CS601C Fall 2024
Professor Paul Dantzig
Due 10/23/24

CS601C First Project

Complete the exercises and answer the questions outlined below. Your submission must be a combination of what the problem statement is you are working on, findings, R code, and plots. Obviously based on previous discussions you must submit an RMD and (PDF, HTML, or Word) documents. Your code must run successfully, and you must answer all the questions to get credit. I.e. The objective is to make it look as closely as possible to real statistics paper / data mining analysis.

Rubik: I have labeled this a combination of midterm and project. This course because it is fully online doesn't really provide an appropriate environment for standard exams or projects. This activity is going to be about 20% of your grade. Just because it has simple requests like "numerical summary of the variables / quantitative columns" doesn't mean you identify one column and give me the mean, median, standard deviation, etc. It means looks at the columns carefully and indicate which columns should be used for your study and why. If you just do the minimum you will end up with a grade like a B. If you want an A each question should be answered in detail. The better your discovery and justification aspects the better the grade including bonus points.

Note: Each of these datasets are widely used in beginning Data Analysis, Statistics, etc. courses. I of course change each of the datasets a little bit each semester nothing major but just enough that if you use a previous classes paper or go out the web and use any of those analysis numbers you will get wrong answers. Note: The easiest dataset is Auto, next is College, and the hardest to make conclusions about is Boston.

1. This exercise relates to the College data set, which can be found in the file College.csv.

It contains a number of variables for different universities and colleges in the US. The variables are:

Private:	Public/private indicator
Apps:	Number of applications received
Accept:	Number of applicants accepted
Enroll:	Number of new students enrolled
Top10perc:	New students from top 10% of high school class
Top25perc:	New students from top 25% of high school class
F.Undergrad:	Number of full-time undergraduates
P.Undergrad:	Number of part-time undergraduates
Outstate:	Out-of-state tuition
Room.Board:	Room and board costs

Books:	Estimated book costs
Personal:	Estimated personal spending
PhD:	Percent of faculty with Ph.D.’s
Terminal:	Percent of faculty with terminal degree
S.F.Ratio:	Student/faculty ratio
perc.alumni:	Percent of alumni who donate
Expend:	Instructional expenditure per student
Grad.Rate:	Graduation rate

(a) Read the data into R. Make sure that you have the directory set to the correct location for the data or use file.choose().

(b) Look at the data using the View() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later.

(c) i. Produce a numerical summary of quantitative attributes in the data set.

ii. Produce a scatterplot matrix of the first ten columns of the quantitative data. Recall that you can reference the first ten columns of a matrix A using A[,2:11].

iii. Produce side-by-side boxplots of Outstate versus Private.

iv. Create a new qualitative variable, called Elite, by binning the Top10perc attribute. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

v. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

vi. Produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) 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.

vii. Continue exploring the data, and provide a brief summary of what you discover. This is where you get to show me what you have learned about correlation, linear regression, and multiple linear regression. First correlate the quantitative variables you think are important, find two attributes that correlate well to do linear regression. See if you can find a third variable to do multiple regression.

2. This exercise uses the Auto data set. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Remove 10 records (your choice which). What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? Did the removal of the records cause any of those calculations to change substantially (more than a few percent).

(e) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. (i.e. use tools such as correlation, linear regression, multiple linear regression). Create some plots highlighting the relationships among the predictors. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

3. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship. Can you find any correlations between per capita crime rate and other quantitative columns? Could use linear regression or multiple regression?

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates?

Pupil-teacher ratios? Comment on the range of each predictor.

(e) How many of the suburbs in this data set bound the Charles River?

(f) What is the median pupil-teacher ratio among the towns in this data set?

(g) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.