Dataset Proposal Submission

Team Members:

Kajal Tiwary (kt755), Clare Garberg (cag199), Clara Richter (cr1100), Elise Rust (er844)

Final Project Topic:

Gender discrimination in STEM fields is a prominent human rights issue that manifests itself via job segregation, employment inequity, the pay gap, and a general lack of freedom in choice of career for women around the world. This topic aims to better understand how gender inequality has evolved over time and by geography. As women in tech, we specifically want to hone in on gender inequities in STEM fields and assess how test scores, school participation, labor participation, salaries, and media portrayal vary by men and women. The following datasets provide the information to thoroughly investigate this subject.

Datasets:

1. **Gender Inequality News Headlines**
    a. **Description:** This dataset contains news article headline text data that pertains to the topics of women in STEM, women in tech, gender equality, and gender inequality. We intend to leverage this dataset to investigate how the media portrays gender inequities and women in tech. Therefore, sentiment will be derived from each headline/description and will be the label or the classification variable in this dataset. Based on how this information is utilized in downstream analysis and in conjunction with other files, the date range may expand and the topics may evolve.
    b. **Source:** News API - https://newsapi.org/
        i. Dataset - https://github.com/garbergc/ANLY-503-Final-Project/blob/main/NewHeadlines_2022_02_13_v2.csv
        ii. Code - https://github.com/garbergc/ANLY-503-Final-Project/blob/main/503_WIT_News_Headlines.py
2. **STEM Occupation Earnings By Gender**
    a. **Description:**
    This dataset, provided by the American Community Survey in 2019, offers a comprehensive overview of median earnings by gender across most STEM fields. Looking at computer occupations, mathematical scientific operations, engineering jobs, life sciences and physical sciences jobs, and more, this dataset does a thorough job comparing earnings across different types of STEM jobs to highlight how the gender pay gap is not homogenous within STEM.
    b. **Source:** United States Census Bureau - https://www.census.gov/data/tables/time-series/demo/income-poverty/stem-occ-sex-med-earnings.html
3. **Gender Representation In Data Science**
    a. **Description:**

This dataset looks at the gender divide in Data Science between 1990 and 2018. Using survey data of 20,000+ respondents, the dataset contains information about men and women's education levels, length of time coding, perceptions of ML/AI, use of different technical tools, and more. We intend to use it to investigate how female representation and engagement with technical fields has evolved since the 1990s.

   b. **Source:** Kaggle - https://www.kaggle.com/martinlbarron/the-gender-divide-in-data-science/data?select=multipleChoiceResponses.csv

4. **Graduates By Gender And Field Of Education**
   a. **Description:**
      This dataset provides information on college graduates' majors by gender, country, and year. This dataset will be used to find a connection between what women study in school and their participation in STEM fields.
   b. **Source:** Organization For Economic Co-Operation And Development (OECD) - https://stats.oecd.org/Index.aspx?QueryId=109881
5. **Fertility And Labor Force Participation**
   a. **Description:**
      This dataset provides information on the percentage of women active in the labor force by country and average births per woman per country for the years 1960 to 2000. It also provides absolute and relative change calculated by country. We will use this dataset to analyze how fertility rates and participation in the labor force may be related to the number of women in STEM in those regions.
   b. **Source:** Our World In Data (OWID) - https://ourworldindata.org/grapher/fertility-and-female-labor-force-participation
6. **Parental Leave By Country**
   a. **Description:**
      This dataset contains information on the length of maternity leave, parental leave, and paid father-specific leave per country. This dataset has a time element and geospatial element (years from 1970-2020 and country names). We will use this dataset to explore the length of maternity leave, parental leave, and paid father-specific leave correspondence with women in STEM in those countries.
   b. **Source:** Organization For Economic Co-Operation And Development (OECD) - https://stats.oecd.org/index.aspx?queryid=54760
7. **Test Scores By Educational Subject And Gender**
   a. **Description:** This dataset contains information on boy and girl math, science, and reading test scores per country dating back to 2003. This dataset has a time element and geospatial element. We will use this dataset to explore how these test scores influence women in STEM in those countries.
   b. **Source:** Organization For Economic Co-Operation And Development (OECD) - https://data.oecd.org/pisa/reading-performance-pisa.htm#indicator-chart
8. **Graduates And New Entrants By Educational And Occupational Fields**
   a. **Description:** This dataset contains information on the percentage of females and males gaining degrees in specific fields and entering the workforce by country

from 2005 to 2019. This dataset has a time element and geospatial element. This will be a great dataset to see the distribution of men vs. women among degrees and fields per country.

b. **Source:** Organization For Economic Co-Operation And Development (OECD) - https://stats.oecd.org/Index.aspx?QueryId=109881