STAT346: Statistical Data Science I

HW #1 – Due: Friday, September 25, 2020 by 6 p.m.

September 11, 2020

Instruction: Answer to the following questions and write your report using R Markdown. You may use the R Markdown template, accompanied with this homework assignment. You should submit two files, through KU Black Board system (https://kulms.korea.ac.kr), which should have the following naming format:

- `stat346_hw1_your_id.rmd`
- `stat346_hw1_your_id.pdf` or `stat346_hw1_your_id.html`

1. This exercise relates to the `College` data set, available in `library(ISLR)`, that is,

```
#install.packages("ISLR")
library(ISLR)
data(College)
```

   It contains a number of variables for 777 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) Use the `summary()` function to produce a numerical summary of the variables in the data set.

   (b) Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or

variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

(c) Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

(d) Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. 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%. 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`.

```
Elite=rep("No",nrow(College))
Elite[College$Top10perc>50]="Yes"
Elite=as.factor(Elite)
College=data.frame(College,Elite)
```

(e) Use the `hist()` function to 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.

(f) Continue exploring the data, and provide a brief summary of what you discover.

2. This exercise involves the `Boston` housing data set.

(a) To begin, load in the `Boston` data set. The `Boston` data set is part of the `MASS` library in R. How many rows are in this data set? How many columns? What do the rows and columns represent?

```
library(MASS)
data(Boston)
```

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

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(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 owneroccupied 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.