General Directions for Analyzing Project Data

This is a general guide to help you get started in analyzing your data.

Work with your group members. Refer to your friendly neighborhood statistics instructor if you have any questions/problems.

Some steps may not apply in your particular case.

Preliminaries

- Choose one member of the group to be the one whose computer will contain the project.
- Start an "RStudio project" for your project
- (If needed) Download the data file from the internet/email to the computer.
- Use the Upload button in Rstudio Files pane (bottom right) to upload your data file into the project folder.
- Click the data file in the Files pane, and choose the import option to import the file.
- Start an RMarkdown report file.
- Delete the material after the first code chunk, then add a code chunk to load the hanoverbase library and another code chunk for the data import code.
- Use the report space to take notes and create graphs. You will clean it up later removing what is not needed and polishing the graphs.
- If you prefer you can instead create a second RMarkdown document, called "journal" or something of the sort, to store the progress in your analysis. Then you will move parts of the journal into your final "report".
- You will need to generate a lot more graphs and compute a lot more numerical summaries than what will go into your final report.
- Save and knit often.

Early Explorations

- The first primary goal is to **own** your data. You should get a grasp for the basic behavior of your data, and know all you can about it, including what your data had for breakfast.
- Compute appropriate numerical summaries and graphs for each individual variable

    - Categorical: Frequency tables, bargraphs, pie charts, dotplots
    - Quantitative: Favstats (Mean, Std.dev, quartiles), skewness, outliers, histogram, boxplot

- Summarize your conclusions in the journal. Don't try to keep important information in memory, write it down in your journal or report.
- Don't worry about finalizing graphs/tables. This is "exploratory": It's for you to learn what is going on with your data.
- Consider interactions between your variables

- **–** Categorical - Categorical: 100% stacked bar charts
- **–** Categorical - Scalar: boxplots, numerical summaries by category
- **–** Scalar - Scalar: scatterplot, "loess" smooth line, linear regression if appropriate, correlation

- Decide on variable transformations

  - **–** Should a scalar variable be "grouped up" to form a categorical variable? Ask me how to do it if you need it.
  - **–** Should some categories in a categorical variable be consolidated to create more balanced groupings?
  - **–** etc

Moving further

- Create a list of questions you would like to try to answer about the data, or just questions that you think are important to address.
- For each question

  - **–** Prepare graphs and numerical summaries that would help answer it. Maybe you already created them above. Now you may have to dig deeper.
  - **–** Consider the effect of other variables to the interaction you observe (for example might a GPA effect based on gender be an effect based on year behind the scenes, because you happened to have too many male seniors in your survey?) You may need to create paneled plots to explore this.
  - **–** Consider the effect of group/sample sizes on your conclusions. Do you have too few people in a group to draw conclusions about that group? Do you have outliers that might affect what you are seeing?

- Keep asking questions and digging deeper
- When you are satisfied, start putting together two things:

  - **–** Your project report (in RMarkdown)
    - ∗ Brief introduction and conclusions sections
    - ∗ Sections for each question/topic you wish to address
    - ∗ Graphs/tables to back up your results
    - ∗ Make sure to customize the graphs/tables, make them **presentable**
  - **–** Powerpoint presentation
    - ∗ Create the graphs in the console. Then click on them and save them or copy/paste them into a Powerpoint presentation.
    - ∗ You may have to edit the graphs separately (they probably require different editing than the document ones).
    - ∗ Graphs should be readable from a distance. Test them!
    - ∗ Avoid long texts in your slides.