**Exercise#2**

Write a two-three-page summary about your approach to the presentation and analysis. Your summary could include some of the following discussions:

1. How did you prepare the data for analysis?

* Read in raw data, using read excel command. For this analysis I used R language.
* get to know the raw data by using summary command (how many rows/columns, how many NAs in rows and columns, what are the column names, column names correspond to which questions in the survey),
* excluded the international student according to survey report and analysis instruction
* clean column names which contain symbols such as …, space, question mark, etc
* exclude those records(rows) that are empty through out all questions and the records answered “I do not wish to complete the XX University Doctoral Exit survey.” In column Q12.
* Now we got the cleaned data call “file” and then divided to two group : the URG(under-represented students) and the Non-URG group(Non under-represented students) based on the ppt of survey report instruction

1. How did you address any missing values?

For the records that are completely empty (that is, if a row or a record fail to answer all the questions), we need to exclude them for the further analysis, and also in column Q12 if they answered they do not wish to complete the XX university doctoral exit survey, they also need to be excluded. When review the data column by column, we also count how many NAs in each column. If some columns’ NA rate is higher, we will drop the column when doing machine learning, if the NA rate in a column is lower, we can use imputation to fill NA out for the further statistics and machine learning calculation. For the methods of imputation, there are lots of imputation methods. For this analysis, we use average imputation.

1. Which statistical methods did you use for your data analysis, and why?

For the faculty interactions core area, when we compare whether there is a significant difference between URG and non-URG students for different categories, we use pairwise t-test method. When p value is larger than 0.05, this means that there is not much significant difference between two different categories.

When I visualized to what extent their experience of different statement (converted extent of strongly disagree, disagree, ambivalent, agree, strongly agree to numbers 1,2,3,4,5), I use basic statistics of mean.

1. Did you determine response rates for the different questions? Why or why not?

Yes

1. Did you consider weighting any of the data? If yes, why?

Yes.

For this data, there are lots of columns. We need to determine how many columns has relatively more significant influence on the overall satisfaction in the survey. Therefore, we need to weighting the data using feature importance analysis.

1. What appropriate tests of statistical significance did you consider using (and used), when evaluating differences across different sub-populations?

ANOVA is suitable to evaluating differences across different sub-populations.