**Stage II: Exploratory Data Analysis**

The second stage of your project is to conduct Exploratory data analysis.

1. Most likely you will need to **“clean” your dataset** first. Make **note of any problematic data** and observations that need to be removed. Consider implications about any decisions you make about missing data. This website shares some simple approaches to missing data (and the relative advantages/disadvantages of each approach): <http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html>
2. Produce **descriptive statistics** (5-number summaries for continuous variables; tables of counts or proportions for categorical variables) for all relevant variables in your data set. **Plot continuous variables using histograms**.
3. Explore the relationships between important pairs of variables both graphically and numerically. Depending on the type of your response and explanatory variables, you may consider graphs such as **boxplots, scatterplots, or segmented bar charts**. You may consider summary statistics (like mean, median, or standard deviation) **by group, correlations, regression equations, and two-way tables with proportions.** At this stage plots can be loose in terms of **titles and labels,** but for your final paper it is essential that your figures have (meaningful) captions and axis labels!

**Your EDA report, to be turned in on Moodle, will meet these guidelines:**

In ***no more than 3 pages*, summarize the main findings** of your exploratory analysis, **referring to specific plots and summary statistics where necessary**. In addition, **describe your plans for building** models to address important research questions, including **which variables will be important to consider** in light of your exploratory analyses. This report should be **meaningful and readable to someone familiar with statistics but unfamiliar with your particular research topic and dataset** (i.e. your professor). Give **concise but precise statements interpreting summary statistics**, etc. – in the context of your data set and research questions you pose. Avoid vague terms like “this data”, “these results”, etc. Also avoid cryptic variable names that you may have used in R. A report like this might be something you’d share with collaborators or store as a reference as you proceed with your analysis.

(a) The Main Body of your EDA report should follow these guidelines:

• No more than 3 pages

• Begin **with a short paragraph introducing your project and primary research questions**. (This introduction will be expanded into several paragraphs for the final paper.)

• Use your **graphical and numerical summaries** to tell a story, **supporting your conclusions with summary statistics**. Weave numerical summaries seamlessly into your text, and refer to graphs where appropriate.

• You do not need to include EVERY plot or table you made in the report. Include at least **2 interesting plots** (if not more!). Name each plot (e.g. Figure 1) so they are easily referred to in your report, and format the figures neatly within your report (without taking up too much space). These exploratory plots don’t have to be perfect in terms of titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels!

• Preview **directions** you plan to go with modeling. What **models** will you begin by fitting, and what variables will be involved.

• Write well! Complete sentences, good flow, proper grammar, etc.

(b) Your EDA report should also include an Annotated Appendix and References section (not included in the 3 page limit) which include these elements:

• Clear definitions of important variables and the (properly cited) source of the data.

• Tables and figures that are informative but were not referenced specifically in the main report. Include a short annotation – one or two sentences on what they show.

• A citation for each reference article (in a standard format) you included in your proposal. Also include a link, if appropriate.

• You should also upload to Moodle:

* An RMarkdown document with R code and (commented) output so that I can trace how you constructed your final data set, what the results of your exploratory data analyses were, and what plots you generated. Comments should be short, but clarify what you’ve done and why.
* The knitted Word document (not html) that was created from the RMarkdown above.
* The data file (as .txt, .csv, or .xls).