

PERRY PLATYPUS

# Gender Wage Gap; Evidence from Glassdoor Dataset

S&DS 563

YALE UNIVERSITY

# Table of Contents

<b>Introduction.....</b>	<b>2</b>
› Background & Motivation .....	2
<b>Design &amp; Research Question .....</b>	<b>2</b>
› Research Question .....	2
› Design .....	3
<b>Data.....</b>	<b>4</b>
› Description .....	4
› Data Collection .....	4
› Limitations .....	4
<b>Summary Statistics.....</b>	<b>5</b>
<b>Multivariate Analysis .....</b>	<b>6</b>
› Discriminant Analysis.....	6
› MANOVA.....	11
› Cluster Analysis.....	15
<b>Discussion .....</b>	<b>18</b>
<b>Conclusion &amp; Points for Further Analysis .....</b>	<b>20</b>
<b>References.....</b>	<b>20</b>

# Introduction

## Background & Motivation

Wage earnings and labour market outcomes have been of great interest to researchers for a long time. The biggest contention in education for a prolonged period of time has been the causal effect of education on wage (Card, 1999). An even more exciting area of research has been socioeconomic factors such as gender, ethnicity, and socioeconomic status of parents and its effect on wage. The gender pay gap – the difference between the earnings of men and women – has barely closed in the United States in the past two decades. In 2022, American women typically earned 82 cents for every dollar earned by men.<sup>1</sup> The situation is much dire in developing countries, for example, in India, according to the Global Gender Gap Report 2021, women, on average, were paid 21% (or almost one-fifth) of the income of men.<sup>2</sup> The situation is further complicated by additional layers of socio-cultural expectations and economic constraints. In many such societies, women's educational and professional opportunities are limited, and there is a high prevalence of informal employment, which often goes unregulated and unprotected, exacerbating income disparities.

The wage gap between gender has been generally attributed to occupational differences (Blau & Kahn, 2016), caregiving responsibility, intra-household resource allocation (Braido, Olinto, & Perrone, 2012) etc. However, these factors do not fully account for the persistent wage gap, suggesting that discrimination and societal norms play a significant role as well. For instance, women are often underrepresented in higher-paying industries and positions, and they are more likely to face career interruptions, which can adversely affect long-term earnings. It is imperative to understand how these factors interact with each other as well, hence this analysis aims to contribute to the broader understanding of economic inequalities and provide actionable insights for policymakers to promote gender equity in the workplace.

## Design & Research Question

### Research Question

This study employs a mixed-methods approach, utilizing multiple statistical techniques to address the underlying factors contributing to gender wage disparities. The primary questions of this research are:

---

<sup>1</sup> <https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/#:~:text=The%20gender%20pay%20gap%20%E2%80%93%20the,every%20dollar%20earned%20by%20men.>

<sup>2</sup> <https://www.weforum.org/publications/global-gender-gap-report-2023/in-full/benchmarking-gender-gaps-2023/>

Does gender play a role in differentiating wages when we account for performance, experience, and education? Following from the results I get from above, I further explore interaction effects through the following question. How does gender, education, and its interaction affect wage and performance evaluations? After which I take a more holistic approach to analyse natural groupings in the data and question whether there are identifiable clusters within the workforce based on gender, performance experience, education, and occupational choices, and if so, how are these clusters characterized by gender proportion.

## Design

To answer my research questions, I use the following methods:

### Discriminant Analysis

- › Variables: Performance Evaluation, Seniority, Age, BasePay, and Bonus.

Discriminant analysis will be used to assess whether these continuous variables can predict gender, thereby understanding if wage determinants contribute directly to gender differentiation in earnings.

### MANOVA (Multivariate Analysis of Variance)

- › Continuous Response Variables: Performance Evaluation (PerfEval) and logarithmic transformation of BasePay (logBasePay).
- › Categorical Predictor Variables: Gender and Education.
- › Continuous Predictor Variable: Age.

MANOVA will help determine if there are significant differences in performance evaluations and pay scales across different genders and education levels, accounting for age.

### Cluster Analysis

- › Variables Used: Gender, Performance Evaluation, Age, Education, Seniority, Job Title, and Department.

Cluster analysis will identify naturally occurring groupings within the workforce based on the specified attributes. This analysis will further explore how these clusters vary in terms of gender proportion.

# Data

## Description

The dataset, sourced from Glassdoor, concentrates on income disparities across various job titles segmented by gender. The dataset includes variables that pertain to education, performance, occupational choices, and experience through the following variables:

- › Categorical Variables: Job Title, Gender, Education, and Department
- › Continuous Variables: Age, Performance Evaluation, Seniority, Base Pay, Bonus

The dataset can be downloaded from Kaggle<sup>3</sup> or Glassdoor website<sup>4</sup> from an article on "How to Analyze Your Gender Pay Gap: An Employer's Guide." It is a subset of the survey of the 2019 Glassdoor Data on the Gender Pay Gap and Salary Transparency and does not include any identifiable information.

## Data Collection

The data was collected through self-reported surveys administered to employees who voluntarily shared their compensation details on Glassdoor. Each entry was anonymized to protect the identity of the respondents.

## Limitations

Before delving into the specifics, it is important to acknowledge several limitations inherent in the dataset used for this study:

- › There could be reporting bias, where individuals may underreport or overreport their income or other demographic details
- › There could be measurement error in reporting
- › Inferences drawn from this may underrepresent gender bias in formal wage as companies that generally publish this data will inherently do so if it is not a major concern to avoid backlash
- › The available dataset is only a subset and we have no sampling methodology, this is singularly the biggest limitation of the inferences from this dataset

---

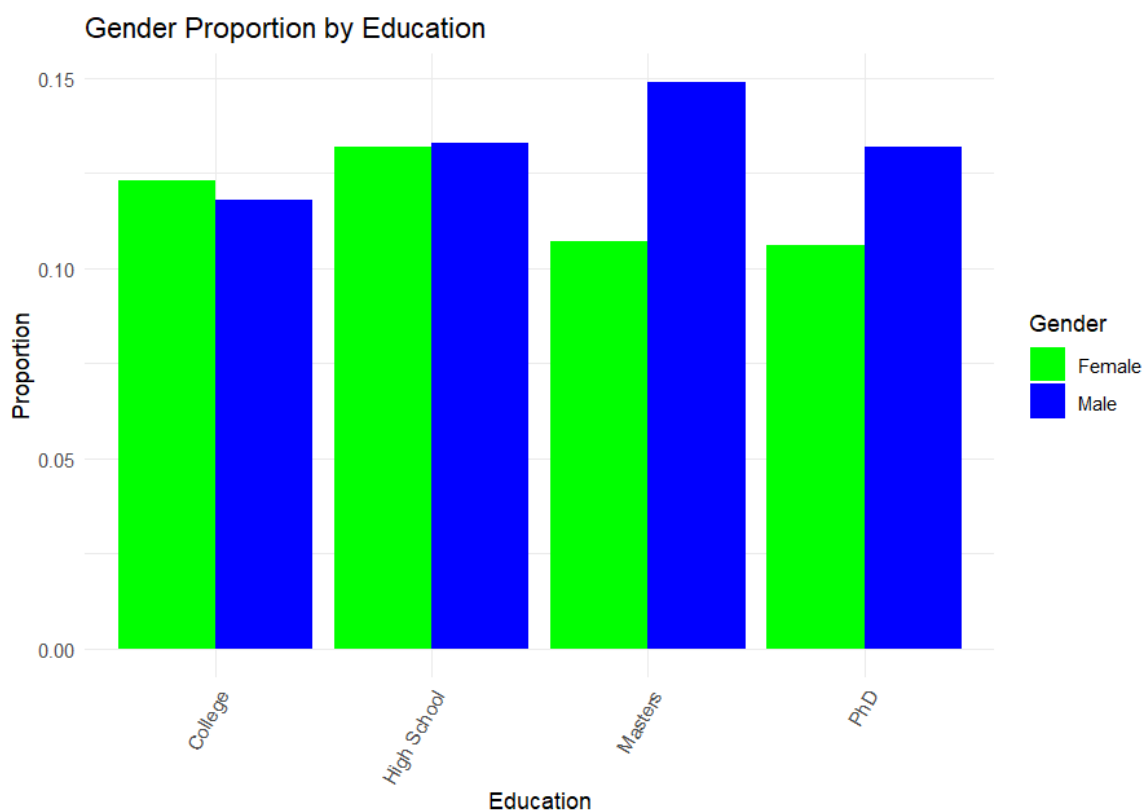
<sup>3</sup> <https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap>

<sup>4</sup> <https://www.glassdoor.com/research/how-to-analyze-gender-pay-gap-employers-guide>

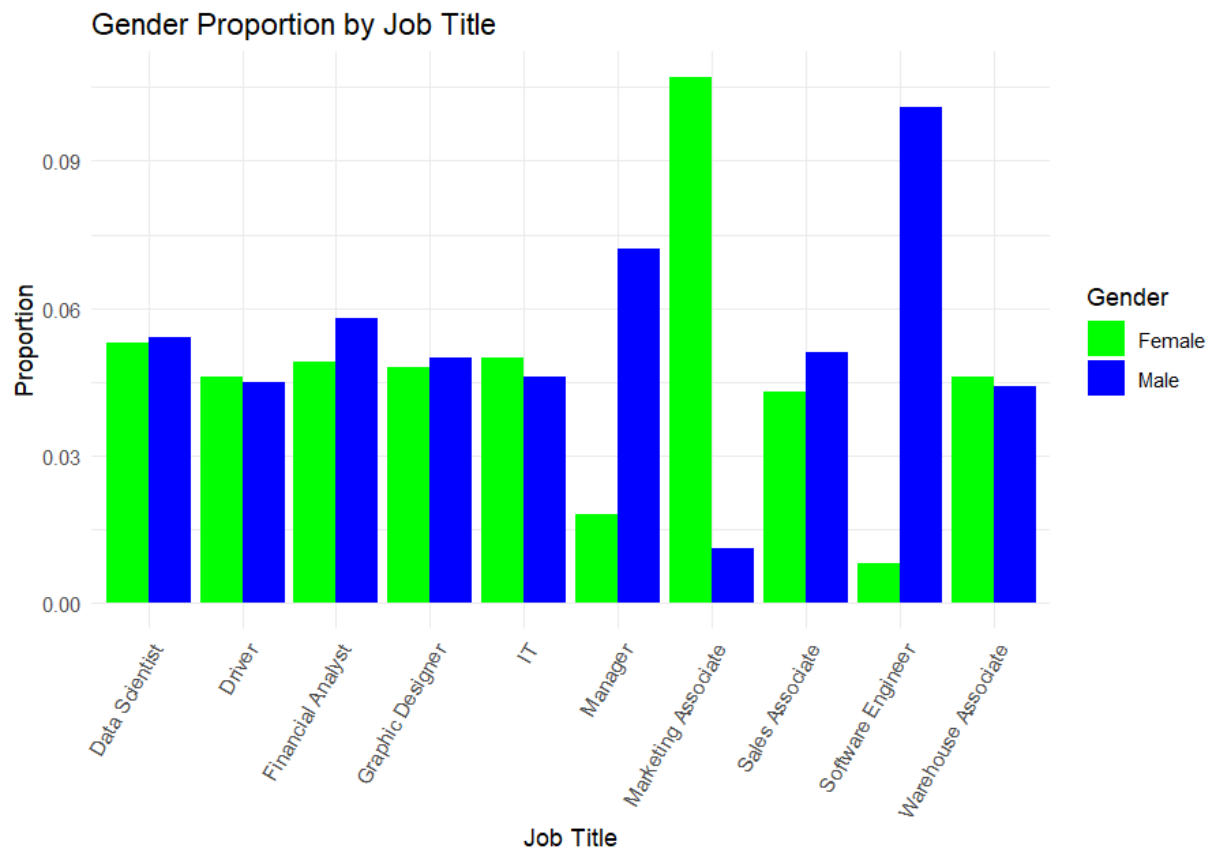
# Summary Statistics

We have a reasonably balanced dataset from the perspective of gender with 468 Females and 532 Male observations. The summary statistics for numeric data is as follows:

Variable	Mean	SD	Min	Max	Median
Age	41.393	14.29486	18	65	41
Performance Evaluation	3.037	1.423959	1	5	3
Seniority	2.971	1.395029	1	5	3
Base Pay (\$)	94,472.65	25,337.49	34,208	179,726	93,327.5
Bonus (\$)	6,467.161	2,004.377	1,703	11,293	6,507



There is a comparable proportion of males compared to females at both the College and High School education levels. At the higher education level is where we see some disparity, the males comprise a higher proportion at the Masters level and the PhD level indicating some hindrance for females to get similar education.



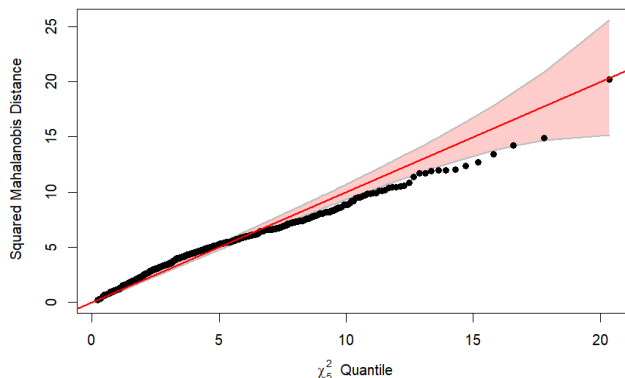
We see that some Job titles have comparable proportions of males and females and any minor difference could be statistically insignificant to draw conclusions. The two jobs that draw our attention are marketing and software which show a trend which can then also be noticed for other columns; we see men dominate in roles that are technical and STEM oriented whereas women are predominantly in client-facing roles. Particularly, we also see a very low proportion of female managers which may indicate some difficulties in women accessing leadership roles. This gives us some insight onto how we can start our analysis. From the education graph and summary statistics, given the balanced dataset, it would be interesting to see if gender plays a role in differentiating employees' wages when accounting for structural parameters that affect wage like performance, experience, and education.

## Multivariate Analysis

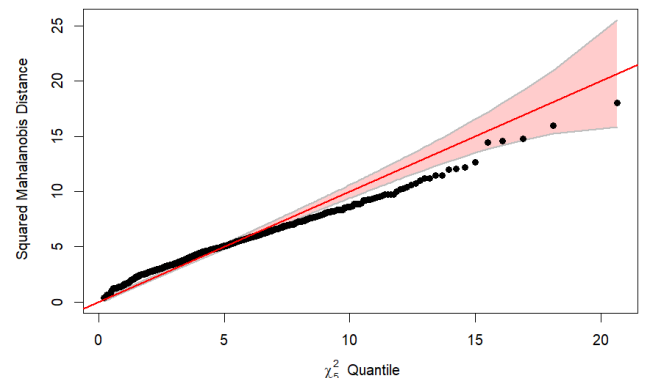
### Discriminant Analysis

First, we evaluate implicit assumptions to discriminant analysis – multivariate normality within each group and similarity of covariances matrices. As we see from chi-square quantile plots, we will need some transformations to adhere to multivariate normality. NOTE: This analysis will only accommodate the continuous variables.

Square Q-Q Plot of ggdata[ggdata\$Gender == "Female", c("Age", "PerfEval", "Sei  
Chi-Square Q-Q Plot of "BasePay", "Bonus")]

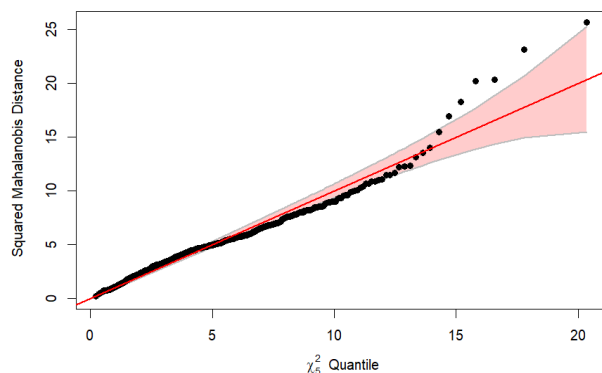


-Square Q-Q Plot of ggdata[ggdata\$Gender == "Male", c("Age", "PerfEval", "Sei  
Chi-Square Q-Q Plot of "BasePay", "Bonus")]

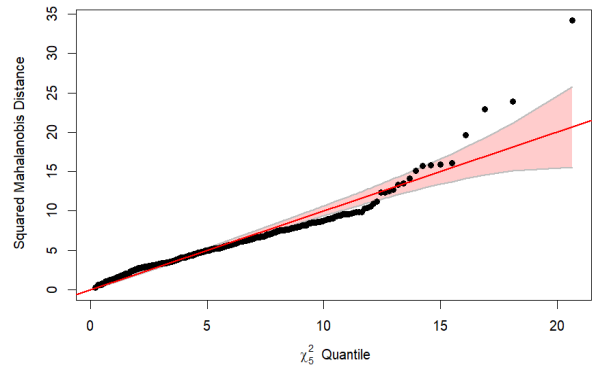


We use log transformations on BasePay and Bonus pay and reevaluate our chi-square quantile plots. Although we still have outliers the mid-section adheres to the line more and is visibly within the border whereas previously it was on/slightly out.

Square Q-Q Plot of ggdata[ggdata\$Gender == "Female", c("Age", "PerfEval", "Sei  
Chi-Square Q-Q Plot of "logBasePay", "logBonus")]



-Square Q-Q Plot of ggdata[ggdata\$Gender == "Male", c("Age", "PerfEval", "Sei  
Chi-Square Q-Q Plot of "logBasePay", "logBonus")]

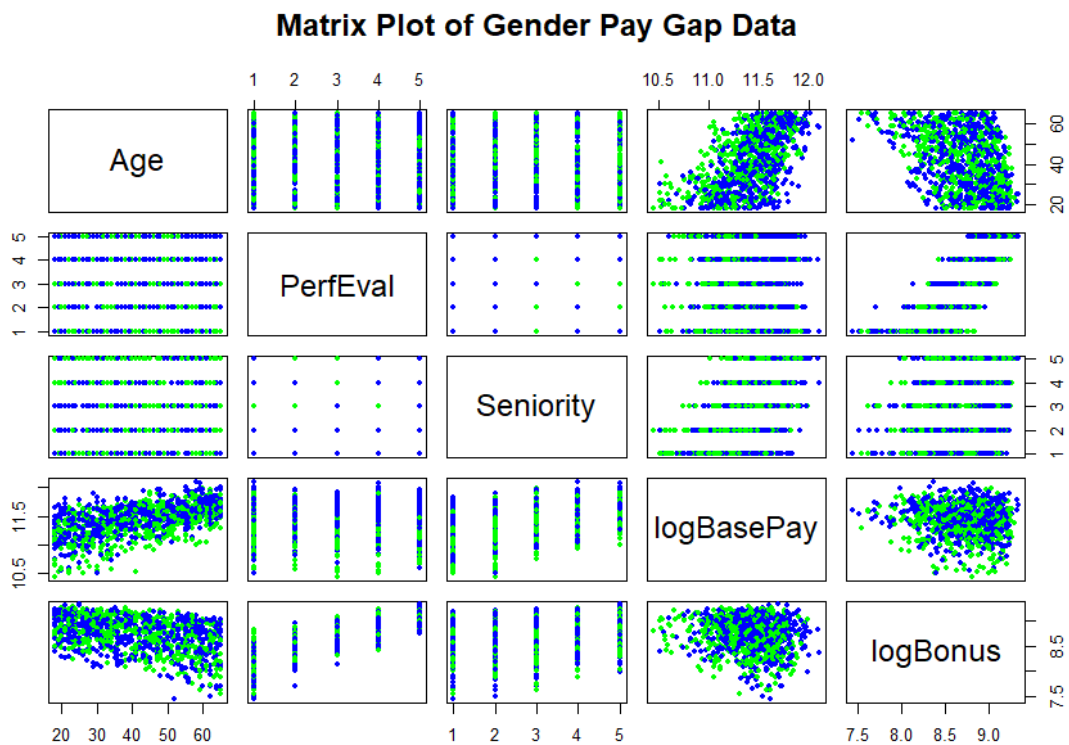


To check for similarity in covariance matrices, I look at covariance matrix and box m statistic.

#### Summary for Box's M-test of Equality of Covariance Matrices

Chi-sq: 27.89831  
df: 15  
p-value: 0.02221

P-Value of 0.02 indicates we reject the null that covariance matrices are similar and will need to explore quadratic discriminant analysis. To take a look at the data we are working with and relationship between variables, we also explore covariance matrix plot.



Trying both LDA and QDA, for linear discriminant analysis, we look at coefficients below:  
 NOTE: Coefficients give direction of maximum discrimination and not magnitude, we see Basepay and Bonus are significant discriminators for LD1 and we inf act just have one discriminating function.

```
-----
> coefficients
      LD1
Age      -0.09920111
PerfEval  0.98441321
Seniority -0.37900707
logBasePay 4.94982190
logBonus  -4.27836932
-----
```

If we'd want to interpret these coefficients, we'd note that, the negative coefficient means that aging has a negative impact on LD1. This indicates that an LD1 score decline is related to age, i.e there is less disparity in gender wage gap amongst people that are older or have more experience. There is a positive impact on LD1 due to performance evaluation but it is not very large. We will explore this further later. The most interesting results we see are that logBasePay is has a significant impact of direction of discrimination whereas logBonus has the opposite in similar magnitude. This would indicate we see wage differences predominantly through logBasePay and not logBonus.

However, as we suspected, the accuracy of prediction of LDA is not too strong about 66%, a little better than guessing. We also calculate accuracy of prediction for quadratic discriminant analysis is slightly better but very similar. To explore if the multivariate means are statistically different, I perform a Wilk's lambda test, which indicates the independent variable (Gender) has a significant multivariate effect on the combined response variables at

0 significance level. To probe into which variables drive this response, I permed ANOVA on each response variable my results were as follows:

- › For age, p-value is not statistically significant, indicating that gender does not significantly affect age in the data, this is good as it reinforces our belief now that any disparity in income is not driven due to less experience. Even when there is data for working professionals at all ages (not very representative of many companies), we still see income differences
- › The p-value is significant, indicating that gender has a statistically significant impact on performance evaluation scores, we will look into this in MANOVA
- › The p-value is not statistically significant, indicating that gender does not affect seniority, affirming our hypothesis on experience levels in the dataset
- › The p-value is highly significant ( $\leq 0.001$ ), indicating that gender has a strong impact on logBasePay, indicating there is in fact a difference due to gender on wage. However, it is not statistically significant for logBonus, now we know which component to look into!

The significant results for Performance Evaluation and logBasePay suggest that these factors may be influenced by gender biases or systematic differences in how genders are treated or paid within the organization. The non-significant results for Age, Seniority, and logBonus suggest that these factors do not differ substantially between genders, indicating either a balanced treatment across these dimensions or that these variables.

I calculate eigenvalues from the model and compute Proportion of Trace explained by each eigenvalue to find that only have one significantly discriminating function and proportion of trace is 1 i.e. my one discriminating function captures all variation between the groups.

I use raw results and the predict function to calculate accuracy of the model, I get that my quadratic model classifies observations correctly 67% of the time. Although this is not ideal, it's indicative of some discriminating ability between the groups. I validate the classification results by cross validation which gives me approximately the same answer 66%.

To determine which of my original variables could explain the discriminating ability better, I look at my standardized discriminant coefficients.

-----  
Female

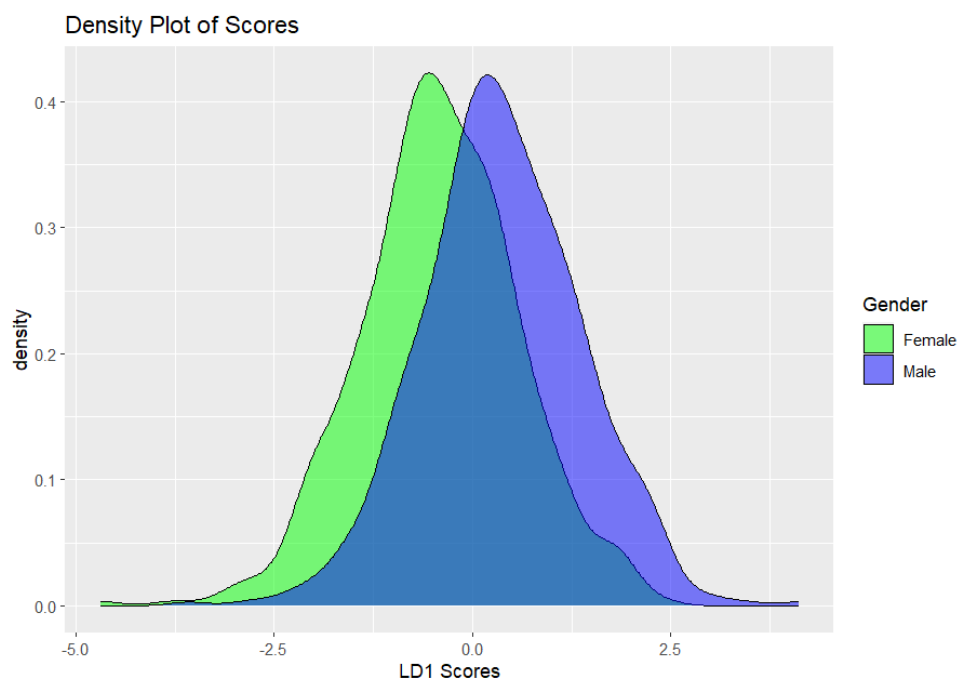
	1	2	3	4	5
Age	1.04	-0.12	-0.05	-0.98	1.19
PerfEval	0.00	-1.00	-0.05	-0.02	-2.26
Seniority	0.00	0.00	-1.01	-0.97	-0.44
logBasePay	0.00	0.00	0.00	1.62	-0.19
logBonus	0.00	0.00	0.00	0.00	2.86

Male

	1	2	3	4	5
Age	0.97	-0.01	-0.01	0.92	1.06
PerfEval	0.00	-1.01	0.01	-0.06	-2.34
Seniority	0.00	0.00	-1.00	0.78	-0.51
logBasePay	0.00	0.00	0.00	-1.66	-0.12
logBonus	0.00	0.00	0.00	0.00	2.77

We examine the absolute values of the coefficients. Larger absolute values indicate stronger discriminatory power. Performance Evaluation (PerfEval), among the original variables, seems to be one of the greatest discriminators for separating gender groups based on the size and patterns of the coefficients. To further set the groups apart, logBasePay and logBonus both exhibit substantial discriminatory power in particular functions. This is somewhat consistent with what we've seen. However, we need to be cautious in interpreting this directly as these variables are correlated.

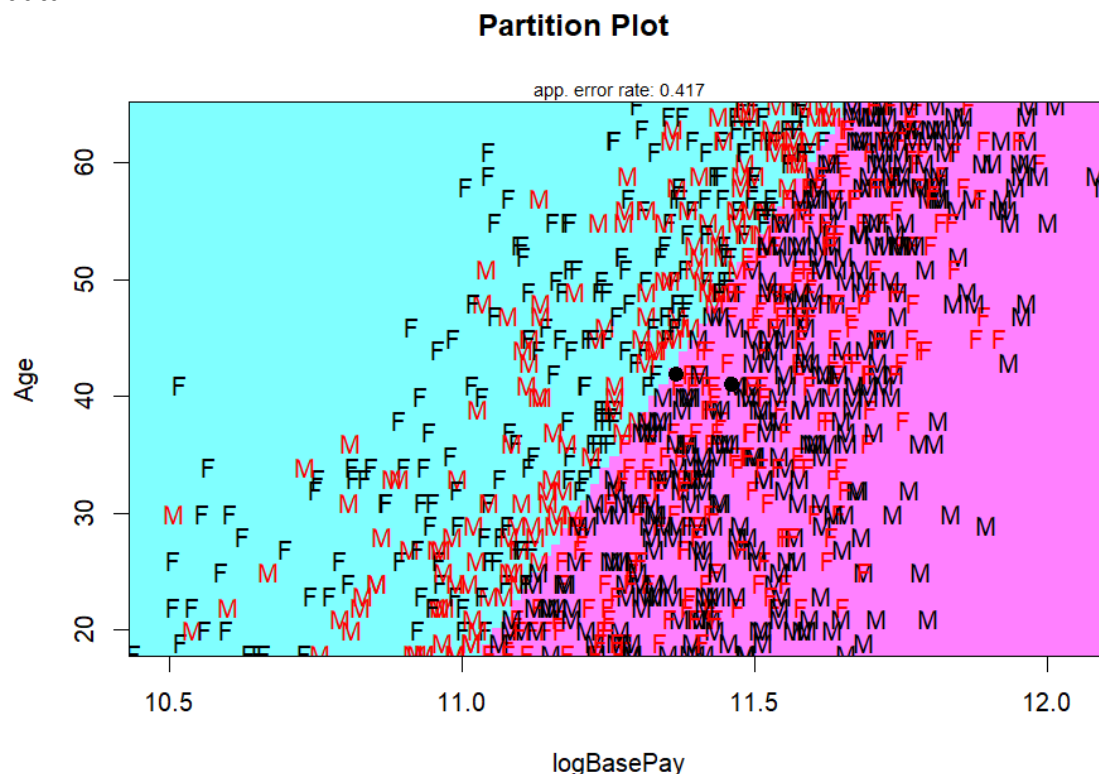
As we only have one discriminating function, we cannot create a score plot of two or more. However, we could visualize our discriminating function better with a regular density plot from the scores we get.



We see some overlap between distributions for males and females and this is expected given the accuracy of classification of our models. Hence although it does distinguish between genders to an extent, there isn't complete separation. Because of this, the discriminant function's capacity to discriminate is limited, which may indicate that other unmeasured factors are more important to explaining wage inequalities or that the gender wage difference in this dataset is very minor.

Lastly, I look at a plot of my data in the space spanned by two of my important

discriminating original variables Age & logBasePay to show which regions are assigned to each group. As we see there is high misclassification and thus all our findings using discriminant analysis have been consistent in presence of weak discriminating ability of the data.

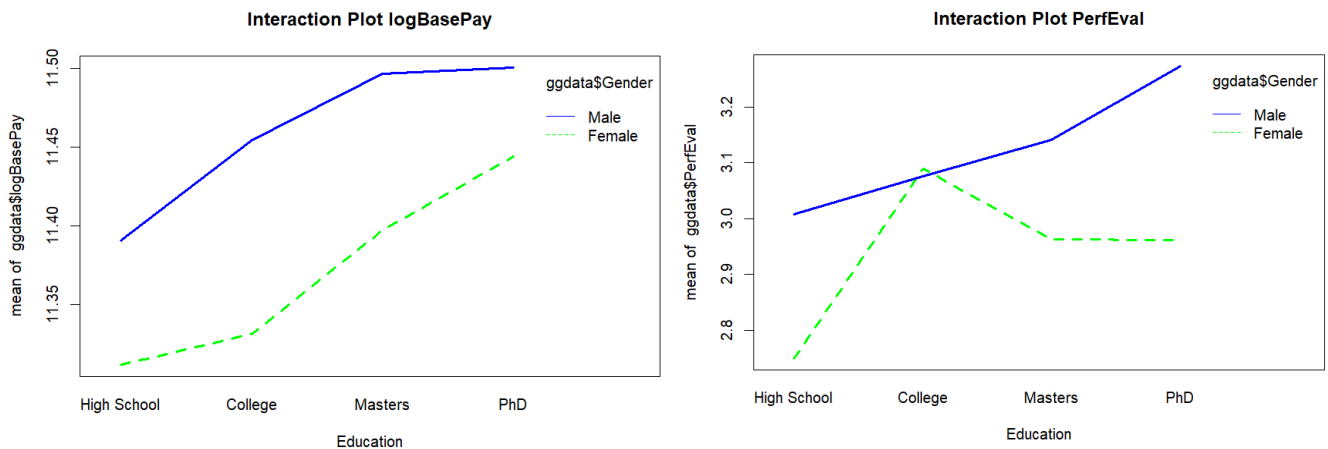


## MANOVA

Following our inferences from discriminant analysis, the difference in means of logBasePay by gender should be further looked into. Additionally, to analyze where this gender bias for wage creeps in we need to include other categorical factors in our analysis such as education or age/experience that determine base pay. Hence, a technique that could accommodate these categorical variables is needed to further explore the patterns in our data.

For MANOVA, I am going to focus on two continuous response variables (PerfEval and logBasePay), categorical predictor variable (Gender and Education), and one continuous predictor variable (Age). I choose education as it is acknowledged to be a driver of wage either through skill, knowledge or signalling effect. Further, we saw a disparity in higher education in our summary statistics which is interesting to explore. I choose age as my continuous variable as it can also account for a demographic characteristic as well as proxy for experience.

To begin with, I explore some interaction plots between my categorical variables and some of my response variables. Firstly, I look at Education, Gender as my predictors with logBasePay then PerfEval.



As we see, for both genders, logBasePay increases generally as education increases. The wage disparity between genders seems to narrow with higher education, starkly after college. An interaction effect between gender and education on log base pay is suggested by the differing slopes of the lines for males and females, however, it needs to be tested. For the interaction plot of performance evaluation, we could clearly predict some interaction effects between gender and education on PerfEval. Here we see very different impact of increased education between genders on performance evaluation. While higher education is associated with higher performance evaluation for males, for females, after college we see a decrease in performance evaluation with increase in education. We could infer a systemic challenge/bias against women at the masters-level and onwards. As we suspect, some interaction in both cases, we run a two-way MANOVA with an interaction term and separate ANOVAs for each dependent variable against independent variables and their interaction.

```
-----
Multivariate Tests: Gender
              Df test stat approx F num Df den Df Pr(>F)
Pillai        1 0.0079223  3.956846      2    991 0.019426 *
Wilks         1 0.9920777  3.956846      2    991 0.019426 *
Hotelling-Lawley 1 0.0079856  3.956846      2    991 0.019426 *
Roy           1 0.0079856  3.956846      2    991 0.019426 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall, there is difference due to gender as for all our multivariate tests, the P value is statistically significant at 0.05 significance level. This suggests that there are gender-based differences in the combined dependent variables (logBasePay and PerfEval).

```
-----
Multivariate Tests: Education
              Df test stat approx F num Df den Df Pr(>F)
Pillai        3 0.0204121  3.409596      6   1984 0.00237584 **
Wilks         3 0.9796436  3.414425      6   1982 0.00234803 **
```

```

Hotelling-Lawley 3 0.0207227 3.419237      6    1980 0.00232064 **
Roy              3 0.0174723 5.777507      3     992 0.00064546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The p-value from all our tests are statistically significant, indicating that different levels of education have a different overall impact on the combined dependent variables.

```

-----
Multivariate Tests: Gender:Education
              Df test stat approx F num Df den Df Pr(>F)
Pillai              3 0.0034610 0.5732162      6    1984 0.75196
Wilks                3 0.9965392 0.5730912      6    1982 0.75206
Hotelling-Lawley    3 0.0034725 0.5729653      6    1980 0.75216
Roy                  3 0.0033936 1.1221358      3     992 0.33905

Type III Sums of Squares
              df logBasePay PerfEval
(Intercept)    1 1.6891e+04 998.2500
Gender          1 4.1166e-01  4.3934
Education       3 1.2825e+00  7.6111
Gender:Education 3 1.4621e-01  3.6413
residuals      992 7.7756e+01 2003.9939

F-tests
              logBasePay PerfEval
(Intercept)  215493.91   494.15
Gender        1.75     0.72
Education     16.36     3.77
Gender:Education 0.62     0.60

p-values
              logBasePay PerfEval
(Intercept)  < 2.22e-16 < 2.22e-16
Gender       0.154996   0.537191
Education    5.6388e-05 0.052537
Gender:Education 0.601011 0.614552

```

Our conclusion from and multivariate results are that are in fact significant effects of gender and education on logbasepay and perfeval but when we look at the interaction term, we see there is no statistical significance. The univariate results are contradictory with only effect of education on logbasepay being significant.

Let's perform multivariate and univariate contrasts to compare levels of Gender with Educations i.e. combinations (Female HS, Female College etc.) to further probe into this and validate our findings from MANOVA.

```

-----
Sum of squares and products for the hypothesis:
              logBasePay PerfEval
logBasePay    1.977131 4.050733
PerfEval       4.050733 8.299114

Sum of squares and products for error:
              logBasePay PerfEval
logBasePay    77.75638 -26.77865
PerfEval      -26.77865 2003.99391

Multivariate Tests:
              Df test stat approx F num Df den Df Pr(>F)
Pillai              1 0.0301657   15.412      2     991 2.5623e-07 ***

```

```

Wilks      1 0.9698343    15.412      2      991 2.5623e-07 ***
Hotelling-Lawley 1 0.0311039    15.412      2      991 2.5623e-07 ***
Roy        1 0.0311039    15.412      2      991 2.5623e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The output from the multivariate contrast indicates the multivariate tests are all significant ( $p = 2.5623e-07$ ), indicating that the combination of gender and education has a statistically significant effect on the combined dependent variables (logBasePay and PerfEval). This means we can reject the null hypothesis that this factor does not affect these outcomes.

#### Linear hypothesis test

Hypothesis:  
 $\text{TRTCOMBMale.High School} - \text{TRTCOMBFemale.College} + \text{TRTCOMBMale.College} - \text{TRTCOMBFemale.Masters} + \text{TRTCOMBMale.Masters} - \text{TRTCOMBFemale.PhD} + \text{TRTCOMBMale.PhD} = 0$

Model 1: restricted model  
 Model 2: logBasePay ~ TRTCOMB

```

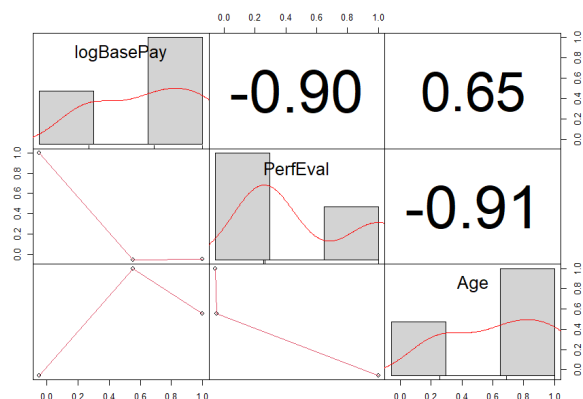
Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      993 79.734
2      992 77.756  1    1.9771 25.224 6.051e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The output from univariate contrast for logBasePay is statistically significant indicating the combination of education and gender DOES in fact effect logBasePay. This affirms our multivariate tests from MANOVA and hence we can proceed.

I now want to add another continuous predictor variable to your model and fit as a multiple-response linear model; adding a continuous variable like Age can help control for its effect and possibly clarify the relationships we find! First, to examine linear relationship between age and our response variables, we make a correlation matrix

We see a decent correlation particularly between age and performance evaluation hence I proceed. I run a two-way MANOVA with interaction between gender and education, I include age as a continuous predictor. The multivariate tests for gender and education are all significant whereas the interaction is not. The multivariate test for Age is also significant!



Term: Age

Sum of squares and products for the hypothesis:

	logBasePay	PerfEval
logBasePay	25.59462	-12.717464
PerfEval	-12.71746	6.319058

Multivariate Tests: Age

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.3294626	243.2138	2	990	< 2.22e-16 ***
Wilks	1	0.6705374	243.2138	2	990	< 2.22e-16 ***
Hotelling-Lawley	1	0.4913411	243.2138	2	990	< 2.22e-16 ***
Roy	1	0.4913411	243.2138	2	990	< 2.22e-16 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

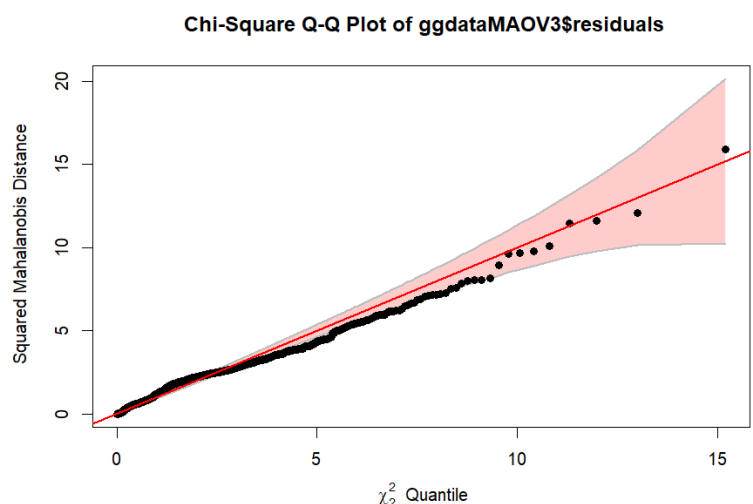
The univariate tests(P-values below) indicate

- Gender: Has a significant effect on logBasePay but a marginal and less significant effect on PerfEval. This means gender differences are more pronounced when it comes to base pay.
- Education: Similarly significant for logBasePay and less so for PerfEval, indicating that educational differences impact base pay more notably.
- Age: Very significant for logBasePay but not for PerfEval, implying that age is a strong predictor for pay but not necessarily for performance evaluations.
- Gender:Education Interaction: Not significant for either logBasePay or PerfEval, in line with the multivariate results.

p-values

	logBasePay	PerfEval
(Intercept)	< 2.22e-16	< 2.22e-16
Gender	7.4849e-10	0.048822
Education	1.0884e-07	0.036136
Age	< 2.22e-16	0.371811
Gender:Education	0.114487	0.616053

To validate our results, let's check model assumptions by making a chi-square quantile plot of the residuals. As we see from the plot, the residuals do approximately follow a multivariate normal distribution and hence we have some validity in our results from MANOVA as the plots indicate the assumptions have not been violated hence GLM is not necessary.



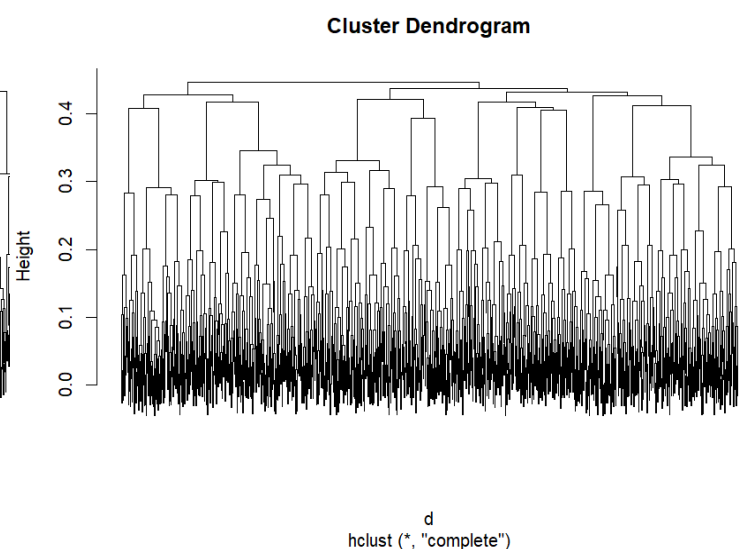
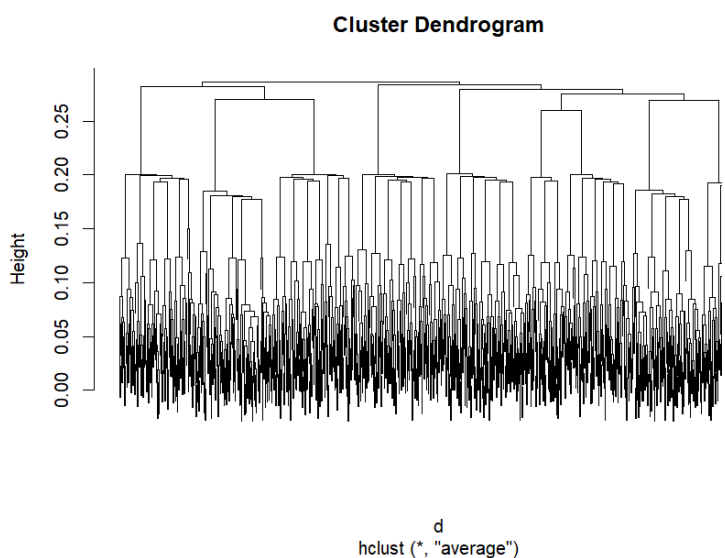
## Cluster Analysis

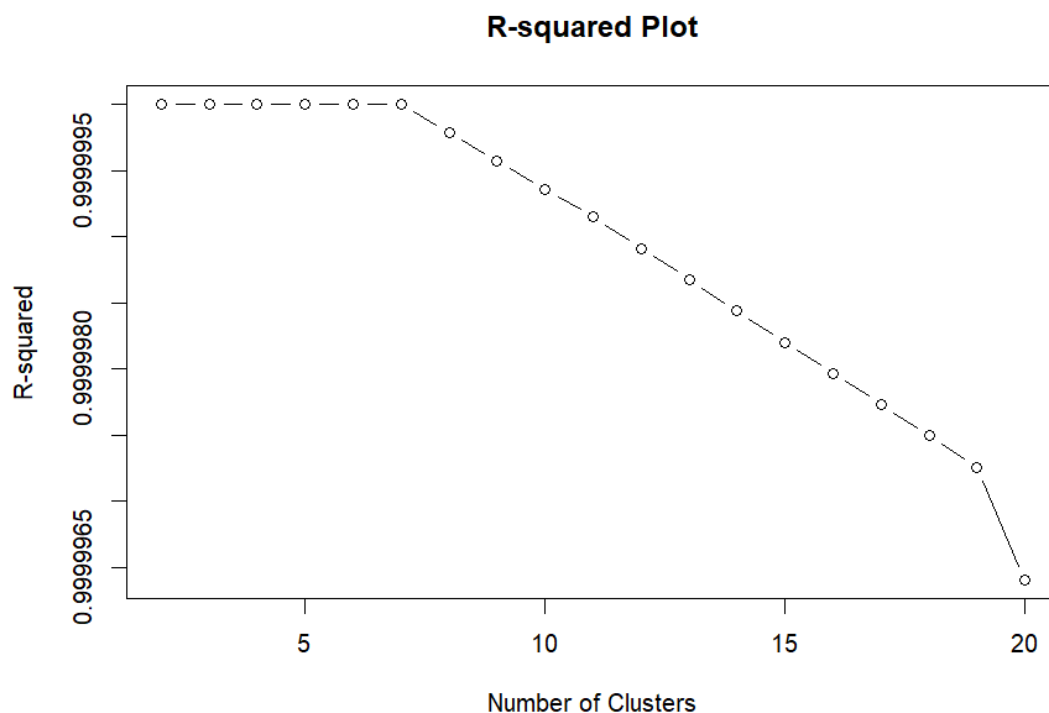
Lastly, a method to accommodate our categorical variables along with continuous variables will be important to accommodate for each factor that can determine a wage difference

between genders. So far, most of our analysis have been limited as they could only factor some variables. Through cluster analysis, my goal is to test naturally occurring groupings within the workforce based on the specified attributes. As our goal with the dataset is to find the drivers of wage difference, I include all factors in my analysis except BasePay and Bonus. I will analyse their wage after the clusters are defined and withing those clusters I will see if we observe patterns in proportion of females.

I use one-hot encoding on categorical variables that have no inherent order such as job titles and departments. This essentially created a dummy variable for each category within the variable. I converted gender into binary variable and used ordinal encoding for education as there is a ranking to exploit there. Now that all our required data is numeric, it is also mixed due to a combination of binary, ordinal, Age etc. hence the distance metric that is most appropriate for it is gower's distance as it is specifically designed to handle this data types.

Hence, I perform hierarchical cluster analysis with gower distance metrics and experiment with two agglomeration methods- average and complete as I want a balanced method to create moderate sized clusters. Visually, we see five or six natural group forming. We get the plot of R squared to decide how many clusters to retain.





We see that about 6 clusters would in fact be ideal. I choose to proceed the analysis and interpretations using average agglomeration as I see six clusters more distinctively for the dendrogram. Further, it allows balanced clusters which accommodates intra-cluster heterogeneity and while considering inter cluster relationships. Complete linkage is sensitive to outliers and hence may not be the best for mixed data types whereas average linkage is generally used for the same. As the goal is to see the average groupings of cluster and evaluate gender proportions, I proceed with this.

Now let's evaluate these clusters for our hypothesis. To do so I first get summary statistics for each cluster BasePay as this is the wage component we have seen to be statistically different across genders.

Cluster	Count	MeanBasePay	MedianBasePay	SD_BasePay
<fct>	<int>	<dbl>	<dbl>	<dbl>
1 1	132	88253.	90770.	22573.
2 2	199	114476.	114621	25079.
3 3	121	90528.	91049	20199.
4 4	133	92828.	92722	22234.
5 5	248	86767.	85478.	24922.
6 6	167	91162.	90780	21803.

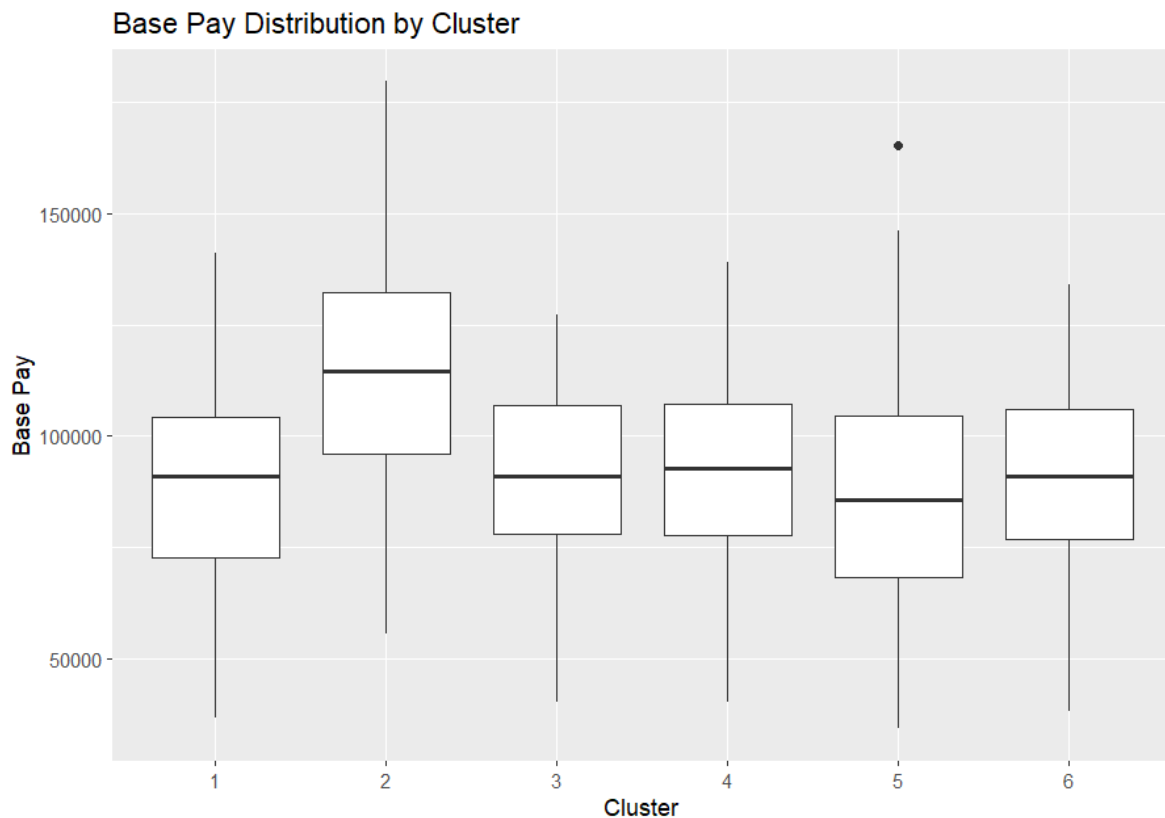
Now let's look at summary statistics of each cluster by gender to see if we see anything definitive above clusters of BasePay and clusters that have different proportions of females.

Cluster	Male	Female	Proportion_Female
<fct>	<int>	<int>	<dbl>
1 1	70	62	0.470
2 2	173	26	0.131
3 3	55	66	0.545
4 4	68	65	0.489

5	5	76	172	0.694
6	6	90	77	0.461

---

To visualize our clusters and BasePay before we interpret the results, let me plot the summary statistics for BasePay by cluster.



Notable patterns in the distribution of gender and BasePay among various groups are shown by the analysis. With only 13% of its members being female, Cluster 2 has a notable gender discrepancy while having the highest mean BasePay (\$114,476.34) and median BasePay (\$114,621). This pattern supports the idea that men continue to dominate high-paying positions and females are underrepresented. Conversely, Cluster 5 shows the highest female representation (69%) but has one of the lowest average salaries, with a mean BasePay of \$86,767.30 and a median of \$85,477.5. Cluster 1,3,4, and 6 are notably a bit more balanced in gender proportion and have average levels of BasePay. This highlights that at tail ends of distribution of income we'll find dominance of one gender whereas in between we see some overlap. This could indicate presence of disparity and not necessarily a gap. Nonetheless, to see these trends by natural forming of the group indicate that issue is not latent. In the next section, I'll discuss how to contextualize our findings altogether and check if it is consistent with economic literature.

## Discussion

The results of our analyses revealed significant patterns concerning the gender wage gap. Discriminant analysis revealed that there exists a discriminating function and that has reasonable discriminating ability between gender based on our continuous factors which

are performance evaluation, age, seniority, BasePay, and Bonus. We saw the function could not fully discriminate between the groups with only 67% accuracy of classification for both LDA and QDA, this could indicate that there isn't an explicit gap in pay between genders or the variables we accounting for cannot fully explain it. For example, a demographic factor like ethnicity could also improve our findings from the data. Using ANOVA and standardized discriminant coefficients, we find significant difference in Performance Evaluation and logBasePay. This suggests that these factors may be influenced by gender biases or systematic differences in how genders are treated or paid within the organization. This section highlighted how wage is affected by bias/discrimination within a work environment leading to a difference in wages across genders.

Using MANOVA we looked at other explanations for wage using predictor like education or age on logBasePay and performance evaluation following the insights we got from discriminant analysis. This could aid our understanding of where the cause of the gap. Our multivariate tests showed statistically significant effect of gender and education on joint dependent variables but not for their interaction. Our univariate tests contradicted this, hence we performed multivariate and univariate contrasts to affirm our results from MANOVA and got mostly consistent results however interaction term was statistically significant here. Further, we added another continuous predictor variable (age) to the model and fit as a multiple-response linear model. We found age to be a strong predictor for pay but not necessarily for performance evaluations. We conclude from this that the interaction between age, gender, and education does largely affect pay. As education does significantly affect pay, discouraging environments for females in higher education, will lead for them to pursue opportunities that do not require it hence creating a gap in income. Age seems to be an obvious predictor of wage as it proxies for experience. (Lazear, 1976)

Lastly, cluster analysis gave us a more holistic sense of the gender wage gap taking into account all our variables that could determine pay. The analysis showed highest wage clusters had disproportionately smaller amount of female whereas lower wage clusters had higher. There were also formations of many "average" clusters that has reasonable split of gender and average wage. Cluster 2 that showed high wage and low female proportion echoes what we saw in our summary statistics where we saw low number of female managers but very high marketing associates. Unlike the other analysis, this took into account occupational choices of job title and department and hence we can account the theory of occupational choice of women (Sloane, 2021) as well as could be indicative of bias against women in leadership managerial positions.

## Conclusion & Points for Further Analysis

To sum up, our in-depth analysis using discrimination, cluster analysis and MANOVA showed significant wage differences across gender, education and age. Discriminant analysis showed key differences between genders in terms of performance assessment and base pay. Cluster analysis showed that high income clusters remain largely male-dominated. Clusters with a higher female representation had lower wages. MANOVA showed education and age strongly predicted base pay, while performance evaluations were less affected by age. These results suggest that closing the gender wage gap requires policies that address not just direct wage discrimination but also systemic biases related to age and education.

Further research could delve into how discrimination happens in workplace or ways in which bias is consequential for a female labour market outcome. It could also assess the factors that women account for while making the decision to work and occupational choice. Lastly, identifying causal effect of these factors on wage could be extremely helpful in taking the next step forward.

## References

- Blau, F. D., & Kahn, L. M. (2016). The Gender Wage Gap: Extent, Trends, and Explanations. *NBER Working Paper No. 21913*.
- Braido, L. H., Olinio, P., & Perrone, H. (2012). GENDER BIAS IN INTRAHOUSEHOLD ALLOCATION: EVIDENCE FROM AN UNINTENTIONAL EXPERIMENT. . *The Review of Economics and Statistics*, 94(2), 552–565.  
<http://www.jstor.org/stable/23262087>.
- Card, D. (1999). The Causal Effect of Education on Earnings. *Handbook of Labor Economics*, Elsevier, Volume 3, Part A, Pages 1801-1863,.
- Lazear, E. (1976). Age, Experience, and Wage Growth. *The American Economic Review*, 66(4), 548–558.
- Sloane, C. M. (2021). College Majors, Occupations, and the Gender Wage Gap. *Journal of Economic Perspectives*, 35 (4): 223-48.