REPORT

Lawsuit

NAME: CHANG JIA JIE
STUDENT NUMBER: 200548740

Executive Summary

Context: A few years ago, the United States District Court of Houston had a case that arises under Title VII of the Civil Rights Act of 1964, 42 U.S.C. 200e et seq. The plaintiffs in this case were all female doctors at Houston College of Medicine who claimed that the College has engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries. The Lead plaintiff in this action, a pediatrician and an assistant professor, was denied for promotion at the College. The plaintiffs had presented a set of data to show that female faculty at the school were less likely to be full professors, more likely to be assistant professors, and earn less money than men, on average.

The complete dataset along with supplementary information and variable descriptions can be downloaded from kaggle at https://www.kaggle.com/datasets/hjmjerry/gender-discrimination

The purpose of this report targets to address real world substantive issue, which is whether there is gender discrimination in the workplace by handling the following questions:

1. Did gender affect salary in the workplace?
2. Did female doctors get lower pay and rank due to gender?

The first question is to investigate how important gender is in differentiating salary. The second question is to investigate how important gender is in differentiating rank.

In this report, k-means clustering, hierarchical clustering, linear regression, logistic regression and classification and regression trees are conducted to address the following two questions. All questions are answered using R. In addition to the report, R code is provided separately in R script. The report includes relevant graphics and tables as part of the answer.

Table of Contents

1 Introduction

This report aims to address two specific questions using R programming languages. Firstly, does gender affect salary in the workplace? Secondly, did female doctors get lower pay and rank due to gender?

Context: A few years ago, the United States District Court of Houston had a case that arises under Title VII of the Civil Rights Act of 1964, 42 U.S.C. 200e et seq. The plaintiffs in this case were all female doctors at Houston College of Medicine who claimed that the College has engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries. The Lead plaintiff in this action, a pediatrician and an assistant professor, was denied for promotion at the College. The plaintiffs had presented a set of data to show that female faculty at the school were less likely to be full professors, more likely to be assistant professors, and earn less money than men, on average.

Link to the dataset is provided here:

https://www.kaggle.com/datasets/hjmjerry/gender-discrimination

2 Variable Description

1 Dept 1=Biochemistry/Molecular Biology
2=Physiology
3=Genetics
4=Pediatrics
5=Medicine
6=Surgery

2 Gender 1=Male, 0=Female

3 Clin 1=Primarily clinical emphasis, 0=Primarily research emphasis

4 Cert 1=Board certified, 0=not certified

5 Prate Publication rate (# publications on cv)/(# years between CV date and MD date)

6 Exper # years since obtaining MD

7 Rank 1=Assistant, 2=Associate, 3=Full professor (a proxy for productivity)

8 Sal94 Salary in academic year 1994

9 Sal95 Salary after increment to 1994

3 Gender discrimination in the workplace

3.1 Investigate Data

The 'Scatterplot Matrix of Student Data' below shows that Sal94 and Sal95 have very high correlation. Promotion info does not exist in the data. Some Dept has consistently high salary. There are fewer Females than Males. Top salaries belong to Males, Clinical Emphasis, and Board Certified.

Due to Sal94 and Sal95 are highly correlated, thus focus on Sal94 only. (Discard ID and Sal95 in analysis)

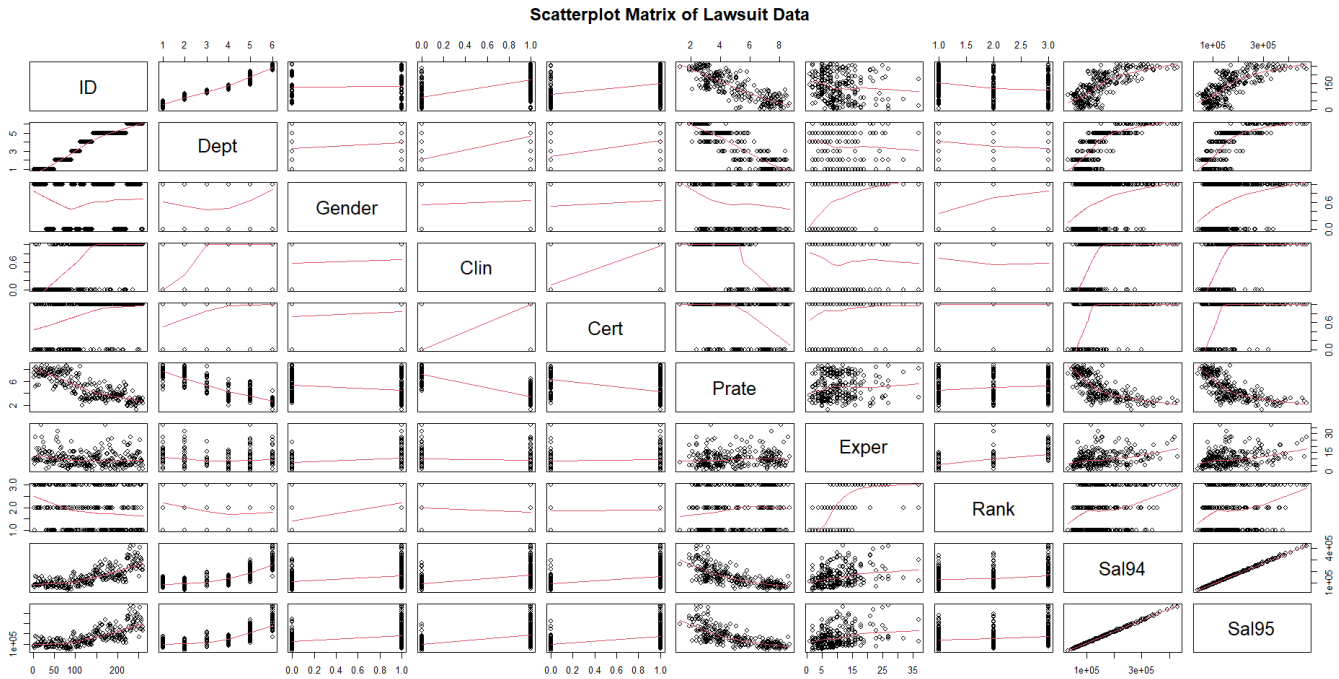

Figure 1: Scatterplot Matrix of Lawsuit Data

3.2 Do Gender affect salary in workplace?

3.2.1 Principal Component Analysis

First two principal components capture 70% of variance. According to pc$rotation, the loadings for "Gender" on PC1 and PC2 are 0.16471and 0.42532, respectively suggesting that Gender is relatively important in PC2 than PC1. Thus, PCA concludes that Gender is an important differentiator.

```
> summary(pc)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.8660 1.4460 0.88847 0.80691 0.67405 0.58629 0.36406 0.23626
Proportion of Variance 0.4353 0.2614 0.09867 0.08139 0.05679 0.04297 0.01657 0.00698
Cumulative Proportion  0.4353 0.6966 0.79531 0.87670 0.93349 0.97645 0.99302 1.00000
```

Figure 2: Summary of Principal Component Analysis

```
> pc$rotation
              PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8
Dept    0.476188677 -0.14590257  0.124550008 -0.02150102 -0.4683424 -0.13834089 -0.502113325 -0.495359627
Gender  0.164710535  0.42532064  0.415886382 -0.75263410  0.2054168  0.08696057  0.004421955 -0.052682830
Clin    0.433400783 -0.13654669  0.110845877  0.31355800  0.6792598  0.17513381  0.176978119 -0.399407471
Cert    0.297042881  0.04235785 -0.866752494 -0.35739490  0.1515528 -0.07882797 -0.041542811 -0.008510228
Prate  -0.502982472  0.14112132 -0.159083060 -0.09168375 -0.1206518  0.03586314  0.303816848 -0.765004825
Exper   0.038463456  0.59957940 -0.149693321  0.30801001 -0.1083032  0.65568323 -0.282609925  0.015112748
Rank    0.003617568  0.59907766  0.003696634  0.30690846  0.1683789 -0.71067900 -0.111143306 -0.033428721
Sal94   0.464284588  0.19766359  0.015809603  0.09778038 -0.4466086 -0.01194350  0.728209938  0.075275167
```

Figure 3: pc$rotation

3.2.2 2-Means clustering

The result of 2-Means clustering shows that cluster 1 has higher salary than cluster 2 with 106467 dollars average. Furthermore, 67% in Cluster 1 are Males, 50% in Cluster 2 are Males.

```
> summary(cluster1$data2.Sal94)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  77087  142249  182945  200912  247953  428876
> summary(cluster2$data2.Sal94)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34514   69702   89956   94445  113207  204279
```

Figure 4: Summary of Gender by Cluster

```
> round(prop.table(table(cluster1$data2.Gender)),2)

   0    1
0.33 0.67
> round(prop.table(table(cluster2$data2.Gender)),2)

  0   1
0.5 0.5
```

Figure 5: Proportion of Gender by Cluster

Perform a goodness of fit test to see whether Cluster 1 is statistically same as Cluster 2 in terms of Gender. Due to P-value(M) < 0.05, reject H0 and conclude that Cluster 1's Gender proportions are different from Cluster 2. 2-Means clustering concludes that Gender is a significant differentiator at 5 % of significant level.

```
        Chi-squared test for given probabilities

data:  M
X-squared = 16.559, df = 1, p-value = 4.717e-05
```

Figure 6: Chi-squared Test

3.2.3 Hierarchical clustering

The result of hierarchical clustering shows that cluster 2 has higher grade than cluster 1 with 93442 dollars. Furthermore, 62% in Cluster 2 are Males, 55% in Cluster 1 are Males.

```
> summary(hc.cluster1$data2.Sal94)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34514   69117   89956   96669  113324  275377
> summary(hc.cluster2$data2.Sal94)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50588  128230  170428  190111  244288  428876
```

Figure 7: Summary of Gender by Cluster

```
> round(prop.table(table(hc.cluster1$data2.Gender)),2)

   0    1
0.45 0.55
> round(prop.table(table(hc.cluster2$data2.Gender)),2)

   0    1
0.38 0.62
```

Figure 8: Proportion of Gender by Cluster

Perform a goodness of fit test to see whether Cluster 1 is statistically same as Cluster 2 in terms of Gender. Due to P-value(M) > 0.05, do not reject H0 and conclude that Cluster 2 Gender Proportions are similar statistically from Cluster 1 Gender Proportions. Hierarchical clustering concludes that Gender is insignificant differentiator between the 2 clusters.

```
> chisq.test(M, p=p.null)

        Chi-squared test for given probabilities

data:  M
X-squared = 3.4807, df = 1, p-value = 0.06209
```

Figure 9: Chi-squared Test

3.2.4 Linear Regression Model

Conduct a linear model to predict Sal94 by all other selected variables excluding ID and Sal95. Summary of the model shows that Gender is statistically insignificant. On the other hand, Dept, Cert, Exper, Clin, Rank are statistically significant. Thus, Gender is not an important variable here.

```
> summary(m.lin)

Call:
lm(formula = Sal94 ~ ., data = data2.dum)

Residuals:
   Min     1Q Median     3Q    Max
-77228 -14563    -99  12044 139758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51112.8    25754.1   1.985  0.04829 *
Dept2       -11965.1     5559.5  -2.152  0.03235 *
Dept3        22370.9     7413.6   3.018  0.00281 **
Dept4        19928.7    10249.9   1.944  0.05299 .
Dept5        72711.9     8654.5   8.402 3.45e-15 ***
Dept6       169259.5    11779.9  14.369  < 2e-16 ***
Gender1      -2801.6     3865.1  -0.725  0.46923
Clin1        15846.2     7863.4   2.015  0.04496 *
Cert1        19562.8     4058.9   4.820 2.51e-06 ***
Prate        -3297.0     3319.1  -0.993  0.32152
Exper         3016.8      348.9   8.646 6.75e-16 ***
Rank2        17218.5     4509.3   3.818  0.00017 ***
Rank3        33875.7     5029.3   6.736 1.13e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25520 on 248 degrees of freedom
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.8994
F-statistic: 194.7 on 12 and 248 DF,  p-value: < 2.2e-16
```

Figure 10: Linear Model Summary

3.2.5 CART Model

Conduct a CART model to predict Sal94 by all other selected variables excluding ID and Sal95. Variable importance shows that Dept, Prate, Clin, Experience, Board-Certified are more important than Gender in explaining Salary. In conclusion, Gender is not a significant predictor of Salary. Hence, no gender discrimination on Salary.

```
> m.cart$variable.importance
       Dept        Prate         Clin        Exper         Cert       Gender         Rank
1.259489e+12 9.327692e+11 4.226641e+11 3.424658e+11 1.868348e+11 1.473891e+11 1.400244e+11
```

Figure 11: CART model's Variable Importance

3.3 Is Distribution of rank associated statistically with Gender in the absence of salary?

3.3.1 Logistic Regression Model

Conduct a logistic model to predict Rank by all other selected variables excluding ID and Salary. Summary of the model shows that Gender is statistically significant in distribution of Rank.

```
> OR.CI <- exp(confint(m.log))
> OR.CI
, , 2

                  2.5 %     97.5 %
(Intercept) 0.0003518374 78.5937183
Dept2       0.6193945404 11.3365317
Dept3       0.1643206899  5.4713884
Dept4       0.1263239500 21.1936291
Dept5       0.0788112083  6.0239662
Dept6       0.0184651939  6.2263795
Gender1     1.2276734403  7.1780576
Clin1       0.0626020926  2.8709586
Cert1       0.1139990085  0.9173602
Prate       0.3227923264  1.5474086
Exper       1.3814158184  1.8211218

, , 3

                  2.5 %    97.5 %
(Intercept) 3.061016e-05 27.248331
Dept2       6.910834e-01 14.252368
Dept3       1.592996e-01  7.214371
Dept4       2.261890e-02  7.074413
Dept5       3.592185e-02  3.829443
Dept6       5.271170e-03  2.959158
Gender1     1.807561e+00 12.948719
Clin1       5.827641e-02  3.729636
Cert1       2.230958e-01  2.095215
Prate       2.840817e-01  1.617619
Exper       1.565478e+00  2.099309
```

Figure 12: Logistic Model Summary

3.3.2 CART Model

Conduct a CART model to predict Rank by all other selected variables excluding ID and Sal94. Variable importance shows that Experience, Prate and Dept are more important than Gender in explaining Rank.

```
> m.cart.rk$variable.importance
     Exper      Prate       Dept     Gender       Cert       Clin
100.238899  47.644912  29.874598  12.474821   7.864484   6.638472
```

Figure 13: CART model's Variable Importance

In conclusion, there is insufficient evidence of gender discrimination on salary. No information to determine Promotion bias as only the current rank is given. In addition, the evidence of gender discrimination on Rank is mixed, so it is not conclusive. However, variables Dept and Experience are far more important than Gender.

5 Appendices

Image used:

Cover Page

Free photo: A Complaint About Using 'Lawsuit' For 'Complaint'.

Available at:

<https://merriam-webster.com/assets/mw/images/article/art-wap-article-main/difference-between-complaint-and-lawsuit-5804-7628bf5a8d39777149a2288cc3d12c20@1x.jpg >

[Accessed: April 2, 2023]