UCB W203, Supplemental Assignment Guidelines (SAG). Oleg Melnikov, 6/27/20

This guideline will help you effectively structure your exciting data science story. Everything you include in your report must be directly relevant and help you build your argumentation. In this process you need to make it easy for your audience to digest your findings. That is you need to accent and interpret interesting (and potentially problematic) patterns in your work. Assume your audience is somewhat familiar with statistical concepts and needs your refresher on some concepts as you tell your story. That is don't assume your audience observes the same interesting patterns that you do in your outputs; and avoid use of words such as "*obvious*", "*clear*", "*trivial*", "*easy to see*", etc. These are usually condescending to your audience. To the contrary, many deductive reasoning steps and implications are not as "*obvious*" as they initially appear to be. In summary, provide reasonable explanations of your work.

Make your story exciting! Make it interesting! Make it dramatic! Make it WOW! Find patterns, outliers, relationships that crush intuitions. Let these insights drive your investigation of the data. Later you will learn to choose feature transformations and machine learning models based on these patterns and surprises in your data.

Disclaimer: This is subject to change, which I will try to keep to a minimum with an advanced announcement, whenever possible.

Rule 0: If in doubt, just ask

Rule OPAC: your work should be **Organized, Precise, Analytical, and Concise**

- For example, "*it's approximately normal*" is vague. State exactly what random variable is approx. normally distributed. I need to see that you differentiate between $X, Y, \bar{X}$, etc. (assuming these are unambiguously defined).
- Assure no errors or warnings in your code. These inject doubts into the correctness of your outputs. Upgrade R and packages accordingly.
- Provide helpful comments for complicated R code

Rule EDA: Exploratory Data Analysis

Consider these components for your EDA (including corresponding analysis of the output, of course) for all your evaluated variables:

1. **Data Structure**: are values real-valued, categorical, factor, ordinal, ...
2. **Interpret a distribution**:
    a. Basic Stats:mean (1st), median, mode, trimmed mean, quartiles, range, SDev (2nd), skew (3rd), kurtosis (4rd), IQR, ...
    b. Visualization (histogram or boxplot, ...)
    c. What theoretical distribution approximates population (Normal, t-student, Bernoulli, Binomial, ...)
3. **Missing values** (counts, or heatmap). Ideate about imputation.
4. **Correlations/relationship** (for multivariate analysis):
    a. correlation matrix or pairplot (= matrix of scatterplots)
    b. Are observations paired (eg. one person feeds two features in the study)?

Rule QQ: Quality over Quantity (see Oleg's Live Session notebooks for formatting examples).

1. Every output should have a purpose and a corresponding analysis/interpretation. Plots/tables without interpretation may result in deductions.
2. Limit your plots and tables to reasonable quantity and sizes.
    a. Size/position your plots optimally to avoid wasting space.
    b. Overlaid and side-by-side plotting allows easier comparison of trends
    c. If you overwhelm your audience with insignificant plots (for the sake of plotting), they will lose their confidence and interest in your work.
        i. You are encouraged to iterate through many-many plots outside of the report. Then choose the worthy plots to include and analyze.
    d. If the original data has very long column names, use meaningful mnemonics instead.
3. No data dumps in the report. The first/last few raw observations should suffice. Typically, there is no reason for printing more than 5-10 rows.
4. Round all your decimal points so that the report is easy to read.
5. Assure the plots are legible, properly labeled, titled, sized. Use plotting space efficiently.
6. Do not overplot!! (learn ways to avoid it!)
7. If printing tables of numbers (correlation matrices), background color gradient is very effective.
8. Less is more! Excessive, disorganized, illegible, unlabeled plotting will result in deductions.

UPDATE 7/1/20: Whenever suitable, please use page numbering, clearly identified sections - to separate solutions for questions, numbered lists (over bullet lists), PDF over HTML - all these make it easier for me (and any of your audience) to navigate around your documents and leave in-document commentaries.