Project Description

Purpose

The purpose of this final project is to implement the tools and knowledge that you gain throughout this course. The rationale for this project includes:

1. It will provide you with more experience using data wrangling 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 glean insights in data.
3. It starts to build your data science portfolio. Establishing a data science portfolio is a great way to demonstrate to potential employers that you have the ability to work with data.
4. Your inquiry will be in line with Cal Poly Humboldt's mission to be renowned for social and environmentally responsibility and action and exemplary partners to our communities, including tribal nations.

The course is structured in a way that allows you to work on your project as you progress through the weeks. I plan to have you work on the project and use some of our in-class time to do peer evaluation of your code and approach. I will also give you feedback on your project prototype so you have time to incorporate that into your final version.

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.

Project Data

You will select one of the prompts below, or suggest your own (with instructor's permission—it should be in line with Cal Poly Humboldt's mission mentioned in item 4 above). Any data sets used will contain key attributes that will demonstrate the data science capabilities that you have learned throughout this course. You may even choose to learn new skills not taught to accomplish your mission. These skills include working with:

- multiple data types (numerical, 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!

You can choose from one of the following prompts, or suggest your own prompt (subject to instructor's permission).

Prompt 1: Homelessness

Each year, Housing and Urban Development (HUD) produces a nationwide estimate of the number of people experiencing homelessness on a single night. This estimate is based on “point in time” counts conducted in January and then typically released the following November. To estimate homelessness **rates,** it is essential to know the relative size of a specific continuum of care (CoC). A CoC is a regional or local planning body that coordinates housing services funding for homeless families and individuals. There are two main measures tracked: the housing inventory count (HIC) and the point-in-time count (PIT). HIC measures the number of beds dedicated to the homeless in the CoC. PIT measures the number of people experiencing homelessness in the CoC on a single night in January. However, the total population of a CoC is not reported by HUD. The US Census Bureau does provide estimates for populations. (In some cases, there will be discrepancies between the geographic boundaries reported by the two government agencies—for this project we will ignore this fact). Consider rental costs (use Zillow’s rent index), median household income, and the percentage of residents living in extreme poverty. Feel free to consider other measures you feel are relevant (for example, does weather correlated with geographic locations with high unsheltered populations?).

What are overall homelessness trends in America over the last decade? What if the population is broken into sheltered versus unsheltered (do the subgroups have different trends over time?)? What about chronic versus nonchronic homelessness (do the subgroups have different trends over time)? After exploring national trends, what are the trends in California over the last decade? Does Humboldt mimic these trends? Focusing on California, what policies and legislation have been enacted in the last decade related to homelessness, and can you see impact in your data from these policies? Are there policies you would suggest the State of California adopt based on your analysis?

Potential data sets to include:

- HUD PIT and HIC Data: https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/
- HUD CoC Shapefiles: https://www.hudexchange.info/programs/coc/gis-tools/
- Zillow Housing Data: https://www.zillow.com/research/data/
- Humboldt PIT info: https://humboldtgov.org/DocumentCenter/View/107770/PIT-Executive-Summary-2022-PDF
- Find your own!

Background Reading:

- https://www.theatlantic.com/magazine/archive/2023/01/homelessness-affordable-housing-crisis-democrats-causes/672224/
- Colburn, Gregg, and Clayton Page Aldern. Homelessness is a housing problem: How structural factors explain US patterns. Univ of California Press, 2022.

Prompt 2: Millenials (the generation born between 1981 and 1996)

It has often been said that Millenials are the first American generation that will do worse than its parents financially. Indeed, Millenials came of age during the wreckage of the Great Recession, have more college debt than previous generations, and face a higher cost of living than previous generations, particularly in terms of housing and homeownership.

In May 2023, The Atlantic published an article called "The Myth of the Broke Millenial." The article claims that as a group, Millenials are not broke and are in fact thriving economically. However, the prosperity within the generation today is not evenly shared.
Use data from the last decade to explore this claim, and to deep dive into the statement that the prosperity is not evenly shared. How is it shared? Consider household income and gender, race, geography, education level, homeownership. How is debt, particularly college loan debt distributed among millennials? In a single household, how has the pattern of earning (for example, between a male and a female partner) changed over the last decade? Can you break this question into subsets of individuals (e.g., by race, education level, type of employment)?

Potential Data Sets to Include:

- United States Bureau of Labor Statistics Public Data API: https://www.bls.gov/developers/home.htm
- United States Census Data Resources: https://data.census.gov
- Find your own!

Background Reading:

- https://www.theatlantic.com/magazine/archive/2023/05/millennial-generation-financial-issues-income-homeowners/673485/
- Twenge, Jean M. *Generations: The Real Differences Between Gen Z, Millennials, Gen X, Boomers, and Silents—and What They Mean for America's Future*. Simon and Schuster, 2023.
- Filipovic, Jill. *OK Boomer, let's talk: How my generation got left behind*. Simon and Schuster, 2020.
- https://www.stlouisfed.org/open-vault/2020/february/millennial-wealth-gap-smaller-wallets-older-generations

Prompt 3: Design your own project.

Your inquiry must be in line with Cal Poly Humboldt's mission to be renowned for social and environmentally responsibility and action and exemplary partners to our communities, including tribal nations. You will need to obtain instructor permission on your topic before beginning.

You will need to import, merge at least two datasets, assess, clean & tidy the data, and then generate your own research questions that you would like to answer from the data by performing exploratory data analysis.

- 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.
- Speaking of insights, keep in mind that your project should follow the chain of data -> insights -> actions. As a data scientist, you will work to create insights that lead to actions. (Don't waste hours on an awe-inspiring visualization that is ignored directly after a presentation and never used again. Make sure visualizations you use help you to tell your story.)
- Simple descriptive statistics can (and usually) yield more of an immediate impact than a complicated model. Brooke Watson gave a compelling presentation at the 2019 RStudio Conference on how the ACLU used exploratory analysis to count immigrant children and reunite

families. (Note: Brooke conducted her analysis in R, so the code will look different than the Python code you have seen, but the steps should be familiar!)
- Do subgroups matter in your data?
- Why are data missing?
- Are trends over time important?

Rubric

Section	Standard	Possible Points
Introduction	**1.1** Provide an introduction that explains the problem statement you are addressing. Why should we 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 that you think will address (fully or partially) this problem. **1.4** Explain how your analysis will help the consumer of your analysis.	10

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.	5

Data Preparation	**3.1** Original source where the data was obtained is cited and, if possible, hyperlinked. **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 (explain 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. Provide a consolidated explanation, either with a table that provides summary information for each variable or a nicely written summary paragraph with inline code.	15

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, 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. Make it easy to see and understand the interesting findings that you uncovered.	15

Summary	**6.1** Summarize the problem statement you addressed. **6.2** Summarize how you addressed this problem statement (the data used and 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	15

Formatting and Other Requirements	**7.1** Proper coding style is followed and code is well commented. **7.2** Coding is systematic – a complicated problem is 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 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.	10

Your report should tell a story with the data, providing a coherent narrative of your findings.