STAT 1601 – Final Project
Fall 2023

1. You will need to select a data set to analyze for the final project. You have the following options:
    a. Choose a data set that has been gathered for us. They are listed on our library guide: https://guides.lib.virginia.edu/stat1601
    b. Find and compile data related to your topic of choice. If you decide to proceed with gathering your own data set, you ***must*** clear the data set through me prior to your analysis. Your data set should have:
        i. A minimum of 20 observations. Your data should be current in the last 5 years, unless you are planning a comparison among years.
        ii. You should have at least 4 variables: at least 2 quantitative and 2 categorical variables.
2. Once you have your data, you will:
    a. Identify research questions.
    b. Conduct data manipulations appropriate for answering your research questions.
    c. Produce several graphical displays of your data to provide visual answers to your research questions.
    d. Produce several numerical summaries about your data to provide answers to your research questions.
    e. Conduct a hypothesis test to answer your research question.
3. Create a video presentation of no more than 20 minutes long.

Elements of the Presentation:

- **Topic and Speaker Introduction (approximately .5 mins):** Choose a suitable topic for your project. List the full names (last, first) of project group members in alphabetical order.
- **Introduction (approximately 3 mins):** Introduce the reader to your research questions
    - Rationale and motivation as to why your questions are relevant
    - Clearly list of your questions
    - Description of relevant variables.
- **Data Preparation and summary: Manipulation and summary of relevant variables (approximately 3 mins):**
    - Describing any manipulations you made to the data. You should be doing some data cleaning. Depending on how clean your dataset is, this might include: renaming columns, removing columns or rows, mutating columns, subsetting (or filtering) the data based on interesting questions to investigate, etc.
        - Subsetting (or filtering) the data to summarize key variables. Think about interesting questions you can investigate through filtering the data.
        - Find appropriate summary measures of key numeric variables (measures of center and spread).
    - A description of each of the variables in your data
        - Name of each variable including your response, and how you will refer to it in your report

        - Written description of each variable (imagine someone knows nothing of your data)
        - Units of quantitative variables, levels of qualitative variables, if relevant
    - Potential issues you have found or are worried about in the data
- **Exploratory Data Analysis (approximately 5 minutes)**: Highlighted numerical and graphical summaries
    - You may select any combination of numerical and graphical summaries that best show the key features of your data You must include at least one summary (numerical or graphical) that relate to each of your research questions in some way. Some specific guidelines include**:**
        - At least one graph for a single categorical variable
        - At least one graph for a single quantitative variable
        - At least one graph with a categorical and quantitative variable
        - At least one graph with two quantitative variables
        - At least one graph that displays three variables
        - At least one well designed graphic that discusses 4 variables.
    - All graphical representations should be created using R.
- **Methods/Analysis for Statistical Inference (approximately 3 minutes)**: Describe the analysis conducted and how it answers your research question. This would include specific techniques and appropriate testing languages.
    - Clearly identify the chosen test.
    - Clearly explain why the chosen test is appropriate for your research question and data.
    - Clearly state the hypotheses that will be tested.
    - Discuss the analysis' conclusions.
- **Conclusions (approximately 2.5 minutes):** Summarize your findings
    - Make general conjectures about your topic and which question might be more most interest moving format with a statistical analysis.
    - Comment on a limitations of this analysis and your data and how you would continue or improve on this research in the future.

Presentation notes:

- A formal presentation entails making a slides to show your talking points, and code outputs.
- The first slide should have the topic of your project as well as the full names (last, first) of project group members in alphabetical order.
- The presentation should be a total of 20 minutes. Use the approximate times provided for each topic above as a guide. It is possible that some teams spend varying times on the topics so feel free to make adjustments to the approximate times provided.
- If you are working in a group, make sure every group member speaks for about the same length of time if possible.
- If you are working in a group and have several video files, combine them into one before uploading.
- Make sure to turn on your video during the presentation. Voice only presentations are not allowed.
- Include a list of all references used at the end of your presentation.
- Overall quality of presentation marks will be awarded for organization, engagement, and effective communication of concepts.
- Keep in mind that this rubric provides a general guide to the final project. The project is open ended (unlike structured assignments and exams) to allow independent thinking and creativity.
- Every result or output or visualization is not expected to be significant or perfect. Present whatever results from you analysis.
- It is not required to have an overarching goal/question for the entire project. Every part could be distinct from other parts.
- Try not to show codes in your presentation unless it is absolutely necessary. Only show the output of your codes, for example, graphs, regression output, wrangling outputs, etc.
- Communicate your results effectively. Imagine you are presenting to an audience with no expertise on data science or coding or statistics.