering them
Auto.sample <- Auto.sample[order(Auto.sample$mpg),]
# plot by using "dotchart"
with(Auto.sample, dotchart(mpg, name, xlab = "mpg"))
with(Auto, plot(originf, mpg), ylab = "mpg")
Auto 392 obs. of 9 variables
college 777 obs. of 18 variables
college 777 obs. of 18 variables
subset 316 obs. of 7 variables
acceptance_rate num [1:777] 0.742 0.88 0.768 0.837 0.756 ...
Elite Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
qualitative_columns int [1:2] 8 9
> #Create some plots highlighting the relationships among the predictors. Comment on your findings
> Auto$cylinders <- as.factor(Auto$cylinders)
> Auto$year <- as.factor(Auto$year)
> Auto$origin <- as.factor(Auto$origin)
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> #2.e) Using the full data set, investigate the predictors graphically,using scatterplots or other tools of your choice.
> #Create some plots highlighting the relationships among the predictors. Comment on your findings
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
weight
mpg


> Auto$cylinders <- as.factor(Auto$cylinders)
> Auto$year <- as.factor(Auto$year)
> Auto$origin <- as.factor(Auto$origin)
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> #2.e) Using the full data set, investigate the predictors graphically,using scatterplots or other tools of your choice.
> #Create some plots highlighting the relationships among the predictors. Comment on your findings
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
> with(Auto, plot(mpg, cylinders))
weight
mpg
cylinders
mpg

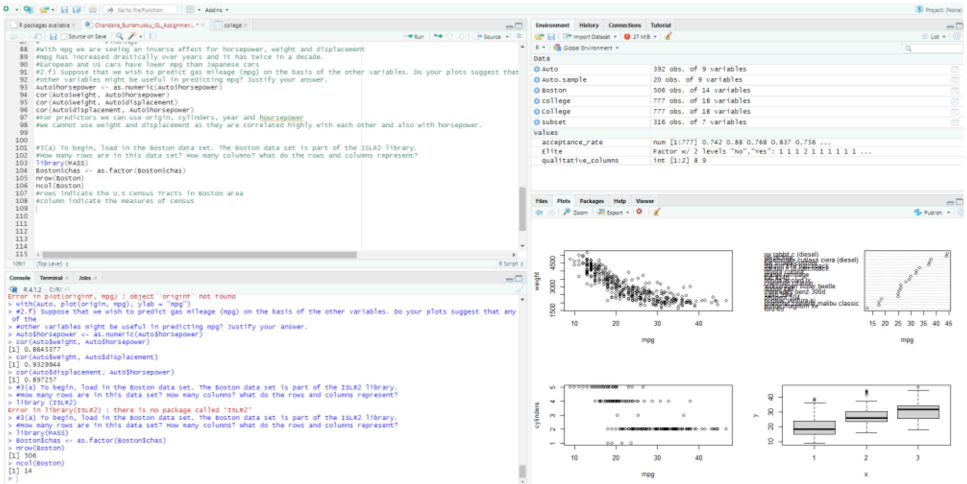

Auto 392 obs. of 9 variables
Auto.sample 20 obs. of 9 variables
college 777 obs. of 18 variables
college 777 obs. of 18 variables
subset 316 obs. of 7 variables
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> #2.e) Using the full data set, investigate the predictors graphically,using scatterplots or other tools of your choice.
> #Create some plots highlighting the relationships among the predictors. Comment on your findings
> pairs(Auto)
Error in pairs.default(Auto) : non-numeric argument to 'pairs'
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
> pairs(Auto[, -qualitative_columns])
> with(Auto, plot(mpg, weight))
> with(Auto, plot(mpg, cylinders))
> # for 20 observations
> Auto.sample <- Auto[sample(1:nrow(Auto), 20),]
> # ordering them
> Auto.sample <- Auto.sample[order(Auto.sample$mpg),]
weight
mpg
cylinders
mpg



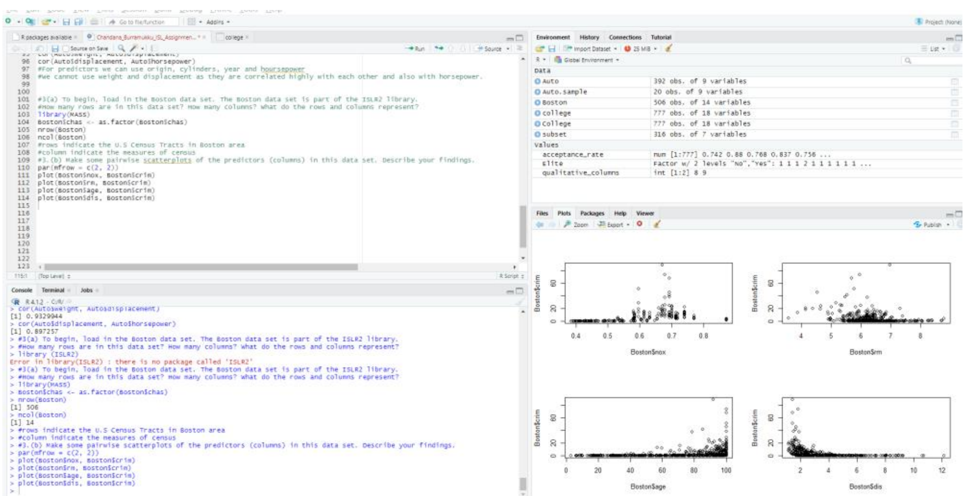
2.f

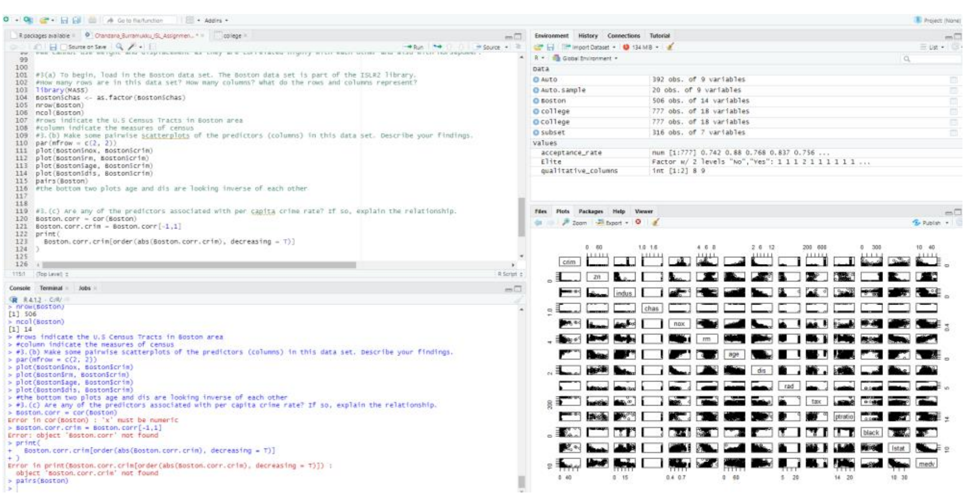


3.a

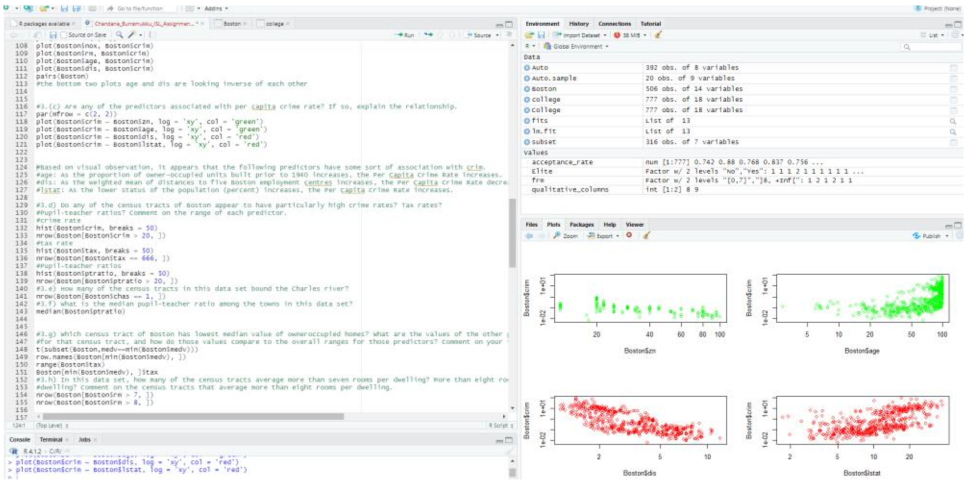


3.b



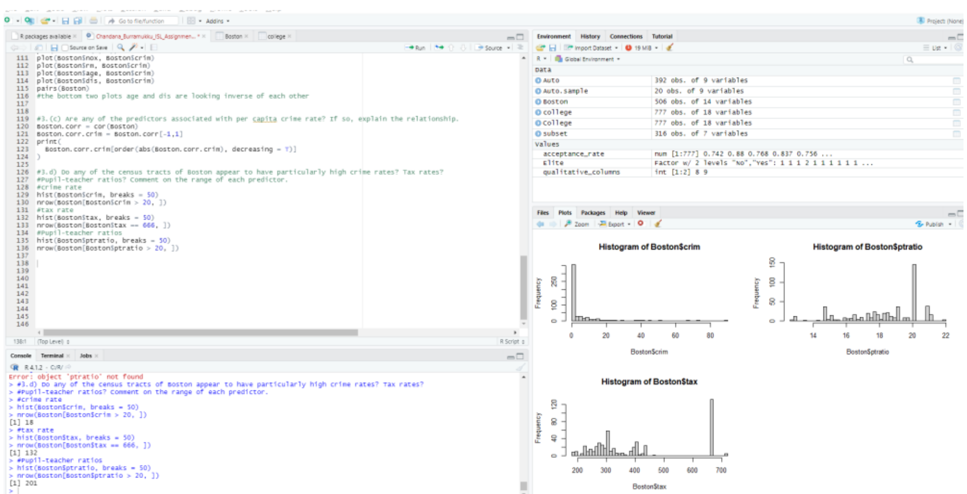


3.c



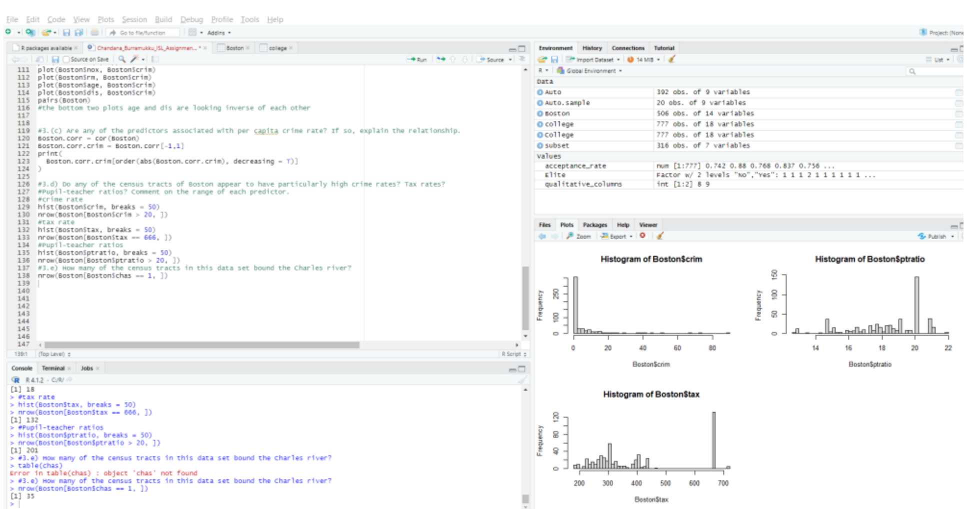
Boston$zn
Boston$age
Boston$dis
Boston$lstat

3.d



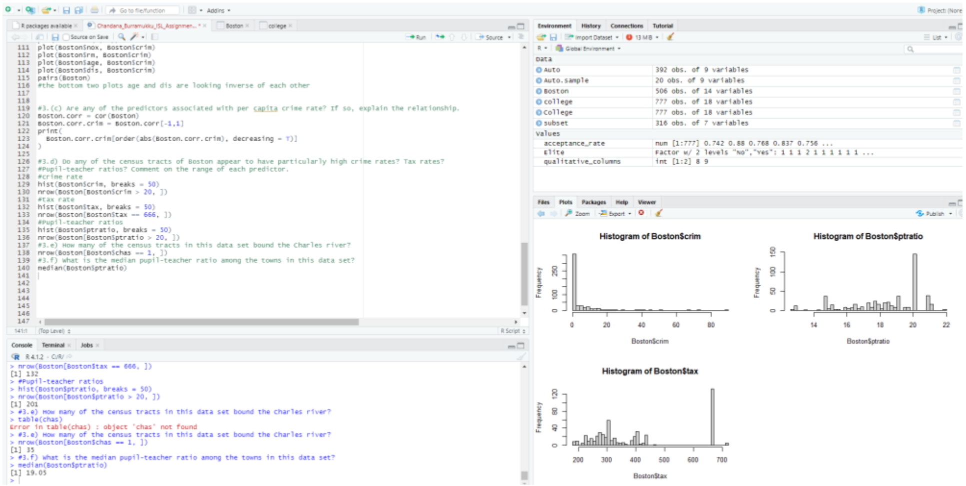
Histogram of Boston$crim
Histogram of Boston$ptratio
Histogram of Boston$tax

3.e

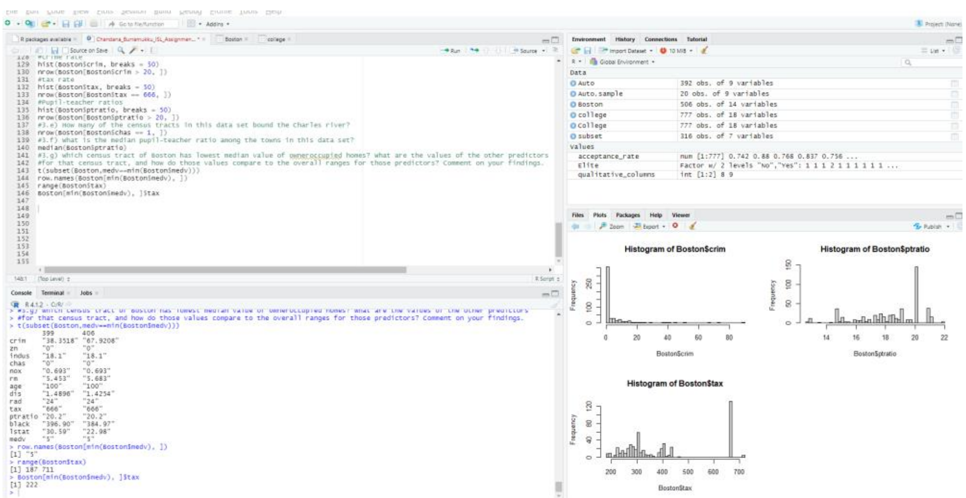


Histogram of Boston$crim
Histogram of Boston$ptratio
Histogram of Boston$tax

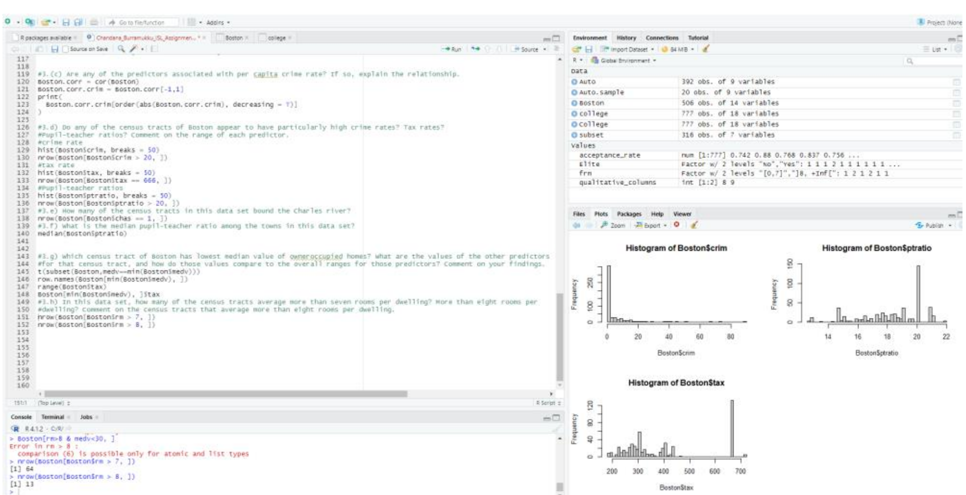
3.f


Histogram of Boston$crim
Histogram of Boston$ptratio
Histogram of Boston$tax
Boston$crim
Boston$ptratio
Boston$tax

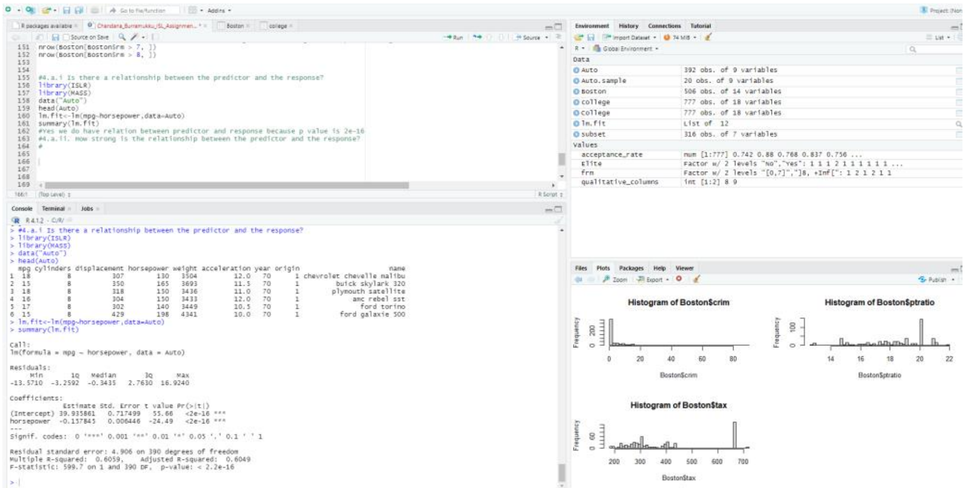
3.g


Histogram of Boston$crim
Histogram of Boston$ptratio
Histogram of Boston$tax
Boston$crim
Boston$ptratio
Boston$tax

3.h


Histogram of Boston$crim
Histogram of Boston$ptratio
Histogram of Boston$tax
Boston$crim
Boston$ptratio
Boston$tax

4.a.i



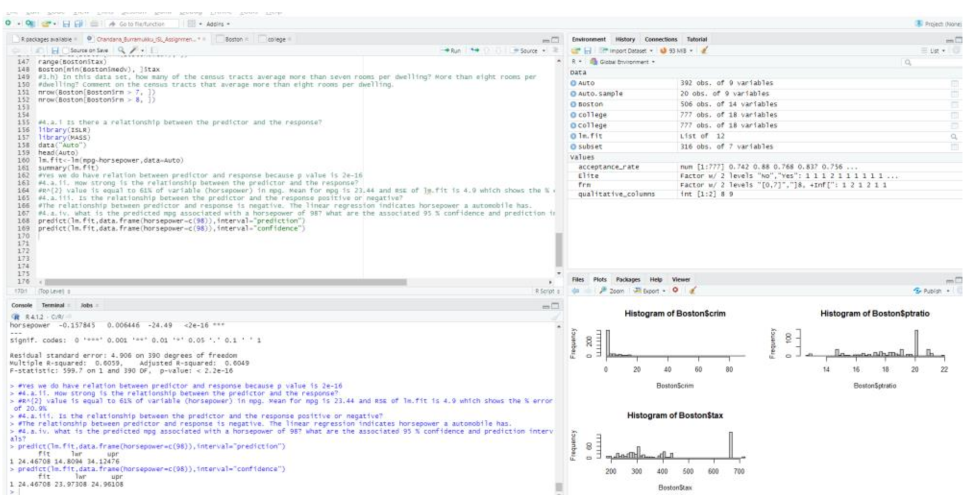
4.a.ii

R file

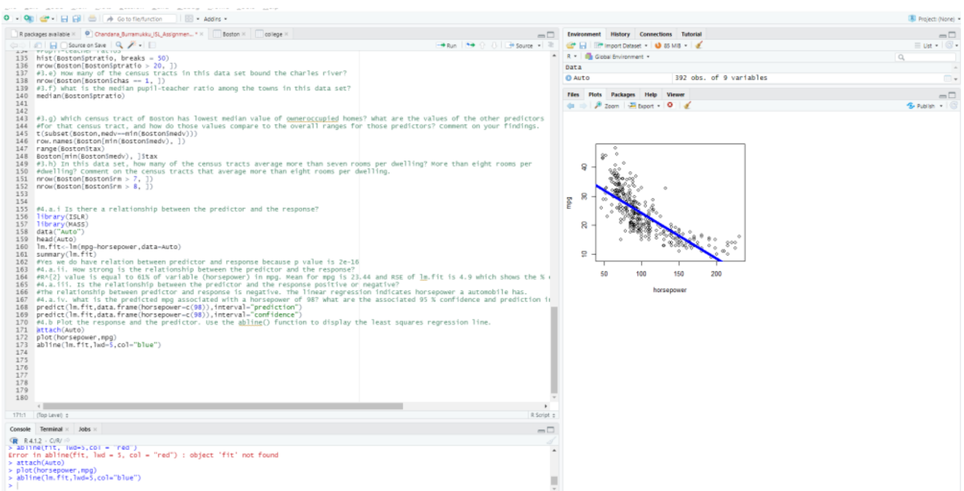
4.a.iii

R file

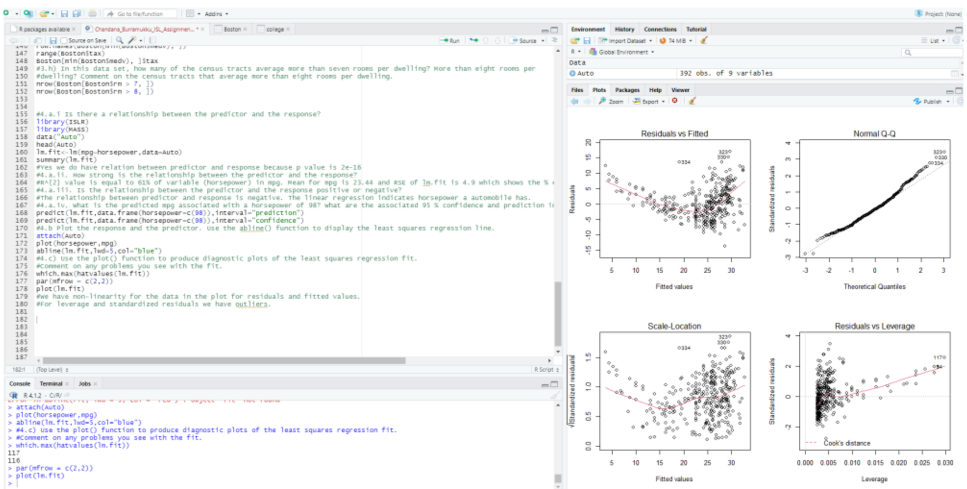
4.a.iv



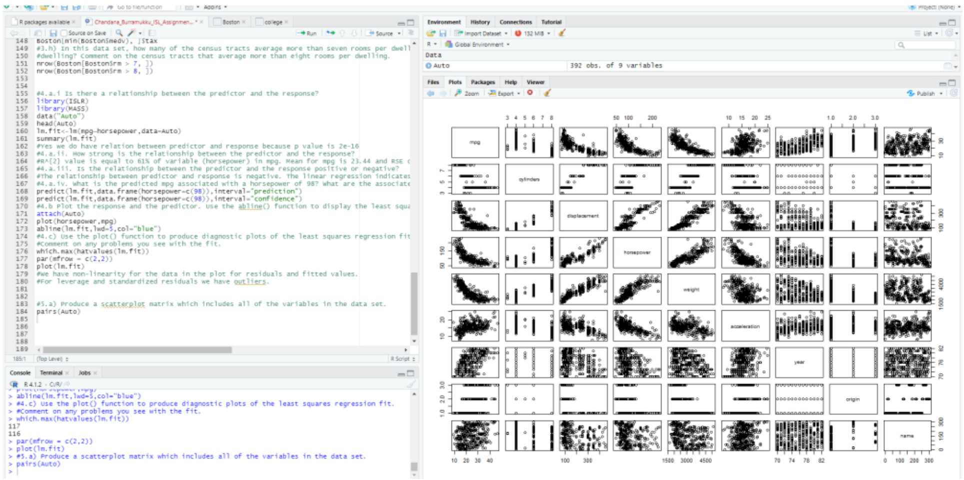
4.b



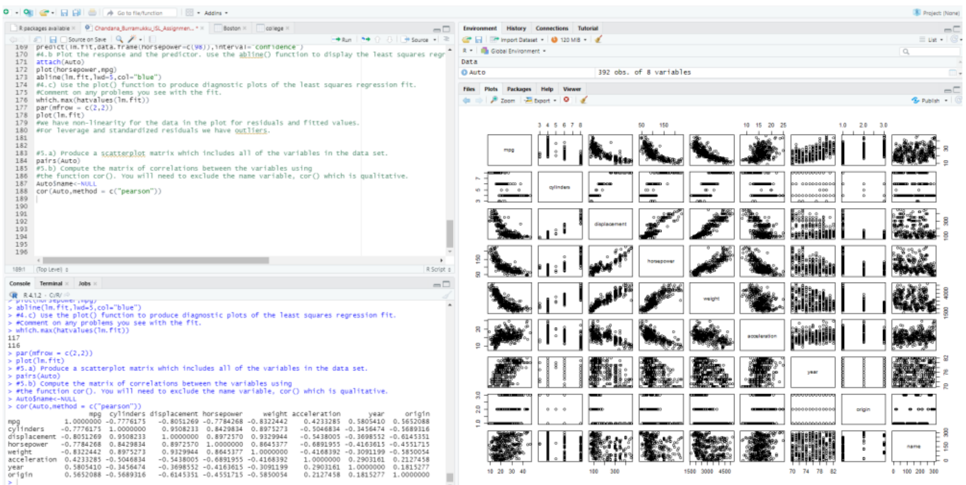
4.c



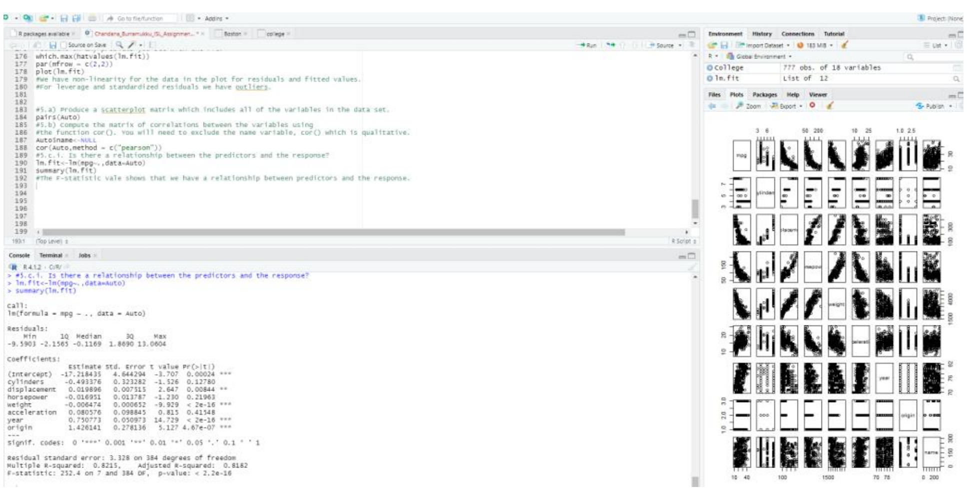
5.a



5.b



5.c.i



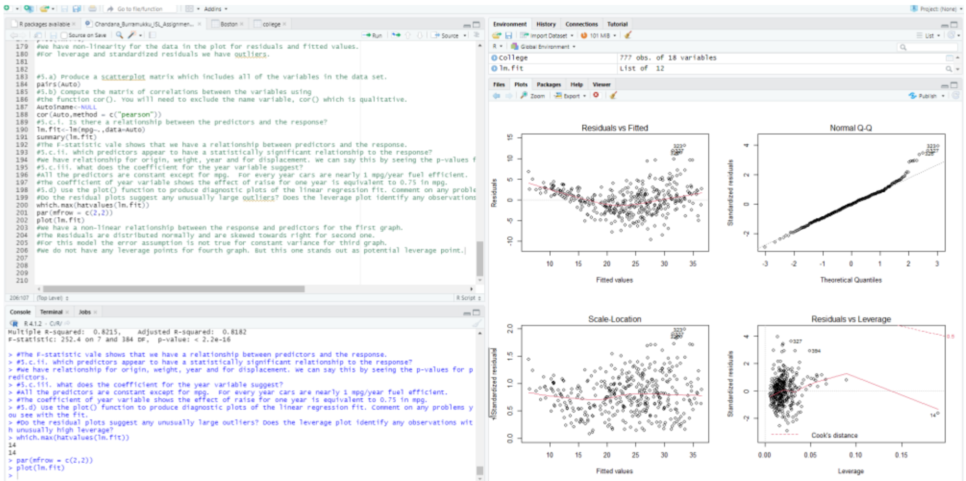
5.c.ii

R file

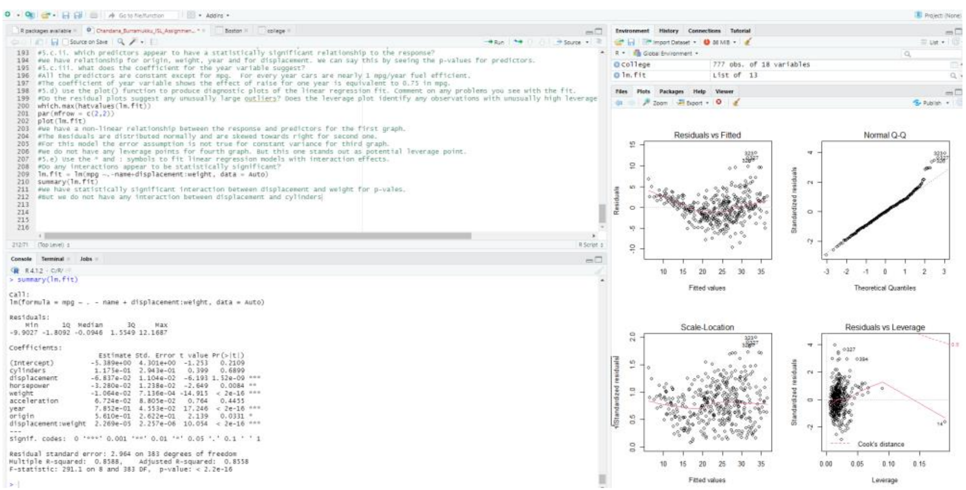
5.c.iii

R file

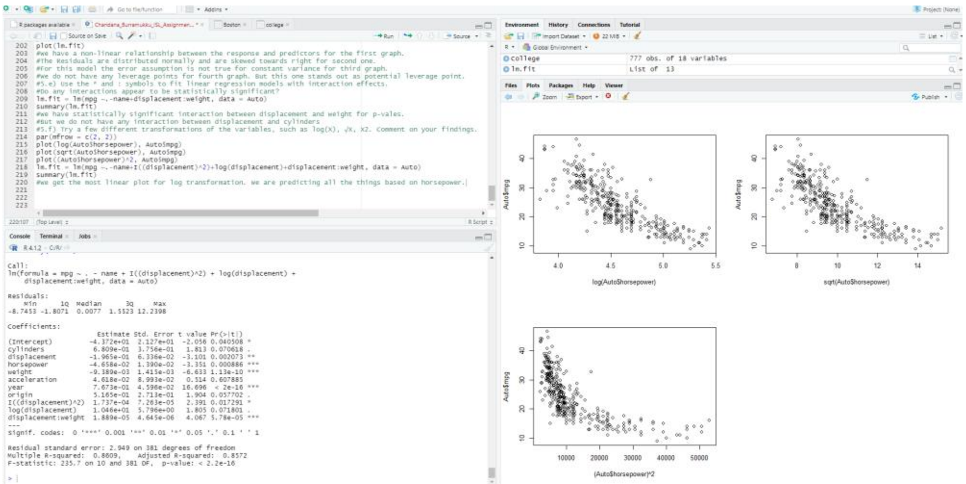
5.d



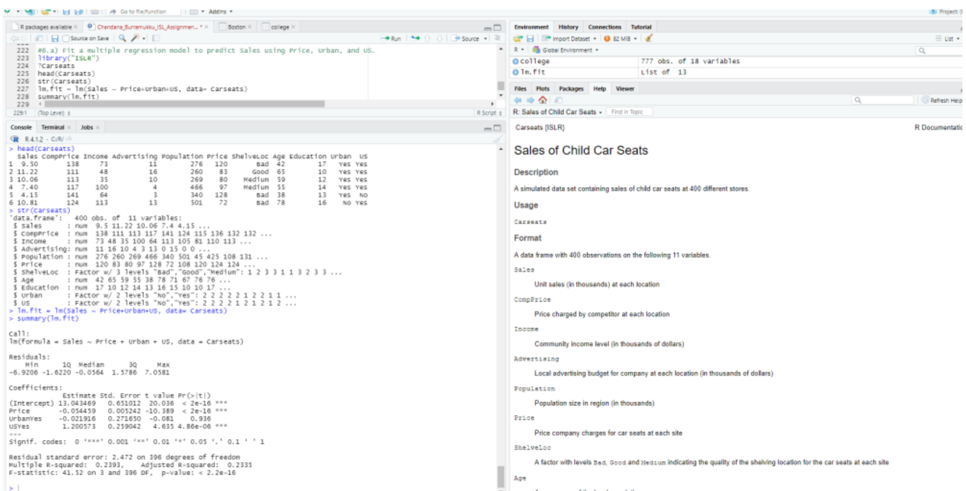
5.e



5.f



6.a



6.b

R file

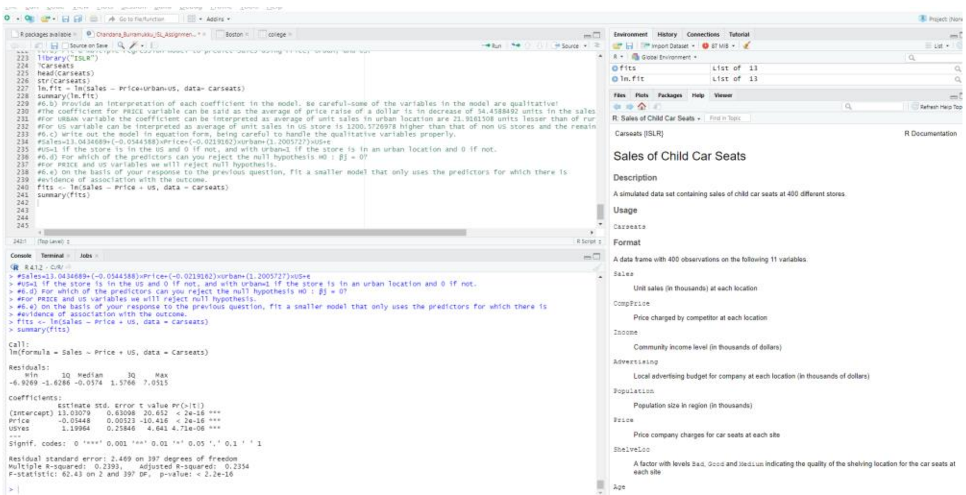
6.c

R file

6.d

R file

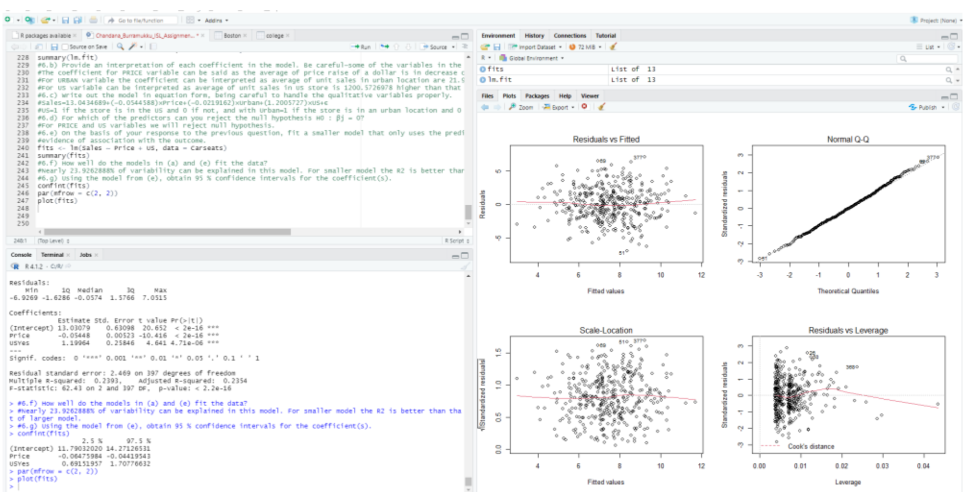
6.e



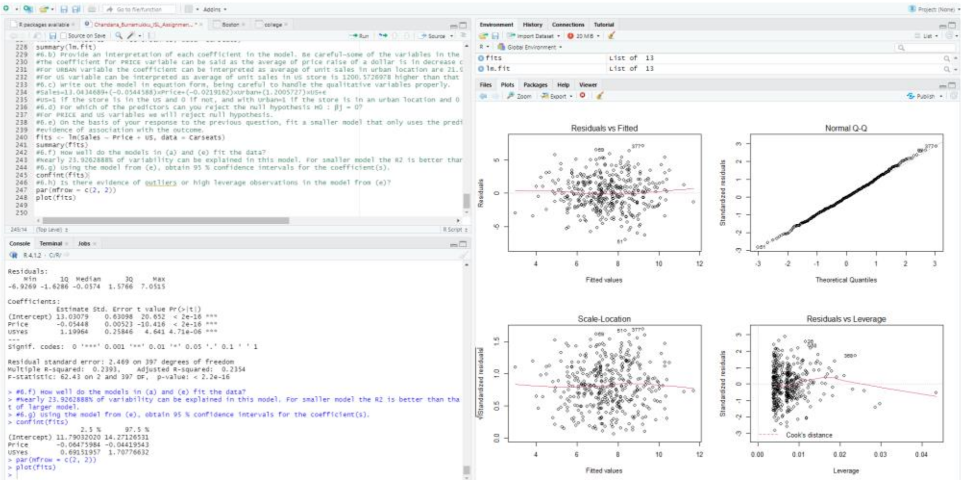
6.f

R file

6.g



6.h



summary(lm.fit)
#6.b) Provide an interpretation of each coefficient in the model. Be careful-some of the variables in the
#The coefficient for PRICE variable can be said as the average of price raise of a dollar is in decrease c
#For URBAN variable the coefficient can be interpreted as average of unit sales in urban location are 21.
#For US variable can be interpreted as average of unit sales in US store is 1200.5726978 higher than that
#6.c) Write out the model in equation form, being careful to handle the qualitative variables properly.
#Sales=13.0434689+(-0.0544588)xPrice+(-0.0219162)xUrban+(1.2005727)xUS+e
#US=1 if the store is in the US and 0 if not, and with Urban=1 if the store is in an urban location and 0
#6.d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?
#For PRICE and US variables we will reject null hypothesis.
#6.e) On the basis of your response to the previous question, fit a smaller model that only uses the predi
#evidence of association with the outcome.
fits <- lm(Sales ~ Price + US, data = Carseats)
summary(fits)
#6.f) How well do the models in (a) and (e) fit the data?
#Nearly 23.9262888% of variability can be explained in this model. For smaller model the R2 is better thar
#6.g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).
confint(fits)
#6.h) Is there evidence of outliers or high leverage observations in the model from (e)?
par(mfrow = c(2, 2))
plot(fits)
Residuals:
Min 1Q Median 3Q Max
-6.9269 -1.6286 -0.0574 1.5766 7.0515
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079 0.63098 20.652 < 2e-16 ***
Price -0.05448 0.00523 -10.416 < 2e-16 ***
USYes 1.19964 0.25846 4.641 4.71e-06 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354
F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16
> #6.f) How well do the models in (a) and (e) fit the data?
> #Nearly 23.9262888% of variability can be explained in this model. For smaller model the R2 is better than that of larger model.
> #6.g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).
> confint(fits)
2.5 % 97.5 %
(Intercept) 11.79032020 14.27126531
Price -0.06475984 -0.04419543
USYes 0.69151957 1.70776632
> par(mfrow = c(2, 2))
> plot(fits)
Residuals vs Fitted
Normal Q-Q
Scale-Location
Residuals vs Leverage
Fitted values
Theoretical Quantiles
Leverage
Residuals
Standardized residuals
Cook's distance