Flavia Jiang (yj472), Rachel Wang (jw879)
INFO2950 Project Phase 1
21 Sep. 2023

First of all, thank Charlie for reviewing this and answering our questions.
Link to our repo: https://github.com/flaviafafaf/info2950_project.git

Dataset Idea 1

Description:
Our first dataset is about college students' dating preferences, which is a fun one. The dataset came from experimental speed dating events from 2002 to 2004 among students in graduate and professional schools at Columbia University, and it was created by Professor Raymond Fisman et al. for their paper "Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment." Variables include demographics, dating habits, beliefs about what they and others find valuable in a mate, etc.

Availability:

- Link to the dataset and the documentation: http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/
- Link to the paper: https://academic.oup.com/qje/article-abstract/121/2/673/1884033?redirectedFrom=fulltext
- Comments: At this stage, we don't think we need to scrape/collect additional data if we choose this topic as our final project topic because the dataset itself already contains sufficient information. We will spend more time interpreting the codebook, cleaning the data, and doing hypothesis testing. We will also make sure our research topic is not exactly the same as the paper's.

Question for reviewers:
As we don't mean to replicate the study but do our own project, do you think we should brainstorm research questions by ourselves first or look at the paper to learn more about the dataset first?

Dataset Idea 2

Stock Prices and Financial Indicators during 2008 Financial Crisis
Description:
This dataset would contain historical stock price data for publicly traded companies, including daily or hourly stock prices, trading volume, and financial indicators like P/E

ratio, market capitalization, and dividend yield. Through this we really want to learn if the financial indicators could predict the future crisis. Can historical stock price data and financial indicators be used to effectively predict financial crises and market downturns, and if so, what patterns and predictive models emerge from the analysis?

Availability:

- There is a dataset from kaggle which retrieved the data from Yahoo finance. It contains prices for up to 01 of April 2020. https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset
- Another dataset is from datahub: https://datahub.io/collections/stock-market-data. This one includes more financial information compared to the previous one.

Question:

There are lots of existing studies analyzing this, how should we deviate from that and be more creative with our study?

Dataset Idea 3

Description:

The third dataset is about population trends by country from 1950 to 2021. Population indicators include total population, age structure, population density, birth rates, and death rates, mostly measured annually. We will have to pick one or two of the indicators to focus on for our project if it is chosen as our topic. If we also want to do multiple regression with the selected population indicator as the response variable, we might need data for economic/social indicators (GDP per capita, Human Development Index, etc.) for our explanatory variables.

Availability:

- Link to the population dataset: https://ourworldindata.org/explorers/population-and-demography (Find "Download" at the bottom of the graph. Choose "Full Data (CSV)" to download the data.)
- Source: The original source is "United Nations, Department of Economic and Social Affairs, Population Division (2022). *World Population Prospects 2022, Online Edition*." The data was collected and visualized by Our World in Data.
- Comments: If we need data for economic/social indicators, we can find it on Our World in Data or from the United Nations, World Bank, etc.

Question for reviewers:

- Will time series analysis be covered in this course? If not, what do students often do with time series data for the project with statistical methods taught in this course?