Project Milestone 1: Data Sources, Relationships, and Plan

For this project, I’m hoping to work with global inflation data to track inflation rates around the world.

* Source 1 (CSV): Annual and Quarterly Inflation Data from the World Bank
  + Found at: <https://datacatalog.worldbank.org/search/dataset/0065307/Inflation-in-Emerging-and-Developing-Economies>
* Source 2 (Web Table):
  + Found at: <https://en.wikipedia.org/wiki/List_of_countries_by_inflation_rate>
  + Tables on this site measure the annual changes in consumer price indices from the World Bank and the UN
* Source 3 (API):
  + Found at: <https://newsapi.org/>
  + News article API for retrieving news articles about inflation.
    - Possibly good for targeting areas with deflation or very high inflation

For my data, I plan on using the World Bank inflation data CSV with its multiple sheets (overall inflation, housing index changes, energy index changes, and food index changes done annually and quarterly for countries around the world) to work with the numbers aspect. The table from wikipedia serves as a second copy of annual World Bank numbers and a set of numbers published by the UN.

I plan on using the News API to search for documents related to inflation in order to gather text data regarding inflation. Its query structure will allow for searches specifying keywords, languages, specific time frames.

The unifying element between all three sets will be countries. The text analysis can target countries during years when they experience high levels of inflation or deflation. Articles generated from those searches can have their texts cataloged and analyzed in an object that can specifically hold counts and other information with the object named for the country and time period itself.

Challenges that will come with this topic could include sifting through inflation rates in different eras and regions having different causes and effects. Working with all of the rates at once when trying to display them would also be an issue since there are records for more than 200 countries. Working with individual regions in defined time spans would likely be best. Additionally, the text analysis could prove to have some challenges since taking the time to read every article retrieved would take entirely too much time and defeat the purpose of trying to mine the text itself anyway. Designing a text analyzer that looks at single word frequency, word-pair frequency, and word-triple frequency, (or separate functions that do each) will likely also take some work but could be done.

It is unlikely that many ethical problems could arise from this analysis. There is no personally identifying data in the dataset. The analysis is intended to be a look at inflation rates over time and inflationary periods tend to have events and other information attached to them. If there happens to be cyclical patterns it could be possible to perform a regression but any predictions made with such a model would definitely not be guaranteed since the context of past events doesn’t necessarily hold for future events.