Ashley Chen
DSCI549
Homework 5

1. This project explores gender disparity in part-time employment across countries using a dataset from UNData. The focus is on identifying which countries have the largest gaps between men and women in part-time work and understanding how these gaps have changed over time. Additionally, the project aims to explore whether national policies or cultural factors are associated with greater or lesser gender disparities in part-time employment. The dataset contains statistics by country and year, including the percentage of adult men and women in part-time employment and the share of part-time workers who are women. These values can be used to calculate a gender disparity ratio, which serves as a key metric throughout the analysis.

   The project will also incorporate external data sources, such as indicators of labor laws, gender equality indices, and family policy data. Combining these sources allows for the use of regression modeling to identify patterns and possible explanations for the observed disparities.

   The analysis will involve several stages: cleaning and preparing the data, calculating new variables, visualizing trends and comparisons using bar charts, line graphs, and maps, and finally, using predictive modeling techniques to explain differences between countries. The dataset is structured and available in formats like CSV and Excel, making it accessible for use with tools such as Python, R, or Excel. This project offers a practical application of core data science techniques, including data cleaning, exploratory analysis, feature creation, data merging, and basic modeling, while highlighting the value of combining datasets to address real-world social questions.

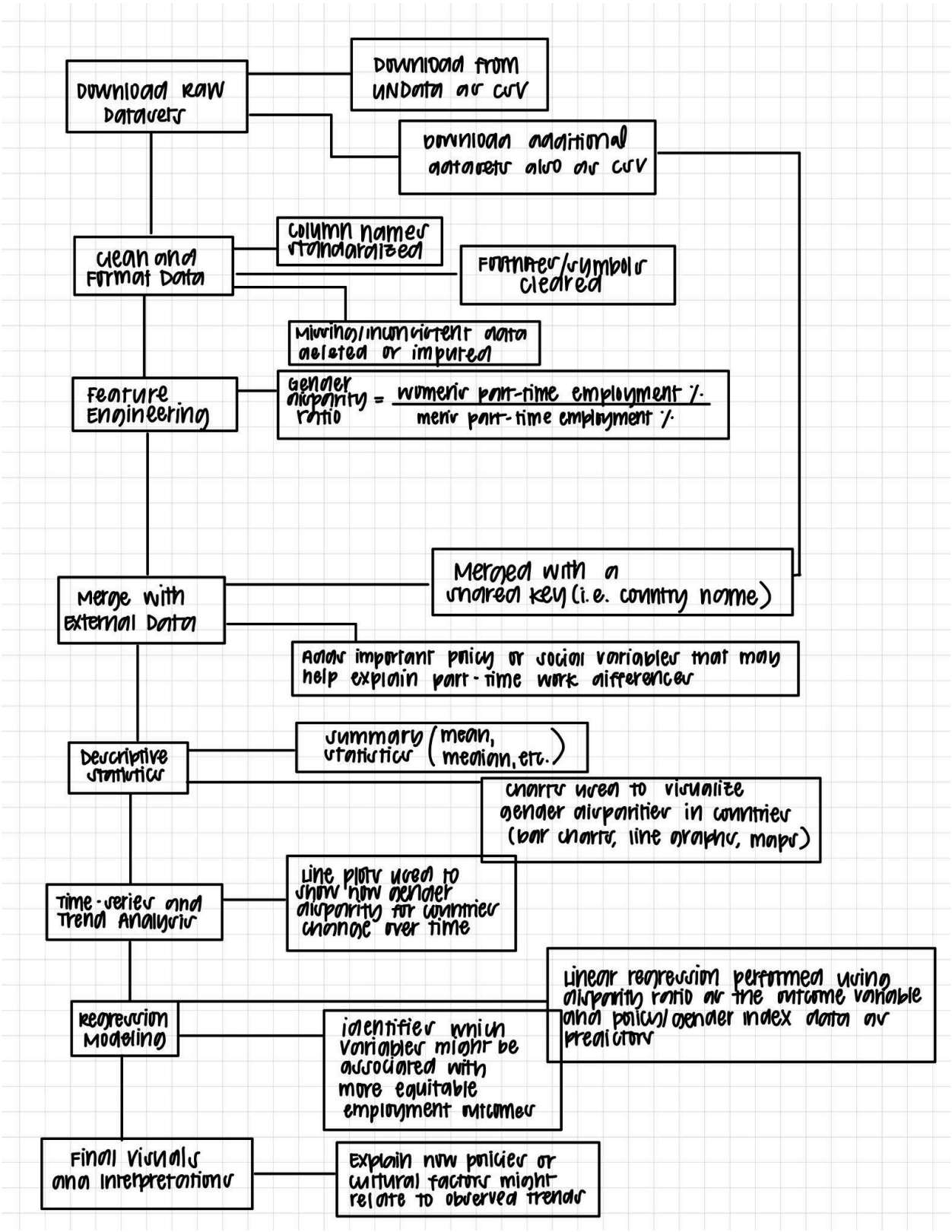
2. The dataset being used is a structured, tabular dataset from UNData. It lists different countries and shows the percentage of adult women and men working part-time in each country, the percentage of all part-time workers who are women, the year the data was collected, and the source of the data. It is a machine-readable dataset, meaning it can be easily downloaded in formats like CSV or Excel and open in programs like Excel, Python, or R for analysis. Each row represents a country in a specific year, and the columns give us numerical text data. Some of the data is missing or inconsistent, such as different years for different countries or small gaps in the numbers, meaning cleaning the data will be an important step before analyzing it. The dataset doesn't include many extra details, so additional data sources will also help to answer some of the bigger questions.

3. The pre-processing steps that I'll follow to prepare the data would include:
   a. Importing the data into a tool like Excel or Python program using libraries such as pandas and NumPy.
   b. Clean the column names so they're consistent and easy to use (like removing special characters)

c. Convert the year column into a number
d. Handle missing values by either imputing them (after deciding which imputation method is best) or skipping those rows
e. Create a new column that shows the gender disparity ratio by dividing the women's part-time percentage by the men's part-time percentage
f. Merge other datasets that have information about labor laws, gender equality, or family policies by country
g. Format everything consistently from percentages in decimal format to consistent country names

4. Analysis Plan
    a. Main Questions
        i. Which countries have the highest and lowest gender disparity in part-time employment?
        ii. How has this disparity changed over time in different regions?
        iii. What policies or country characteristics might explain these differences?
    b. Variables/What Will Be Analyzed
        i. Gender disparity ratio (from main dataset)
        ii. Country names and years
        iii. Additional data from external sources, such as gender equality rankings or childcare policy data
    c. Main Analysis Steps
        i. Explore the data with charts (bar graphs, maps, line charts)
        ii. Compare gender disparities across countries and over time
        iii. Join in external data sources for a more complete picture
        iv. Use regression modeling to test if certain policies (like better family leave) are connected to smaller gender gaps
        v. Create visualizations that explain our findings in a clear and simple way

5. See below


Download raw datasets
Download from UNData as CSV
Download additional datasets also as CSV
Clean and Format Data
Column names standardized
Formatter/symbols cleared
Missing/inconsistent data deleted or imputed
Feature Engineering
Gender disparity ratio = women's part-time employment % / men's part-time employment %
Merge with External Data
Merged with a shared key (i.e. country name)
Added important policy or social variables that may help explain part-time work differences
Descriptive statistics
summary statistics (mean, median, etc.)
charts used to visualize gender disparities in countries (bar charts, line graphs, maps)
Time-series and Trend Analysis
Line plots used to show how gender disparity for countries change over time
Regression Modeling
Linear regression performed using disparity ratio as the outcome variable and policy/gender index data as predictors
identifies which variables might be associated with more equitable employment outcomes
Final Visuals and Interpretations
Explain how policies or cultural factors might relate to observed trends

6. Appendix:

a. Dataset:
    i. The dataset below displays information on part-time employment statistics, listing them by country and year. It also breaks down the percentage of men and women in part-time employment, while specifically listing women's share of part-time work, allowing for gender-based analysis of labor trends across different nations.

Country or area	Year		Percentage of adult employment that is part-time		Women's share of part-time employment	Source
			Women	Men	%	
Albania	2001	a,b,c,d	50.2	46.1	39.7	O
Argentina	2011	e,f,g,h	31.3	11.9	64.8	LFS
Armenia	2008	b,i,j	30.1	19.0	56.2	LFS
Aruba	1994	e,k,j	12.3	4.0	71.0	HS
Australia	2011	l,b,f,m,n	38.5	13.2	70.9	LFS
Austria	2011	b,f,d,o	32.8	7.0	80.2	LFS
Azerbaijan	2003	l,b,f,d	24.3	13.3	56.6	HS
Bahamas	2002	l,b,k,j	11.9	9.6	53.9	HS
Bahrain	2004	e,i,j	2.8	2.2	25.1	LFS
Belgium	2011	b,f,d	32.4	7.0	79.9	ELFS
Belize	1999	p,c,d	25.5	12.0	48.9	HS

    ii. The dataset is provided in a table format, making it easy to read and download, satisfying the accessible criteria. It is also free to use for analysis as it is from an open-access source (UNData), meaning it complies with the open-license criteria. And finally, the dataset is tabular and structured with defined columns, which is suitable for direct use with data processing tools such as Python, R, or Excel for statistical analysis and visualization, making it machine processable.
    iii. This "Part-time employment" dataset was found from UNData, an international database, where it can be downloaded as CSV, Excel, or JSON files, which are compatible for machine processing.
b. Questions/Outline:
    i. What is the gender disparity in part-time employment across countries and regions?
        1. This question could be answered with the dataset above by calculating the ratio of women's percentage in part-time employment to men's percentage for each country. Visualizations

like bar charts or a heatmap would also give a clear visual of the disparities geographically. The output would be a clear visualization of countries with the highest and lowest gender disparity in part-time employment.

2. No additional data is needed.

ii. Has the gender disparity in part-time employment changed over time in specific regions or globally?

1. Use time-series analysis, line graphs or trend charts, to examine changes in gender disparity over years within regions or countries. The output would be trend charts showing trends in gender disparity for part-time employment.
2. Because the dataset covers multiple years, no additional data is needed.

iii. What factors might explain the differences in part-time employment by gender between countries?

1. Conduct a regression analysis to identify the influence of cultural or policy factors on women's share of part-time employment, including variables such as labor laws and cultural gender equality indices. The output would be a predictive model identifying the most significant factors explaining differences in gender representation in part-time employment. Going beyond descriptive statistics, a multivariate regression analysis using supplementary data sources could be conducted. This analysis would examine the relationship between women's share of part-time employment and external variables such as labor laws, national childcare support policies, and cultural attitudes toward gender equality
2. Additional data on labor laws, childcare policies, and cultural gender norms would be needed. Collaborating with a sociologist or labor economist to contextualize findings would also assist in this analysis.
3. Since the UN dataset alone provides limited variables, the dataset could be merged with publicly available cross-country data on labor regulations, social policy, and cultural indicators. This would allow us to build a predictive model that identifies which contextual factors are most strongly associated with gender disparity in part-time employment rates. Additionally, to strengthen the theoretical foundation of the analysis, we could consider consulting relevant literature or experts in gender studies or labor economics to guide variable selection and interpret findings more effectively.