Stat 313: Project 2

Write-Up Guidelines

You will be summarizing your results in a (guided) written report following the **Project Writing Guidelines** posted on Canvas. The results **must** be written in the RMarkdown template provided. You can include all pertinent plots inline in the typed document, rather than in the appendix. Your group will be submitting both the RMarkdown file used to generate your report and the knitted HTML file to Canvas by **Sunday, March 14 at 11:59pm**.

Begin by downloading the RMarkdown and data files from Canvas and saving them on your computer in the **same** folder. Then open the RMarkdown file in RStudio and click "knit" to be sure that RStudio knows where to look for your data. If you encounter an error, make sure that the data are located in the same folder as the RMarkdown file!

Data Background

For this project we will look at the profile content of 10,933 Okcupid users. The data were acquired from the Okcupid website on July 1, 2012 through a method called "scraping" (link to Wikipedia description). Okcupid is a free dating website, designed by mathematicians, which today has over 112,000 users, boasts that over 7.3 million messages per day, and prides themselves on the inclusive nature of their website (all users are welcome).

"Find the relationship of your dreams, a one-night stand, a sham marriage (we're not judging). In fact, we don't care what you do, just don't be a jerk about it."

The dataset provides an assortment of self reported (user input) variables,

- age
- body type
- education
- ethnicity
- height (in inches)
- income
- job type
- the date they were last online (prior to July 1, 2012)
- sexual orientation

- religion
- zodiac sign, and
- marital status.

Data analytics at Okcupid has asked you to answer a few questions they are interested in, which could impact the way they make “matches.” Specifically, the analysts are **only interested** in the profiles of individuals who identified as “male.” They are interested in the following questions:

- Do self reported heights of male users differ by marital status.
  - If so, how do the statuses differ?
- Does the relationship between a male user’s self reported height and marital status differs by sexual orientation.

Introduction (5 pts)

- (2 pts) Give a brief background of the research problem and how the data were collected.
- (3 pts) Clearly outline the question(s) of interest you will address with your statistical analysis. The more specific you define the question of interest here, the easier the rest of the analysis and report will be. The research questions should be as specific as possible! Your *Summary of Statistical Findings* should directly answer the question you pose here.

Statistical Methods (20 pts)

This section should lay out the steps, decisions, and logic leading to the statistical model you will use to answer the research question of interest.

- (5 pt) Describe the response and explanatory variables.
  - Provide a table summarizing the number of observations for every combination of marital status and sexual orientation (i.e. available & bisexual, married & gay, etc.).
  - Provide a description of who is the most represented group in the data and who is the least represented group.

- (3 pts) Provide a data visualization of the relationship between height and marital status.
  - Discuss the relationship seen in the plot, making direct reference to aspects of the plot.

- (3 pts) Provide an additional data visualization of the relationship between height and marital status, differentiated by sexual orientation.
  - **Note:** Keep in mind the different methods for adding a second categorical explanatory to your plot (colors or facets)!

- (5 pts) Describe the appropriate statistical method(s) you will use to answer the questions of interest that you stated previously.
  - Be specific about why the method(s) being used are appropriate for the investigation at hand (e.g. types of variables).
  - Also keep in mind that, like multiple linear regression, there are two different ways we can include two categorical variables in our statistical model.

- (4 pts) Check all model conditions of the statistical method you used.
  - Describe what each condition is **in the context of these data**.
  - Reference and include appropriate plots necessary for checking the model conditions.
  - Justify your conclusions regarding the conditions!

Summary of Statistical Findings (15 pts)

In this section you will write up your findings for the questions of interest.

- (5 pts) What is your conclusion for the questions of interest? Namely, "Do self reported heights of male users differ by marital status?".
  - Base your conclusion on the visualization you created and the one-way ANOVA model you fit to these data.
  - When communicating the results of the statistical test make certain you include the value of the test statistic and the distribution it follows under the null hypothesis.
  - Arrive on a conclusion based on the strength of evidence provided by the p-value.
    - If you make a rejection decision, you **must** specify an α !
    - Do not use the word "significant"!

Example for an ANOVA: There is strong evidence that at least one credit rating has a different mean income (p-value = 0.027, F-stat = 3.067 on an F(3, 1616) distribution).

- (5 pts) Provide estimates and interpretations of the differences in self-reported height for each marital status.

- (5 pts) What is your conclusion for the second question of interest? Namely, "Does the relationship between a male user's self-reported height and marital status differs by sexual orientation?".
  - Base your conclusion on the visualization you created and the one-way ANOVA model you fit to these data.
  - When communicating the results of the statistical test make certain you include the value of the test statistic and the distribution it follows under the null hypothesis.
  - Arrive on a conclusion based on the strength of evidence provided by the p-value.
    - If you make a rejection decision, you **must** specify an α !
    - Do not use the word "significant"!

Data Ethics (5 pts)

These data were acquired by a user scraping the information of other users from the Okcupid website. The data were “cleaned” so user’s profile names were removed, however there are names identified in the responses to the `essay0` variable (link to issue).

- (2.5 pts) What are the ethical implications of scraping users information from the internet without their permission? Are these issues remedied by “de-identifying” the data?

The selections for gender in 2012 were “male” and “female,” and options of “gay”, “straight”, and “bisexual” were provided for sexual orientation. Today, there are far more options (link to Okcupid options).

- (2.5 pts) What are the ethical implications of providing users with the limited options of gender and sexual orientation? How has Okcupid addressed this issue?

Scope of Inference (4 pts)

Write a brief Scope of Inference statement. Specifically, answer these two questions and comment on their implications:

- (2 pts) Based on the sampling method used, who can you infer the results of your statistical findings to?
- (2 pts) Based on the study design, are cause-and-effect statements justified? If so, why? If not, what conclusions can be reached?

Make sure you write the scope of inference **in the context of these data**, not just generic statements!

Project Presentation (3 pts)

- (1 pts) Your report should not have any spelling errors! To check for spelling errors in RStudio, click the green check mark button next to the “Knit” button.
- (2 pts) Your report should look as neat and professional as possible. Make sure that your figures don’t end up in the middle of your paragraphs, and that your sections have headings.
  - Your plots should be included in the section in which they are discussed!
  - The output from your statistical model should look as nice as possible. Output from `summary()` looks terrible in a statistical report! I would highly recommend using the `tidy()` function from the **broom** package to tidy up your ANOVA table.

Group Evaluation (3 pts)

Each member of your group will fill out a group evaluation form detailing each member's contributions, cooperation, communication, and participation.

- It is not expected that every group member is an expert on these topics.
- Rather, it is expected that every group member articulates what they are and are not comfortable contributing.

- Every member of the group can (and should) contribute to proof reading your final report!

If you take control of your group's project and do not let others contribute, your grade will be deducted 20%.

If you fail to contribute to your group's project, your grade will be subject to my discretion.