Capstone Project 1 Proposal

Title: Motherhood in the workforce

Problem: One of the societal norms is for people to have kids as they get older. Women seem to carry most of the costs of family. There have also been multiple studies proving that gender pay gap exists. Is gender the main variable or are there other factors such as family that play into it? With that in mind, are there characteristics of companies that are more friendly towards woman, particularly those with families?

Who might care?
Individuals who are looking to pursue a career in STEM (anyone from high schoolers to college students who are thinking about their future as well as individuals who are thinking of a career change). Nonprofits who are interested in this realm of data and information. Companies who are looking to increase the gender diversity of their workforce and stay competitive in the business world. Recruiters and job hunting agencies/resources (such as LinkedIn, Indeed, etc.) would be interested in this project as they work to sell individuals on the companies they work for/are recruiting for. HR departments would also be interested as they seek to change their company's work culture and create space for the ever diversifying workforce.

Questions to analyze?
Is there a gender pay gap in the STEM workforce? Does gender play a role in becoming a supervisor?
Does the presence of family or children play a role in these differences? If so, how?
What are the characteristics of employers who are more women friendly? Can we make an employer suggestion for women based on whether they want to have kids or not?

Data: The survey data will be acquired from the IPUMS-Higher ED (https://highered.ipums.org/highered-action/variables/group), which focuses mainly on scientists and engineers. The variables I am choosing contains information about: age, gender, race/ethnicity, total number of children, labor force status, employer sector, size of employer, start-up business within last 5 years, summarized primary work activity, supervisory work, salary, and family-related reasons for changing employer or job. I will also choose these data from both year 2013 and 2003 to compare how things have changed within 10 years.

Approach to solving this problem:
Python numpy and pandas packages will be used for data wrangling. Both datasets will be merged together then cleaned and checked for any inconsistencies or missing/duplicate data.

Matplotlib and seaborn packages will provide exploratory data analysis (EDA) through plots and visual representation of statistical values. Correlations and comparisons will be viewed here. Most questions will be answered here and will provide insight to the predictive model.

Not sure what to use here for predictive model to answer the main/final question.

Deliverables:

1. Python code on GitHub:
    a. Data wrangling
    b. Data exploration analysis
    c. machine learning model development
2. Report on capstone project