COLLEGE OF ENGINEERING AND COMPUTER SCIENCE

VINUNIVERSITY

[COMP4010] Data Visualization
Project 2 Write up

Bachelor of Science in Computer Science

Prepared by

Hoang Khoi Nguyen	**SIS ID: 202000222**
Nguyen Duy Anh Quan	**SIS ID: 202000284**
Khau Lien Kiet	**SIS ID: 202000119**

Under the supervision of Prof. Le Duy Dung

1. Introduction

1.1. High-level Goal

In this project, we explore the different facets of gender pay gaps in the United Kingdom. Despite advancements in workplace equality, wage disparities between genders remain a significant challenge across various sectors. Utilizing data from the UK Pay Gap initiative, our objective is to create an interactive dashboard that not only highlights these disparities but also provides insights into their dynamics over time and across different industries. This visual tool aims to foster a deeper understanding of the underlying factors contributing to pay inequality.

1.2. Description

- **Objective**: The primary goal is to construct a comprehensive interactive dashboard that visualizes different aspects of the gender pay gap across various sectors—ranging from technology and finance to education and healthcare. This dashboard will not only display current disparities but also track changes over time, providing a temporal dimension to our analysis.
- **Approach**: We intend to use Shiny framework to present this complex data in an accessible and engaging manner, which is a dashboard. Through interactive charts and graphs, users will be able to explore different layers of the data, such as comparisons by sector and changes over the years.
- **Impact**: The intended impact of our visualization is to increase awareness and understanding of the pay gaps within different sectors of the UK economy.
- **Innovation**: Our approach is innovative in its use of interactive elements to engage users actively. By allowing users to navigate through different aspects of the data, users will have a more profound engagement with the issues and develop better understanding of how pay gaps affect various sectors differently. In addition, the product is accessible, thereby user could interact with this innovation at any time with any devices.

1.3. Variables used

Field	Description
EmployerName	The name of the employer at the time of reporting
DiffMeanHourlyPercent	Mean % difference between male and female hourly pay (negative = women's mean hourly pay is higher)
MaleLowerQuartile	Percentage of males in the lower hourly pay quarter
FemaleLowerQuartile	Percentage of females in the lower hourly pay quarter

MaleLowerMiddleQuartile	Percentage of males in the lower middle hourly pay quarter
FemaleLowerMiddleQuartile	Percentage of females in the lower middle hourly pay quarter
MaleUpperMiddleQuartile	Percentage of males in the upper middle hourly pay quarter
FemaleUpperMiddleQuartile	Percentage of females in the upper middle hourly pay quarter
MaleTopQuartile	Percentage of males in the top hourly pay quarter
FemaleTopQuartile	Percentage of females in the top hourly pay quarter
EmployerSize	Number of employees employed by an employer

Table 1: Definition of used variables

2. Approach

2.1. Data Preprocessing – Categorizing by Sector

As we want to do our analysis broadly by sectors, we will create a new data frame to filter out companies belonging to targeted sectors. Potential notable sectors may include, but not limited to Technology, Finance, Education, Healthcare, Retail, Entertainment, Travel, Consulting, Banking, Marketing. This new data frame consists of three columns: `EmployerName`, `EmployerID`, and the `Sector` it belongs to. Companies that cannot be categorized into our target sectors will be categorized as "Others".

Our filtering and categorizing plan involves three steps:

- **Step 1:** Group by EmployerID, store SIC code, EmployerID, EmployerName, count occurrence. The raw pay gap data frame has record of company over many years, so grouping them reduces repeated categorizing. Also, we want to only analyze companies that have complete records from 2018 to 2022. Assuming that companies only submit one report each year, remove companies that have < 6 count.
- **Step 2:** In code, employ keyword-based filtering. Create a list of keywords that each company belonging to a sector may have in their name and categorize accordingly. For instance, Employer with EmployerName == Bank of England belongs to Banking sector.
- **Step 3:** For the remaining uncategorized companies, we will merge by SIC code with `SIC07_CH_condensed_list_en` data frame to get `description`, which describes the employer's purpose and sectors of work at the time of reporting, to manually categorize. Recursively add new keywords representing sectors that we may have missed in step 2.

2.2. Question for Analysis

- **Theme:** On each sector, what are the paygap, paygap change over year, and distribution of population among gender?
- **Mini Question 1:** How is the difference in paygap with respect to the hourly pay in these sectors?
    - **Variables:** `current_name`, `diff_mean_hourly_percent`
    - **Approach:**
        - Utilize a diverging bar chart on each sector
        - Create a data frame for all companies in the same sector. Place `diff_mean_hourly_percent` on x_axis and `current_name` known as the company name on y_axis. This could be achieved by using ggplot2.
        - Ensure the chart provides a consistent theme among each sector's plot and use a contrast color scheme to distinguish between the genders.
- **Mini Question 2:** How do the percentage of females vs males in an hourly pay quartile range change from 2018 to 2022?
    - **Variables:** pairs in `MaleLowerQuartile`, `FemaleLowerQuartile`, `MaleLowerMiddleQuartile`, `FemaleLowerMiddleQuartile`, `MaleUpperMiddleQuartile`, `FemaleUpperMiddleQuartile`, `MaleTopQuartile`,`FemaleTopQuartile`, `EmployerSize`
    - **Approach:**
        - Utilize a dumbbell chart
        - Faceting for quartile ranges.
        - For each hourly pay quartile range, the y axis is the year, and the x axis are the female and male percentage in that quartile range.
        - Represent the value of female and male percentage by dots with two contrasting color, with a line connecting two dots within same year.
- **Mini Question 3:** What is the distribution of population across income groups by gender?
    - **Variables:** `MaleLowerQuartile`, `FemaleLowerQuartile`, `MaleLowerMiddleQuartile`, `FemaleLowerMiddleQuartile`, `MaleUpperMiddleQuartile`, `FemaleUpperMiddleQuartile`, `MaleTopQuartile`,`FemaleTopQuartile`, `EmployerSize`
    - **Approach:**
        - Utilize a bar chart with two genders side-by-side
        - As we have the percentage of the population by pay quartile, we can plot the distribution of population across the income groups by gender.

2.3. Build a dashboard

As our team's approach is to build a dashboard, it is necessary to use the **`shiny`** package, in which a need on the design of UI and its server logic. In addition, to get the product accessible to many users, we would deploy the app by installing the **`rsconnect`** package in RStudio and set

up the account information to link RStudio with shinyapps.io. Afterwards, our team would have the public domain to host the Shiny app from local to website server.

3. Implementation

3.1. Components

The dashboard has the User Interface (UI) and its logic server (Back-end).

- **UI**
  - Tab: Introduction, Overall Analysis, Sector Analysis, Team
  - Select option
    - Select tab
    - Select sector
    - Select year
- **Server**
  - The database: `sector_data_frames`
  - The plot functions: `plot_diverging_bar_chart`, `plot_dumbell_chart`, `plot_gender_distribution`

3.2. User flow

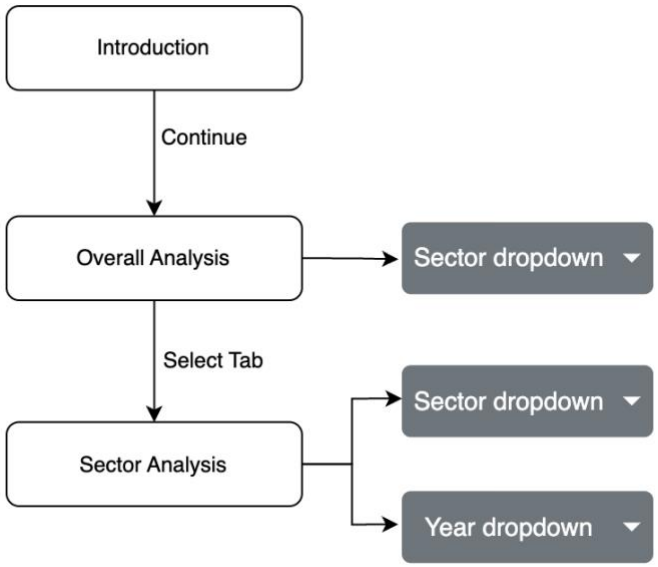

Figure 1: Product User Flow

[1] When user accesses the product, the default page is on Introduction tab.

[2] When clicking on `Continue` button, it navigates to the `Overall Analysis` tab.

[3] In this tab, user could choose any sector among 10 listed ones.

[4] After having the full picture of the sector from 2018 to 2022, they could look on details of the sector on that particular year by navigating to `Sector Analysis` tab.

[4] User selects the sector and observe the particular year.

[5] After choosing, the tab is shown two following graphs:

[5.1] Average Percentage by Gender Across Quartiles: Let user observe how disparity among different types of quartiles

[5.2] Diverging Bar chart of Sector Pay gap: Let user know on each single company having its percentage of how difference among male and female hourly pay and how much.

3.3. User manual

- Since the product had been deployed in the shinyapp.io. For accessing this dashboard, user could visit this URL https://paygapuk.shinyapps.io/dashboard/ to view and not for editing.
- Lesson plan: When user access the dashboard, we would provide the general illustration of the graph. Importantly, to satisfy the user-friendly product, having description on how to read the graph to let user get the intuitive thought before reading the chart.

3.4. Feature analysis

No.	Feature	Description
1	Tab	There are 4 tabs to serve for each purpose: ▪ `Introduction` tab ▪ `Overall Analysis` tab ▪ `Sector Analysis` tab ▪ `Team` tab
2	Option selection	At the `Overall Analysis` tab, user could select among 10 sectors At the `Sector Analysis` tab, user could select among 10 sectors, in which years ranging from 2018 to 2022
3	Story	The story would follow from the beginning to the end, adapted with each tab:

		▪ `Introduction` tab: The product provides the overview and the goal of the product ▪ `Overall Analysis` tab: The product provides the notations under the graph to help user with an intuitive understanding. ▪ `Sector Analysis` tab: The product provides the notations under the two graphs to help user with an intuitive understanding.
4	Interactive plot investigation	The plot provides a functionality that user can zoom in/zoom out to investigate more details of the illustration from different views. Additionally, user could hover to each chart component to get the accurate percentage or number for any detail research.

Table 2: Feature description

4. Discussion

Mini Question 1: How is the difference in paygap with respect to the hourly pay in these sectors (i.e. Technology, Airline, Pharmaceutical/Medical, Financial, Bank, Hotel, Consulting, Accounting, Marketing, School/Education, Travel, etc.)?

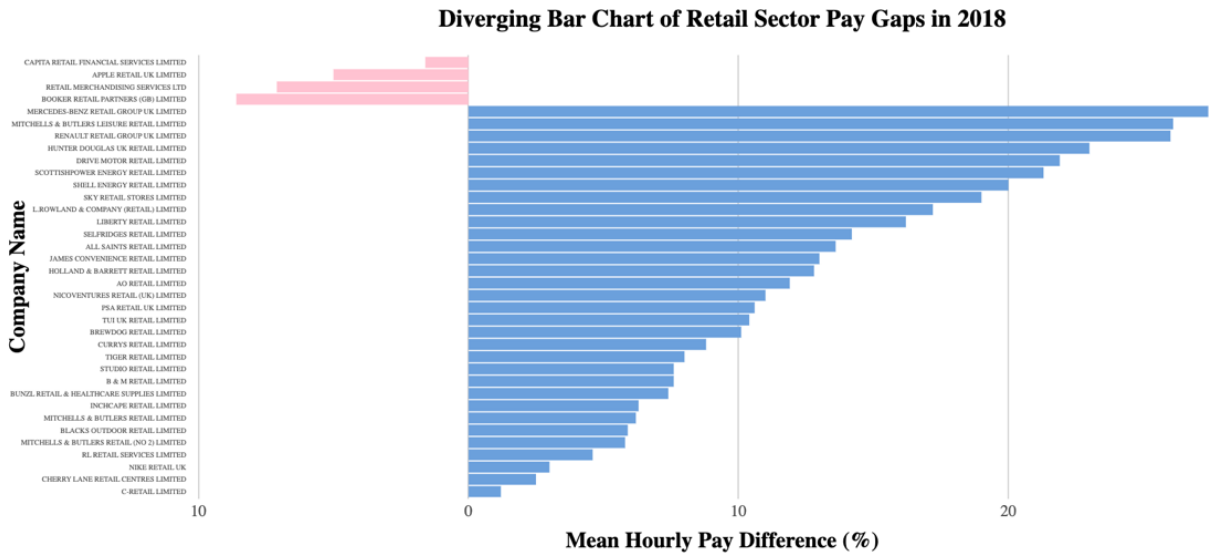

Figure 2: Diverging Bar Chart of Pay Gaps in Retail Sector in 2018

- From the graphs of mini question 1, it could be seen that for every sectors across the 5 years, the majority of companies would have males being better paid than females. There are very few instances where we would be able to observe companies having females being paid higher than males.

Mini Question 2: How do the percentage of females vs males in an hourly pay quartile range change from 2018 to 2022?

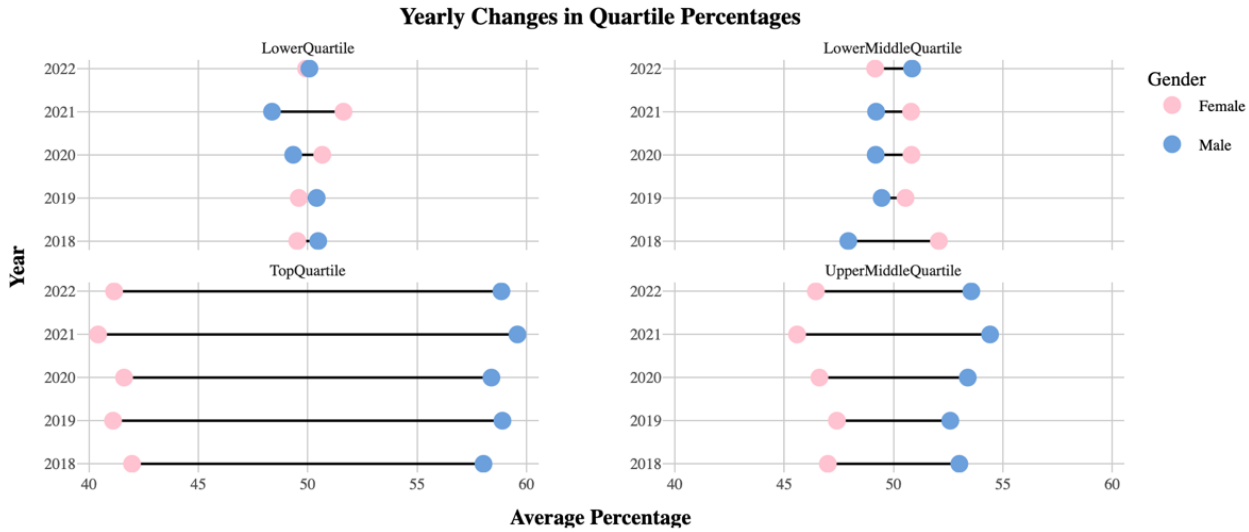

Figure 3: Average Percentage by Gender in Quartile Percentages in Retail Sector from 2018 to 2022

- For the Healthcare and Education sector, females are consistently paid higher than males across all the pay quartile and this train remains consistent throughout the 5 years.
- For the Consulting sector, males are paid higher than females across all pay quartiles except for the Lower Quartile. Moreover, there seem to be no clear trends of changing in the pay gaps across the pay quartiles.
- For the Travel sector, for the Top Quartile, males are paid higher than females and the gap continues to widen. For the Lower Middle Quartile, despite females being paid higher than males, the gap is narrowing.

Mini Question 3: What is the distribution of population across income groups by gender?

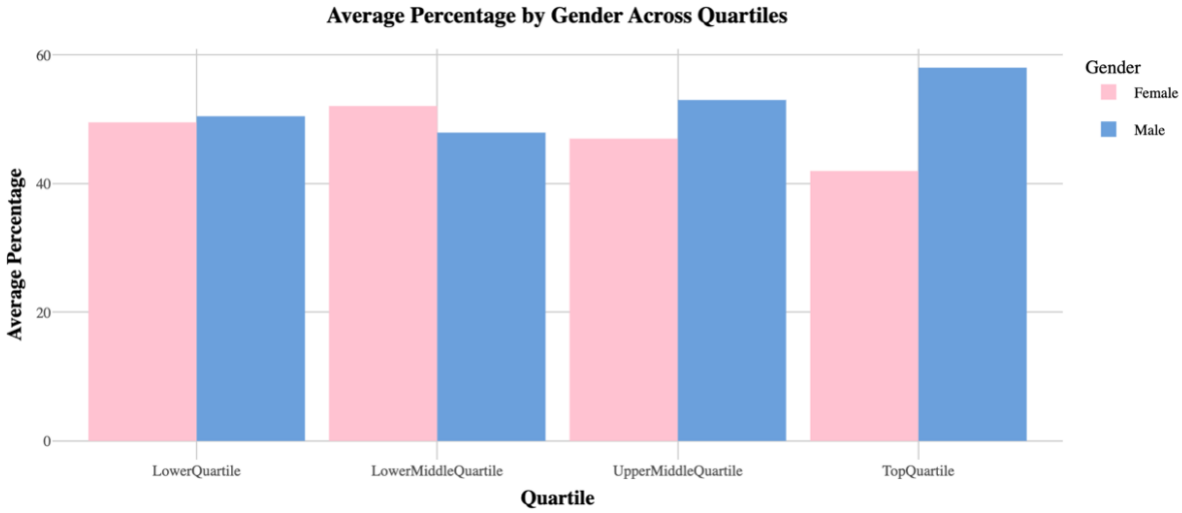

Figure 4: Average Percentage by Gender of Retail Sector Across Quartiles

- For the Consulting sector, females are being paid fairly equal to males starting from the Lower Quartile. However, the gaps became increasingly large as the quartile progresses in favor of males. This is true throughout the 5 years of sample data.
- For the Entertainment sector, females are being paid higher than males starting from the Lower Quartile. The gaps, however, began to close as the quartile progresses.

5. Limitations

- The data used in this project is from the UK Pay Gap initiative. This means that our findings are not generalizable to other countries or regions.
- The data volumn is very large. As a result, we have to make a trade-off between effective visualization and comprehensive data representation.

6. Future Directions

- The scope of the dashboard could be expanded to include data from other countries or regions. This would allow for a more global perspective on gender pay gap.
- In addition to hourly wages, we could also include information on salary, bonuses, and benefits. This would provide a more comprehensive picture of the gender pay gap.
- The dashboard could also incorporate data on other factors that contribute to the gender pay gap, such as education, experience, and career breaks. This would allow for a more nuanced understanding of the issue.

References

- https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/articles/findoutthegenderpaygapforyourjob/2016-12-09
- GitHub: https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-06-28
- Dataset: https://gender-pay-gap.service.gov.uk/viewing/download