Final Project

Purpose

The purpose of this **individual/group** final project is to put to work the tools and knowledge that you gain throughout this course. This provides you with multiple benefits.

1. It will provide you with more experience using data cleaning tools on real life data sets.
2. It helps you become a self-directed learner. As a data scientist, a large part of your job is to self-direct your learning and interests to find unique and creative ways to find insights in data.
3. It starts to build your data science portfolio. Establishing a data science portfolio is a great way to show potential employers your ability to work with data.

The course is structured in a way that allows you to work on your project as you progress through the weeks. Thus, you *should* not have to cram during the last two weeks of the term to complete your project. Rather, I plan to have you work on the project and use some of the in-class time to do peer evaluation of your code.

Project Goal

The principal goal of this project is to import a real life data set, clean and tidy the data, and perform basic exploratory data analysis; all while using R Markdown to produce an HTML or PDF report that is fully reproducible.

Project Data

You will need to select one data set from the four that I have supplied below. All four data sets contain key attributes that will demonstrate the data science capabilities that you have learned throughout this course. You may even need to learn new skills not taught to accomplish your mission. These include working with:

- multiple data types (numerics, characters, dates, etc)

- non-normalized characteristics (may contain punctuations, upper and lowercase letters, etc)
- data sets that need to be merged
- unclean data (missing values, values that do not align to the data dictionary)
- variables that need to be created (i.e. the data may contain income and expense variables but you want to analyze savings such that you need to create a savings variable out of the income and expense variables)
- data that needs to be filtered out
- and much more!

Available data sets include:

You can choose from one of the following four data sets posted on Canvas. Each dataset has its own challenges and strengths.

- Dog Data
- Lodge Data
- NFL Data
- Global Music Data

Note: Your homework group members may or may not all select the same dataset. If members in your peer group select the same dataset, your work should reflect an individual/pair effort.

Project Report

You will write an R Markdown HTML or PDF report that provides the sections in the grading rubric below. You will need to import, assess, clean & tidy the data, and then come up with your own research questions that you would like to answer from the data by performing exploratory data analysis (if you'd like to perform a predictive model to answer your hypothesis that is fine but it is not required). Some thoughts to help you:

- Make a storyboard. Your project should be a logical, cohesive story–not simply a bunch of graphs created for the sake of making them. The story may change as you dive deeper into the data and find insights, but a storyboard gives you direction and purpose for developing insights. Clear writing means a clear mind, and a storyboard is vital to producing a good story.

- Speaking of insights, keep in mind that your project should follow the chain of data -> insights -> actions. As a future data analyst (or data scientist), you work to create insights that lead to actions, not to waste 40 hours on a awe-inspiring visualization that is ignored directly after a presentation and never used again.
- Simple descriptive statistics can (and usually) yield more of an immediate impact than a complicated model.
- Do subgroups matter in your data?
- Why are data missing?
- Are trends over time important?

Although each data set’s data dictionary contains some additional questions worth pursuing, try to be creative in your analysis and investigate the data in a way that your classmates most likely will not. Creativity is an essential ingredient for a good data scientist!

Section	**Standard**	**Possible Points**
Introduction	**1.1** Provide an introduction that explains the problem statement you are addressing. Why should I be interested in this? **1.2** Provide a short explanation of how you plan to address this problem statement (the data used and the methodology employed) **1.3** Discuss your current proposed approach/analytic technique you think will address (fully or partially) this problem. **1.4** Explain how your analysis will help the consumer of your analysis.	5
Packages Required	**2.1** All packages used are loaded upfront so the reader knows which are required to replicate the analysis. **2.2** Messages and warnings resulting from loading the package are suppressed. **2.3** Explanation is provided regarding the purpose of each package (there are over 10,000 packages, don't assume that I know why you loaded each package).	5
Data Preparation	**3.1** Original source where the data was obtained is cited and, if possible, hyperlinked.	10

Section	Standard	Possible Points
	3.2 Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.). **3.3** Data importing and cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process. **3.4** Once your data is clean, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible. **3.5** Provide summary information about the variables of concern in your cleaned data set. Do not just print off a bunch of code chunks with str(), summary(), etc. Rather, provide me with a consolidated explanation, either with a table that provides summary info for each variable or a nicely written summary paragraph with inline code.	
Exploratory Data Analysis	**4.1** Uncover new information in the data that is not self-evident (i.e. do not just plot the data as it is; rather, slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information). **4.2** Provide findings in the form of plots and tables. Show me you can display findings in different ways. **4.3** Graph(s) are carefully tuned for desired purpose. One graph illustrates one primary point and is appropriately formatted (plot and axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.). **4.4** Table(s) carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features. Size of table is appropriate. **4.5** Insights obtained from the analysis are thoroughly, yet succinctly, explained. Easy to see and understand the interesting findinsg that you uncovered.	10
Summary	**6.1** Summarize the problem statement you addressed. **6.2** Summarize how you addressed this problem statement (the data used and	5

Section	Standard	Possible Points
	the methodology employed). **6.3** Summarize the interesting insights that your analysis provided. **6.4** Summarize the implications to the consumer of your analysis. **6.5** Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.	
Formatting & Other Requirements	**7.1** Proper coding style is followed and code is well commented (see section regarding style). **7.2** Coding is systematic - complicated problem broken down into sub-problems that are individually much simpler. Code is efficient, correct, and minimal. Code uses appropriate data structure (list, data frame, vector/matrix/array). Code checks for common errors. **7.3** Achievement, mastery, cleverness, creativity: Tools and techniques from the course are applied very competently and, perhaps,somewhat creatively. Perhaps student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, unusually sophisticated application of tools from course. **7.4** .Rmd fully executes without any errors and HTML produced matches the HTML report submitted by student.	15

Total possible points: 50
Due no later than: Thursday, March 11, 2022, 5:59PM PT

I expect your report to tell a story with the data. I do not want you to just report some statistics that you find but, rather, to provide a coherent narrative of your findings. Here is an example of the type of report that I am looking for:

- AirBnB user pathways

You need to submit the HTML or PDF file and the .Rmd file that produced the HTML or PDF report, your data, and any other files your .Rmd file leverages (images, .bib file, etc.). Your submitted files should be named with year, course number, lastname, first & middle initial, and then “finalproject.” For example my file name would be: 2022_DSCI353_paparasa_finalproject.Rmd. I expect to be able to fully reproduce your report by knitting your .Rmd file.

Any additional details regarding the final project will be provided in class.