STA 303H1S / 1002 HS -Winter 2019 Assignment # 1
“Points”

Posted: Wednesday, January 16, 2019 (minor updates made by 10am, January 17)
Due: In Crowdmark by 10pm on Monday, January 28, 2019.
Late assignments will be subjected to a penalty of 20% per day late.

Instructions:

- Use R (or R Studio) to do the analysis for the following questions.
- Use a benchmark significant level of 5%.
- Compile your solution as a PDF document (Word, LaTeXor Rmarkdown can be your base).
- Presentation of solutions is very important. Your assignment should have two main sections-Solutions and Appendix. Include relevant plots, and quote relevant numbers from your R output for your solutions. Unless asked otherwise, include all R codes and output in your Appendix. Marks will be awarded for excellent presentation.
- Write and submit **your own work**. For instance, personalized your code as much as possible, using your first name. **All plots produced must be given a title with the last 4 digits of your student number**.
- Where appropriate, your answers are expected to be written in plain English.

Grading: There are 2 main questions. The grand total for this assignment is 100 marks. A general marking scheme for each part is given below:

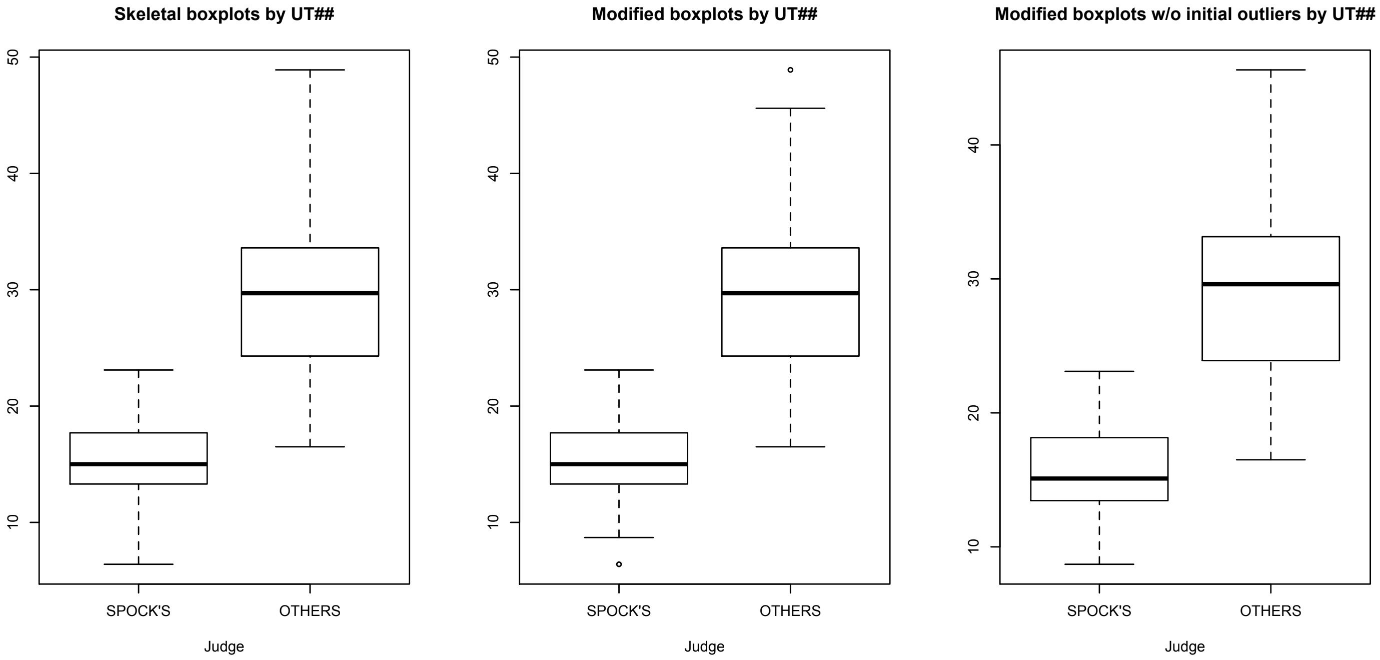
Per Question Part

- 100%: complete and correct answers
- 80%: answers with minor problems
- 60%: good answers that are unclear, contain some mistakes, missing components
- 40%: poor answers with some value
- 20%: irrelevant answers
- 0: unanswered questions

Presentation and Appendix

- 10 points: well presented, easy to read, proper English used, R code and extra output in Appendix
- 6 points: good presentation, some R code in main write-up.
- 2 points: poor presentation, handwritten, hand-drawn diagrams, unnecessary R code in main section.
- 0 point: illegible, missing R-codes/output

1. (30 marks) Consider the 3 pairs of side-by-side box plots given below, which were drawn in R. The data for this question is in the file “juries.csv”.



(a) (10 marks) Recreate the side-by-side box plots and include the last four digits of your student number in the title. For each pair of box plots, provide the single line of R code, beginning with the R function- `boxplot` used to recreate it.

(b) (15 marks) For each box plot specify the values of the following - the first quartile, the second quartile, the third quartile, the end points of the two whiskers, the extreme (outlier) points (identified as small circles), if any, and the limits of the 1.5IQR Rule Range.

(c) (5 marks) Compare the three pairs of box plots. Which pair best represents the data and why?

2. (60 marks) Consider the data, “bbw99.csv” based on the birth weights of 99 babies, along with the smoking status of their mothers. Answer the questions that follow. The variables in the dataset are:

- `id`- an identification number from 1 to 99
- `bwt`- baby birth weight in ounces
- `smoke`- smoking status of mother (coded as 0, if she was a non-smoker, and 1, otherwise)

(a) (5 marks) Which variables are categorical? Name the levels of each categorical variable.

(b) (20 marks) Conduct an appropriate hypothesis test to determine whether there is a difference in the mean birth weight between babies born to mothers who were smokers and babies born to mothers who were nonsmokers. Include the following:

i. Side-by-side boxplots

ii. Null and Alternative Hypotheses

iii. A test statistic and it’s distribution

iv. Test assumptions

v. Test diagnostics (checking model assumptions)

vi. P-value

vii. Results (brief discussion and conclusion)

(c) (5 marks) Name two(2) statistical methods which are equivalent to your method used in part (b) above.

(d) (25 marks) *Create a subset of the data by removing the row of observations whose 'id' matches the last 2 digits of your student number*. For instance, this can be done in R by `shivon.subset` $<-$`shivon.data`$[-1,]$ if my student number ends with '01'.
Then redo the analyses of part (b) above with your data subset.

(e) (5 marks) Compare your results of part (b) and part (d). Do you think that the observation removed was influential?