IDS 702 Team Project Overview

Fall 2022

Purpose

The purpose of the team project is to provide an opportunity to complete a data analysis project from start to finish. This includes:

- Selecting data
- Writing research questions
- Performing extensive exploratory data analysis and data cleaning
- Writing a statistical analysis plan
- Developing appropriate statistical models
- Communicating results through a written report and presentation
- Working with a team

Teams

Please see Sakai for your group assignment. Teams have been assigned primarily based on interests as indicated in the pre-course survey. You are responsible for dividing the work equitably among team members. Every team member is required to contribute to each portion of the project: coding, writing, and presenting. Learn from each other and, most of all, be kind. You will receive individual grades for the project that incorporate feedback from fellow team members.

Data

Your first step is to select the dataset that you are most interested in analyzing. The IDS 702 teaching team has gathered a few datasets that we recommend you use. You do have the option to use data outside of these options but **it must be approved by the instructor within one week (i.e., by Thurs, Sept 29)**. Send an email or slack message with a brief overview of the dataset (source, number of observations and variables, proposed outcome variables).

1. 2021 CDC Behavioral Risk Factor Surveillance System Survey

2. World Inequality Database

3. Spotify data from Web API

4. Uber pickups in NYC

5. Motor vehicle collisions in NYC

Note: One option is to combine 4 and 5 and assess a research question relating uber trips to motor vehicle collisions

6. Tennis: ATP, Tennis: WTA

Note: You may choose either tennis dataset, you do not have to use both.

You are welcome to specify a subset of a particular dataset (e.g., only North Carolina in the CDC BRFSS). Be sure to look at the data to understand its scope prior to writing your research questions. If you have questions about accessing any of the datasets, contact the teaching team (instructor and TAs).

Proposal: due Tues, Oct 4, 11:55 PM

In the proposal, list the following:

- Team member names
- Selected dataset
- Two proposed research questions

You are required to write **two research questions** to answer with your data. The research questions must adhere to the following guidelines:

- The questions must pertain to two distinct types of outcomes (i.e., cannot be answered by the same type of model). For example, you might select a continuous outcome for one question and an ordinal outcome for the other question. The outcomes of interest should be clear from the questions.
- You can propose a combination of prediction or inference research questions, but at least one must pertain to inference.

I recommend taking a look at the "research questions" resource on Sakai (in the Resources folder). The page presents various pitfalls when writing a research question and lists several examples of strong questions.

You are more than welcome to send me your proposal early for feedback. The proposal does not need to be produced in R Markdown.

Part 1: due Fri, Oct 21, 11:55 PM

For part 1, you will conduct exploratory data analysis on your selected dataset. You are required to produce a report of your exploratory data analysis findings in R Markdown. The report should be at most five pages. Tables and figures should be well formatted with clear labels and descriptions. You can organize the report as follows:

- **Data Overview**: Provide the chief characteristics of your data, including sample size, number of variables, and source. Include your research questions here.
- **Primary relationship of interest**: Present descriptive statistics and exploratory plots in whichever format you think is best (tables, figures) for your primary relationship of interest (dependent variable and primary independent variable, if applicable). Describe your findings.
- **Other characteristics**: Briefly describe other variables in the data. If there are many, do not list them all. Rather, describe the types of variables that are present (e.g., "demographic information").
- **Potential challenges**: Describe aspects of the data that may present challenges in the modeling stage. For example, might certain categorical variables need to be collapsed? Is there a lot of missingness? Could the size of the dataset present model selection challenges?

Part 2: due Fri, Nov 4th, 11:55 PM

In part 2, you will produce a statistical analysis plan. The plan should be at most 2 pages, produced in R Markdown, and include the following:

- **Overview**: Briefly describe the dataset and research questions.
- **Models**: Describe the type of model you will use for each research question. Justify your selections.
- **Variable selection**: How will you perform variable selection for your models? For each research question, list any variables that you are selecting a priori to be included in your model.
- **Challenges**: How do you plan to address the challenges you presented in part 1?

An example of a statistical analysis plan is posted in Lessons on Sakai. The "primary analysis" and "secondary analysis" sections can serve as guides for how to discuss what kinds of models will be fit and why.

Part 3: Report and Presentations

November 22 in class: Draft of reports due for peer review

November 29 and December 1 in class: Team presentations

December 2: Final reports due

Report: Your report will be an 8-10 page self-contained document describing your analysis. It should be written as a professional document that can be understood by someone with limited statistics background (e.g., a client). **You are also required to submit an RMD file that includes your code for the EDA and analysis.** The report should be organized as follows:

- **Abstract**: A few sentences describing the purpose of the analysis, the data, and key results
- **Introduction**: Provide more background on the data and research questions. Be sure to cite the data and background information appropriately (APA style is fine)
- **Methods**: Describe the process you used to conduct analysis. This includes EDA and any relevant data cleaning information (e.g., did you exclude missing values? If so, how many? Did you collapse categories for any variables?) Then describe the models you fit, and any changes you made to improve model fit (e.g., did you exclude any influential points? Did you do have to address multicollinearity issues? Did you transform any variables?). Also describe model diagnostics. The organization of this section may depend on your particular dataset/analysis, but you may want to break it into subsections such as "Data," "Models," and "Model assessment." Note that you **do not** present any results in this section.
- **Results**: Here you should present results for all aspects of the analysis. The structure of this section should mirror the structure of the methods section. For example, you can start with a few key EDA results (e.g., a table of descriptive statistics), then present model results, then address assessment. This is the section where you will primarily refer to tables and figures. You should have at least 1 figure for each research question that illustrates a key result of the analysis.
- **Conclusion**: Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.

A few things to keep in mind:

- You should never refer to actual variable names in the text, tables, or figures. For example, if a variable for height is called "ht___cm," you should always say "height," and the first time you mention it you should state that it is measured in cm. In plots and tables, it should say "height (cm)"
- The report should be produced in R Markdown and knit to PDF. This may mean you need to create tables "manually" with knitr. I recommend this anyway because you can customize the labels and formatting.
- Someone should be able to read the abstract and look at the tables and figures and have a pretty good idea of 1) the goals of your analysis, and 2) the key results.
- I recommend using colorblind-friendly color palettes in your figures. It can be even better to differentiate with line types or symbols instead of relying on color.

- Keep you audience in mind! A non-statistician should be able to read your report and have a good idea of what you did.
- You can have an appendix if tables or figures are too large to fit into the main text. For example, if you have several predictors, you may want to put a table of model results in the appendix.

Presentations: Your team will give a 12 minute presentation (with an additional 3 mins for Q&A/transition) on your analysis. All team members are required to present. The presentation should be organized as follows:

- **Background**: Includes motivation, data source, research questions
- **Methods**: Briefly describe the models you used to answer your research questions
- **Results**: What did you find? (This should be the majority of your presentation)
- **Conclusion**: Present limitations and future directions

Things to keep in mind:

- Each team member must present
- The presentation should be focused on the results of your analysis rather than data cleaning or technical details of the model. Prioritize creating clear plots/visuals that communicate your message.
- Focus on storytelling. Why is it important/interesting to answer these research questions? What did you find that is compelling? How might the work be continued in the future?
- You can use any program you'd like to create your slides (powerpoint, keynote, R Markdown, etc.)
- Plan to spend a lot of time creating nice slides. This is not something that should be thrown together at the last minute. I am happy to review slides and offer feedback.

Team Assignments

Green team: Dany Jabban, Scott Lai, Zhonglin Wang

Blue team: Elisa Chen, Ahmed Ibrahim, Genesis Qu, Pomelo Wu

Gray team: Nick Carroll, Song Oh, Emmanuel Ruhamyankaka, Jiaxin Ying

Yellow team: Susanna Anil, Sukhpreet Sahota, Xianchi Zhang, Yuanjing Zhu

Red team: Aditya John, Heather Qiu, Isha Singh, Jenny Shen

Turquoise team: Bella Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

Purple team: Chloe Liu, Ruixin Lou, Paul Mckee, Krisi Ann Van Meter

Black team: Pragya Raghuvanshi, Lorna Aine, Eric Rios Soderman, Emma Wang

Orange team: Echo Chen, Pooja Kabber, Andrew Kroening, Dingkun Yang

White team: Kashaf Ali, Ya-Yun Huang, Xiaoquan Liu, Zhanyi Lin