MAS456 Final Project Report

20200639 Woojin Chae

December 17, 2023

1 Introduction

This paper examines the patterns of discrimination faced by wage workers, utilizing data from the KLIPS survey, which includes responses from 3,576 individuals. The main objective of this research is (1) to identify which factors are strongly associated with experience of hiring discrimination, and (2) to check whether such discrimination stands out in a particular group. In the following section, this paper will address the following key questions:

Research topics

1. What are the important variables that are associated with the experience of hiring discrimination? How are those variables related to the experience of hiring discrimination?
2. What are the important principle components (PC) that explain a large portion of variation in the following 12 explanatory variables (gender, age, education level, employment status, income, self-rated health, disability, and experiences of discrimination in receiving income, training, getting promoted, being fired, and obtaining higher education)? How would you interpret those PCs?
3. Identify subgroups (clusters) based on the 12 variables you use for PCA to answer question 2. Also, identify subgroups based on the important PCs you find to answer question 2. Compare the clustering results
4. Is there a difference in under-reporting of hiring discrimination between males and females?
5. Is there an association between the experience of hiring discrimination and health?

2 Research Questions

2.1 Identifying and Analyzing Key Variables in Hiring Discrimination Experiences

Logistic regression analysis is an effective method for identifying key variables strongly associated with 'disc_hire'. The coefficients derived from this model reflect the change in log odds of the 'disc_hire' occurrence for each unit increase in a specific variable. These coefficients' magnitudes measure their importance, while their signs indicate the direction of correlation. Prior to model fitting, it is essential to standardize the data due to the differing scales of the features. This standardization is crucial because coefficients from variables on different scales are not directly comparable. Once the model is appropriately adjusted for these scale differences, the coefficients can be obtained and their significance evaluated by considering their absolute values. The results obtained from this process are as follows.

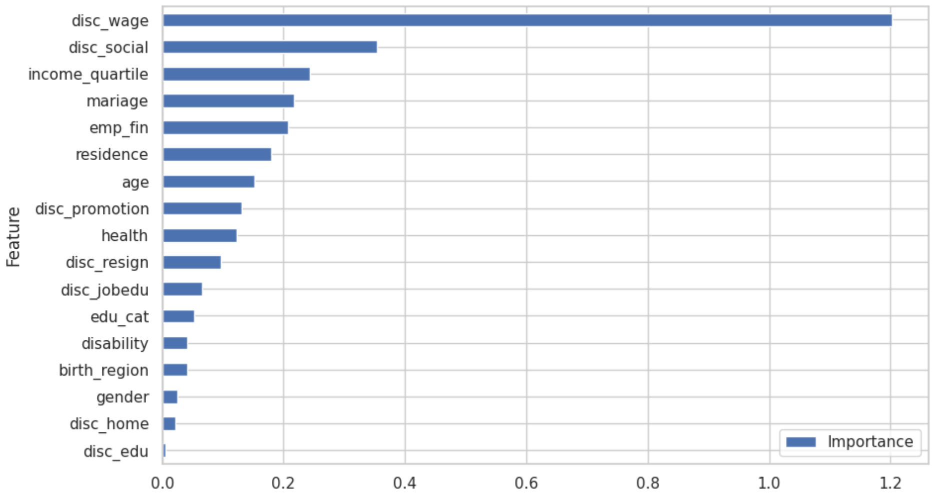
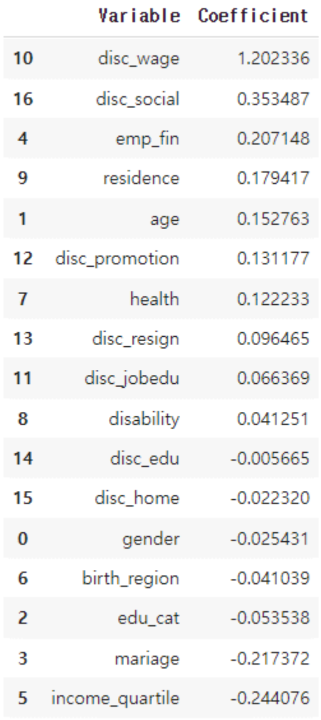


Figure 1: feature importance obtained from logistic regression



	Variable	Coefficient
10	disc_wage	1.202336
16	disc_social	0.353487
4	emp_fin	0.207148
9	residence	0.179417
1	age	0.152763
12	disc_promotion	0.131177
7	health	0.122233
13	disc_resign	0.096465
11	disc_jobedu	0.066369
8	disability	0.041251
14	disc_edu	-0.005665
15	disc_home	-0.022320
0	gender	-0.025431
6	birth_region	-0.041039
2	edu_cat	-0.053538
3	mariage	-0.217372
5	income_quartile	-0.244076

Figure 2: logistic regression coefficients

The results indicate a strong association between ‘disc_wage’ and ‘disc_hire’(1). Additionally, ‘disc_social’ and ‘income_quartile’ also show significant associations with ‘disc_hire’. It is important to note that the coefficient for ‘income_quartile’ is negative(2), in contrast to the positive coefficients of ‘disc_hire’ and ‘disc_social’. This suggests that an increase in income quartile is associated with a reduced likelihood of experiencing hiring discrimination.

However, further investigation is needed. In the variable ‘disc_wage’ and ‘disc_social’, the survey scheme assigns 0 to ‘No’, 1 to ‘Yes’, and 2 to ‘Not Applicable’. This categorization could potentially introduce bias in the logistic regression analysis. Specifically, the high coefficient observed for ‘disc_wage’ and ‘disc_social’ might be influenced more by the ‘Not Applicable’ responses (coded as 2) rather than the affirmative responses (coded as 1). This possibility necessitates a more detailed examination to ensure the accuracy and reliability of the findings.

Fortunately, the instances where ‘disc_wage’ and ‘disc_social’ is coded as ’Not Applicable’ (2) are relatively few in comparison to the total sample size (2). This suggests that any potential bias introduced by this coding is likely to be minimal. Consequently, the impact on the overall analysis is expected to be insignificant. Therefore, the initial conclusions drawn from the logistic regression analysis regarding ‘disc_wage’ and ‘disc_social’ and its association with ‘disc_hire’ appear to be valid and can be maintained.

Selecting variables with extreme coefficients is a common practice in logistic regression analysis, but this method alone may not be comprehensive due to the potential correlation among these variables. Highly correlated features, if all included, might lead to redundancy and multicollinearity in the model. It is, therefore, crucial to assess the inter-relationships between these variables to ensure the robustness of the model. To address this, Principal Component Analysis (PCA) will be employed in the following section. PCA is an effective approach for reducing dimensionality and mitigating issues of multicollinearity by identifying the principal components that capture the most variance in the data.

2.2 PCA of 12 Explanatory Variables Influencing Discrimination Experiences

We proceeded to conduct Principal Component Analysis (PCA) on a set of 12 explanatory variables to explore their relevance to ‘disc_hire’. These variables include gender, age, education level, employment status, income, self-rated health, disability, and experiences of discrimination in various contexts—specifically, in receiving income, during training, in promotion opportunities, in instances of being fired, and in obtaining higher education. The aim of this PCA is to uncover the underlying patterns and relationships among these variables and to determine how they collectively contribute to the phenomenon of hiring discrimination.

In this analysis, PCA was performed with the objective of capturing 80% of the variance in the existing data. To determine the number of Principal Components (PCs) needed to achieve this threshold, a scree plot was utilized. The scree plot, which visually represents the variance captured by each component, indicated that 8 PCs were sufficient to encompass 80% of the data’s variance. Subsequently, these 8 PCs were extracted and analyzed. The results of this analysis are presented below.

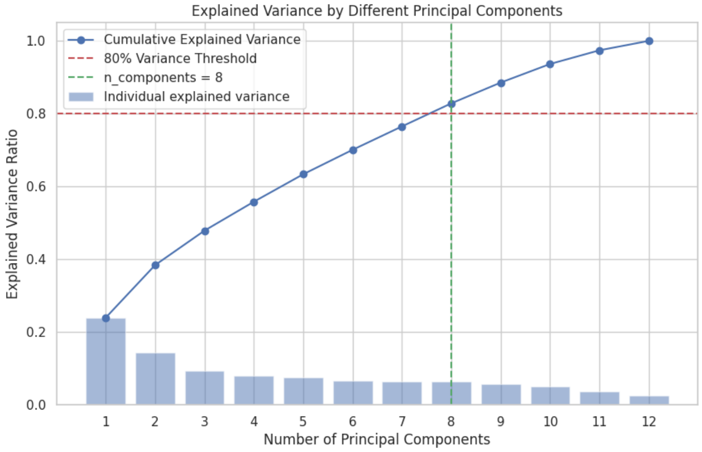
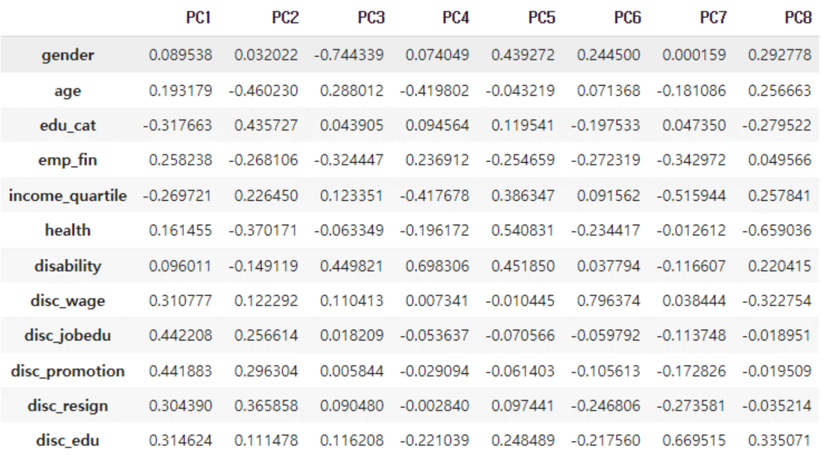


Figure 3: Scree plot

Figure 3 comprehensively shows how much variance each PC captures, how variance accumulates as the number increases, and how many PCs are used to cross the target threshold. One can easily find that 8 PCs are enough to represent 80% variance of whole data. Based on this finding, the subsequent analysis will be conducted using these 8 PCs.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
gender	0.089538	0.032022	-0.744339	0.074049	0.439272	0.244500	0.000159	0.292778
age	0.193179	-0.460230	0.288012	-0.419802	-0.043219	0.071368	-0.181086	0.256663
edu_cat	-0.317663	0.435727	0.043905	0.094564	0.119541	-0.197533	0.047350	-0.279522
emp_fin	0.258238	-0.268106	-0.324447	0.236912	-0.254659	-0.272319	-0.342972	0.049566
income_quartile	-0.269721	0.226450	0.123351	-0.417678	0.386347	0.091562	-0.515944	0.257841
health	0.161455	-0.370171	-0.063349	-0.196172	0.540831	-0.234417	-0.012612	-0.659036
disability	0.096011	-0.149119	0.449821	0.698306	0.451850	0.037794	-0.116607	0.220415
disc_wage	0.310777	0.122292	0.110413	0.007341	-0.010445	0.796374	0.038444	-0.322754
disc_jobedu	0.442208	0.256614	0.018209	-0.053637	-0.070566	-0.059792	-0.113748	-0.018951
disc_promotion	0.441883	0.296304	0.005844	-0.029094	-0.061403	-0.105613	-0.172826	-0.019509
disc_resign	0.304390	0.365858	0.090480	-0.002840	0.097441	-0.246806	-0.273581	-0.035214
disc_edu	0.314624	0.111478	0.116208	-0.221039	0.248489	-0.217560	0.669515	0.335071

Figure 4: Principle Components

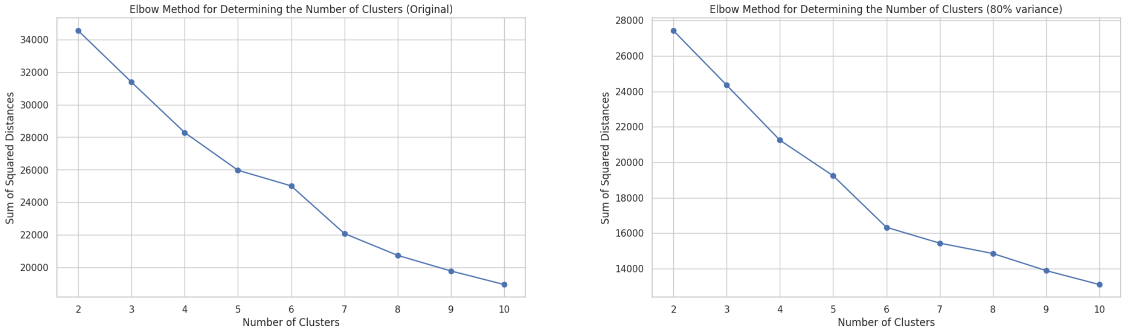
Figure 4 presents the eight principal components that were derived from the analysis. While each component is a mix of various variables, it is possible to discern distinct interpretations for each of them. These interpretations help in understanding the underlying patterns and relationships represented by the components, providing insights into the different dimensions captured by the PCA. The following are the interpretations for each of the principal components:

- **PC1** appears to predominantly capture elements related to various forms of perceived discrimination in the workplace. This is evidenced by its large loadings on several specific variables: 'disc_jobedu' (discrimination in job-related education), 'disc_promotion' (discrimination in promotion opportunities), 'disc_wage' (discrimination in wages), 'disc_edu' (discrimination in educational settings), and 'disc_resign' (discrimination leading to resignation). These loadings suggest that PC1 is a significant composite indicator of workplace discrimination experiences.

- **PC2** reveals a notable contrast between the variables 'age' and 'education level', suggesting a reflection of varying socio-economic or generational experiences. This contrast may indicate that these two variables represent different dimensions of the data, capturing the diverse backgrounds and life stages of the individuals in the study. The inverse relationship between age and education level in PC2 could be indicative of changing educational trends over generations or differences in career stages influenced by age and education.

- The subsequent principal components can be interpreted as above.

2.3 Comparative Analysis of Subgroup Clustering Based on Original Variables and Principal Components

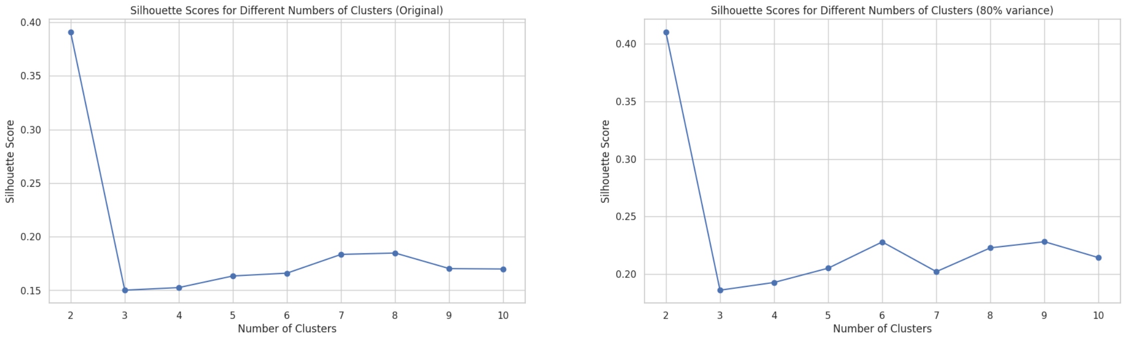


(a) Sum of squared distances (original)

(b) Sum of squared distances (PCs)

Figure 5: Elbow method

In this section, we compare clusters derived from the original 12 explanatory variables with those obtained from the principal components (PCs) extracted from these variables. To determine the optimal number of clusters, we initially employed the Elbow method. This method involves identifying a point where the overall rate of decrease in the sum of squared distances from each point to its assigned cluster diminishes significantly. However, as we can see from Figure 5, the resulting graph from the Elbow method did not provide a clear indication of the best number of clusters to use.

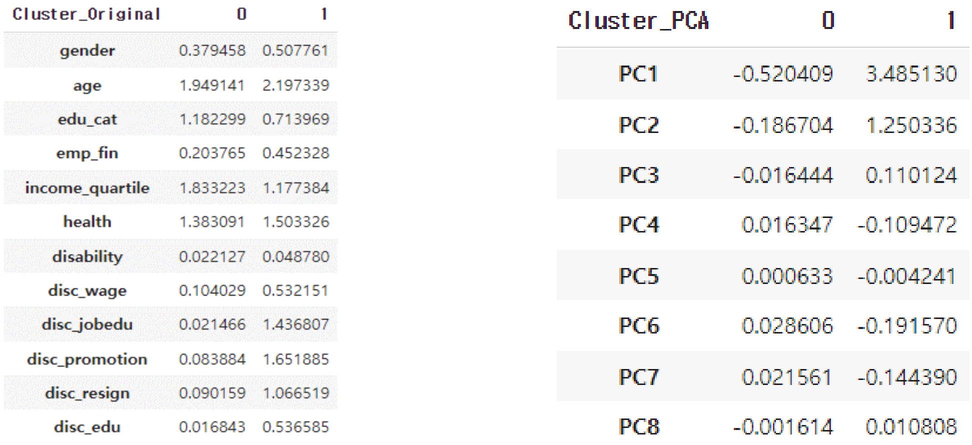


(a) Silhoutte scores (original)

(b) Silhoutte scores (PCs)

Figure 6: Silhoutte score criterion

Consequently, we turned to an alternative criterion, the silhouette score (6), which offers a more definitive approach. The silhouette score assesses how well each data point fits within its cluster by measuring both cohesion within the cluster and separation from other clusters. This method provides a quantitative assessment, as opposed to the more heuristic nature of the Elbow method. Applying the silhouette score criterion, we found that the optimal number of clusters is 2 for both the original data and the data represented by the PCs. Clustering each data into two groups leads as follows.



Cluster_Original	0	1
gender	0.379458	0.507761
age	1.949141	2.197339
edu_cat	1.182299	0.713969
emp_fin	0.203765	0.452328
income_quartile	1.833223	1.177384
health	1.383091	1.503326
disability	0.022127	0.048780
disc_wage	0.104029	0.532151
disc_jobedu	0.021466	1.436807
disc_promotion	0.083884	1.651885
disc_resign	0.090159	1.066519
disc_edu	0.016843	0.536585

(a) Subgroups obtained from original data

Cluster_PCA	0	1
PC1	-0.520409	3.485130
PC2	-0.186704	1.250336
PC3	-0.016444	0.110124
PC4	0.016347	-0.109472
PC5	0.000633	-0.004241
PC6	0.028606	-0.191570
PC7	0.021561	-0.144390
PC8	-0.001614	0.010808

(b) Subgroups obtained from PCs

Figure 7: Cluster results

Figure 7 depicts the centroids of the two clusters formed by the original data (as shown in 7a) and by the principal components (PCs) (as shown in 7b), respectively. Examining 7a, we observe that Cluster 1 is characterized by higher average values in 'disc_wage', 'disc_jobedu', 'disc_promotion', and 'disc_resign', indicating a higher level of perceived discrimination in these areas. Additionally, this cluster tends to include an older age group with lower education levels, relatively lower income, and poorer health status. In contrast, Cluster 0 exhibits lower average values in the discrimination-related variables, implying less perceived discrimination. This cluster is associated with a younger demographic, higher education levels, and better income and health status.

Regarding 7b, which represents clustering based on the principal components, Cluster 1 shows higher values on the PCs, suggesting a stronger presence of the features these components represent. This could indicate a higher level of perceived discrimination and certain socio-economic factors. Conversely, Cluster 0 demonstrates lower values on the principal components, pointing to lower levels of the features captured by these components, such as perceived discrimination and socio-economic variables.

Considering this aspects, one can conclude as follows.

- The clustering based on the original data provides a detailed view of each group's characteristics, focusing on specific variables like age, education, and discrimination experiences.
- Clustering based on PCA-reduced data offers a more abstract view, focusing on underlying patterns represented by the principal components.
- Both approaches identify distinct groups within the data. However, the PCA-based clustering is less direct in terms of original variables but captures broader patterns in the data, while the original data-based clustering provides more direct interpretability.
- **In both cases, the clusters seem to differentiate between individuals with higher and lower levels of perceived discrimination and socio-economic status.**

2.4 Under-Reporting of Hiring Discrimination in terms of Gender

To investigate the existence of gender-based differences in the under-reporting of hiring discrimination, one employed two prediction models: logistic regression and random forest. Each model was tuned to maximize AUC score. The optimal parameters identified through cross validation were as follows: for the random forest classifier, a maximum depth of 5 and 200 trees; for the logistic regression, an L1 penalty of 1. Under these parameters, the models achieved AUC scores of 0.901 and 0.879, respectively (5). These results indicate a high level of predictive accuracy in both models, with the random forest classifier showing a slightly better performance.

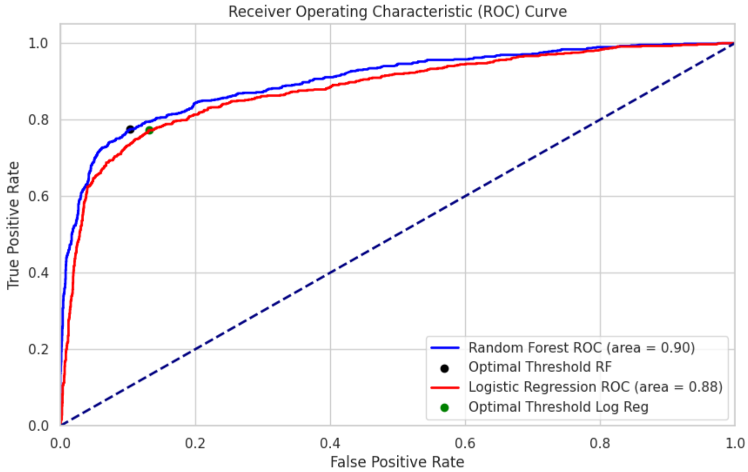
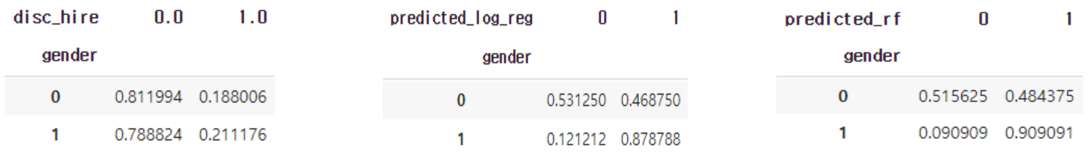


Figure 8: ROC curve with knotted optimizer

Once the models with the highest Area Under the Curve (AUC) scores were established, the next step was to fine-tune them by determining the optimal threshold for binary prediction. This threshold was identified using the ROC curve, and it is evaluated by maximizing the sum of sensitivity (true positive rate) and specificity (true negative rate). After evaluation, threshold for logistic regression, random forest predictor were 0.147, 0.218 respectively (6).

Using this optimizer, one conducted prediction for NaN-respondents, and the result was follows.



disc_hire	0.0	1.0
gender		
0	0.811994	0.188006
1	0.788824	0.211176

(a) Existing respondents

predicted_log_reg	0	1
gender		
0	0.531250	0.468750
1	0.121212	0.878788

(b) Non-respondents (pred by logreg)

predicted_rf	0	1
gender		
0	0.515625	0.484375
1	0.090909	0.909091

(c) Non-respondents (pred by r.f)

Figure 9: Distributions of ‘disc_hire’ between ‘gender’

Upon analyzing Figures (9b) and (9c), common patterns emerge in the model’s predictions regarding hiring discrimination experiences between males and females. The predictions for males show a relatively even distribution between ’True’ (indicating discrimination) and ’False’ (indicating no discrimination), which suggests a less distinct pattern of hiring discrimination. Conversely, for females, both models consistently predict a higher rate of hiring discrimination experiences. This pronounced pattern of discrimination against females, as indicated by the models, stands in contrast to the trends observed in the original data (Figure 9a). The difference between the models’ predictions and the original data is particularly noteworthy, as it suggests that the underlying patterns of discrimination may not be immediately apparent from the original data alone. This underscores the importance of advanced analytical models in uncovering deeper insights that might otherwise remain hidden in the data.

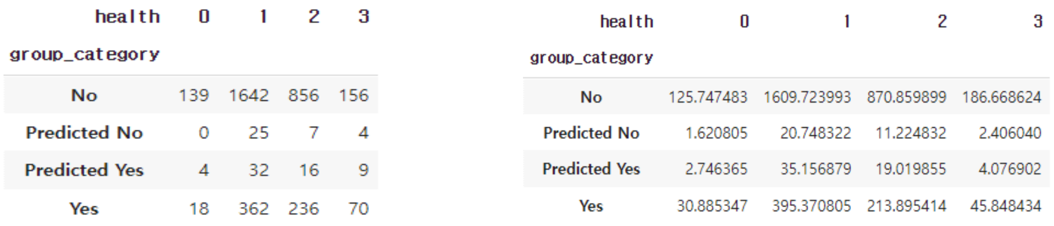
2.5 Link Between Hiring Discrimination Experiences and Self-Rated Health Scores

In this section, we compared the distribution of self-rated health among four distinct groups categorized based on their responses to ‘disc_hire’. These groups are: (1) those who responded ’No’, (2) those who responded ’Yes’, (3) those who responded ’NA’ but were predicted as ’No’, and (4) those who responded ’NA’ but were predicted as ’Yes’, using the random forest model developed earlier. The distribution of self-rated health for each group is depicted in Figure 10a.

To determine whether there is a relationship between self-rated health and ‘disc_hire’, we conducted an overall test of independence. The hypothesis under test is whether these categories are all independent of each other. If they are indeed independent, we would expect the distribution of self-rated health to be similar across each category of ‘disc_hire’. In other words, the expected frequency in each cell of our contingency table (which cross-tabulates these two variables) should be roughly equivalent. Therefore, (i, j)th cell can be evaluated as follows.

$$E[(\text{Group Category i, Health j})] = \frac{\text{\# of ppl in category i}}{\text{total number of ppl}} \times \text{\# of ppl with health score j}$$

Then, the expected frequency under null hypothesis can be depicted as follows.



health	0	1	2	3
group_category				
No	139	1642	856	156
Predicted No	0	25	7	4
Predicted Yes	4	32	16	9
Yes	18	362	236	70

(a) Observed frequency

health	0	1	2	3
group_category				
No	125.747483	1609.723993	870.859899	186.668624
Predicted No	1.620805	20.748322	11.224832	2.406040
Predicted Yes	2.746365	35.156879	19.019855	4.076902
Yes	30.885347	395.370805	213.895414	45.848434

(b) Expected frequency

Figure 10: Contingency table

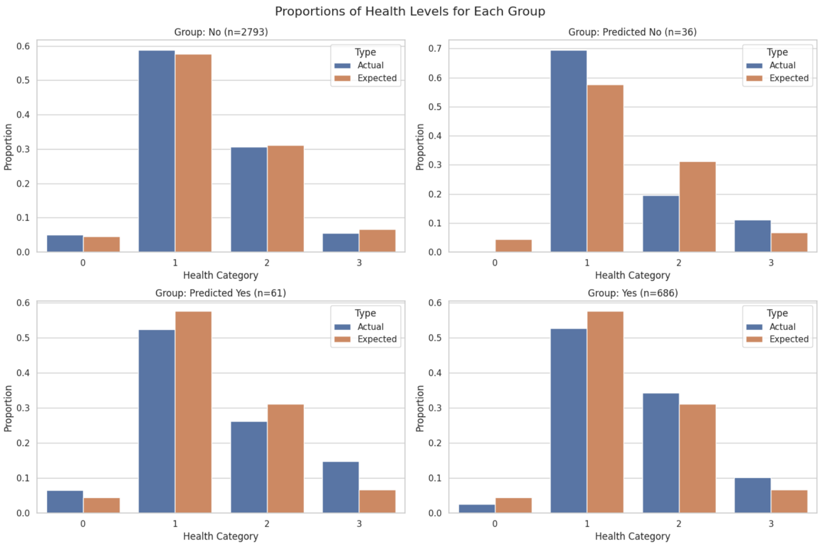


Figure 11: Expected frequency (TL: No, TR: Predicted No, BL: Predicted Yes, BR: Yes)

To verify this, chi-square test can be a good approach for it. If the two variables are independent, we can establish the test-statistics shown below and measure the p-value to determine whether null hypothesis can be accepted. The test statistics are calculated as follows.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{ with } df = (r-1) \times (c-1) \qquad O\text{: observed, } E\text{: expected}$$

$$TS = \frac{(139 - 125.747)^2}{125.747} + \cdots + \frac{(70 - 45.848)^2}{45.848} = 42.953 \text{ with } df = (4-1) \times (4-1)$$

$$\text{p-value} = 2.198 \times 10^{-6}$$

Given that the obtained p-value from the chi-square test is low, we reject the null hypothesis, indicating that there is a statistically significant association between self-rated health and 'disc_hire'. To further explore this relationship, it is necessary to conduct pairwise comparisons between the different groups to identify which specific pairs show distinct distributions in terms of self-rated health. Since there are four groups, a total of six pairwise comparisons are required (each group compared with every other group)

To control for the Familywise Error Rate (FWER) in these multiple comparisons and to maintain the overall Type 1 error rate at 0.05, we employed the Bonferroni correction. In this case, each pairwise test will be evaluated at a significance level of 0.05/6, which helps to mitigate the risk of false positives that increases with multiple comparisons. One can see its results below (12).

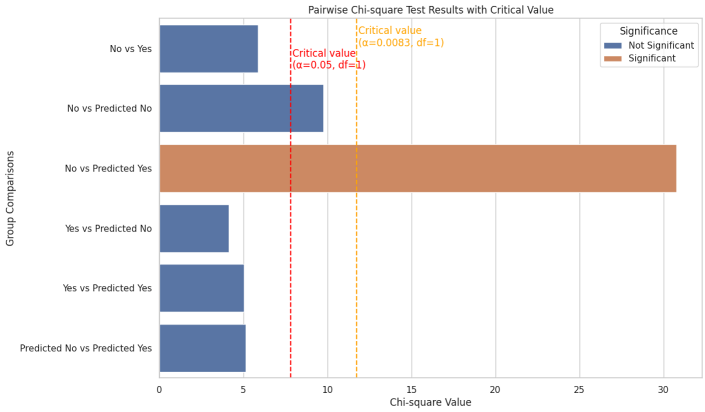
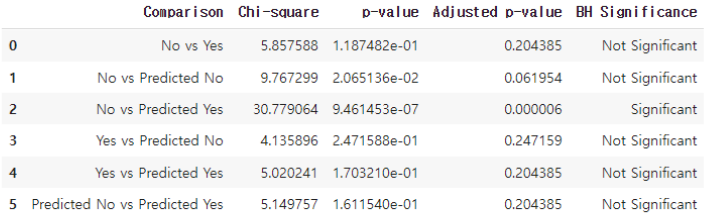


Figure 12: Pairwise comparison using bonferroni criterion

According to (12), the test comparing the 'No' group versus the 'Predicted Yes' group was the only one to be rejected. Interestingly, this result would have been different had the Bonferroni correction not been applied; the 'No vs Predicted Yes' test was close to being significant without this conservative adjustment. Recognizing that the Bonferroni correction is a particularly conservative method for controlling the Familywise Error Rate (FWER), we proceeded to apply the Benjamini-Hochberg procedure as an alternative. This procedure aims to control the False Discovery Rate (FDR), offering a less stringent approach compared to Bonferroni. The goal was to determine if other tests, which were not rejected under the Bonferroni criterion, might be found significant under the FDR control criteria.



	Comparison	Chi-square	p-value	Adjusted p-value	BH Significance
0	No vs Yes	5.857588	1.187482e-01	0.204385	Not Significant
1	No vs Predicted No	9.767299	2.065136e-02	0.061954	Not Significant
2	No vs Predicted Yes	30.779064	9.461453e-07	0.000006	Significant
3	Yes vs Predicted No	4.135896	2.471588e-01	0.247159	Not Significant
4	Yes vs Predicted Yes	5.020241	1.703210e-01	0.204385	Not Significant
5	Predicted No vs Predicted Yes	5.149757	1.611540e-01	0.204385	Not Significant

Figure 13: Pairwise comparison using benjamini-hochberg procedure

However, the Benjamini-hochberg procedure also yield the same result. Therefore, we can conclude that under the 95% significance level, it has the same distribution except for No vs Predicted Yes.

3 Summary

3.1 Key Variables regarding 'disc_hire'

As a result of searching for variables showing a high relationship with disc_hire using logistic regression analysis, it was confirmed that **disc_wage**, **disc_social**, and **income_quartile** were high. In the case of income_quartile, there was a negative direction, which is acceptable as it means that the income level is high and there is no experience of discrimination in the employment process. Although the survey data had room for bias, the effect was found to be insignificant, and the above conclusion was maintained.

3.2 Interpretation of PCs obtained through PCA

Using PCA, it was possible to lower the dimension of the data and maintain the explanatory power of the entire data. Although the variables were mixed in this process, there was no significant difficulty in grasping the meaning of the principal component. In this study, PCA was conducted with the aim of capturing up to 80% of the existing data variance, and as a result, it was confirmed that 8 principal components were sufficient. As a result of carefully examining the first main component among the principal components obtained through this, it was confirmed that the characteristics that could appear in common (disc_wage, disc_jobedu, disc_promotion, disc_resign, disc_edu) among those who have experienced discrimination in hiring were grouped in common.

3.3 Cluster comparison between clusters from original data and PCs

One clustered the data using the KMeans algorithm based on the principal components obtained above, as well as original data, then compared the results. The optimal number of clusters in both cases were chosen based on silhoutte score. Fortunately, the optimal number of clusters were the same with two, making it easy to compare. When the results of clustering with existing data and when using as the principal component were checked, it was confirmed that a qualitatively similar cluster was formed. Comparing based on the centroid of each cluster, one could find that both cases clustered data into two groups having similar properties; (1) workers exposed to discrimination and (2) workers who were not qualitatively clustered. From this result, one could strongly affirm that the meaning of the original data could be well maintained even if clustering was performed using the main components obtained by PCs.

3.4 Under-Reporting Hiring Discrimination in terms of Gender

Since the existing data included questionnaires who did not respond to questions related to employment discrimination experiences, a prediction model was constructed based on the respondent data to estimate the expected responses of the questionnaire to see if there was any underestimated part in the existing data. At this time, logistic regression and random forest models were used as prediction models, and the hyper-parameter was tuned so that the highest auc score could be obtained from each model. Among the obtained models, the model was advanced by finding the threshold probability that allows the sum of specificity and sensitivity to be the highest. Finally, as a result of predicting the expected responses of non-respondents using the above model, there was a difference in employment experience between men and women, which was not revealed in the existing data. Both men and women experienced employment discrimination at a higher rate than the existing data, and it was confirmed that the degree was particularly large for women. As a result, it was confirmed that the existing data may be somewhat underestimated in relation to employment discrimination between men and women. At the same time, it should be noted that only samples that are likely to produce these results may have been cherry-picked. Samples may be configured so that those who have experienced employment discrimination are more likely to come out, such as whether they are reluctant to respond for a series of reasons. Therefore, it would be better to somewhat reserve responses to the analysis.

3.5 Dependence on hiring discrimination and self-rated-health

To examine whether it is related to employment discrimination experience and self-rated health level, data was divided into a total of 4 groups and categorized according to health level within them. Then, as a result of performing the chi-square test under the assumption that it is not related (independent), it was concluded that all groups did not have the same distribution of health levels.

Therefore, pairwise comparison between groups was performed to identify groups with differences in distribution. At this time, the bonferroni method, which adjusts the fdr more liberally, was applied as a method for controlling the FWER in multiple testing. As a result, it was found that there was no significant difference in all except for the case of 'No Vs Predicted Yes'. In this case, it may be possible to conclude that it is related, but as mentioned in 'Gender Difference in Under-Reporting Hiring Discrimination', it should also be noted that among non-respondents, the sample may have been configured to show a high correlation between self-rated-health and experience of discrimination in hirring. For example, there may be many people in poor health in the sample who have been discriminated against for a series of reasons. In this case, it is impossible to conclude prematurely because the sample cannot be considered to be representative of the original population.by the existing data.

4 Appendix

4.1 Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report , confusion_matrix
import numpy as np

# Filtering the dataset for ’disc_hire’ values of 0 or 1
filtered_data = data[data[’disc_hire’].isin([0, 1])]

# Splitting the data into features (X) and target (y)
X = filtered_data.drop(’disc_hire’, axis=1)
y = filtered_data[’disc_hire’]

# Standardize predictors
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Creating the logistic regression model
logreg = LogisticRegression()

# Fitting the model with the training data
logreg.fit(X_scaled , y)

# Plot importance of the coefficients
coefficients = logreg.coef_[0]
feature_importance = pd.DataFrame({’Feature’: X.columns , ’Importance’: np.abs(coefficients)})
feature_importance = feature_importance.sort_values(’Importance’, ascending=True)
feature_importance.plot(x=’Feature’, y=’Importance’, kind=’barh’, figsize=(10, 6)).legend(loc=’lower right’)
```

Listing 1: Logistic Regression

```
print(len(data[(data[’disc_wage’]==2) & (data[’disc_hire’].isna()==False)]))
print(len(data[(data[’disc_social’]==2) & (data[’disc_hire’].isna()==False)]))

% 11
% 18
```

Listing 2: the number of noise among 3479 data

4.2 PCA

```
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

# Selecting only the relevant columns for PCA
pca_columns = [’gender’, ’age’, ’edu_cat’, ’emp_fin’, ’income_quartile’,
               ’health’, ’disability’, ’disc_wage’, ’disc_jobedu’,
               ’disc_promotion’, ’disc_resign’, ’disc_edu’]
pca_data = filtered_data[pca_columns]

# Standardizing the data - important for PCA
scaler = StandardScaler()
pca_data_scaled = scaler.fit_transform(pca_data)

# Creating a PCA object
pca = PCA()

# Fitting PCA on the standardized data
pca.fit(pca_data_scaled)

# Explained variance ratio
explained_variance_ratio = pca.explained_variance_ratio_

# Choosing the number of components that captures at least 80% of total variance
cumulative_variance = explained_variance_ratio.cumsum()
n_components = (cumulative_variance < 0.80).sum() + 1

# Creating a PCA object with the selected number of components
pca_selected = PCA(n_components=n_components)
pca_selected.fit(pca_data_scaled)

# Principal components
principal_components = pca_selected.components_
```

Listing 3: PCA Analysis

4.3 KMeans

```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Function to perform KMeans clustering and calculate silhouette scores
def perform_kmeans(data, n_clusters_range):
    silhouette_scores = []
    kmeans_models = []

    for n_clusters in n_clusters_range:
        kmeans = KMeans(n_clusters=n_clusters, n_init=’auto’, random_state=42)
        kmeans.fit(data)
        silhouette_avg = silhouette_score(data, kmeans.labels_)
        silhouette_scores.append(silhouette_avg)
        kmeans_models.append(kmeans)

    return kmeans_models, silhouette_scores

# Defining the range of possible clusters
n_clusters_range = range(2, 11)

# Performing KMeans clustering on original data
kmeans_models_original, silhouette_scores_original = perform_kmeans(pca_data_scaled,
     n_clusters_range)

# Performing KMeans clustering on the data reduced by PCA
# Using the PCA-transformed data with the number of components that explain 80% of the variance
pca_transformed_data = pca_selected.transform(pca_data_scaled)
kmeans_models_pca, silhouette_scores_pca = perform_kmeans(pca_transformed_data, n_clusters_range)

# Finding the optimal number of clusters based on silhouette scores
optimal_clusters_original = n_clusters_range[silhouette_scores_original.index(max(
     silhouette_scores_original))]
optimal_clusters_pca = n_clusters_range[silhouette_scores_pca.index(max(silhouette_scores_pca))]

# Descriptions of the clusters will be based on these optimal numbers
optimal_kmeans_pca = kmeans_models_pca[optimal_clusters_pca - 2]

(optimal_clusters_original, optimal_clusters_pca)
% (2, 2)
```

Listing 4: Clustering using KMeans

4.4 Logistic Regression Predictor and Random Forest Predictor

```python
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import roc_curve
# ...omit...
# Setting up cross-validation and model training
# Random Forest
rf_param_grid = {’n_estimators’: [100, 200, 300], ’max_depth’: [5, 10, 15]}
rf_grid = GridSearchCV(RandomForestClassifier(random_state=42), rf_param_grid, cv=5, scoring=’
     roc_auc’)
rf_grid.fit(X_train_scaled, y_train)

# Best models and their parameters
best_rf_model = rf_grid.best_estimator_

# AUC scores for the best models
rf_auc = roc_auc_score(y_train, best_rf_model.predict_proba(X_train_scaled)[:, 1])

# Setting up cross-validation and model training
# Logistic Regression with Lasso (L1) penalty
log_reg_param_grid = {’C’: [0.01, 0.1, 1, 10, 100]} # smaller C denotes stronger regularization
log_reg_grid = GridSearchCV(LogisticRegression(penalty=’l1’, solver=’liblinear’, random_state=42),
                            log_reg_param_grid, cv=5, scoring=’roc_auc’)
log_reg_grid.fit(X_train_scaled, y_train)

# Best models and their parameters
best_log_reg_model = log_reg_grid.best_estimator_

# AUC scores for the best models
log_reg_auc = roc_auc_score(y_train, best_log_reg_model.predict_proba(X_train_scaled)[:, 1])

(rf_grid.best_params_, log_reg_grid.best_params_, rf_auc, log_reg_auc)

```

```
% ({'max_depth': 5, 'n_estimators': 200},
%  {'C': 0.1},
%  0.9011937903901778,
%  0.8797141228748673)
```

Listing 5: Prediction using Logistic Regression and Random Forest

```
# Compute predicted probabilities for both models
y_pred_probs_rf = best_rf_model.predict_proba(X_train_scaled)[:, 1]
y_pred_probs_log_reg = best_log_reg_model.predict_proba(X_train_scaled)[:, 1]

# Compute ROC curve for Random Forest
fpr_rf, tpr_rf, thresholds_rf = roc_curve(y_train, y_pred_probs_rf)

# Compute ROC curve for Logistic Regression
fpr_log_reg, tpr_log_reg, thresholds_log_reg = roc_curve(y_train, y_pred_probs_log_reg)

# Function to find the optimal threshold
def find_optimal_threshold(fpr, tpr, thresholds):
    sum_sensitivity_specificity = tpr + (1 - fpr)
    optimal_idx = np.argmax(sum_sensitivity_specificity)
    optimal_threshold = thresholds[optimal_idx]
    return optimal_threshold

# Optimal thresholds
optimal_threshold_rf = find_optimal_threshold(fpr_rf, tpr_rf, thresholds_rf)
optimal_threshold_log_reg = find_optimal_threshold(fpr_log_reg, tpr_log_reg, thresholds_log_reg)

# Use these thresholds for making final predictions on the training set
optimal_predictions_rf = np.where(y_pred_probs_rf >= optimal_threshold_rf, 1, 0)
optimal_predictions_log_reg = np.where(y_pred_probs_log_reg >= optimal_threshold_log_reg, 1, 0)

(optimal_threshold_rf, optimal_threshold_log_reg, optimal_predictions_rf,
    optimal_predictions_log_reg)

% (0.2183535486789095,
%  0.1474335246161669,
%  array([1, 0, 0, ..., 0, 0, 0]),
%  array([1, 0, 0, ..., 0, 0, 0]))
```

Listing 6: Finding threshold which maximizes specifity + sensitivity

4.5 Multiple and Pairwise comparison using chi-square test

```
from scipy.stats import chi2_contingency

# Creating a new column to categorize respondents into the four groups
data['group_category'] = data.apply(
    lambda x: 'No' if x['disc_hire'] == 0 else (
        'Yes' if x['disc_hire'] == 1 else (
            'Predicted No' if x['disc_hire_predicted'] == 0 else 'Predicted Yes'
        )
    ),
    axis=1
)

# Creating a contingency table for health across the four groups
contingency_table = pd.crosstab(data['group_category'], data['health'])

# Performing the chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

(chi2, p, dof)
% (42.953423080743505, 2.1984561785251035e-06, 9)
```

Listing 7: Chi-square test for overall data

```python
# Function to perform pairwise chi-square tests
def pairwise_chi2_test(table, groups, corrected_alpha):
    results = []
    for i in range(len(groups)):
        for j in range(i + 1, len(groups)):
            sub_table = table.iloc[[i, j]]
            chi2, p, _, _ = chi2_contingency(sub_table)
            result = {
                'Group 1': groups[i],
                'Group 2': groups[j],
                'Chi-square': chi2,
                'p-value': p,
                'Significant at alpha 0.0083': p < corrected_alpha
            }
            results.append(result)
    return results

# Pairwise comparisons
groups = ['No', 'Yes', 'Predicted No', 'Predicted Yes']
corrected_alpha = 0.05 / 6 # Bonferroni Correction
pairwise_results = pairwise_chi2_test(contingency_table, groups, corrected_alpha)

from statsmodels.stats.multitest import multipletests

# Extracting the p-values from the pairwise_results
p_values = [result['p-value'] for result in pairwise_results]

# Applying the Benjamini-Hochberg procedure to adjust the p-values
adjusted_p_values = multipletests(p_values, alpha=0.05, method='fdr_bh')[1]

# Adding the adjusted p-values to the pairwise_results DataFrame
pairwise_df['Adjusted p-value'] = adjusted_p_values
pairwise_df['BH Significance'] = np.where(pairwise_df['Adjusted p-value'] < 0.05, 'Significant', 'Not Significant')

# Displaying the updated DataFrame with adjusted p-values
pairwise_df[['Comparison', 'Chi-square', 'p-value', 'Adjusted p-value', 'BH Significance']]
```

Listing 8: Pairwise multiple chi-square test using bonferroni method and benjamini-hochberg procedure