

An example document looking into the gender pay gap in UK companies*

Andrew Moles

Anton Boichenko

21 November, 2022

Abstract

The gender pay gap has been an issue for generations, but how are UK companies doing in the current climate?

*Replication files are available on the [author's Github account](#)

Table of contents

Introduction	3
Methods	4
Results	5
Discussion	8

List of Figures

1	The gender pay gap bot in action	3
2	Median percentage difference in hourly pay by postcode area and company size . . .	7

List of Tables

1	Table to show the sectors with, on average, the largest percent gap in men's wages compared to women's wages.	5
---	---	---

Introduction

International women's day in 2022 was full of drama in the Twitter world. This was caused by the *Gender Pay Gap Bot*. Whenever a company listed on the [governments gender pay gap service](#) tweets about international women's day, the bot will automatically respond with that companies median gender pay gap.

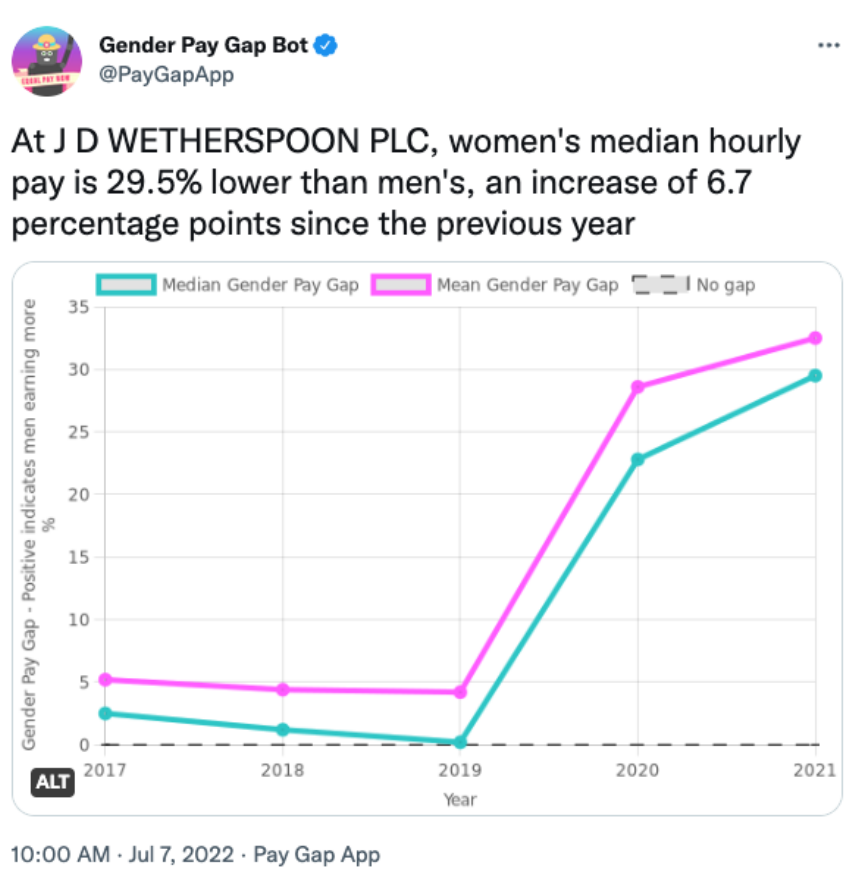


Figure 1: The gender pay gap bot in action

The bot caused a lot of companies to remove their tweets and even post some amendments. The point of the project was to call out companies that talked a good talk but had very poor gender pay equality.

This report wanted to look more into the governments gender pay gap service data to see if we could see any extra insights or interesting patterns.

Methods

The data used for this report is from the UK [governments gender pay gap service](#). This data has been helpfully hosted and joined together by [the TidyTuesday project](#).

There are several variables of interest which we can use in an analysis of this data:

DiffMeanHourlyPercent, DiffMedianHourlyPercent, PostCode, SicCodes, and EmployerSize.

DiffMeanHourlyPercent and DiffMedianHourlyPercent are the mean or median % difference between male and female hourly pay (negative = women's mean/median hourly pay is higher). PostCode is each companies postal code. SicCodes are used to describe the employer's purpose and sectors of work at the time of reporting, e.g. the company is in the finance sector. EmployerSize indicates the number of employers which is grouped into bands such as *250 to 499* or *5000 to 19,999*.

The formula used to calculate the DiffMeanHourlyPercent and DiffMedianHourlyPercent columns is as follows:

$$\frac{value_{original} - value_{new}}{value_{original}} \times 100$$

If females are paid better, the data will show a minus value for that column. This is because the original value in the formula is male as shown below:

$$\frac{male_wage - female_wage}{male_wage} \times 100$$

In the analysis of the data several R packages were used which are: the tidyverse [[Wickham et al., 2019](#)], tidytext [[Silge and Robinson, 2016](#)], ggtext [[Wilke, 2020](#)], patchwork [[Pedersen, 2020](#)], geogrid [[Bailey, 2018](#)], rmapshaper [[Teucher and Russell, 2022](#)], and sf [[Pebesma, 2018](#)].

Results

We can first look at the sector averages, here we are looking at the top 10 sectors that have, on average, the largest percent gap in men's wages compared to women's wages. To get the sectors, Tokenisation¹ of Standard industrial classification of economic activities (SIC) codes was used in an attempted to simplify the results.

sector	percent increase in mens wages compared to womens
primary	0.2835
secondary	0.2494
education	0.2383
financial	0.2112
construction	0.1978
technology	0.1932
information	0.1892
head	0.164
offices	0.164
technical	0.1615

Table 1: Table to show the sectors with, on average, the largest percent gap in men's wages compared to women's wages.

Ideas for this analysis were taken from from this blog post by [Julia Silge](#).

A more elaborate analysis we can do is to look at the gender pay gap by company size and by postcode area. Figure 2 gives us an overall general picture of this difference by area. The clearest

¹Tokenisation is a way of separating text into smaller units called *tokens*, which can be either words, characters or sub words.

outcome here is that most of the grids for all employee sizes are green or black, the black hexagons indicating missing data for the postcode area. There are seemingly some areas where this isn't so much the case, which seem to be around Lancashire and Northumberland.

The code for Figure 2 came from [Georgios Karamandis](#) as a Tidy Tuesday contribution.

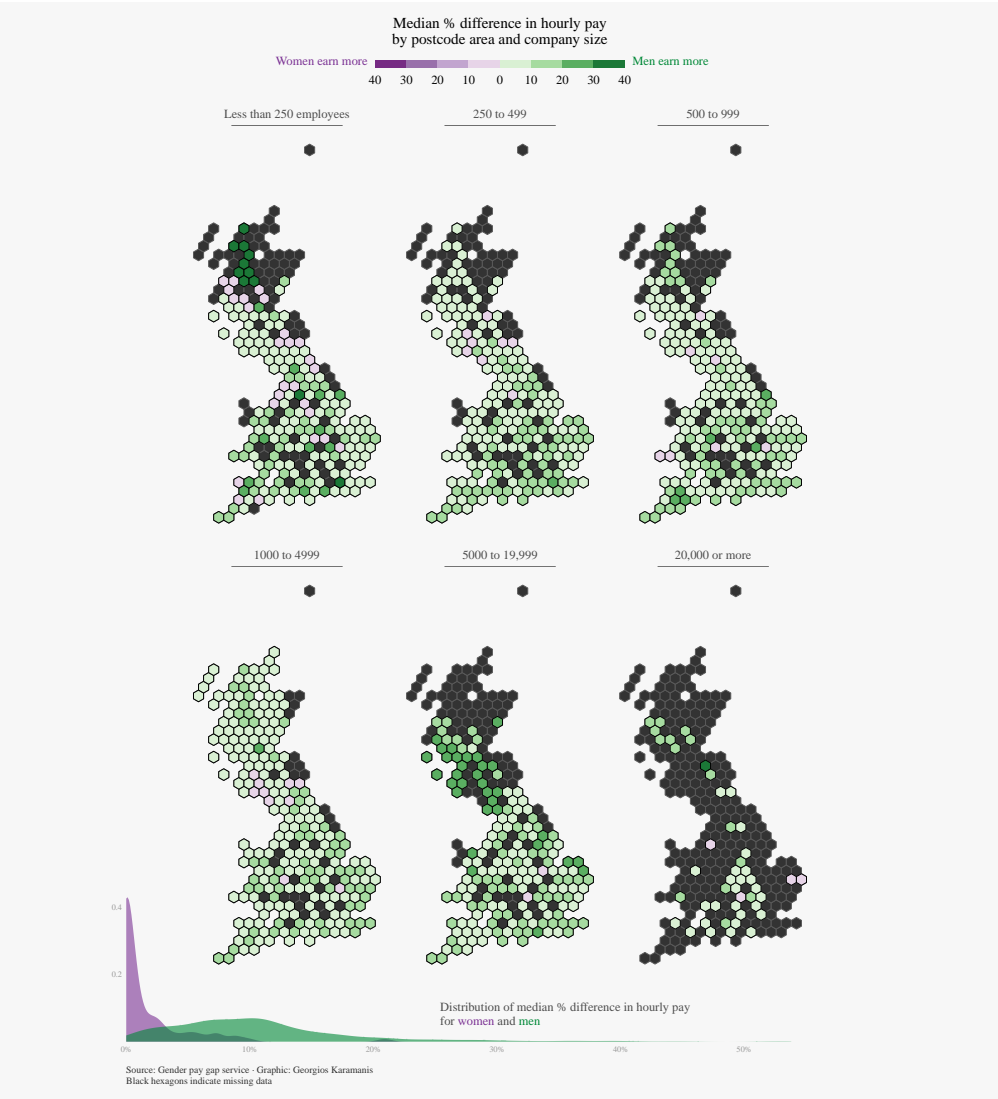


Figure 2: Median percentage difference in hourly pay by postcode area and company size

Discussion

Theresa May, in her first statement as Prime Minister highlighted that:

“If you’re a woman, you will earn less than a man.”

Our results show that on average, this is still the case but there are a lot of caveats.

Pay between genders is complicated as it does vary over a lifetime. It can differ depending on whether you are looking at full-time or part-time working (or both), what age you are, what ethnicity you are and your seniority within an occupation. For example there may be a wider (or narrower) GPG in those earning a higher salary as an actor than in those earning a lower salary.

It’s also important to consider the ratio of men to women in each occupation and whether any difference is due to one gender being disadvantaged in this field or whether they are just naturally more or less inclined go into that type of work².

The reasons for the gender pay gap are complex and overlapping:

- girls do well at school but often choose occupations or sectors that offer narrower scope for financial reward - many of the highest paying sectors are disproportionately made up of male employees.
- a proportion of the gap may be due to the negative effect on wages of having worked part-time or having taken time out of the labour market to look after family.
- women may not progress in work at the same rate as men due to cultural attitudes, lack of flexible working and stereotyping.
- some older women may need to learn new skills to take advantage of employment opportunities in growing sectors; others may have increased caring responsibilities for

²This text was taken from [ONS gender pay gap web pages](#)

partners, grandchildren or ageing parents.

References

- Joseph Bailey. *geogrid: Turn Geospatial Polygons into Regular or Hexagonal Grids*, 2018. URL <https://CRAN.R-project.org/package=geogrid>. R package version 0.1.1.
- Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL <https://doi.org/10.32614/RJ-2018-009>.
- Thomas Lin Pedersen. *patchwork: The Composer of Plots*, 2020. URL <https://CRAN.R-project.org/package=patchwork>. R package version 1.1.1.
- Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016. doi: 10.21105/joss.00037. URL <http://dx.doi.org/10.21105/joss.00037>.
- Andy Teucher and Kenton Russell. *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations*, 2022. URL <https://CRAN.R-project.org/package=rmapshaper>. R package version 0.4.6.
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.
- Claus O. Wilke. *ggtext: Improved Text Rendering Support for 'ggplot2'*, 2020. URL <https://CRAN.R-project.org/package=ggtext>. R package version 0.1.1.