A close-up of a certificate

Description automatically generated with medium confidence

A picture containing text, screenshot, font, number

Description automatically generated

A picture containing text, screenshot, font, black and white

Description automatically generated

# GitHub URL

https://github.com/[MareseF/UCDPA\_Marese-Feeney\_2023 (github.com)](https://github.com/MareseF/UCDPA_Marese-Feeney_2023)

# Abstract

This project has been completed as a part of a project submission for UCD Specialist Certificate in Data Analytics Essentials. The objective of the project is to review and compare the gender pay gap of males vs females based on a Glassdoor real-world dataset obtained from Kaggle.com. The purpose is to gain insights into the gender pay variation of males vs females based on their gender, age, education, performance, operational area and their job titles.

As detailed, the project uses a Glassdoor dataset obtained from Kaggle.com which was downloaded as 2 CSVs and imported into the Jupyter Notebook for review. After analysis of the data, a number of insights were gained, and these are outlined in the Visualisation, Insights and Results section. The implementation uses iterators, merging of the datasets, dropping of duplicates, regex, reusable code, Numpy, dictionary, lists, machine learning and graphic visualizations to carry out the interrogation of the data.

The dataset had more males than females. Average pay was $100,047, with 287 men and 202 women earning above the mean. The average pay for men ($104,918) was higher than for women ($96,416), but other factors should be considered due to uneven gender distribution. More females went to college (26%) compared to males (22%), while more males had a PhD (24.8%) than females (22.6%). The common job titles were Associate and Marketing, while Manager and Warehouse were less common. Gender, age, education, and performance impacted pay for Marketing Associates and Managers. Graphs showed more men in higher salary ranges and more women in lower ranges. The linear regression model fit perfectly (accuracy 1.0), while the random forest classifier made no correct predictions (accuracy 0.0). The XGBoost model had a reasonable RMSE of approximately 847, suggesting decent performance. Further analysis and cross-validation are recommended.

# Introduction

Gender pay gap refers to the difference in average earnings between men and women in the workforce. It is an ongoing issue that affects many countries and industries. Ireland has only in recent years implemented reporting requirements through the Gender Pay Gap information Bill 2021.

Analyzing gender pay data can shed light on the magnitude and drivers of the gender pay gap and help identify potential solutions. As it is only a recent requirement in Ireland for employers with >250 employees to publish information on the gender pay gap, it meant that local data was not available to interrogate for this project. For this reason, I used a Glassdoor dataset available on Kaggle.

As a Diversity & Inclusion Champion, it is important to me that we ensure all employees are paid fairly regardless of gender. Analyzing gender pay data can help identify any disparities that may exist and provide a basis for addressing them.

# Dataset

As detailed above, the Glassdoor data set was chosen from Kaggle as there was no available dataset for Ireland. Due to the limitations with the data set, i.e. one data set, I spilt the data into two files and overlapped some of the entries to ensure that I could clearly demonstrate the merging and removal of duplicates.

This project uses a real-world dataset from Glassdoor - The source of the data is [Glassdoor Gender Pay \_ Kaggle](https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap). The data was downloaded as a CSV file and then split into two datasets to allow for demonstration of merging of the data.

In total, once cleaned, there are 1,000 rows of data captured under the following headings:

Table . Data Explained

A picture containing text, screenshot, font, number

Description automatically generated

# Implementation Process

Within this section, details of how the data was collected, cleansed and analysed is illustrated. This is done through the use of Regex, iterators, lists, dictionaries, Numpy and the application of a custom function and machine learning.

## 5.1 Data Collection

### 5.1.1 Importing Libraries

Prior to importing the datasets, the relevant libraries were imported as illustrated in Figure 1 below:

A screenshot of a computer program

Description automatically generated with medium confidence

Figure . Libraries imported

### 5.1.2 Importing the Data

The data was imported from the two CSV files:

* Glassdoor Gender Pay Gap\_1.csv
* Glassdoor Gender Pay Gap\_2.csv

Two dataframes were created for the data sets:

* Glassdoor Gender Pay Gap\_1.csv = “pay\_gap1”
* Glassdoor Gender Pay Gap\_2.csv = “pay\_gap2”

|  |  |
| --- | --- |
| A picture containing text, font, screenshot, line  Description automatically generated  Figure . Importing Glassdoor Gender Pay Gap\_1.csv | A picture containing text, font, screenshot, line  Description automatically generated  Figure . Importing Glassdoor Gender Pay Gap\_2.csv |

In the first instance, Glassdoor Gender Pay Gap\_1.csv was imported and reviewed before the Glassdoor Gender Pay Gap\_2.csv file was imported.

## 5.2 Analysing data

### 5.2.1 Merging the Data sets

The two datasets outlined in Figures 2 and 3 were merged, and a new data frame name was assigned – “pay\_gap”

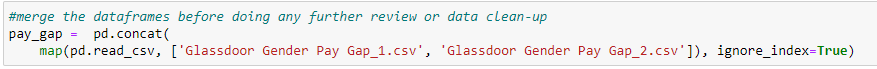


Figure . Merged data set "pay\_gap"

### 5.2.2 Replace missing values or drop duplicates

Prior to merging the two datasets, the data was analysed for null values but none were identified:

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated  Figure . Check for Null values in pay\_gap1 | Figure . Check for Null values in pay\_gap2 |

The data was analysed to identify if there were any duplicates in the data. 4 items were found an these were removed from the dataset as illustrated in Figures 7 below:

A screenshot of a computer

Description automatically generated with medium confidence

Figure . Dropping Duplicates

To assist with analysis of the data, a new column “TotalPay” was created by combining the columns “BasePay” and “Bonus”.

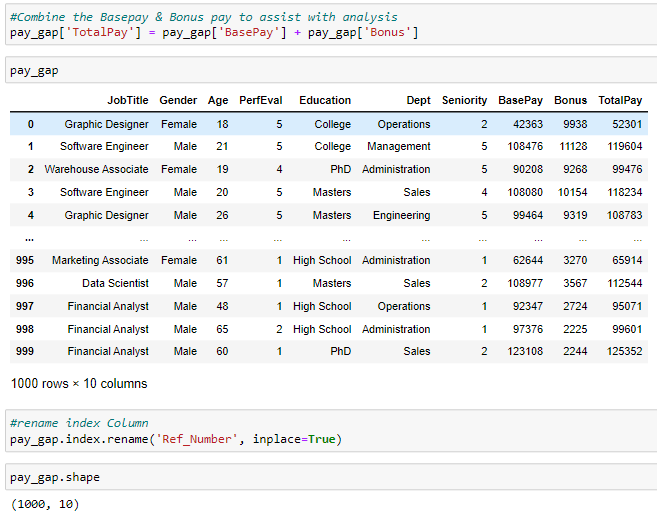


Figure . Addition of TotalPay Column

### 5.2.4 Use Regex to extract a pattern in data

Regular expression was used to extract patterns in the data. The process used is outlined in Figure 9 below whereby the unique job titles were identified and used to indicate the number of individuals with that title.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure . Use Regex to extract a pattern in data

### 5.2.6 Use of iterators

The iterator groupby() function was used to group 'Gender', 'Age', 'Education', 'Seniority'; 'Gender', 'JobTitle', 'Age', 'Education', 'TotalPay' and 'Gender', 'JobTitle', 'Age', 'Education', 'PerfEval' and calculate the average pay for each group. An example of this is illustrated in Figures 10 below.

A screenshot of a computer

Description automatically generated with medium confidence

Figure . Use of iterator to calculate the average pay for each group

## 5.3 Python

### 5.3.1 Define a custom function to create reusable code

A calculation was carried out to identify the average pay by gender in the context of the “TotalPay”, “Bonus” and “BasePay”:

A screenshot of a computer

Description automatically generated with medium confidence

Figure . Use of custom code to review average pay by gender

### 5.3.2 NumPy

Numpy function was used to review the Mean, Median and Standard Deviation of Pay.

A screenshot of a computer

Description automatically generated with medium confidence

Figure . Mean, Median and Standard Deviation of Pay

Further analysis was carried out to identify the number of Males and Females getting above or below the mean and median pay. ¶

A screenshot of a computer program

Description automatically generated with low confidence

Figure . Identification of the number of Males and Females getting above or below the mean and median pay¶

### 5.3.3 Use of Dictionary or List

List usage was demonstrated as part of Regex function, so use of dictionaries is detailed here. The dataframe “pay\_gap” was converted to a list of dictionaries in order to identify the average pay for male vs female respondents.

A screenshot of a computer code

Description automatically generated with low confidence

Figure . Use of dictionaries to identify average pay for males vs females

## 5.4 Machine Learning

As detailed in the course material, Machine Learning, is often described as predicting the future – whilst not 100% accurate it is certainly useful for making predictions. For the dataset used in this project Linear Regression, Random Forest and Boosting were used.

## 5.4.1 Hyper Parameter Tuning

### 5.4.1.1 Linear Regression

Linear regression was used to predict the total pay. The LinearRegression() class was used to train a linear regression model on the training set, and then the predict() method was used to obtain the predicted values of the target variable for the test set. Finally, the r2\_score() function was used to calculate the R^2 score of the model on the test set. Model accuracy: 1.0

A screenshot of a computer program

Description automatically generated with low confidence

Figure . Predicting Total Pay using Linear Regression

The insight from this is that the trained linear regression model achieves a perfect accuracy score of 1.0, indicating that it fits the data perfectly and can accurately predict the TotalPay values based on the provided features. This suggests a strong linear relationship between the selected features and the target variable in the dataset.

### 5.4.1.2 Random Forest

Radom forest classifier was used to predict gender pay. The data was split into training and test sets using the train\_test\_split function. The test\_size parameter specified that 20% of the data will be used for testing, while the remaining 80% will be used for training. The random\_state parameter ensured reproducibility of the split.

A random forest classifier was used and the model was trained on the training set using the fit method, where X\_train represents the input features and y\_train represents the corresponding target labels.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure . Predicting Gender pay using Random Forest

The insight from this is that the trained random forest classifier has an accuracy of 0.0 on the test set, indicating that it is not making any correct predictions. This suggests that the model might be underfitting or there could be issues with the data or the model's parameters that need to be addressed. Further analysis and investigation are required to understand the reasons behind this low accuracy and to improve the model's performance.

## 5.4.2 Boosting

An XGBoost regression model was used to predict the “TotalPay” variable. The model's performance is evaluated using the root mean squared error (RMSE) metric, which measures the average deviation between predicted and actual values. The obtained RMSE value of 847.5122 indicates the model's accuracy in predicting 'TotalPay', with lower values indicating better performance.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure . Predict the 'TotalPay' variable using XGBoost model

The insight from this is that the trained XGBoost model achieves an RMSE of approximately 847.5122 when evaluated on the test set. The RMSE represents the average difference between the predicted and actual values, and in this case, it indicates the level of error in the predictions made by the model. The RMSE value being smaller than the mean total pay suggests that the model is performing reasonably well in predicting the total pay values.

To visualize the predicted values against the actual values, a scatter plot was generated. Furthermore, the mean absolute error and the R-squared score were calculated.

|  |  |
| --- | --- |
| A picture containing text, screenshot, line, plot  Description automatically generated  Graph . Scatter Plot Actual Vs. Predicted Values | Figure . Mean Absolute Error |
| Figure . R-squared score |

The above results indicate that the model can make accurate predictions and has a strong fit to the data. However, further analysis of the results is required.

# Results

Highlighted in this section are various charts illustrating the gender pay gap analysis that was carried out including: No. Males Vs Females, Education distribution by Gender, distribution of TotalPay by Gender and Seniority, Job Title by Gender and TotalPay by Age & Gender.

|  |  |
| --- | --- |
| A picture containing screenshot, rectangle, square, design  Description automatically generated  Graph . No. Female and Males | A close-up of a pie chart  Description automatically generated with low confidence  Graph . Highest Education by Gender (%) |

|  |  |
| --- | --- |
| A picture containing text, screenshot, line, parallel  Description automatically generated  Graph . JobTitle by Gender | |
| A picture containing text, circle, colorfulness, compact disk  Description automatically generated  Graph . JobTitles by Gender (%) | A picture containing circle, text, compact disk, colorfulness  Description automatically generated  Graph 6. Seniority by Gender (%) |
| A picture containing text, screenshot, diagram  Description automatically generatedGraph . Age Vs TotalPay by Gender | A picture containing text, screenshot, diagram  Description automatically generatedGraph . Distribution of TotalPay by Gender |

# Insights

* As highlighted in the data, there are 532 Males Vs 468 Females. Section 5.3.1, identifies the mean pay was 100,047. When further analyzed as per Section 5.3.2, it was found that:
  + 287 Men were paid above the mean compared with 202 females
  + 248 men were paid under the mean compared with 266 females

Section 5.3.3 further analyses the data and indicates that the average pay for men is 104,918 Versus 96,416 for women. It is worth noting that there were 46 more men than women captured in the data and for this reason, it is better to look at the Male Vs Female gender pay in the context of more than just gender.

* As per Graph 3, Highest Education by Gender (%) – 26% of the total females went to College versus 22% of all males. In addition, the total number of females to complete a PhD is much lower than the total number of males – 22.6% Vs 24.8%. Education by Gender as illustrated in line 26 of the Jupyter notebook indicates that this variance equates to a difference of 26 men having completed a PhD.
* The most common Job Titles were Associate (302) and Marketing (118) whilst the least common were Manager(90) and Warehouse(90).As per the iterator function outlined in Section 5.2.6, it was determined that the average pay for a Marketing Associate varied on Gender, Age, Performance. For example A 65 year old Female with a Masters had an average pay of 111,933 compared with a 51 year old Male, with a PhD who had an average pay of 66,900. In addition, on review of a the title Manager, it was found that there was again, little variation as illustrated below:
  + Male, Age: 23, Education: PhD, Average Pay: 99,501
  + Female, Age: 24, Education: PhD, Average Pay: 112,658
  + Male, Age: 62, Education: College, Average Pay: 165,250
  + Female, Age: 62, Education: Maste rs, Average Pay: 163,011

Further data and data analysis is required with an equal number of male and female respondents to fully understand the gap and to further interrogate the data.

* Graph . Age Vs TotalPay by Gender clearly illustrates that there are more men than women in the higher salary range and more women than men in the lower totalpay scale. This is more clearly illustrated in Graph 7 Distribution of TotalPay by Gender. Further insights by job can be taken from Graph . JobTitle by Gender which clearly illustrates that there are more men than women in Software engineer roles and the reverse is the case for Marketing Associate Roles. With regards to the marketing roles the top 3 highest paid are as follows based on line 43 of the Jupyter notebook:
  + Female, , Age: 62, Education: PhD, PerfEval: 3, Average Pay: 128695.0
  + Male, Age: 40, Education: PhD, PerfEval: 2, Average Pay: 122350.0
  + Female, Age: 54, Education: PhD, PerfEval: 3, Average Pay: 119967.0

As per Table 1, a PerfEval of 3 is better than one of 2 yet the second highest earner is not the best performer. Also, looking at the top 3 highest earners for Sofware engineers is:

* + Male, Age: 59, Education: College, PerfEval: 1, Average Pay: 163390.0
  + Male, Age: 59, Education: Masters, PerfEval: 4, Average Pay: 162935.0
  + Male, Age: 60, Education: Masters, PerfEval: 4, Average Pay: 162262.0

Whilst there are only 8 females in this profession, the top earning female is 36th out of all of the individuals in this filed. What’s also interesting here is that the highest paid individual, a has a poor performance score compared to others in the top 3.

* Insights regarding the machine learning are captured under Section 5.4 and indicate that the trained linear regression model achieved a perfect accuracy score of 1.0, indicating that it fits the data perfectly and can accurately predict the TotalPay values based on the provided features. In relation to the random forest classifier an accuracy of 0.0 on the test set was reported indicating that it is not making any correct predictions. However, the XGBoost model achieved an RMSE of approximately 847. The RMSE value being smaller than the mean total pay of 100,047 suggests that the model is performing reasonably well in predicting the total pay values. Further cross-validation and analysis would be beneficial to ensure a more comprehensive understanding of each of the model’s performance.

# References

* Kaggle. (n.d.). Glassdoor - Analyze Gender Pay Gap. [Online]. Available at: https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap (Accessed: 12 April 2023).