**Project Instructions**

*Due on Sakai by 11:59 PM on December 3, 2021.*

**Please also upload a CSV or Excel file with your data. If you are using a 538 dataset, you can upload a link to the GitHub page with the data.**

*Assignment Details*

For your final project, you and your group will analyze your dataset in R and discuss your results. Specifically, you will run a linear regression model (or several) to test expectations you make about the patterns you expect to observe in the data. Then, after conducting the analyses in R, you will discuss your findings. You should include an RMD file with the code you use, an Excel/CSV file with your final data that you use to run the models and a knitted html file with your project write-up. It is important to write in a manner that is clear, concise, and statistically correct. You should revise the first part of your project that you already turned in and received feedback on.

Your project write-up should include:

* **Before Chunk One:** An introduction where you introduce your topic to the reader and briefly discuss your expectations based upon your research (estimate: approx. 300-400 words- these may vary)
* **Chunk One:** A discussion your variables of interest. (approx. 100-150 words)
  + After this chunk, please discuss measures of central tendency and spread for your main dependent and independent variables.
* **Chunk Two**: A visualization of your variables of interest.
  + Construct at least three graphs in R that visualize your data effectively for the reader. These can be any of the plots we have learned in class or others (not pie charts) that you believe effectively visualize the data for the reader.
  + After this chunk, please provide additional context to help the reader better understand the visualizations. (approx. 100-150 words)
* **Chunk Three:** Conduct a simple hypothesis test (e.g., difference of means test, Chi-Square Test). After this chunk, please discuss and interpret your results. (approx. 200-250 words, 10 points.)
* **Between Chunks Three and Four:** Discuss the control variables you will include in your regression model. You should include at least three control variables beyond the focal independent variable(s) in the model. Then, please tell the reader what evidence you need to support your expectations. (approx. 300 words, 5 points)
* **Chunk Four:** Please run your model. After running this model, please discuss your results. Did you find evidence in support of your expectations? In discussing your results, please interpret the coefficients and p-values and discuss your R-squared value. Also, please discuss whether there are any influential points or collinear predictors. (approx. 500-600 words, 20 points)
* **After Chunk Four:** Write a brief conclusion where you discuss the implications of your results and any next steps you would take in the analysis. (approx. 150-200 words, 5 points)
* After Chunk Four, please include a statement of 2 or 3 sentences discussing how each group member contributed to the project.

**You should also include (15 points):**

* An RMD file with all the code you used to conduct the analysis and plenty of comments so that the replicator can see what you did.
* Please also include your data as an Excel file or CSV when you upload on Sakai.

*Logistics (15 points, 10 for writing quality and 5 for other logistics)*

Your final draft section should:

* incorporate the feedback from the proposal
* include an RMD file, Excel/CSV file with your data, and a knitted html file
* be between 1800 and 2100 words long (hard cap at 2500 words); please cite any works you include and include a word count with your submission
* be well-written; [here](https://www1.cmc.edu/pages/faculty/JPitney/writing.htm) is a link to a guide to writing that Professor Jack Pitney of Claremont McKenna College designed.
* use a consistent citation style of your choice throughout

Please note, here are some common issues from previous semesters to keep in mind as you are working:

* When making a title, make sure to put a space after the #.
* Don’t show tons of output, especially with loading of packages, set code chunks to include=FALSE if this is the case.
* If you loaded tidyverse, don’t also load dplyr and ggplot2
* Put Works Cited at the end like a standard paper or project.
* Don’t put narrative inside code chunks.
* If you used the R Markdown standard template for the project, please remove template language/code from the file before turning in.
* Use measures of center & spread that best fit your variables’ distributions
* Make axis labels clear. (especially important with names)
* Don’t have overlapping tick marks on axes.

*Rationale*

This assignment will bring together the statistical concepts that we have learned over the course of the semester and give you practice honing your statistical writing skills. You will also have the opportunity to revise your project based on my feedback on your proposal and your group’s suggestions as you work together in class and outside of class.

*Rubric*

* 15: points: revision of previous sections
* 40 points: Project write-up sections (point breakdown above).
* 15 points: Inclusion of RMD file with comments and Excel/CSV file.
* 15 points: Engagement in the peer review process
* 10 points: Writing quality and the overall readability and flow of the project.
* 5 points: Following of other above logistics.

Possible Data Sources

* Some datasets:
  + The *General Social Survey* tracks American behaviors, opinions and demographics.
    - <https://gss.norc.org/get-the-data>
  + The V-Dem Institute has data that conceptualizes democracy throughout the world for the past several centuries.
    - <https://www.v-dem.net/en/>.
  + The US Census has lots of data on its website on numerous topics relating to the American people.
    - <https://data.census.gov/cedsci/>
  + *Daily Kos Elections* has a treasure drove of data on topics ranging from candidate gender and race to presidential vote and demographics by district.
    - <https://www.dailykos.com/stories/2018/2/21/1742660/-The-ultimate-Daily-Kos-Elections-guide-to-all-of-our-data-sets>.
  + Gary Jacobson at UCSD has put much of his congressional elections data online.

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/29559>.

* + The Cooperative Congressional Election study has put data about voters in congressional elections online.
    - <https://cces.gov.harvard.edu/data>.
  + The American National Election Study has put data about voters in presidential (and some midterm) elections
    - <https://electionstudies.org/data-center/>.
    - I will put the most recent ANES from 2016 on Sakai; if you need help downloading an earlier version, please let me know.
    - There are also elections studies from other countries (e.g., Britain, Israel) if you would like to examine a different county’s electoral politics.
  + Scott Adler at CU-Boulder has put data on historical congressional district demographics online.
    - <https://sites.google.com/a/colorado.edu/adler-scott/data/congressional-district-data>.
  + Voteview has data on congressional voting records, among other variables (<https://voteview.com/>).
  + Harvard’s dataverse has many datasets; if nothing else seems interesting so far, you could try searching here.
    - <https://dataverse.harvard.edu/>.
  + Princeton’s library has made guides with many data sources.
    - <https://libguides.princeton.edu/?b=g&d=a>
    - Here is their source for elections: <https://libguides.princeton.edu/elections>.
  + If you are interested in non-political topics, Kaggle is a source of many datasets: <https://www.kaggle.com/datasets>.
  + The R Datasets package has lots of datasets: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>.
  + The FiveThirtyEight package has lots of data, both political and non-political: <https://cran.r-project.org/web/packages/fivethirtyeight/vignettes/fivethirtyeight.html>.

Other useful resources include:

* [R Data Sources for Regression Analysis](https://rfun.library.duke.edu/blog/data-sources-for-regression-analysis/)
* [TidyTuesday](https://github.com/rfordatascience/tidytuesday)
* [World Health Organization](https://www.who.int/gho/database/en/)
* [The National Bureau of Economic Research](https://data.nber.org/data/)
* [International Monetary Fund](https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42&sId=1479329328660)
* [United Nations Data](http://data.un.org/)
* [United Nations Statistics Division](https://unstats.un.org/home/)
* [U.K. Data](https://data.gov.uk/)
* [U.S. Data](https://www.data.gov/)
* [European Statistics](https://ec.europa.eu/eurostat/)
* [Statistics Canada](https://www.statcan.gc.ca/eng/start)
* [Pew Research](https://www.pewresearch.org/download-datasets/)
* [UNICEF](https://data.unicef.org/)
* [CDC](https://www.cdc.gov/datastatistics/index.html)
* [World Bank](https://datacatalog.worldbank.org/)