Project Description

Project overview

This project will be an opportunity to use the skillz you have been acquiring in this course and apply them to a data analysis problem on a topic of your interest. You will gain experience with these aspects of the data analysis process:

- Formulating a problem, with well-defined questions that you would like to answer with the help of data
- Acquiring, cleaning, and organizing data
- Data exploration and visualization
- Predictive modeling
- Interpreting your results
- Communicating your findings to both a technical and non-technical audience

You will be responsible for submitting the following:

- Project proposal with data
- Technical report, for an audience that has a technical background
- Executive summaries, for an audience that does not have a technical background

Data sources

You will need (at least) one data set for your project. Data sets should include several variables (5 or more). If you are doing regression or classification, there should be at least one outcome. As far as the number of observations, you should try to aim for "hundreds" of observations as the minimum, though in some cases a smaller number of observations may be ok too. It is recommended that you choose data related to something that you are interested in, as that will likely make the project more interesting to you.

Several places to find data are given here: https://www.gtpm.ai/data/. For those interested in sports data, I also have several ideas for data. Some other ideas:

- COVID-19 (a range of possibilities here, either US-focused or international)
- Climate change (hot topic, no pun intended)
- Diving: One group gathered results from 15-20 recent competitions, and there could certainly be some more interesting work with the benefit of additional data
- Sports in general (a range of possibilities, including some possible projects organized by our visitor – Elliot – which would involve proprietary data and a non-disclosure agreement)
- "Most Americans Today Believe the Stock Market Is Rigged, and They're Right" (Bloomberg Businessweek). See: https://apple.news/AijUGHJk0QeK7vx1EUYf0yw

If there's something in particular you're interested in, let me know and we can try to find some data.

Project proposal

Your project description should contain a 1-2 paragraph description of the project. Additionally, you should describe your data set and the source of the data, and specify the website URL or R package from which you will be obtaining your data set. You will be assessed on the following:

- how well-written your project description is in terms of content, style, flow, and grammar.
- how well you describe the questions you would like to answer, the real-world applicability of the questions you would like to answer, the analysis you plan to do, and the data you will use
- whether or not the data sets make sense for your analysis

Project writeups

A project template and further guidelines for your project writeups are in the `Project/` folder on Canvas, but here are the basics of what your project will contain:

- Abstract
- Introduction
- Data exploration and visualization
- Modeling/Analysis
- Visualization and interpretation of the results
- Conclusions and recommendations
- Executive Summary

Submission

Please name your documents like this

- `ObtainData.R` - code used to get the raw data (if applicable)
- `Rawdata.csv` or `Rawdata.rds` - raw data before cleaning. If your data is huge, `.rds` files will be noticeably faster.
- `PrepData.R` - code that cleans and organizes your data and prepares it for analysis.
- `Data.csv` or `Data.rds` - data that has been prepared for analysis
- `Analysis.R` - code that performs the analysis, etc. Everything except data cleaning
- `Report.Rmd` - written report
- `Report.pdf` - written report, with code chunks suppressed (`echo=FALSE`). Only include text and outputs of the code chunks, but don't show code chunks, so that it looks like a more formal report.
- `ExecutiveSummary.Rmd`
- `ExecutiveSummary.pdf`

Please make sure a classmate can run your code on their own machine, and knit your report, before you submit it. Points will be deducted if I can't run your source code or knit the document without modification. You will need to avoid using absolute file names like

```
d = readRDS('/Users/YourUserName/Data.rds')
```

Put the file in the folder that your .Rmd file is in, and say

```
d = readRDS('Data.rds')
```

There are several reasons for doing this:

a. your code and data are in the same place

b. it's less confusing if you come back and look at it later
c. It makes it easier to share code and data with others. They can put the code and data in any folder on their computer that they choose, and run it without any modifications. They do not need a folder called `/Users/YourUserName/`, which they most likely are not going to have unless they happen to have exactly the same username as you.