Final Project

Project Overview

The goal of this project is to apply what you have learned in this course to conduct statistical analysis. Multiple data frames are available, but your group will choose only one to use for the project. After choosing the data frame, your will need to brainstorm on several issues about the data such as the sample and sampling methods that might have been used to gather the data, a possible target population, variables and their types, among others. Your group will then come up with two research questions that interest you and then use the statistical techniques learned in the course (and beyond, if you like) to analyze data and write a report to communicate your findings. As part of answering the research questions, you will perform both exploratory and inferential analyses. Exploratory analysis involves exploring patterns and trends in the data using visualization tools as well as numerical summary statistics. Inferential analysis on the other hand, involves performing hypothesis tests and confidence intervals.

Research Topic

Your research topic is a short statement that describes the general area of interest that your research questions will be addressing. It should be broad enough to allow for multiple research questions to be asked, but specific enough to give a clear idea of what you are interested in.

Research Questions

You should have at least two research questions but no more than four. Each research question should be related to the research topic and should be answerable using the data you have chosen. Each research question should also be answered using a different inference tool (e.g., linear regression, inference for single proportion, difference in means, etc.). The research questions should be specific and clearly defined. They should also be interesting and relevant to the research topic.

Data

Multiple data sets are available online for you to choose from. If your team decides that you want to collect your own data, please contact me so we can talk about that. The following are possible sources of data sets:

- Openintro website: Most of these data sets can be loaded into your work space as long as the openintro package is installed and running.
- Kaggle: This is a platform for predictive modeling and analytics competitions in which companies and researchers post data and statisticians and data miners from all over the world compete to produce the best models.
- TidyTuesday: Tidy Tuesday is a public and free platform where data science enthusiasts share data and analyses. A new dataset is posted every Monday morning to a public data repository. The data is in a standardized format to help participants analyze and visualize it. The data can be downloaded from their GitHub repository or accessed through the tidytuesdayR package.
- FiveThirtyEight: This is an American website that does opinion poll analysis on politics, economics, and sports in the United States. 538 shares their data sets, related articles, and code.

If your team decides to use existing data set, it is important to understand that the data may have been collected for a different purpose and may not be perfect for your research questions. You will need to examine the data to see if it is appropriate for your research questions. If need be, you can modify your research questions appropriately for the purposes of this project.

Your chosen data set should meet the following criteria:

- Must have at least 100 observations (cases) and 5 variables (columns)
- At least 2 of the variables must be useful and unique predictor variables:
  - Variables such as Property address, Social security number, name, among others may not be useful predictors.
  - If you have multiple columns with the same information (e.g., "state abbreviation" and "state name"), then they are not unique predictors.
- Have both numerical and categorical variables that can be used as predictors in regression analysis (if you choose to use regression).
- Your data set must have at least one variable that can be identified as a unique response variable. This can be numerical (for linear regression) or categorical (for logistic regression). If you find it reasonable to create a categorical variable from a numerical variable, feel free to do so but be sure to provide the rationale for the same.

- If you use any of the data sets we have used in the weekly labs, be sure to use different variables from the ones used for the labs. Your questions must not be similar to those explored in the labs.

If you are unsure whether your chosen data set meets the above criteria, please check with me before setting on a data set.

Project Requirements

Your finished project will have the following sections. Read carefully to understand what is expected in each section. The project should be written in a report format and should be at least 8 pages long. The report should be well-organized and clearly written. You should use headings and subheadings to organize your report. You should also use figures and tables to present your data and results. The report should be written in a professional manner and should be free of grammatical errors.

1. **Abstract**: This section should provide a brief overview of the research topic, questions, data, and analyses you are conducting. It should also provide a brief overview of the results you have obtained. Think of the abstract as an executive summary of your work that someone can read and understand without reading the rest of the report. You should write the abstract last, after you have completed the rest of the report.
2. **Introduction**: You will write two to three paragraphs introducing your topic and justifying why it is important to study. A good introduction gives the reader a broad background of the research problem and contextualizes it. You should examine the data carefully (e.g., location, target population, variables etc.) as you write your introduction and articulate the problem under investigation. Claims made in the introduction as part of justifying the importance of your study should be justified by way of citation. One of the easiest ways to find articles to cite is to search the Google Scholar website: https://scholar.google.com/
3. **Research Questions and Hypotheses:** Clearly state the two research questions that you seek to answer. As indicated earlier, each research question should be answered using a different inference tool (see project overview). For each question, state the hypotheses in words and in symbols.
4. **Literature Review**: This section should provide a brief overview of the literature related to your research topic. You should cite at least 5 sources and synthesize their methods and findings in relation to your own research. Guidelines for writing an effective literature review can be found here.
5. **Methods**: This section should describe the data you are using and the analyses you plan to conduct. It should include the following subsections:

- **Data Description**: This section should describe the data you are using. This should include the source of the data, the number of observations and variables, and a brief description of the variables and their types (e.g., numerical, categorical). You should also include a brief description of the data collection process. Note that the data collection process may not be available for all data sets. In this case, you are allowed to speculate (reasonably) on the data collection process and be sure to mention that in your proposal.
- **Research Questions and Analyses**: This section should list the research questions you are addressing. You should also describe the variables of interest in your data set that you analyzed to answer the questions. Discuss the methods used for each question and why you think they are appropriate.
- **Data Cleaning** (optional): This section should describe the steps you took to clean the data. This may include removing missing values, transforming variables, etc. Depending on your data set, you may not need to do much data cleaning. If this is the case, you should still describe the steps you took to clean the data, even if they are minimal.

6. **Data Analysis and Results**: This is one of the most important and heavily weighted parts of your project. The analyses procedures should be sound and justified appropriately (under methods). Here, you just need to implement the analyses described under the methodology section of your paper. Your analyses should demonstrate a thorough understanding of statistical concepts learned in this course (and beyond, if you like). Your data analyses must include both Exploratory Data Analysis and Inferential analysis (see analysis plan under milestone 1 for details). Each part of the analysis should be accompanied by detailed interpretations in a coherent manner and in the context of the research questions.
7. **Conclusion**: This section should summarize the research questions you addressed, the data you used, the analyses you conducted, and the results you obtained. Findings of your study must be tied back to those reported in the literature review; for example, do your findings concur with previous studies or do they contradict them? It should also discuss the implications of your results and suggest directions for future research. Be sure to state any limitations of your study and what you would do differently if you were to do the study again.
8. **References**: This section should list all the sources you cited in your report. Use APA format for your references.

Grading Rubric

Points will be earned each section according to the following guidelines:

- ***Abstract (2 pts)*** - Does the abstract provide a brief overview of the research topic, questions, data, and analyses you are conducting?
- ***Introduction (3 pts)*** - Does the introduction articulate the research topic and questions you are addressing? Does it provide a brief overview of the data? Does it provide a couple of citations to help you make the case for your chosen topic of study?
- ***Research Questions and Hypotheses (2 pts)*:** Are the questions answerable as stated using the data? Are the hypotheses stated clearly in words and symbols?
- ***Literature Review (4 pts)*** - Does the literature review provide relevant in-text citations? Does it examine and synthesize the sources in context of the current study?
- ***Methods (5 pts)*** - Does the methods section describe the sample, the data, and the analysis methods used? Does it justify the choice of analysis methods? Does it describe the data cleaning processes? Does it provide both EDA and Inferential analysis?
- ***Data Analysis and Results (6 pts)*** - Are the analyses implemented correctly in R? Are the interpretations for both research questions/hypotheses done correctly and in context of the study? Do the analyses demonstrate a thorough understanding of statistical concepts learned in the course?
- ***Conclusion (2 pts)*** - Does the conclusion summarize the research questions addressed, the data you used, the analyses you conducted, and the results you obtained? Are the findings linked back to the cited literature? Does it discuss the implications of your results and suggest directions for future research?
- ***References (1 pt)*** - Are all the sources you cited in your report listed in the references section? Are they in APA format?

Submission

Submissions of the final written report should be made through Canvas. You should submit both the pdf and the source code (.qmd) versions of your project.