S&DS 23eData Analysis
Final Project Guidelines
Due Friday, 8.13.21, 11:59pm, uploaded to CANVAS as PDF or DOC AND RMD

Overview
Analyze a dataset of your choice and write a 10-20 page report of your findings. This report must be created in RMarkdown and you'll submit both a knitted PDF/doc file and the raw Rmarkdown code. Your goal is to demonstrate your ability to code in R, to clean data, to use appropriate graphical and statistical techniques in R, and to interpret your results.

Groups
You are encouraged but certainly not required to work in groups. Groups can be up to 4 students. Everyone in the group gets the same grade.

Data
You should choose a dataset that is interesting to you, OR you may use one of three datasets provided by myself. The dataset should have at least 10 variables and at least 50 observations. You must have at least two continuous variables and at least two categorical variables. Some datasets will have hundreds of variables and more than 100,000 observations. Getting the cleaning the data may be the most difficult part of your project. **YOU ABSOLUTELY SHOULD DISCUSS YOUR DATA WITH MYSELF OR A TA BEFORE TURNING IN YOUR PROJECT.**

There are many online sources for data – you can just go to Google and search for a subject and then add 'data'. You can also scrape data off a website.

Here are some good sites:

- ICPSR https://www.icpsr.umich.edu/icpsrweb/landing.jsp. More than 10,000 datasets here
- Kaggle https://www.kaggle.com/datasets
- The Census Bureau (http://www.census.gov/)
- NOAA (http://www.nodc.noaa.gov/)
- The US Environmental Protection Agency (http://www.epa.gov/epahome/Data.html).

Other ideas:

- Use your web scraping tools to get data on all roll call votes in the 116th Senate (2nd session, 2020)

You should **NOT** choose a dataset that has already been extensively cleaned and analyzed (i.e. from a textbook or 'nice example' website). However, if there is minimal cleaning to do, then put more effort into something else.

You do **NOT** need to use all the variables in your dataset; indeed, you may end up cleaning/analyzing only 6 to 10 variables. Your goal is not be comprehensive, but to demonstrate what you've learned.

If you decide not to find your own data, you can use one of the following three datasets, all available on CANVAS under `Files → Final Project Information`. Dataset information on variables and collection methods are also provided.

- World Bank Data from 2016
- Environmental Attitudes from the General Social Survey of 2000
- Food Choices (we looked briefly at a few variables in class) : https://www.kaggle.com/borapajo/food-choices

Format

Your project should be presented as a report; it should have appropriate RMarkdown formatting and discussions should be in complete sentences. There is no minimum length (brevity and clarity are admired), and your knitted report should not be more than 15 pages long, including graphs and relevant output (just suppress irrelevant output). You should NOT have pages of output that you don't discuss. You also don't need to have RMarkdown show every last bit of output your code creates. It should feel more formal than a homework assignment, but you should be extremely concise in your discussion.

Sections of the Report

- Introduction (Background, motivation) – not more than a short paragraph.
- DATA: Make a **LIST** of all variables you actually use – describe units, anything I should know. Ignore variables you don't discuss.
- Data cleaning process – describe the cleaning process you used on your data. Talk about what issues you encountered.
- Descriptive Plots, summary information. Plots should be clearly labeled, well formatted, and display an aesthetic sense.
- Analysis – see below
- Conclusions and Summary – a short paragraph.

Content Requirements

Your report should include evidence of your ability in **each** of the following areas:

1) **Data Cleaning** – demonstrate use of find/replace, data cleaning, dealing with missing values, text character replacement, matching. It's ok if your data didn't require much of this.
2) **Graphics** – show appropriate use of at least ONE of each of the following – boxplot, scatterplot (can be matrix plot), normal quantile plot (can be related to regression), residual plots, histogram.
3) **Basic tests** - t-test, correlation, AND ability to create bootstrap confidence interval for either a t-test or a correlation.
4) **Permutation Test** – include at least one.
5) **Multiple Regression** – use either backwards stepwise regression or some form of best subsets regression. Should include residual plots. A GLM with a mix of continuous and categorical predictors is fine here.

6) **AT LEAST ONE OF THE FOLLOWING TECHNIQUES** – ANOVA, ANCOVA, Logistic Regression, Multinomial Regression, OR data scraping off a website.

Additional Comments

Please do NOT have appendices – unlike a journal article, include relevant plots and output in the section where you discuss the results (more of a narrative). This said, you should ONLY include output that is relevant to your discussion. I can always look at your RMarkdown code if I have questions. It is fine to suppress both long output and parts of your R code.

As you work on this project, I expect you will regularly pester myself and TA's.

Submission - Please read this carefully

1) ONLY ONE person in a group should upload a copy of the final project (i.e. if there are three people in a group, only one person needs to upload the files.
2) BE SURE to put all members' names on your project documents.