**Data Science PDA 7 Project Planning Stage**

Introduction

In today’s society, developed countries seem to have disguised the prejudices abhorrently apparent in many developing countries, namely sexism. Has the subsiding patriarchy, the remnants of male dominated religions, fully transited to a point where male and females are equal in nowadays society or does it seep unnoticed from the foundations. Let’s look at the gender pay gap to discover more about equal treatment in the UK today.

The Problem

Aim:

The gender pay gap – mostly went unnoticed and only in 2017 was *Gender Pay Gap Reporting* added to the Equality Act of 2010 but this only applies to companies with 250+ employees[1], leaving small companies unchecked.

The aim of this project is to discover and analyze the gender pay gap since its inclusion into the Equality Act.

* Since 2017 has the pay gap been decreasing throughout the UK?
* Are there any places where the pay gap has remained unchanged or has widened?
* How does Scotland compare to England?
* How do rural areas compare to those that are more urbanized?

The Audience

The visualizations and aggregated data will be presented to a think tank from which they will suggest ways of increasing equality between genders in society. The medium for text related documents will be MS Word and the visual aspect will be presented in Jupyter Notebook[12] . As a contingency a PDF version of the notebook will be attached alongside to ensure successful delivery.

Methodology

(A) To give a more holistic representation of the pay gap throughout the UK, we must first identify industries and sectors that have the highest and lowest pay gap between genders and a sample industry which sits within the median. These 3 industries will allow us to see the overall trend in the UK so as to draw a better conclusion on whether or not the pay gap is lessening.

The benefits of analyzing the pay gap in this way, will allow us to be less biased about our conclusion and will help identify areas in the UK where the gap has not improved. If such places exist then we will be able to ask “Why?” and explore whether it is an outlier or if it is just dominated by a sector which ranks at the bottom of improving the pay gap.

(B) A problem that I foresee is that due to companies not having to report if they have less then 250 employees, then rural areas having a smaller population and smaller companies might show a biased representation of how much the pay gap persists. In which case finding data pertaining to those businesses with less than 250 employees or a data-set which includes smaller companies should prevent such bias.

The problem with the aforementioned is that the data-sets that I intended to use [2] are primarily aggregated and so most don’t show the size of the companies that make up the gender gap statistics. However, the comma separated files[3] have the raw data which might help in identifying if those companies with less than 250+ employees are included in the aggregated scores.

The Data-sets

The Data-sets described above on the Gov.uk website, one being a zip file consisting of ‘xlsx’ files[2] that are aggregated, another being a set of ‘csv’ files[3] representing each year since 2017 .

The zip file contains numerous xlsx files but those of interest are listed below:

* Home Geography Table 8.12
* Occupation Table 14.12
* Work Geography Table 7.12
* Industry Table 16.12
* Work Region Industry Table 5.12

However, due to the official analysis on the gov.uk website[4] which states that between 2022 and 2023 the largest difference in occupational groups (skilled trade) has decreased, taking the mode of the lowest, highest and median occupational groups since 2017 might be a fairer and more representative analysis of the pay gap rather than just using 2023 as a marker of where each occupation falls in the pay gap measure. In which case several zip folders with the above files will be needed, as each zip folder accounts for 1 year.

The xlsx workbooks are split into 3 different sheets (not include the notes worksheet) “All”, “Full time”, “Part time” – For this analysis, since we are not differentiating between the number of women in higher/lower paid jobs and solely focusing on the difference in pay, it makes sense that we use “All’ data, to have a more inclusive overview regardless of position or contractual hours.

Each sheet is split into “Description”, “Code”, “Gender Pay Gap Median” and “Gender Pay Gap Mean”. The Gender gap median is the difference in the average hourly earnings between men and women (not including overtime) and the mean is the mean of the given sample.

For this analysis, the median is going to be used as a measure of Gender Pay Gap, as we are going to use the xlsx files to identify occupations and counties and contrast with the data in the csv files, thus the mean would be an unrealistic measure.

The CSV files, consist of a few difference columns, none that adequately identifies the company sector. However further reading[5] suggests that SIC codes can be used to identify a company sector, this column is present in the current CSV file which means maybe grouping by SIC before identifying the industry from the list, mentioned in the above link.

The problem is whether, the list should just be parsed using “BeautifulSoup4”[6] at the time of interpretation or whether the list should be transformed into a CSV file before analysis for ease.

The CSV file from[3] also includes postcodes as a column, this can be used with “GeoPy”[7] to identify location but again having a CSV file containing the UK postcodes in a CSV format.

Other notable columns is are:

* EmployerSize so that those companies less than 250 people can be accounted for.
* DiffMeanHourlyPercent – Mean percent between male and female pay
* DiffMedianHourlyPercent -Median percent between male and female pay

A negative value on either the DiffMean/DiffMedian HourlyPercent columns represents women’s pay being higher than men’s. Since we are interested in only those sectors where the gender pay gap is higher for men than women, to highlight job sectors or cities where inequality persists, the negative values should be removed when cleaning the data-set.

After parsing the CSV files, the merged/joined data-frame should contain approx (10,000\*(2023-2017))

The Plan

This section will outline the proposed direction and implementation of the analysis and visualizations, answer some of the questions outlined above and loosely estimate the amount of time that will be needed to produce and evaluate the material hereto described.

Duration 1 to 2 days – Transform and Clean Data, Source additional data.

* **Download Postcode CSV file -** In order to connect each row of the CSV data with a city the postcode column must be used. Instead of geo-locating the city using GeoPy - as described in the previous section, a CSV file of postcodes and their corresponding city can be found at ONS Postcode Directory[8]
* **Download SIC code CSV file –** In order to connect each row to an industry, Standard Industrial Codes must be used. Previously it was suggested that they could just be parsed using BeautifulSoup but a CSV file containing the needed information has been found[9]
* **Download and load the CSV and xlsx data-sets** – Once downloaded, the data will be loaded using Pandas[10]. A Python[11] library.
* **Merge related Data Frames** – Each year from the CSV files will be merged into one data-frame, containing 70,000+ rows and a subset of columns
* **Clean the data-frame**
* **Load additional CSV files –** Load additional CSV files such as postcodes and SIC codes into a data-frame
* **Transform the data-frame -**  Transform the data-frame, working out city from postcode and industry sector from SIC codes, using the additional data-frames.
* **Unzip and load xlsx files –** Unzip and load the xlsx files
* **Identify Lowest, Median and Highest occupational sectors from xlsx file**

Duration 1 to 2 days – Aggregate data

* **Aggregate gender pay gap based on job sector**
* **Aggregate gender pay gap based on city and distinguish between urban/ rural based on population.**
* **Aggregate gender pay gap based on country (England, Scotland, Wales, Northern Ireland)**

Duration 1 to 2 days – Visualize data

* **Compare rural cities to urban cities (multiple plots)**
* **Compare the gender pay gap in Scotland to that in England (multiple plots)**
* **Compare the lowest, median and highest sample job sectors, identified above and compare the trend from 2017 to 2023**
* **Inspect data and visualizations to identify any other necessary plotting**

Duration 1 to 2 days – Evaluation

* Write an evaluation of the development stage

From the above estimations, it should take approximately 8 days to completion.

1. Transform and Clean Data, Source additional data (02/01/24 - 04/01/24)
2. Aggregate data (04/01/24 - 06/01/24)
3. Visualize data (06/01/24 - 08/01/24)
4. Evaluation (08/01/24 – 10/01/24)

If the development stage begins on the 02/01/24 as expected, it should be complete by the 08/01/24

Reference List

1. ICAEW (2023) Gender Pay Gap reporting: the regulations, Available at <https://www.icaew.com/technical/trust-and-ethics/information-law-and-guidance/gender-pay-gap-reporting-the-regulations> (Accessed 31st December 2023).

2. Government Digital Service (2023) Gender Pay Gap, Available at <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/annualsurveyofhoursandearningsashegenderpaygaptables> (Accessed 31st December 2023).

3. Government Digital Service (2023) Gender Pay Gap Service, Available at <https://gender-pay-gap.service.gov.uk/viewing/download> (Accessed 31st December 2023).

4. Office For National Statistics (2023) Gender Pay Gap In The UK 2023 <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/genderpaygapintheuk/2023> (Accessed 01st January 2024)

5. Companies House (2023) Nature of Business: Standard Industrial Classifications <https://resources.companieshouse.gov.uk/sic/> (Accessed 01st January 2024)

6. Leonardo Richardson (2023) BeautifulSoup4 <https://pypi.org/project/beautifulsoup4/> (Accessed 01st January 2024)

7. PyPi (2023) Geopy <https://pypi.org/project/geopy/> (Accessed 01st January 2024)

8. Office For National Statistics (2023) Postcode Directory <https://geoportal.statistics.gov.uk/datasets/ons::ons-postcode-directory-may-2023/about> (Accessed 01st January 2024)

9. Companies House (2023) Condensed SIC List In Condensed Format <https://assets-origin.publishing.service.gov.uk/media/5a7f8639e5274a2e87db65e1/SIC07_CH_condensed_list_en.csv/preview> (Accessed 01st January 2024)

10. The Pandas Development Team (2023) PyPI <https://pypi.org/project/pandas/> (Accessed 01st January 2024)

11. Python Org (2023) <https://www.python.org/> (Accessed 01st January 2024)

12. Jupyter (2023) <https://jupyter.org/>(Accessed 01st January 2024)