CAP 5300, Project: What Is Expected during Presentation

In your presentation, you will need to address the following tasks:

1. Conduct some descriptive analysis on the data, figuring out:

   - distributions of variables,
   - variables that appear to be strongly related with each other (using appropriate methods to quantify the relationships based on whether variables are numerical or categorical).

2. Conduct one-sample statistical inference for your response variable of interest, e.g.

   - Construct confidence interval for its population mean value. Make sure to check the assumptions, and interpret the confidence interval accordingly.
   - Conduct a hypothesis test against a certain value. Interpret its results.

3. Conduct statistical inference (construct a confidence interval, conduct hypothesis test against a pre-specified constant) for categorical explanatory variable of interest, e.g. for a population proportion for a certain category of that variable.

4. Conduct two-sample statistical inference to address a question like: "Is there a significant difference between two groups of observations according to some categorical variable?". For example, if you had a *Salary* response variable, and a *Sex* categorical variable, then "Is there a difference in salary between males and females?" Check the assumptions, interpret the results.

5. In addition, proceed to conduct **one** of the following tasks:

   (a) Figure out the explanatory variable(s) that appear to be the best predictor for your response variable of interest (here you may either use a bunch of simple linear regressions and evaluate their quality of fit, or you can just run multiple linear regression, despite it not being covered in the class yet).
   (b) For confidence interval and hypothesis test conducted in part (2), run "sanity check" simulations as follows: take random sub-samples of fixed size from your full data set, calculate the confidence intervals and hypothesis tests against the true full sample mean, record the coverage and type I errors obtained. Comment on whether those are as expected.
   (c) Construct confidence interval for some arbitrary population parameter that involves your response variable of interest. (anything **but** sample mean or proportion; e.g. you can do median, quantile, correlation, etc) E.g. "population median salary", or "population correlation between salary and GPA". **Note**: Be able to walk us through the process of building that confidence interval.
   (d) Conduct some alternative task of comparable complexity to parts $(a) - (c)$ that pertains to the material covered in class, having discussed it with instructor beforehand.