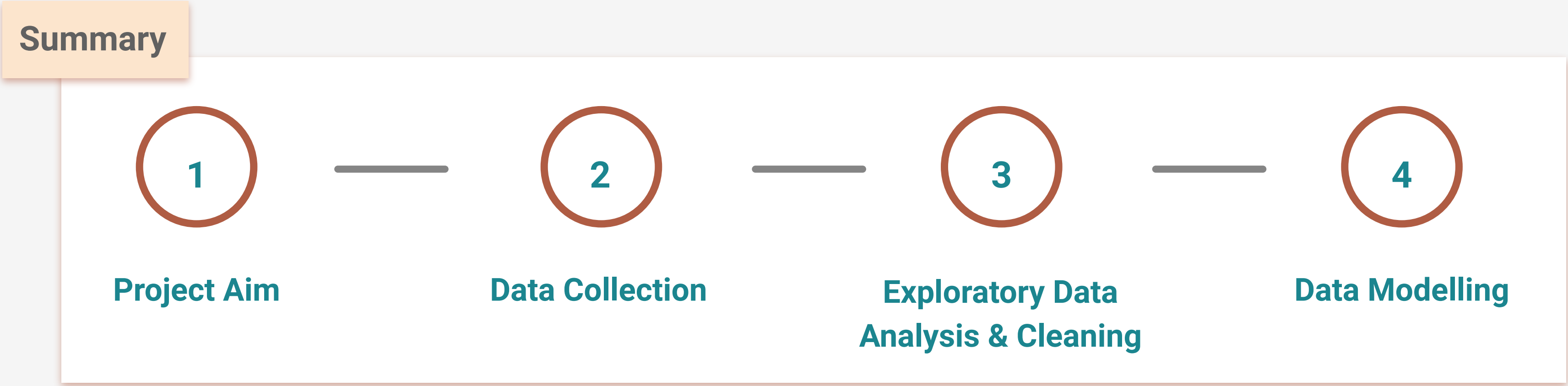
Predicting Gender Pay Gap Statistics

Gita Soni



Summary

1 Project Aim

2 Data Collection

3 Exploratory Data Analysis & Cleaning

4 Data Modelling

Project Background and Rationale

Background

The gender pay gap is the difference in the average hourly wage of all men and women across a workforce. If women do more of the less well paid jobs within an organisation than men, the gender pay gap is usually bigger.

Need

Reducing the gender pay gap is vital to achieving gender equality.

Project Goal

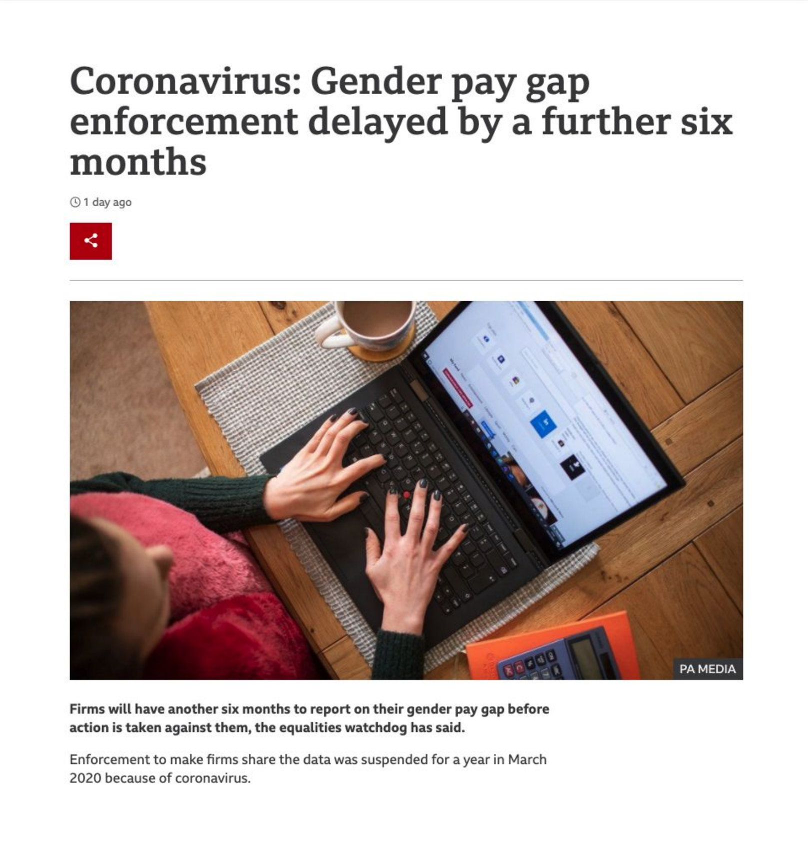
The aim of this project was to determine which factors best predict whether a company will have better or worse gender pay gap statistics, in order to understand which areas need to be targeted to help achieve gender equality.

In the News...

Previous legal requirement for companies with over 250 employees to report gender pay gap statistics.

As a result of Coronavirus, gender pay gap reporting was not required in 2020 or 2021.

This model uses data from the last full year where complete data was captured was 2019, which included data from March 2017 - April 2018.



Coronavirus: Gender pay gap enforcement delayed by a further six months

1 day ago

PA MEDIA

Firms will have another six months to report on their gender pay gap before action is taken against them, the equalities watchdog has said.

Enforcement to make firms share the data was suspended for a year in March 2020 because of coronavirus.

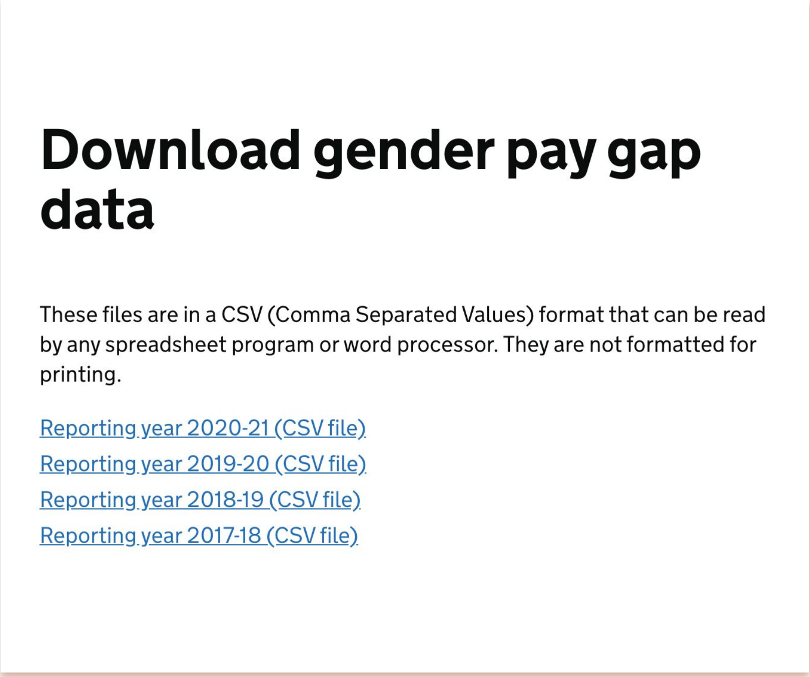
Data Collection

Dataset

The central data set came from the UK Government website - referred to in this project as the government gender pay gap data.

This includes a list of companies, their gender pay gap results and some basic company information:

- Employer Name
- Employer Sector
- Address
- Company Number
- Reported Gender Pay Gap Statistic
- Company Website Link
- Responsible Person
- Employer Size
- Submitted Date



Download gender pay gap data

These files are in a CSV (Comma Separated Values) format that can be read by any spreadsheet program or word processor. They are not formatted for printing.

Reporting year 2020-21 (CSV file)

Reporting year 2019-20 (CSV file)

Reporting year 2018-19 (CSV file)

Reporting year 2017-18 (CSV file)

Further Data

Along with the government gender pay gap data set, further company information was obtained from two key sources:

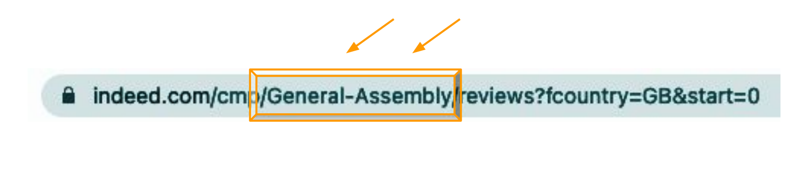
01 **Indeed.com**

- Company information such as ratings: Overall star rating, categorical rating in work life balance, pay and benefits, job security and advancements, management and culture.
- Reviews including list of pros and cons.

02 **Companies House**

- Further information on company status, type, location
- Number of female officers.

Indeed Web Scraping

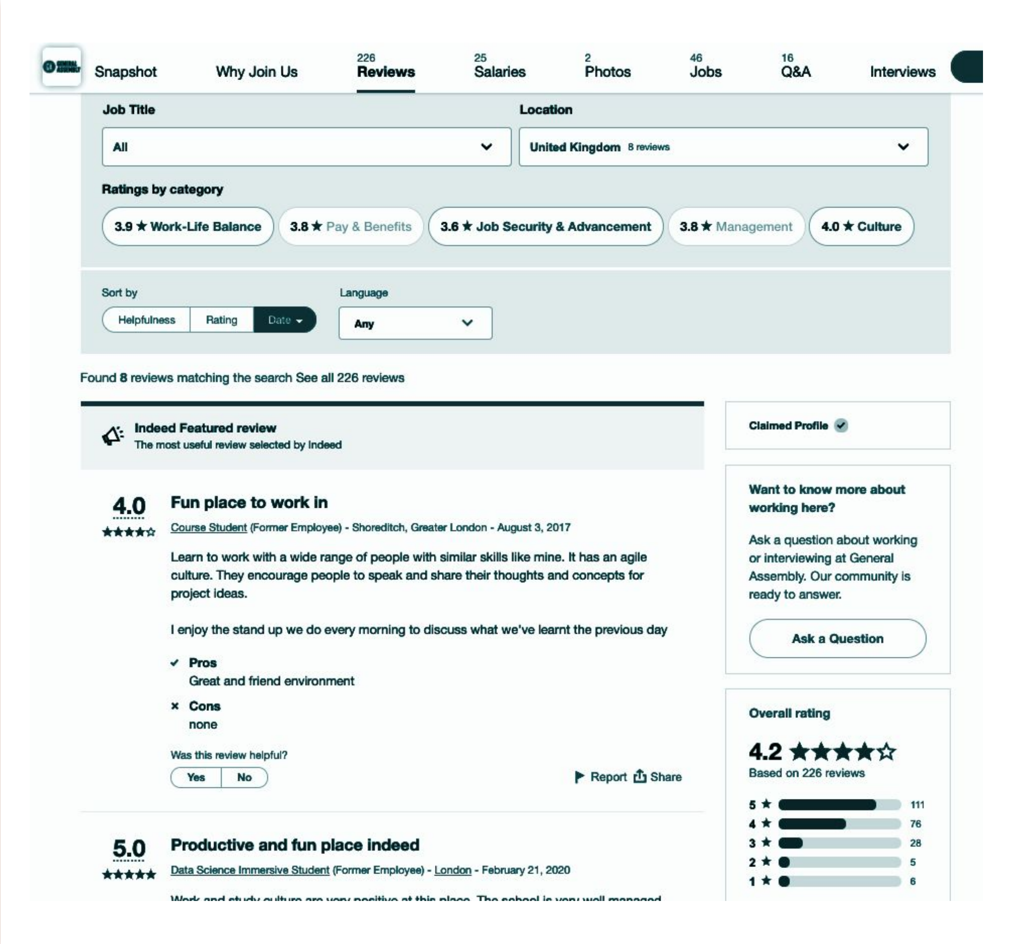


indeed.com/cmp/General-Assembly/reviews?fcountry=GB&start=0

Companies were listed in the government data set by company number or full company name but did not appear in the same way on the Indeed website.

The first word of the full company name was inserted in the URL to try and get a match.

This reduced the dataset by about 20% - partly because some companies did not have an Indeed company page and partly due to the above.



Job Title: All

Location: United Kingdom 8 reviews

Ratings by category

3.9 ★ Work-Life Balance | 3.8 ★ Pay & Benefits | 3.6 ★ Job Security & Advancement | 3.8 ★ Management | 4.0 ★ Culture

Sort by: Helpfulness | Rating | Date

Language: Any

Found 8 reviews matching the search See all 226 reviews

Indeed Featured review
The most useful review selected by Indeed

4.0 ★★★★☆

Fun place to work in

Course Student (Former Employee) - Shoreditch, Greater London - August 3, 2017

Learn to work with a wide range of people with similar skills like mine. It has an agile culture. They encourage people to speak and share their thoughts and concepts for project ideas.

I enjoy the stand up we do every morning to discuss what we've learnt the previous day

✓ Pros
Great and friend environment

× Cons
none

Was this review helpful?
Yes | No

Report | Share

5.0 ★★★★★

Productive and fun place indeed

Data Science Immersive Student (Former Employee) - London - February 21, 2020

Work and study culture are very positive at this place. The school is very well managed

Claimed Profile ✓

Want to know more about working here?

Ask a question about working or interviewing at General Assembly. Our community is ready to answer.

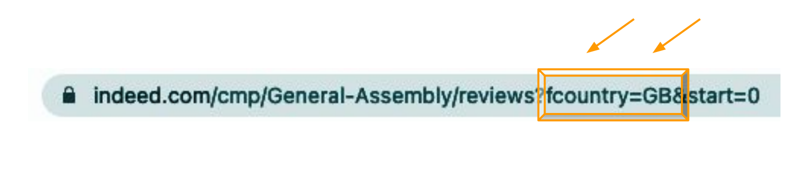
Ask a Question

Overall rating

4.2 ★★★★☆
Based on 226 reviews

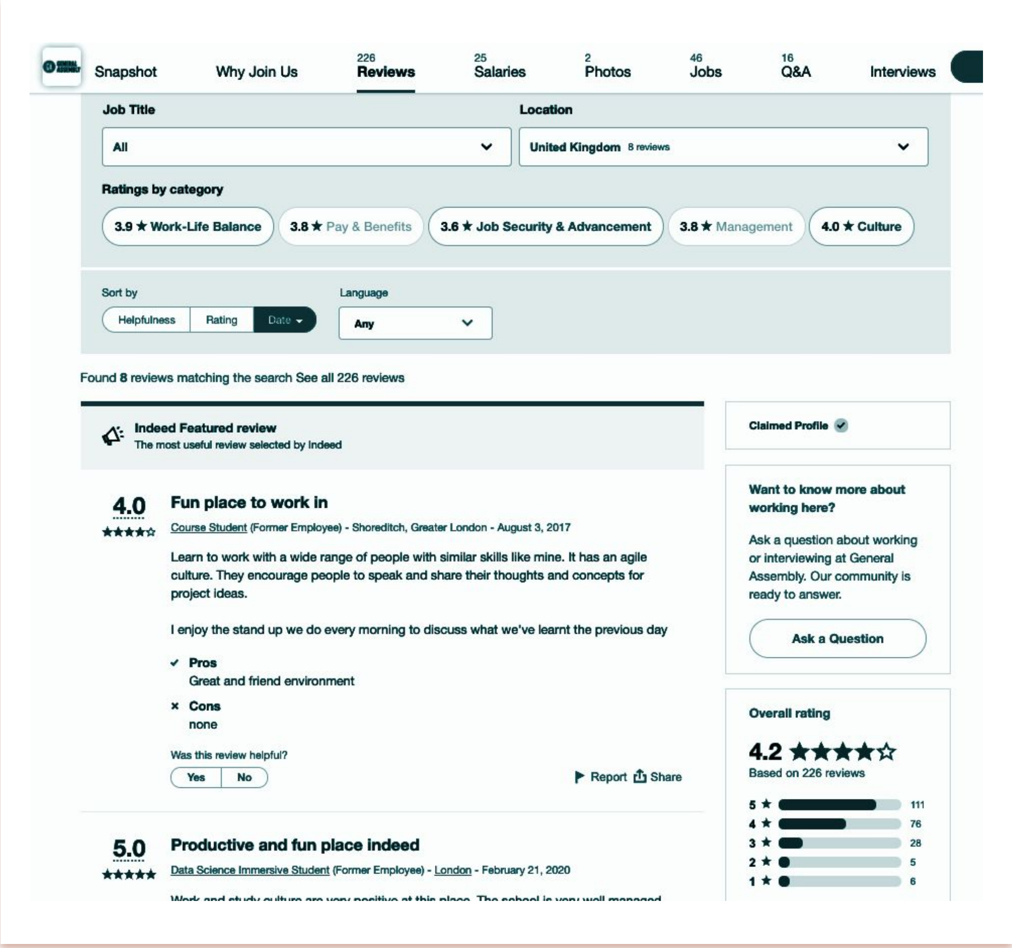
5 ★	111
4 ★	76
3 ★	28
2 ★	5
1 ★	6

Indeed Web Scraping



indeed.com/cmp/General-Assembly/reviews?fcountry=GB&start=0

To get the reviews for the UK only the filter 'country=GB' was applied in the URL.



Job Title: All

Location: United Kingdom 8 reviews

Ratings by category

3.9 ★ Work-Life Balance | 3.8 ★ Pay & Benefits | 3.6 ★ Job Security & Advancement | 3.8 ★ Management | 4.0 ★ Culture

Sort by: Helpfulness | Rating | Date

Language: Any

Found 8 reviews matching the search See all 226 reviews

Indeed Featured review
The most useful review selected by Indeed

4.0 ★★★★☆

Fun place to work in

Course Student (Former Employee) - Shoreditch, Greater London - August 3, 2017

Learn to work with a wide range of people with similar skills like mine. It has an agile culture. They encourage people to speak and share their thoughts and concepts for project ideas.

I enjoy the stand up we do every morning to discuss what we've learnt the previous day

✓ Pros
Great and friend environment

✕ Cons
none

Was this review helpful?
Yes | No

Report | Share

5.0 ★★★★★

Productive and fun place indeed

Data Science Immersive Student (Former Employee) - London - February 21, 2020

Work and study culture are very positive at this place. The school is very well managed

Claimed Profile

Want to know more about working here?

Ask a question about working or interviewing at General Assembly. Our community is ready to answer.

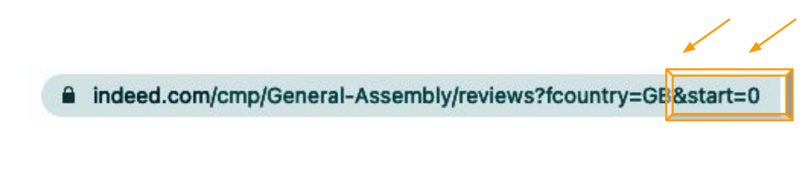
Ask a Question

Overall rating

4.2 ★★★★☆
Based on 226 reviews

5 ★	111
4 ★	76
3 ★	28
2 ★	5
1 ★	6

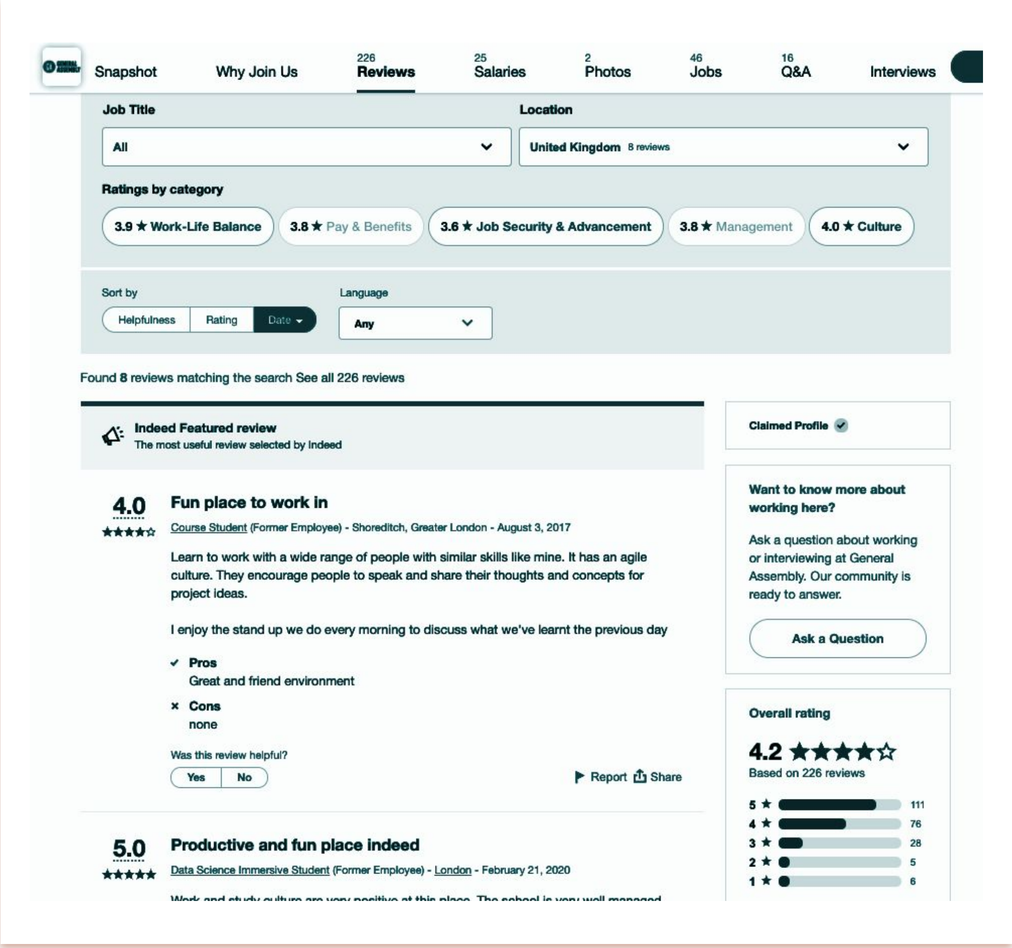
Indeed Web Scraping



indeed.com/cmp/General-Assembly/reviews?fcountry=GB&start=0

Each company had 20 reviews per page, and I needed to ensure there were no duplicates so had to tailor the code to check how many reviews a company has and then apply an appropriate filter.

To filter the results by date, the date of the review was converted to a DateTime feature, filtering on reviews after March 31 2018 - which is the date the government gender pay gap data related to.



Job Title: All

Location: United Kingdom 8 reviews

Ratings by category

3.9 ★ Work-Life Balance | 3.8 ★ Pay & Benefits | 3.6 ★ Job Security & Advancement | 3.8 ★ Management | 4.0 ★ Culture

Sort by: Helpfulness | Rating | Date

Language: Any

Found 8 reviews matching the search See all 226 reviews

Indeed Featured review
The most useful review selected by Indeed

4.0 ★★★★☆

Fun place to work in

Course Student (Former Employee) - Shoreditch, Greater London - August 3, 2017

Learn to work with a wide range of people with similar skills like mine. It has an agile culture. They encourage people to speak and share their thoughts and concepts for project ideas.

I enjoy the stand up we do every morning to discuss what we've learnt the previous day

✓ Pros
Great and friend environment

× Cons
none

Was this review helpful?
Yes | No

Report | Share

5.0 ★★★★★

Productive and fun place indeed

Data Science Immersive Student (Former Employee) - London - February 21, 2020

Work and study culture are very positive at this place. The school is very well managed

Claimed Profile ✓

Want to know more about working here?

Ask a question about working or interviewing at General Assembly. Our community is ready to answer.

Ask a Question

Overall rating

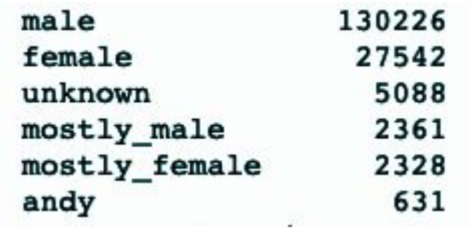
4.2 ★★★★☆
Based on 226 reviews

5 ★	111
4 ★	76
3 ★	28
2 ★	5
1 ★	6

Calculating Proportion of Female Officers

A gender classifier was used to classify each officer as male or female and calculate an overall company score for the proportion of female officers.

Some of the categories appear as ‘mostly_male’ or ‘mostly_female’, for these officers, their middle name was reviewed, if they had a male middle name and mostly_male first name the gender was manually input as male and similarly if they had a female middle name and mostly_female first name ender was manually input as female.



```
male             130226
female            27542
unknown            5088
mostly_male        2361
mostly_female      2328
andy                631
```

Calculating Proportion of Female Officers

527/ 7,810 which is 6.7% of the full list of companies had 50% or more female officers.

Exploratory Data Analysis

Feature Variables

The feature variables used in the final model were:

1 Employer Size

2 Sic Codes (Sector code)

3 Submitted after the deadline

4 Company Status

5 Jurisdiction

6 Registered address - country

7 Company Type

8 Locality

9 Proportion of Female Officers

10 Company Link

11 Responsible Person

12 Review Cons - Text Column

13 Review Ratings

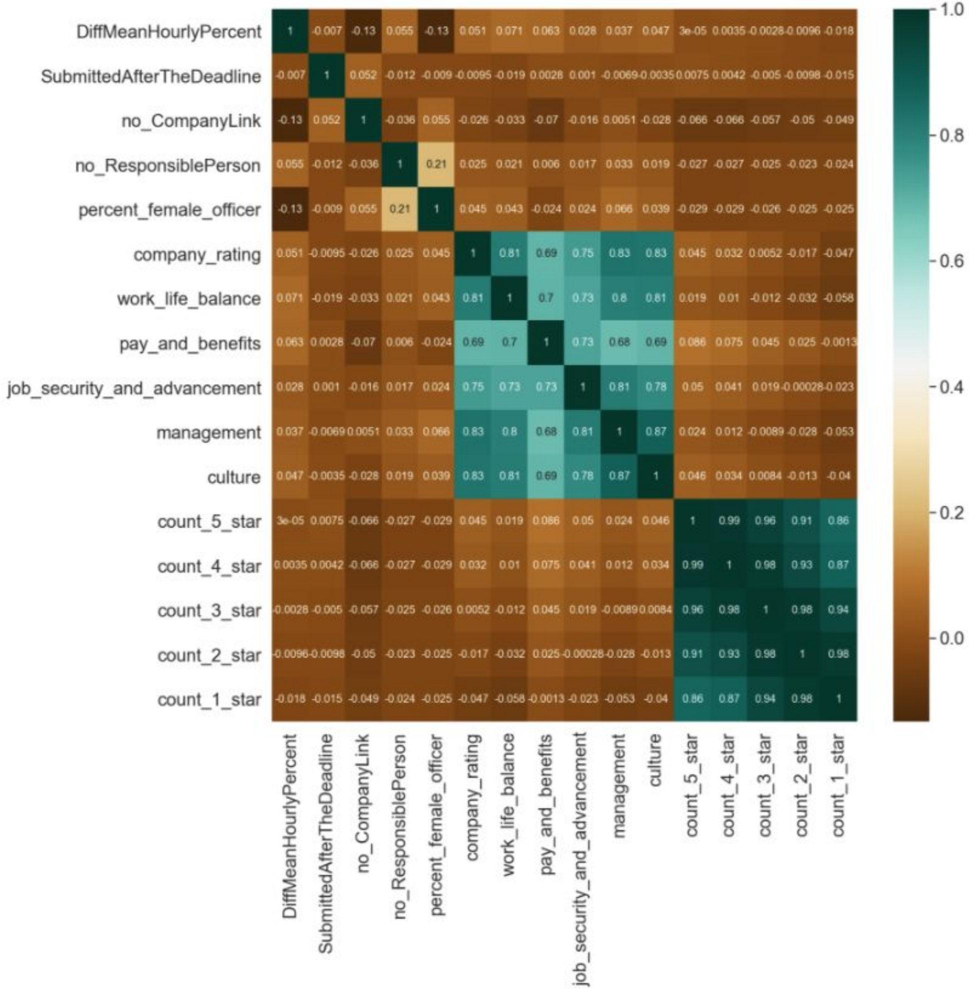
14 Count of star reviews

Heatmap

Companies who provided a company link to detailed pay gap reporting, as well as those with high % female officers had lower differences in pay.

Correlated ratings - people tend to rate consistently, all high or all low.

Correlated star count - larger companies have more reviews, small companies have fewer.

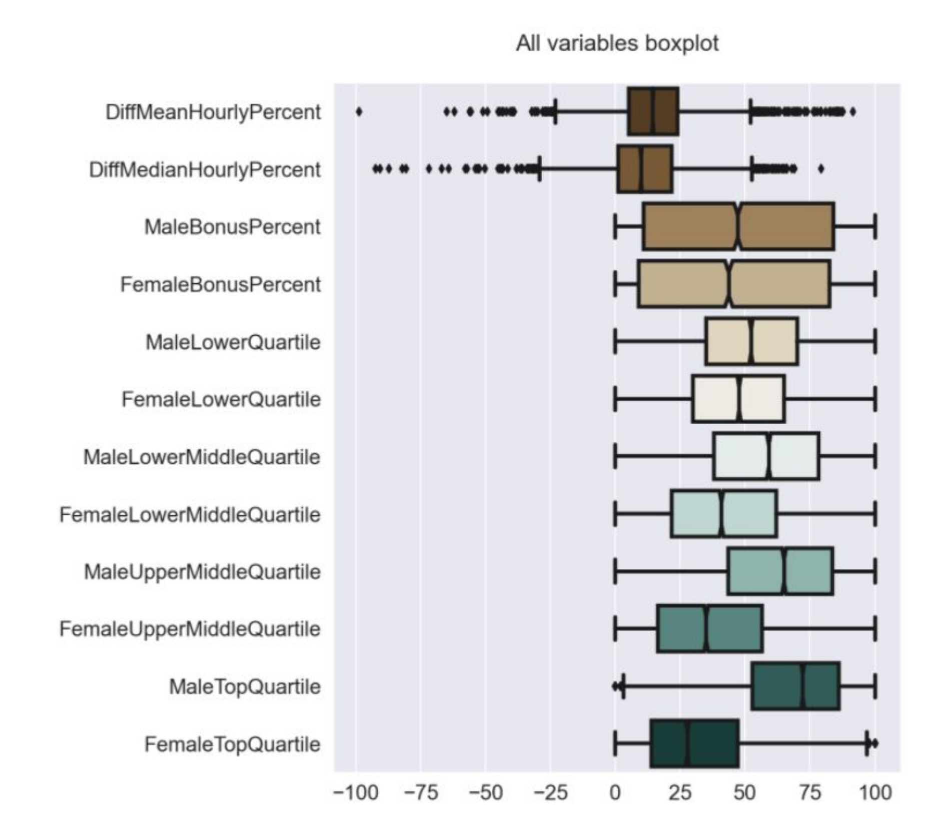


Selecting the Target Variable

The most salient statistics were:

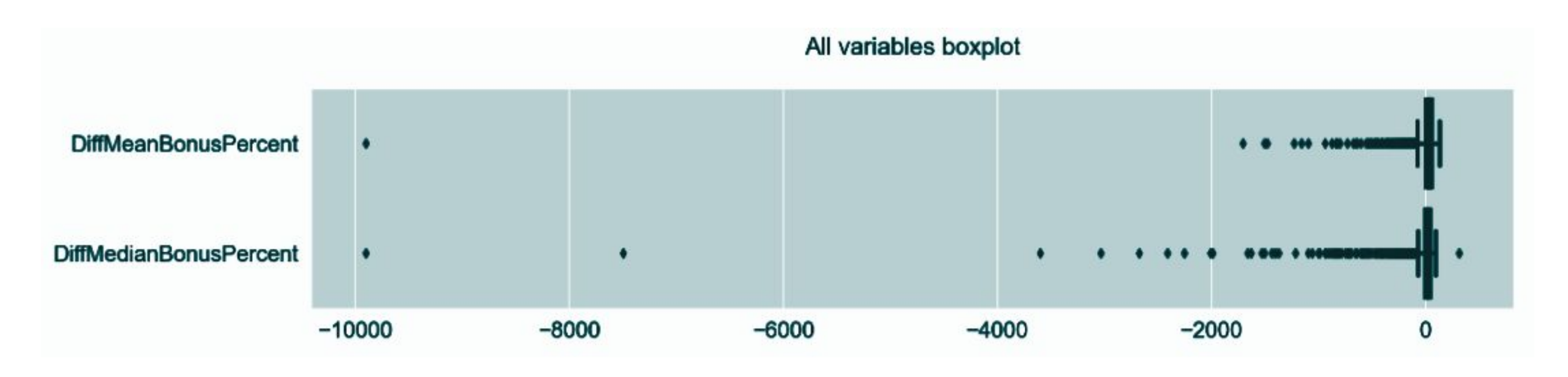
- ➔ % difference in mean hourly wage
- ➔ % difference in mean bonus pay
- ➔ Difference in the % men who received a bonus vs women

Correlated star count - larger companies have more reviews, small companies have fewer.

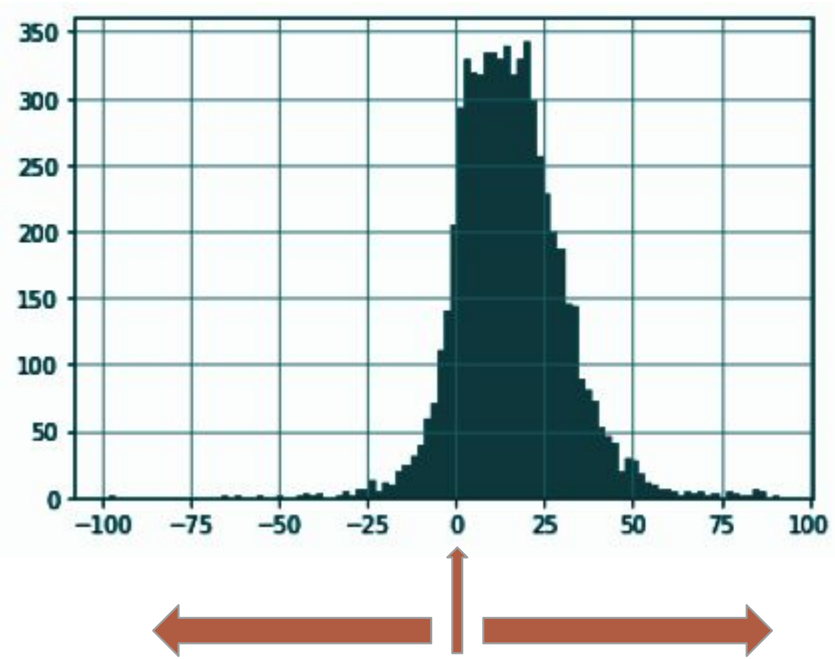


Selecting the Target Variable

The bonus pay contained a few outliers, the minimum value for the mean bonus pay was -9,900 which means than for that company, women's mean bonus pay was 9,900 times larger than men's.



Mean Pay Gap



Women Paid More

Men Paid More

Data Cleaning

The majority of data munging took place whilst scraping the data.

The main things required at this stage in terms of cleaning were checking all scraped data was as expected, and converting the company sector category codes to interpretable sector groups.

1

The first part involved ensuring data types were correct and all numerical values were in an appropriate range.

2

The second part involved converting the category codes into descriptions and then grouping into more general sector categories - these categories are pre-defined here: **resources.companieshouse.gov.uk/sic**

Data Modelling

Continuous or Categorical Target Variable

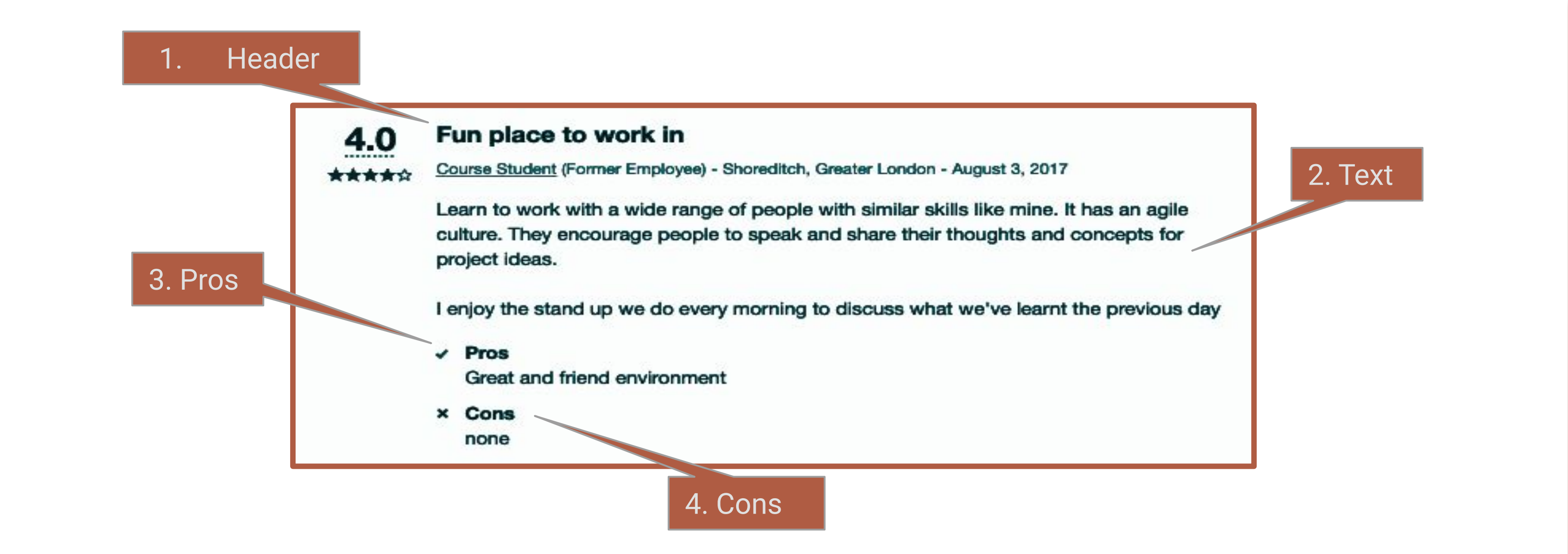
The first model reviewed was linear regression with a to predict the mean pay gap as a continuous variable. This gave a mean CV score of 0.208 - a little over bassline.

The target variable of mean pay gap was converted to a categorical variable with two categories:

- Above Median ‘Mean Gender Pay Gap’ Results (Referred to from now on as ‘above median’ gender pay gap for simplicity)
- Below Median ‘Mean Gender Pay Gap’ Results (Referred to from now on as ‘below median’ gender pay gap for simplicity)

Natural Language Processing (NLP)

The first step when modelling was to apply NLP to the review text that was scraped from the Indeed website.



1. Header

2. Text

3. Pros

4. Cons

4.0

★★★★☆

Fun place to work in

Course Student (Former Employee) - Shoreditch, Greater London - August 3, 2017

Learn to work with a wide range of people with similar skills like mine. It has an agile culture. They encourage people to speak and share their thoughts and concepts for project ideas.

I enjoy the stand up we do every morning to discuss what we've learnt the previous day

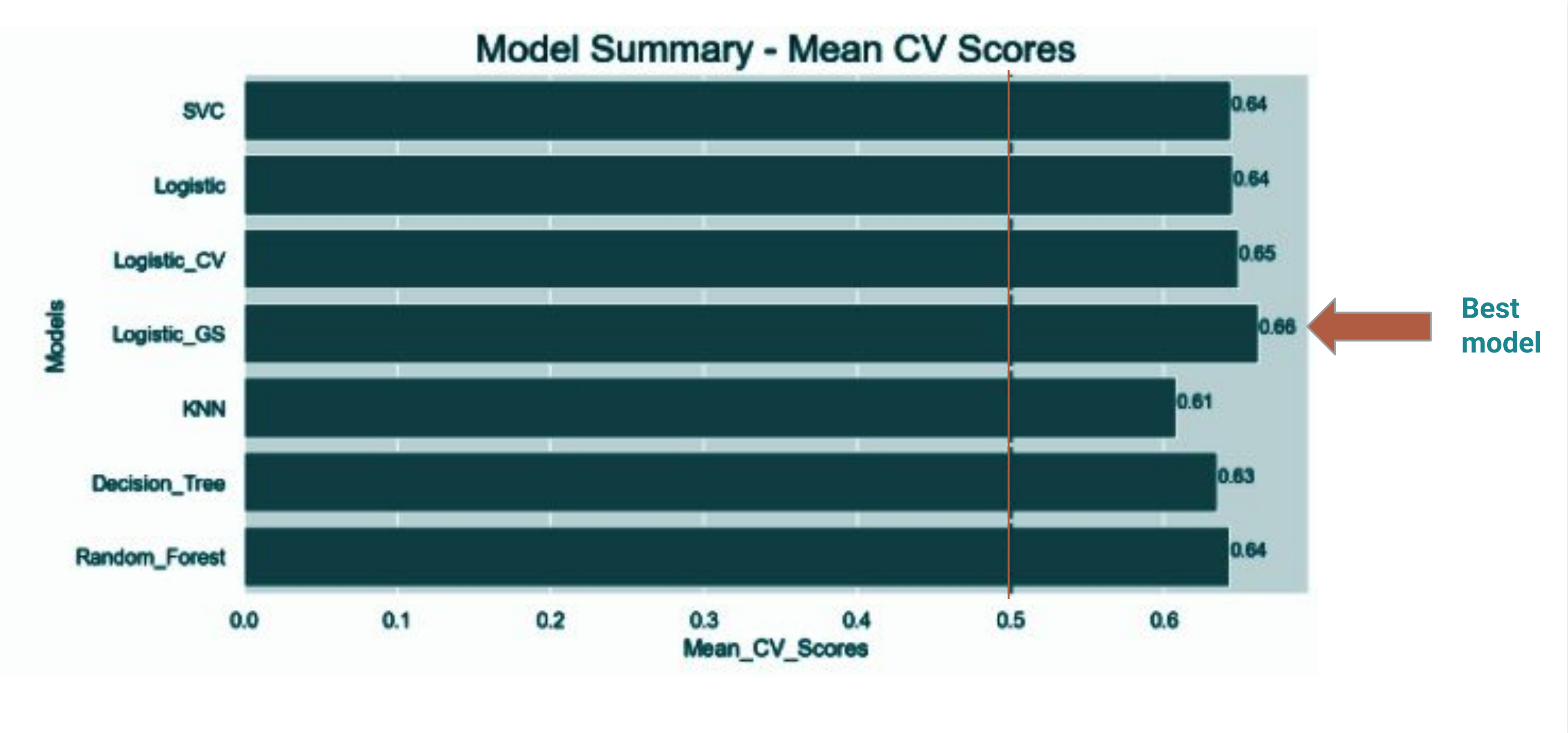
✓ **Pros**

Great and friend environment

× **Cons**

none

Summary of Model Performance



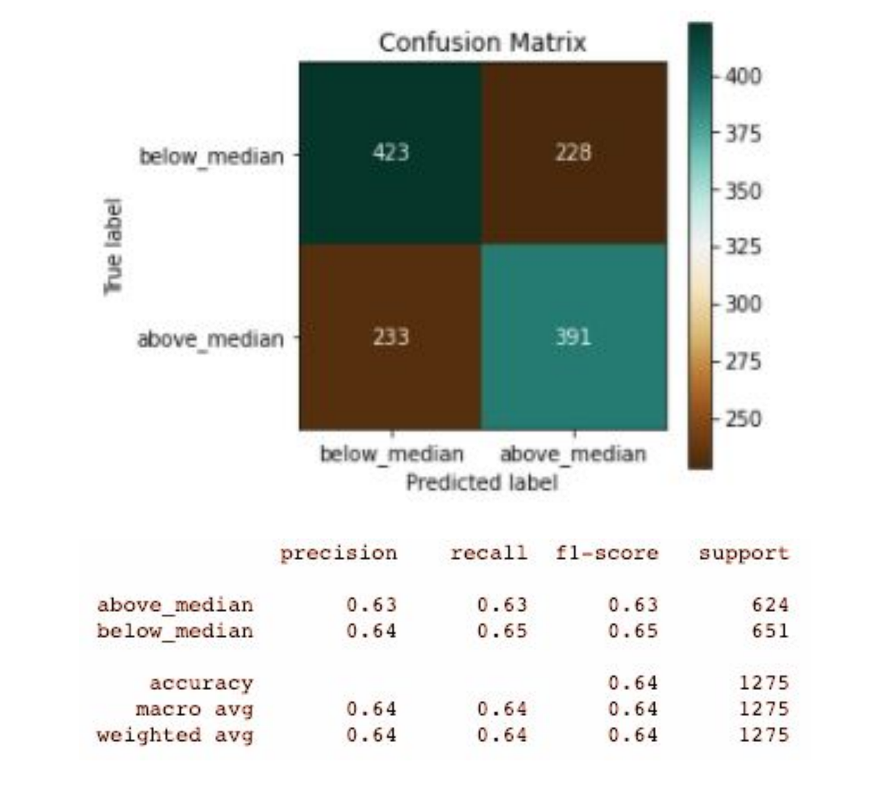
Final Model - Logistic Regression

Model CV Score = 0.66

Confusion Matrix and Classification Report

A precision score of 0.63 for above median and 0.64 for below median shows the model was slightly better at not falsely predicting a below median gender pay gap results as above median than the reverse.

Similarly the recall score of 0.63 for above median and 0.65 for below median shows that the classifier was better at correctly identifying all below median results than above median.



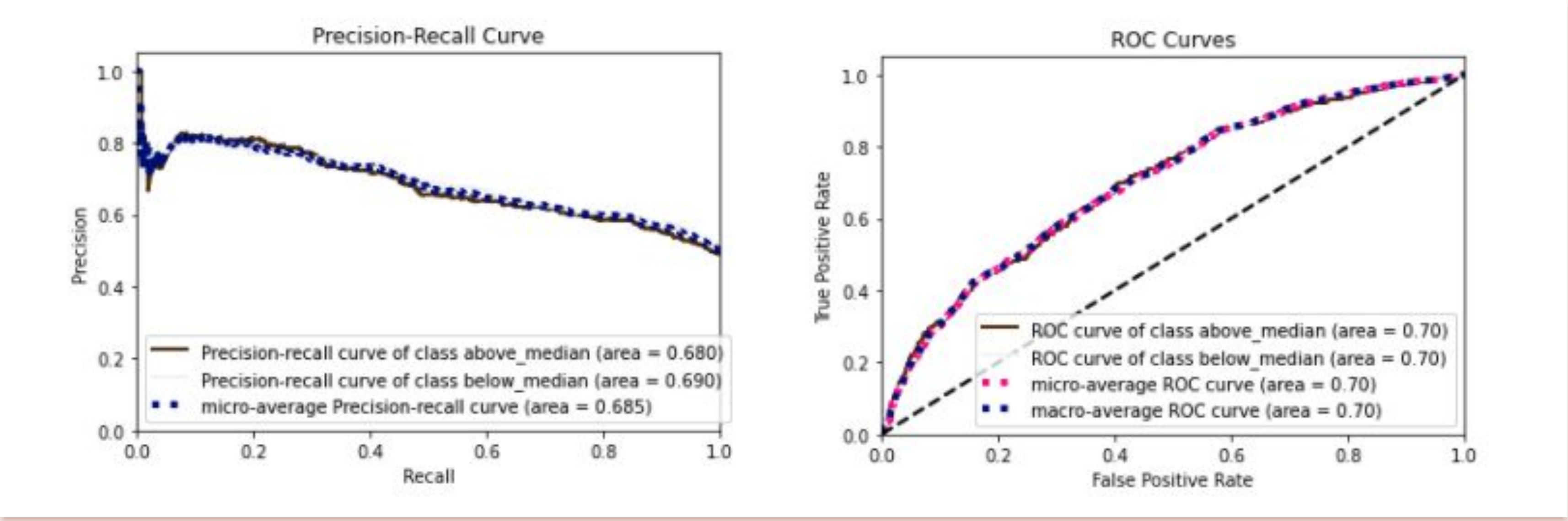
```
              precision    recall  f1-score   support

above_median       0.63      0.63      0.63       624
below_median       0.64      0.65      0.65       651

    accuracy                           0.64      1275
   macro avg       0.64      0.64      0.64      1275
weighted avg       0.64      0.64      0.64      1275
```

Precision Recall and ROC Curves

Looking at the precision recall curve and area under the ROC curve shows that the model skill is better than bassline, but not perfect at correctly predicting the results.



Coefficients Summary

Each category of coefficients was reviewed and evaluated individually to understand which factors most impact gender pay gap. The results were filtered on variables with a coefficient greater than 0.1 or less than -0.1.

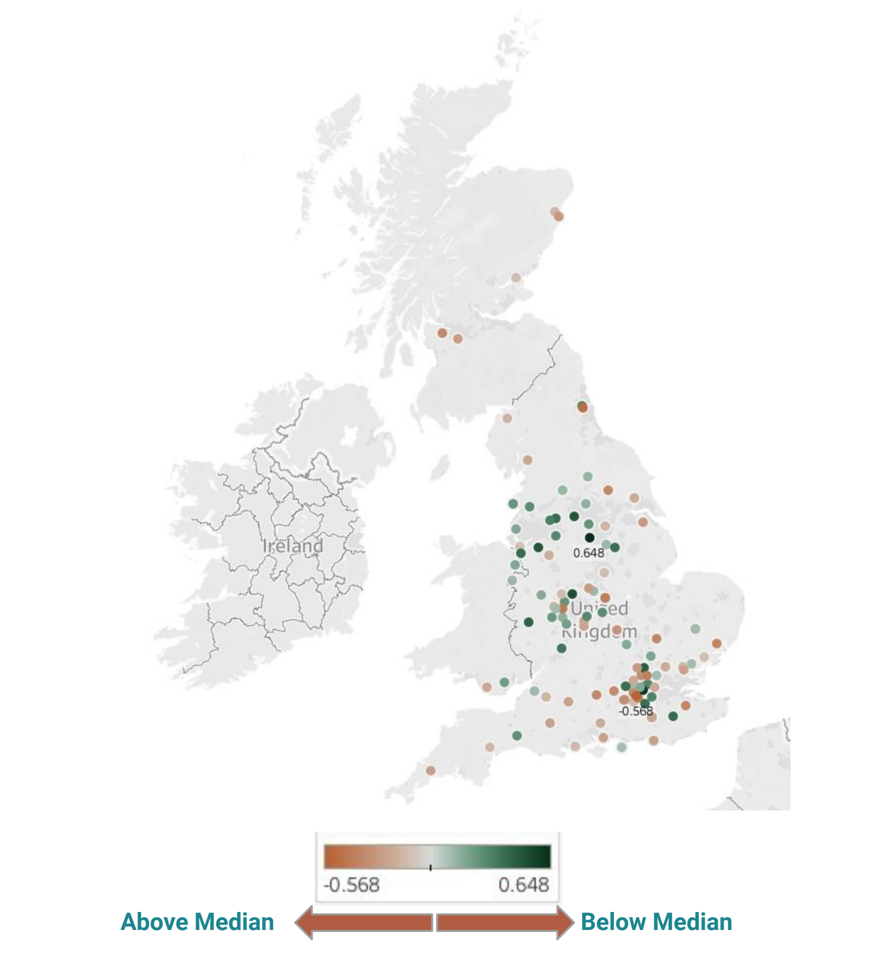
The categories reviewed in detail include:

- ➔ Location
- ➔ Company Size
- ➔ Company Sector
- ➔ Company Assigned Responsible Person
- ➔ Proportion of Female Officers
- ➔ Reviews

Location

The Midlands had many towns with high positive coefficients, this means that the model predicts below median gender pay gap scores for towns in this area - in other words, there is a lower rate of inequality in the Midlands.

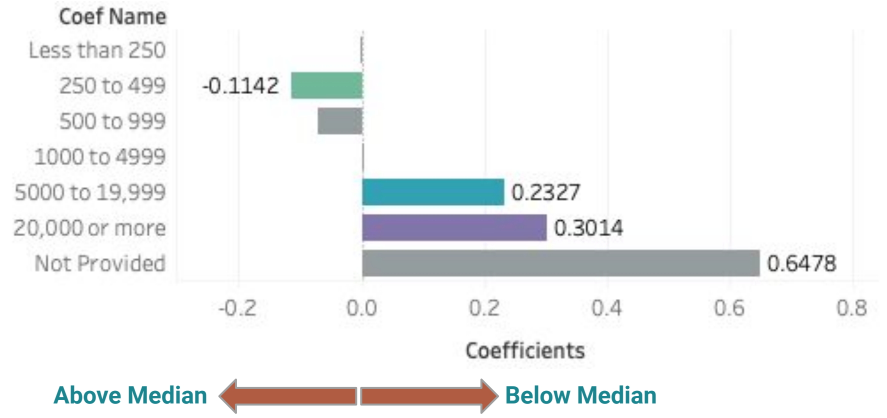
Interestingly, London had a coefficient of -0.2 which means that the model predicts above median scores for companies in London, and the gender pay gap scores in london show a higher rate of inequality. This is generally the case in the areas surrounding London as well, as can be seen on the right.



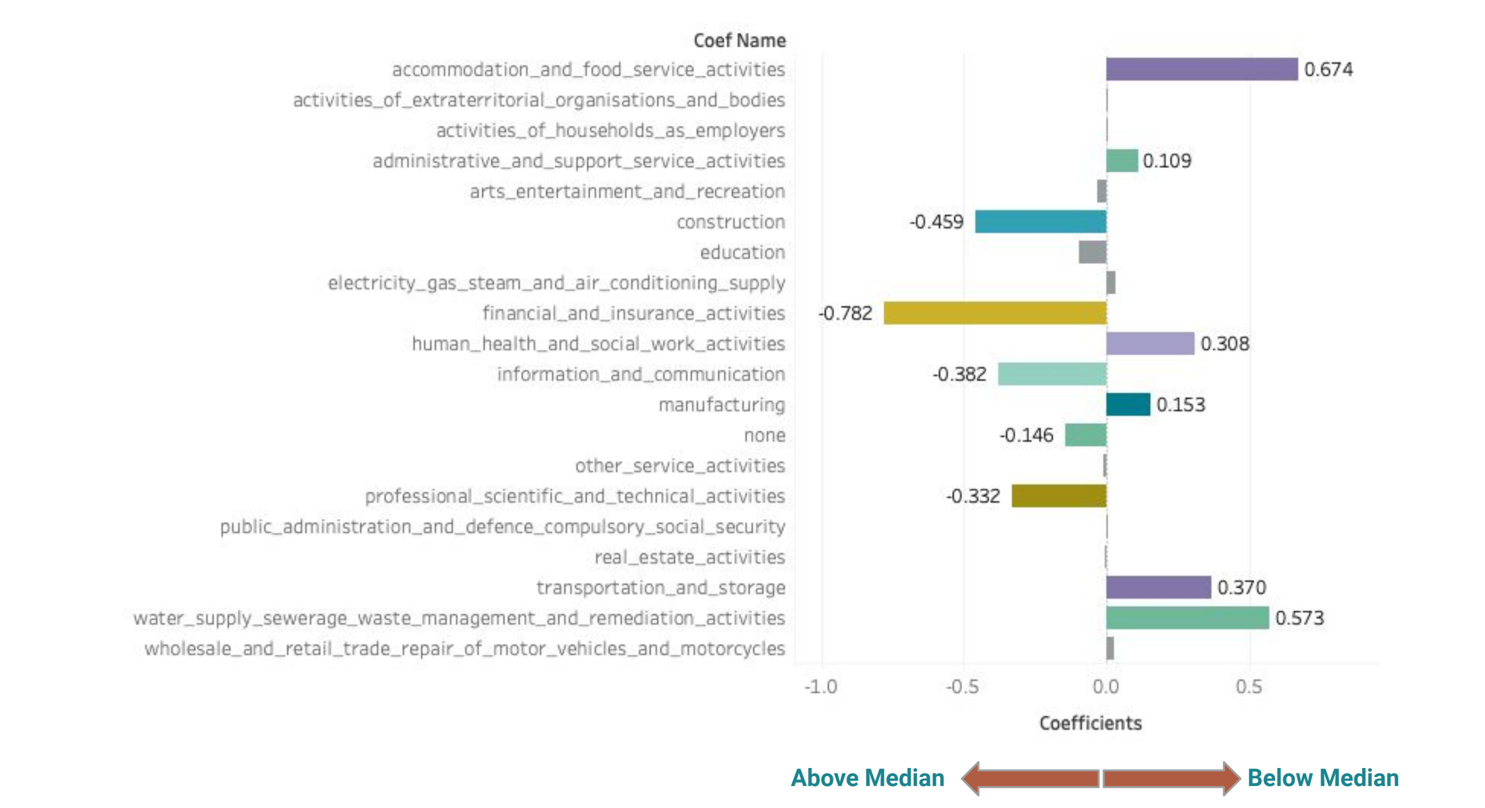
Company Size

Smaller companies were associated with above median pay and larger companies with below median. This would suggest that larger companies are better at providing more equal pay to men and women.

Eight companies did not provide this information and are listed as ‘Not Provided’. However, the ‘Not Provided’ companies also have a high positive coefficient which implies that those who didn’t include the company size tended to have below median gender pay gap stats.



Company Sector



Company Sector

- The 2018 educational enrollment rates (all levels) split by gender were reviewed, for each sector in order to determine any links This information was obtained from: **stats.oecd.org/#**

- Some of the sectors could not be isolated from the educational database, therefore no conclusion was reached as to whether or not they impact gender pay gap.

- However it does demonstrate that educational enrollment rates play a part in gender pay gap - since those with higher male enrollments tended to have above median scores meaning that the model predicted these sectors to have gender pay gap scores above median, meaning that males are paid more the females.

Company Sector

Women enrolled in sectors associated with below median scores:

NA	Accommodation and food service activities
NA	Accommodation and food service activities
75%	Human health and social work activities
NA	Manufacturing
25%	Transport and storage and water supply
NA	Sewerage waste management and remediation activities

Men enrolled in sectors associated with above median scores:

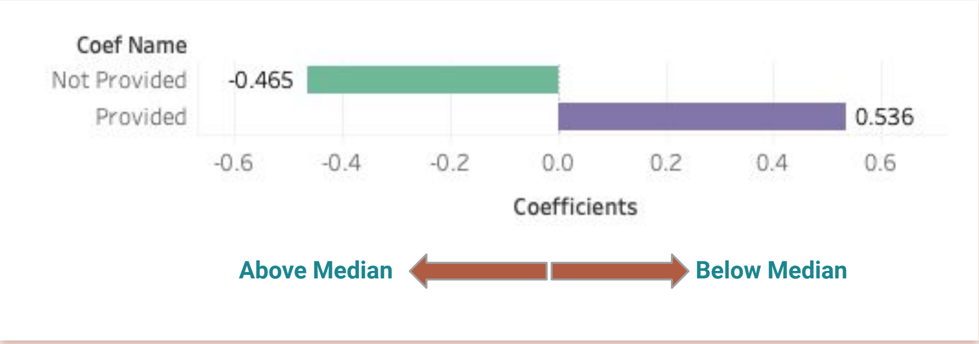
68%	Construction
NA	Financial and insurance activities
83%	Information and communication
45%	Professional, scientific and technical activities

technical activities

Company Assigned Responsible Person

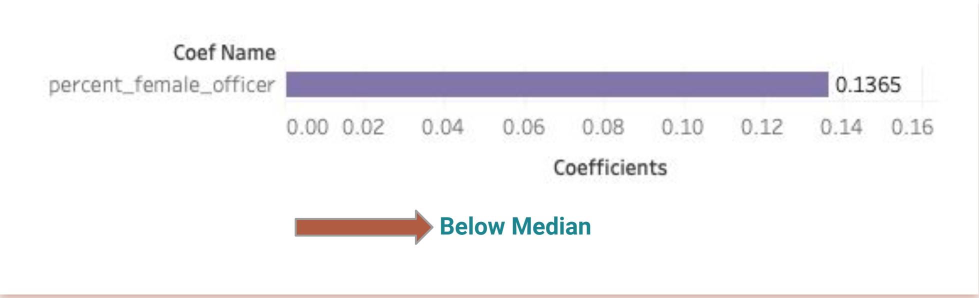
Companies that assigned a responsible person to their gender pay gap report were associated with lower gender pay gap (below median results) than those who did not list a responsible person.

This is unsurprising if you assume that those who did not provide a responsible person do not have a dedicated person to take ownership of gender pay gap within their organisation.



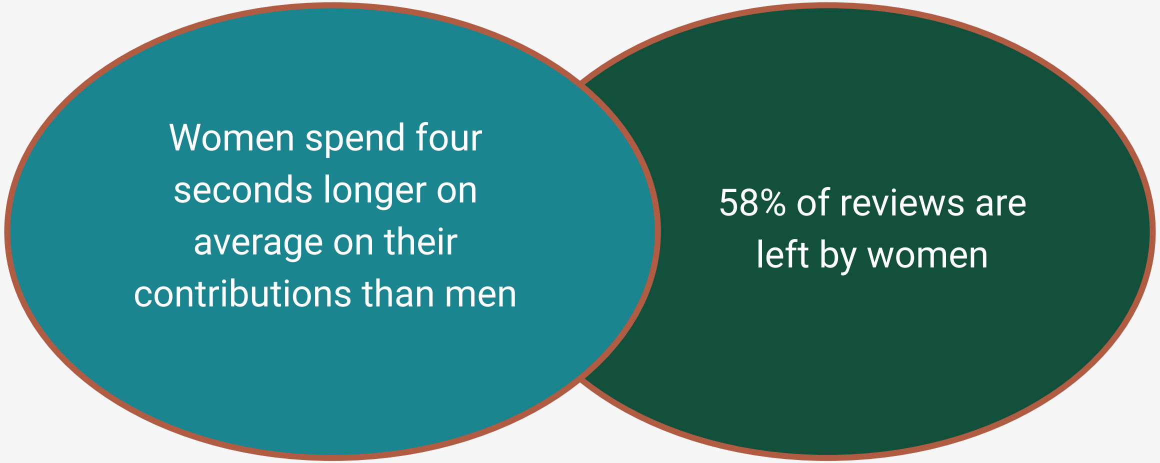
Proportion of Female Officers

A higher proportion of female officers was associated with below median gender pay gap. This also aligns with the heat map shown above, that suggested the proportion of female officers was correlated to the mean gender pay gap.



Reviews Associated with Below Median Pay Gap

According to rating and review platform Feefo



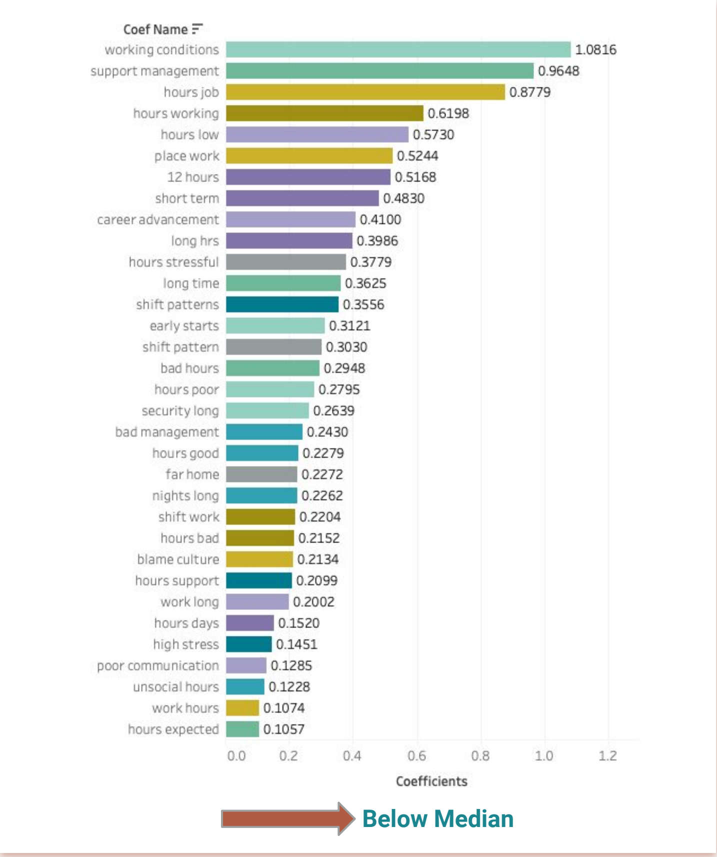
Women spend four seconds longer on average on their contributions than men

58% of reviews are left by women

Reviews Associated with Below Median Pay Gap

The ‘review cons’ themes that were most associated with below median gender pay gap were:

- ➔ Negative Culture, e.g. ‘blame culture’, ‘high stress’, ‘poor communication’
- ➔ Long hours, e.g. ‘nights long’, ‘early starts’, ‘hours stressful’
- ➔ Poor Management, e.g. ‘working conditions’, ‘bad management’

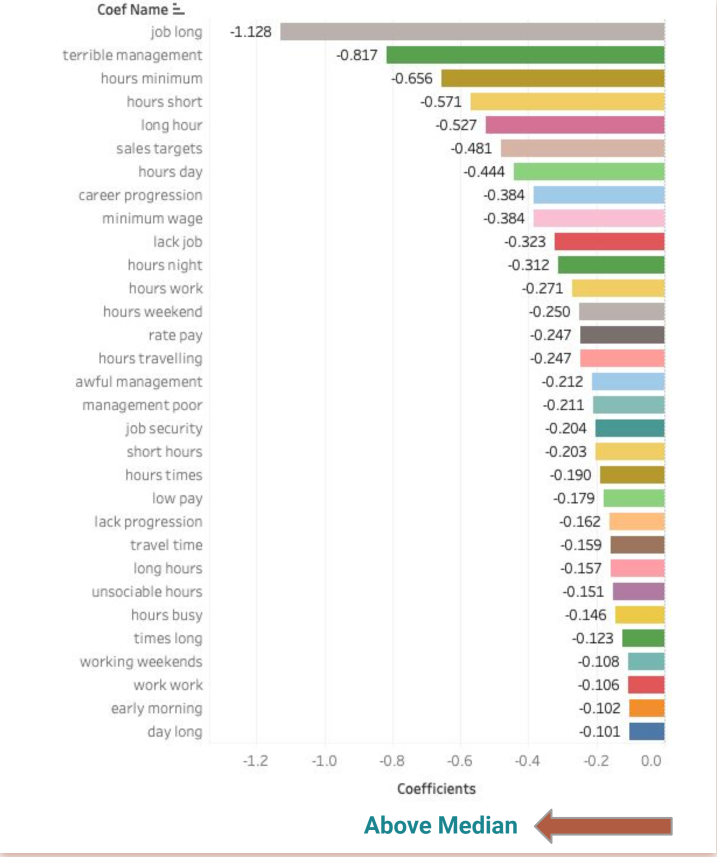


Reviews Associated with Above Median Pay Gap

The 'review cons' themes that were most associated with above median gender pay gap were:

- ➔ Career Progression, e.g. 'Career Progression', 'lack progression',
- ➔ Poor Pay, e.g. 'Minimum Wage', 'Low Pay'
- ➔ Long Hours, e.g. 'Job long', 'times long', 'working weekends'
- ➔ Poor Management, e.g. 'Terrible Management', 'awful management'

Whilst it is not known whether the reviews were written by a male or female, it is interesting that a lack of career progression and poor pay are themes associated with above median gender pay gap.



Conclusion and Additional Resources

- ➔ Sector, Location and Review Cons were the most influential predictors of whether the mean pay gap would be above or below median.
- ➔ Construction and Information and Communication were sectors that the model predicted would be associated with above median pay gap, meaning men typically paid more than women. The majority of individuals enrolling in courses related to these fields were male.
- ➔ The Midlands typically had a lower pay gap compared to the South of England.
- ➔ Reviews indicating a lack of career progression or low salary are linked with an above median gender pay gap.
- ➔ Size, Proportion of Female Officers and whether or not the company had an assigned responsible person were also important features.

Three useful articles to understand more about gender pay gap:

1) ***ig.ft.com/gender-pay-gap-UK/***
2) ***mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace***
3) ***indeed.com/lead/women-in-tech-report***

Thank You!