Project Specifications

Description:

While in-class exercises and homework assignments can develop your understanding and skills, they cannot fully prepare you for working with data "in the wild." The goal of the final project is to apply your skills in data wrangling with Python to real-world data to explore a problem facing business or society. The project gives each team the freedom to expand on the concepts learned in lecture and lab and work in a domain that they are passionate about.

- This project may be completed individually or in groups of two. You do not need to be in the same lab section to be partners. If you work in pairs, only ONE person should turn in the deliverables. Make sure you include both partner's names on the work.
- The project must leverage at least two sources of real (not simulated) data. The datasets must be able to be integrated into a single data frame (vertically or horizontally merged). At least one dataset must be obtained by you via web scraping and/or APIs. The other can be a pre-existing dataset downloaded from a site like Kaggle or Data.gov. (Some possible data sources are listed in "Open Data Sources").
- Groups will perform descriptive analytics using the integrated data, including calculating summary statistics, hypothesis testing, fitting models, and creating visualizations. While you are strongly encouraged to generate your own topic, some example topics include:
    - Analyze Congress members' voting records considering the population demographics of their home states (NY Times API, Census data)
    - Rank a group of movies by box office revenues, accounting for inflation (Wikipedia data, Inflation data)

Grading:

The final project is worth 100 points total (17% of your final grade). The points are assigned to each deliverable as shown below.

Submissions will be assessed based on the following criteria:

- Novelty: What interesting data sources are used? What type of unique analysis and/or insights does your project provide?

- Difficulty: How challenging is the implementation of your project? How difficult is it to retrieve, process, and integrate the datasets? Does your code extend beyond skills covered in course materials?
- Accuracy: Has the data been cleaned and integrated in a way that does not introduce errors or bias? Are the visualizations and analytical methods appropriate for the data and/or research questions?
- Clarity: Does your report clearly describe the motivation, research questions, data sources, analyses, and results of your project? Is the report written in a professional (error-free) style?

Item	Points
Proposal	40
Check-In	20
Project Report, Data & Code	40

Project Proposal:

Each team must submit a proposal to ICON by Sunday, October 20, 2024 @ 11:59pm. The proposal should include:

- Introduction: Provide background information on the context so that a non-expert can understand. Describe the problem that your descriptive analysis is meant to explore. Make sure to include links/citations for any facts and figures that are not common knowledge. For example: "The University of Iowa is a college in Iowa City" is common knowledge. "Jeff Bezos has a current net worth of $140 billion" is a fact that should be cited.
- Data: Clearly describe the source of your data. By the time of the project proposal, you should have identified at least one of your data sources. Include a description of how you plan to collect at least one additional data source (e.g., scrape data from a website, collect data from a cryptocurrency API). Make sure to include a link/citation for all data sources. Create a "data dictionary," a table that lists fieldnames, data types, and descriptions (example below).

Example data dictionary (college football data)

Field	Type	Description
College	Text	College name
City	Text	College city
State	Text	College state
Division	Text	Football division (e.g., FBS, FCS)
Conference	Text	Football conference (e.g., PAC12)
AP	Numeric	Team preseason AP poll ranking
WP	Numeric	Team winning percentage over last 5 years
Bowl	Logical	Team played in postseason bowl in previous season?
Draft	Numeric	Number of players taken in NFL draft over last 5 years

- Proposed Analysis: Present at least three potential research question(s). For some questions, you may have a hypothesis (expected result) in mind, but others may simply be exploratory. The final project must include multiple methods from the course (e.g., a combination of summary statistics, visualization, and statistical tests/models). One question may be answered from a single data set. The other questions must be answered from your combined data sets.

Project Check-In:

Each group will receive feedback on their project proposal. After that point, each group must have an in-person project check-in with Mike between Monday, November 18 and Friday, November 22, 2024. Check-in will happen in-person in my office (PBB-S292).

I will use Calendly so you can make an appointment. The project check-in is intended to resolve any questions you may have about proposal feedback or your proposed analysis. Each group should have **all** of their data collected by the time of the project check-in.

Project Final Report:

Your final report should be a clearly and professionally written document that includes all your design and implementation details as well as data analysis and results. Specifically, you need to organize your report into four sections as follows (can have additional sub-sections):

- **Introduction**: Provide background information on the domain and motivation for study
- **Data**: Describe the sources of all your data (including citation/link). Create a data dictionary for the final, merged dataset
- **Analysis**: Present your research questions and results. You should interpret the results of your analysis considering your initial hypotheses and the broader context. Be careful not to extrapolate or over-generalize
- **Conclusion**: Briefly present your overall conclusions, limitations, and suggestions for future work. Each group must submit the project report (.docx or .pdf), Python notebooks (.ipynb), and data files (.csv or .txt) to ICON by 5:00 pm on Tuesday, December 17, 2024.

Project Data/Code:

All work for the project must be completed in Python, and all code should be provided in the form of Python notebooks (Jupyter). To ensure that your work is portable and reproducible, it is highly recommended that you separate your work into multiple notebooks.

For example, you could create one notebook for Web scraping, and then save the data in a file. Then a second notebook for data integration, cleaning, and saving the final, merged data. A third notebook might contain all descriptive analysis. Ensure your scripts are appropriately named so we can determine the run order.

Each group must submit all Python notebooks, downloaded data files (.csv or .txt), and a copy of your raw, scraped data in .csv or .txt format. We should have everything needed to replicate your results both by running your scraping files or by using the raw, scraped data you provide.