Perceived Discrimination in the European Labor Market: Demographic Determinants and Cross-Country Variations

Introduction

The participation of diverse labour force is crucial for the economic growth and innovation development of a country as a wider range of perspectives, knowledge, skills and experience can be brought into the country. Migrants, who are also part of the labour market, often facing unique challenges in their host countries, including legal barriers, discrimination and language proficiency. According to the international migration stock data (2024), nearly 87 million international migrants lived in Europe. Out of the 87 million migrants, there were around 44 million who were born within Europe, but living elsewhere in the region and around 40 million of non-European migrants resided in European regions (International Organization of Migration, 2020). Despite their significant presence, migrant employees experience sizeable employment gaps compared to the native employees.

The study by Giang Ho and Rima Turk-Ariss on the labour market integration of migrants in Europe (2018) found that employment opportunities for migrants would gradually converge to that of natives, but full coverage has not been observed even after 20 years. The persistent employment gap highlights migration integration complexity and suggests that discrimination, along with other structural barriers, may continue to affect the experience in workplace by the migrants. Another interesting perspective to investigate is the so called "Integration paradox". The integration paradox suggests that the higher-educated and more integrated migrants reports that they have experienced higher level of discrimination than those who are less integrated (Verkuyten, 2016). Furthermore, international migration trend shows that nowadays almost as many females as males migrate to and within Europe and females follows the new trend of migrating on their own for the purpose of searching for jobs, in contrast to the purpose of family unification as in the past (Cortinovis et al., 2020).

Therefore, this study aims to first examine the demographic factors influencing the discrimination perception among the labour force participants in Europe, focusing on the role of gender, age, educational attainment and country of birth. Specifically, the study will first examine the gender-level of inequality in terms of perceiving discrimination and then assess whether certain groups such as highly skilled migrants are more likely to over-report or under-report their discrimination perception in order to provide further evidence to the integration paradox. The study will evaluate the predictive power of the demographic factors on the likelihood of perceiving discrimination and how perceptions of discrimination vary across different hosting countries. By exploring these relationships, the objective of the study is to provide insight into the underlying factors contributing to the perceived discrimination which can be used to offer policy implications applicable to certain countries that can potentially help minimize the employment gap.

Hypotheses

Despite the statistics shows that there are almost as many female as male migrants in Europe, the study also shows that, as the female migrants first enter the hosting countries, their proficiency in hosting country language as well as familiarity in social norms impose a barrier for them to find the jobs in those fields

(Schieckoff, Sprengholz, 2021; Das, Kotikula, 2019; Raijman, Semyonov, 1997). Therefore, below is the first hypothesis relating to gender perceived discrimination:

- **H1**:Female migrants are more likely to perceive discrimination compared to the male migrants, especially the discrimination relating to "Lack of qualification", "Lack of language skills" and "No suitable job available".

The argument that supports the integration paradox suggests that highly educated migrants have more social exposure to mainstream society and thereby have more opportunity to face discrimination (Schaeffer, Kas, 2023). Another argument is that highly educated migrants might be more likely to interpret barriers or challenges as discrimination as they feel that they should have access to equal opportunity as the natives. Therefore, it is hypothesized that:

- **H2**:Highly educated migrants who have experienced tertiary education are more likely to over-report discrimination compared to migrants who are less educated.

After understanding how different factors would affect the perceived discrimination, the final step would be to assess the predictiveness of these factors on the likelihood of perceiving discrimination by the migrants among different countries. Since the factors such as educational attainment, country of birth, age, hosting country and gender all have effect on leading the migrants to experience discrimination, the third hypothesis is the following:

- **H3**:Demographic factors such as educational attainment, country of birth, age, hosting country and gender significantly predict the likelihood of migrants perceiving discrimination in Europe.

Despite that EU countries have agreed to use certain common immigration and permit rules, there are other aspects which each country can develop their own rules (European Commission, 2025). For instance, each country can decide on the number of migrants who can be admitted to the country. This implies that immigration patterns are not identical among all the European countries and the extent of discrimination perceived could be different among all the countries. Therefore, in order to see whether there are any potential opportunities to develop suitable policies and rules for addressing migration discrimination among different subsets of countries. Below hypothesis is made:

- **H4**:There are distinct groups of European countries where the migrants share simliar discrimination perception based on the demographical backgrounds of migrants in those countries.

Data description

Data

This study will utilize the dataset from Eurostat, which was collected in year 2021 and encompassed 29 countries including both EU members and non-EU members such as Switzerland, Spain, Italy and Austria through the European Union Labour Force Survey (EU LFS). The survey is performed by randomly rotating samples of persons from private households aged between 15 and 74. The dataset provides the number of migrants measured in thousand persons who fall under different combinations of gender, age, country of birth, hosting country, educational attainment and the perception of types of discrimination experienced by the labour force. After excluding the combination without number of migrants, the dataset comprises only 27 countries which include Spain, Italy, Hungary, Netherlands, Austria, Belgium, Switzerland, Greece, Croatia, France, Luxembourg, Cyprus, Czechia, Germany, Portugal, Finland, Estonia, Norway, Slovenia, Sweden, Denmark, Slovakia, Ireland, Lithuania, Malta, Poland, Latvia. There were 79697.9 thousands of observations under this data.

Description of response variable

The response variable is the discrimination that the respondent perceives, which in total consists of 9 types. While most of the types are easily to interpret based on its name, for better clarification, discrimination type such as "No suitable job available" primarily refers to job mismatch between the respondent's skills and job availability and "Never sought work or never worked" refers to no work experience.

Below shows the total observations for each type of perceived discrimination. There are 55.54% of migrants reported that they have not experienced any discrimination, and 22.43% reported they have experienced only one of the discrimination types and 20.03% reported they have experienced more than one discrimination types. Majority of the respondents who have responded experiencing discrimination reported that they have experienced multiple discrimination, in particular the age interval between 25 and 54 years who represents the core working population. Younger individuals (15-24 years) are more likely to have never sought work or worked, which is likely due to ongoing education or lack of work experience. Meanwhile, older individuals (55-74 years) appear to face difficulties in finding suitable jobs, possibly due to skill mismatches or age-related employment challenges.

Table 1: Response variables with associated number of observations (in thousands person)

Response_variable	Total_observations_in_thousands
Citizenship or residence permit	356.6
Discrimination due to foreign origin	500.1
Lack of language skills	2363.5
Lack of recognition of qualifications	1346.0
Language skills, qualifications, citizenship, foreign origin, job and other barriers	15959.9
Never sought work or never worked	9270.6
No suitable job available	1263.0
None	44268.0
Other	4370.2

Predictors

In this study, there are five predictors which are all categorical and their levels are as follows:

Table 2: Predictor levels

Predictor	Levels
Age	15-24 years old, 25-54 years old, 55-74 years old
Gender	Female, Male
Educational_attainment	Less than primary, primary and lower secondary education (Level 0-2), Tertiary education (Level 5-8), Upper secondary and post-secondary non-tertiary education
Country_of_birth	EU27 except reporting country, Foreign country, Non-EU27 countries (from 2020) nor reporting country
Hosting_country	Austria, Belgium, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland

Methodology

Since the response variables are categorical and with unordered categories, and in order to preserve the nature of data, this study will employ both the multinomial logistic regression and latent class analysis, which are the statistical methods for analyzing categorical data.

Before starting the modelling process, the correlation between each pair of predictors is checked to make sure that the model is not affected by collinearity. The method that will be employed in this case will be Cramer V due to the nature of the predictors.

Multinominal logistic regression

In order to test the hypotheses relating to the effect of predictors such as gender, educational attainment on the perceived discrimination, how the effect of one predictor varies across the levels of another, as well as the predictiveness of the variables, multinominal logistic regression is specified as follows:

let X_1=Gender,X_2=Age,X_3=Educational attainment,X_4=Country of birth, X_5=Hosting country be the predictors and the response variable with 9 nominal outcomes. The multinomial logistic regression model is presented as follows:

$$logit(Y_j) = ln(\frac{P(Y = j|X)}{P(Y = J|X)}$$

$$= \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5 + \beta_{j6}X_2X_3 + \beta_{j7}X_3X_5 + \varepsilon_j$$

Where $j = 1, 2, ..., J - 1$ and we have set the reference category for each predictor and the response variable as the following:

- Females for *Gender*, Less than primary, primary and lower secondary education (Level 0-2) for *Educational attainment*, Non-EU27 countries (from 2020) nor reporting country for *Country of birth*, Austria for *Hosting country*, None for *Perceived discrimination* and 15-24 years old for *Age*
- β_{j0} is the intercept for outcome j relative to the baseline category J
- $\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4}, \beta_{j5}$ are the coefficients of main effects and X_2X_3, X_3X_5 are the interaction terms
- ε_j is the error term

This model would be used to 1) Estimate the effect of gender on the perceived discrimination while taking into account of the interaction term between education and gender to determine if education level has also contributed to the likelihood of perceiving discrimination; 2) Estimate the effect of educational attainment on the probability of perceiving discrimination in order to validate the "integration paradox"; 3) Test the overall predictive power of the predictors. This will involve assessing how well the model fits the data and its ability to capture the relationship between the response variable and the predictors. The performance will be evaluated by using Fisher exact test and calculating the change in Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for every additional variable added to the model.

PCA and K-mean clustering

Since the goal is to find the distinct European countries which share similar discrimination perception based on the demographical background, data at the individual level does not provide direct insights into country-level patterns and broad trends. Therefore, in order to address this, the data will first be transformed into proportions by aggregating it at the country level. The proportions would be calculated for each category of the variables within each hosting country as follows:

$$proportion for each level of categories = \frac{Number of migrants in the levle}{Total number of migrants in that Hosting country}$$

This will therefore transform the categorial dataset to be numerical. Despite the transformation would result in the loss of some individual-level information such as specific combination of demographical profiles of individuals, it preserves the composition of migrant population in each country and their aggregated patterns at the country-level, such as proportion of migrants in each level of age and educational attainment, gender, different country of birth.

The transformation of categorical levels into proportions would change the dimension of the original dataset. The variables for the transformed dataset are likely to be dependent due to the calculation of proportion based on the same total observation in each country. This introduces the necessity of performing dimension reduction technique before we can proceed with clustering techniques. This study will primarily use Principal Component Analysis (PCA) due to its efficiency in transforming correlated variables into uncorrelated ones.

After obtaining the desired uncorrelated variables, this study will continue by employing the K-mean clustering algorithm in order to find the optimal clusters of countries. The optimal number of clusters will be determined using the Elbow Method, which will evaluate the within-cluster sum of squares for each value of K and the resulted clusters will be assessed using metrics such as silhouette scores.

Results

Below shows the result of Cramer V regarding the correlation between each pair of predictors. The result shows that overall, the predictors exhibit either negligible or weak associations (between 0 and 0.2), implying that the predictors are largely independent of each other and there is not a concern of collinearity. As shown by the Cramer V result, "Educational attainment" appears to have weak association with other predictors "Age" and "Hosting country". Therefore, in order to avoid potential collinearity, the interaction terms between these predictors should be included in the regression model.

Table 3: Cramer's V scores for each pair of predictors

	Country_of_birth	Educational_attainment	Age	Sex	Hosting_country
Country_of_birth	1.0000000	0.0371805	0.0233559	0.0363961	0.1326292
Educational_attainment	0.0371805	1.0000000	0.1606308	0.0308024	0.1326317
Age	0.0233559	0.1606308	1.0000000	0.0504320	0.1541855
Sex	0.0363961	0.0308024	0.0504320	1.0000000	0.0678646
Hosting_country	0.1326292	0.1326317	0.1541855	0.0678646	1.0000000

The effect of gender on the perceived discrimination was extracted from the model results. The result shows that the effect of gender is highly statistically significant (p value < 0.005) for all types of discriminations. As shown by the odds ratios which are mostly less than 1 except for "Citizenship or residence permit", females are more likely to perceive discriminations compared to the male migrants. The less than 1 odds ratios indicate that females face greater discrimination when they lack qualifications, have no prior job experience, or experience a mismatch between their skills and the job requirements compared to the males. Particularly, females are more likely to perceive "Never sought work or never worked" as indicated by the lowest odds ratio. This provides evidence to the first hypothesis that female migrants experience greater discrimination compared to the male migrants, especially due to the result of not being able to work due to family responsibility.

In order to test whether higher educational attainment would lead to over-reporting of discrimination, the effect of tertiary education is extracted below. As shown by the p value, the effect of tertiary education is not statistically significant as p value are mostly greater than 0.005 except for "Never sought work or never worked" and "Other" reason. This implies that most of the population with high education level is not statistically different from those with low education level in terms of perceiving discrimination.

Table 4: Intercept terms, p-values and odds-ratio for Sex variable (Female as reference category)

y.level	term	estimate	std.error	statistic	p.value	odds_ratio
Citizenship or residence permit	SexMales	0.0655570	0.0895751	0.7318664	0.4642501	1.0677536
Discrimination due to foreign origin	SexMales	-0.3594915	0.0808392	-4.4469958	0.0000087	0.6980312
Lack of language skills	SexMales	-0.4293914	0.0432072	-9.9379553	0.0000000	0.6509051
Lack of recognition of qualifications	SexMales	-0.7536204	0.0576410	-13.0743823	0.0000000	0.4706595
Language skills, qualifications, citizenship, foreign origin, job and other barriers	SexMales	-0.3558047	0.0195262	-18.2219501	0.0000000	0.7006094
Never sought work or never worked	SexMales	-1.4154465	0.0291794	-48.5084710	0.0000000	0.2428172
No suitable job available	SexMales	-0.5390726	0.0575099	-9.3735616	0.0000000	0.5832890
Other	SexMales	-0.2945810	0.0338589	-8.7002450	0.0000000	0.7448436

Furthermore, the less than 1 odds ratios reveals that the population with no education are more likely to report discrimination. Hence this shows that migrants with high education attainment do not over-report discrimination perceived.

Table 5: Intercept terms, p-values and odds-ratio for Educational attainment interaction terms (Less than primary, primary and lower secondary education (levels 0-2) as reference category)

y.level	term	estimate	std.error	statistic	p.value	odds_ratio
Citizenship or residence permit	Educational_attainmentTertiary education (levels 5-8)	-1.5216114	1.7947412	-0.8478166	0.3965401	0.2183597
Discrimination due to foreign origin	Educational_attainmentTertiary education (levels 5-8)	-2.1351583	2.2065849	-0.9676303	0.3332291	0.1182259
Lack of language skills	Educational_attainmentTertiary education (levels 5-8)	-0.4545019	0.6165285	-0.7371952	0.4610036	0.6347641
Lack of recognition of qualifications	Educational_attainmentTertiary education (levels 5-8)	-2.6730556	4.0555060	-0.6591177	0.5098202	0.0690409
Language skills, qualifications, citizenship, foreign origin, job and other barriers	Educational_attainmentTertiary education (levels 5-8)	-0.2822824	0.2542471	-1.1102678	0.2668836	0.7540607
Never sought work or never worked	Educational_attainmentTertiary education (levels 5-8)	-4.2199249	0.3983035	-10.5947466	0.0000000	0.0146997
No suitable job available	Educational_attainmentTertiary education (levels 5-8)	-1.9834970	1.8230383	-1.0880172	0.2765875	0.1375872
Other	Educational_attainmentTertiary education (levels 5-8)	-3.1006747	1.4822578	-2.0918592	0.0364511	0.0450188

Despite the result obtained above shows that high education level does not influence the perceived discrimination significantly, it is still important to consider the interaction between educational attainment and other predictors to see if its effect varies cross different levels of other predictors. Below results shows that for population aged between 25 and 74 years old with high educational level are more likely to experience more than one type of discrimination and "Never sought work or never worked" compared to those with age below 25. For the reason "Never sought work or never worked", it could be due to populations pursue further studies or take retirement or career breaks. It is also interesting to see that the population in Spain with high education level are more likely to perceive "Never sought work or never worked" than those who are in Austria.

Following the analysis of some key predictors such as educational attainment and Gender, the focus now shifts to evaluating the overall performance of the model. The result of Fisher's exact test shows that the p-value of the overall model is 0.00009, which suggests that there are statistically significant relationships between the predictors and response variable. In addition, for the purpose of assessing the model fit, AIC and BIC are calculated each time when a new predictor is added to the model, starting with the null model. The results are being plotted below. From the plot we can see that AIC and BIC both demonstrate a steadily decrease before the addition of "Hosting_country* Educational_attainment" interaction term. When the interaction term is added, BIC appears to increase but AIC remains decreasing. This reveals that BIC has imposed strong penalty on the complexity of model while AIC continues to favour the more complex model. After another interaction term between education and age being added, BIC starts to decrease again as the model fit has improved significantly which outweighs the penalty on model complexity. Therefore, the combination of these results suggests that the predictors are meaningful and the model is not overfitted.

Table 6: Intercept terms, p-values and odds-ratio for Educational attainment variable (Less than primary, primary and lower secondary education (levels 0-2) as reference category)

y.level	term	estimate	std.error	statistic	p.value	odds_ratio
Language skills, qualifications, citizenship, foreign origin, job and other barriers	Educational_attainmentTertiary education (levels 5-8):AgeFrom 25 to 54 years	1.0736839	0.2184530	4.914942	0.0000009	2.926139
Language skills, qualifications, citizenship, foreign origin, job and other barriers	Educational_attainmentTertiary education (levels 5-8):AgeFrom 55 to 74 years	1.3077426	0.2238707	5.841508	0.0000000	3.697817
Never sought work or never worked	Educational_attainmentTertiary education (levels 5-8):AgeFrom 25 to 54 years	3.6645967	0.3482606	10.522571	0.0000000	39.040386
Never sought work or never worked	Educational_attainmentTertiary education (levels 5-8):AgeFrom 55 to 74 years	3.6332845	0.3503233	10.371233	0.0000000	37.836886
Never sought work or never worked	Educational_attainmentTertiary education (levels 5-8):Hosting_countrySpain	0.8730675	0.2154782	4.051767	0.0000508	2.394244
Other	Educational_attainmentTertiary education (levels 5-8):AgeFrom 25 to 54 years	3.4340825	1.1411907	3.009210	0.0026193	31.002953

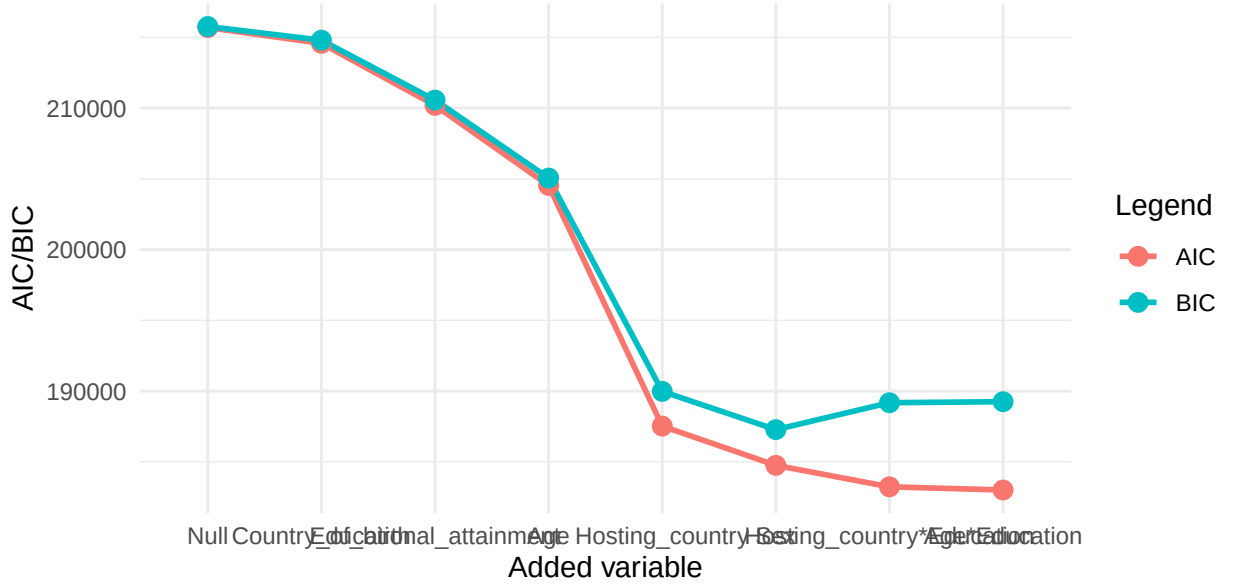


Figure 1: Change in AIC/BIC for additional predictor added once at a time

After the transformation of the dataset, the dimension of the dataset has changed to have only 27 observations (one for each country) and 20 variables (3 for Educational attainment, 2 for Gender, 9 for Perceived discrimination, 3 for Country of birth and 3 for Age). Since the variables in the tranformed dataset are the levels of the categories in the original dataset with the proportions sum to 1, this violates the assumptions of Principal Component Analysis as the variables are interdependent and data lies in a constrained space. By transforming the dataset again using log ratio, it allows the values to be unbounded and the distance between the data points becomes Euclidean. By evaluating the proportion of variations explained by each principal components, the first 3 principal compnents have been selected which explains 80% of the variation. The Elbow Method is then run for different value of k for the K-mean clustering using PCA with 3 components and the plot below shows that the optimal number of clusters is 4. The average silhouette score yielded is 0.54, which shows a reasonable clustering structure, but two of the clusters are less distinct. Below has also provided the countries under each cluster.

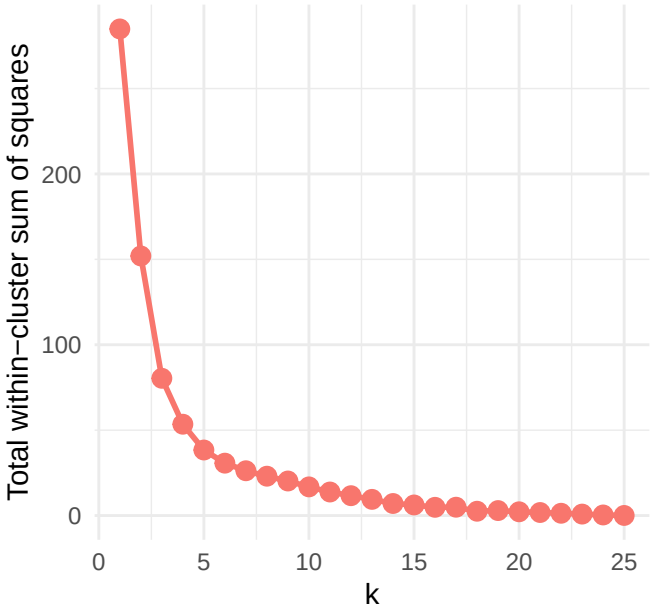
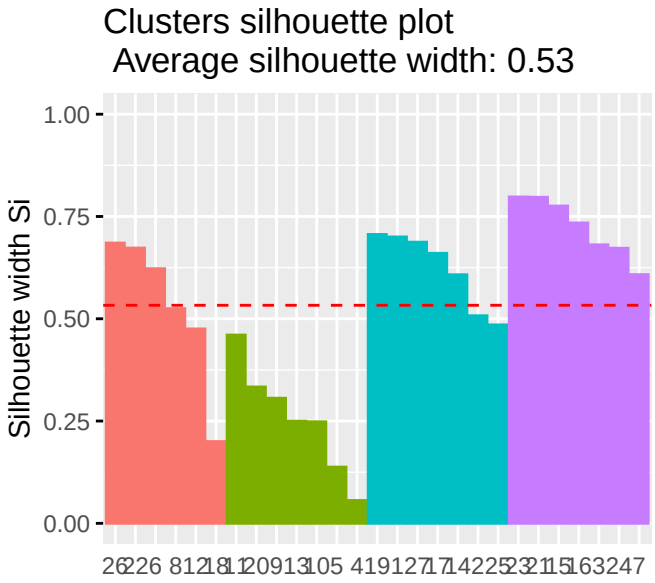


Figure2: Change in total within cluster sum of squares



Silhouette Analysis for K−Means Clustering

```
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```

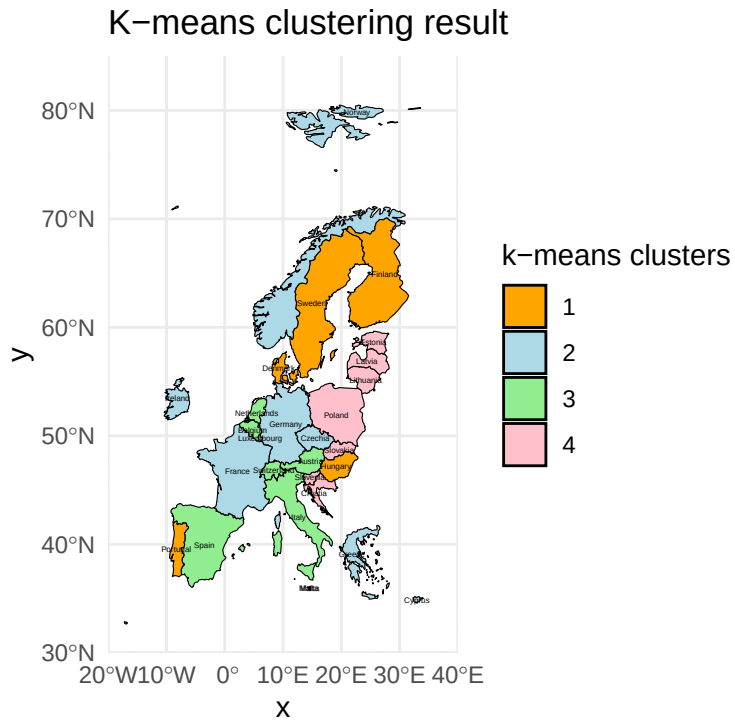


Table 7: Assignment of countries to the clusters

cluster	countries
1	Denmark, Finland, Hungary, Malta, Portugal, Sweden
2	Cyprus, Czechia, France, Germany, Greece, Ireland, Norway
3	Austria, Belgium, Italy, Luxembourg, Netherlands, Spain, Switzerland
4	Croatia, Estonia, Latvia, Lithuania, Poland, Slovakia, Slovenia

Limitations

There are several limitations to be considered in this study. Firstly relating to the structure of dataset. The categorical structure of the dataset has imposed difficulty in modelling. While the method of PCA and K-means have been chosen to test the final hypothesis, the overall performance of the modelling performance still needs to be improved due to the small size of observations and large number of variables resulted from the data transformation. Further study on dealling with categorical variable dataset will be done to improve the modelling process again and make inference on the clusters (which I have not done yet and need helpful suggestion)

References

Cortinovis et al., 2020. *Gendered migrant integration policies in the EU: Are we moving towards delivery of equality, non-discrimination and inclusion?* Available: https://cdn.ceps.eu/wp-content/uploads/2023/03/ITFLOWS_Report-on-gendered-migrant-integration-and-outcomes.pdf [10 March, 2025].

Das S., Kotikula A., 2019. *Gender-based employment segregation: understanding causes and policy interventions.* Available: https://documents1.worldbank.org/curated/en/483621554129720460/pdf/Gender-Based-Employment-Segregation-Understanding-Causes-and-Policy-Interventions.pdf [8 March, 2025].

European Commission, 2025. *EU Immigration Portal.* Available: https://immigration-portal.ec.europa.eu/general-information/who-does-what_en [10 March, 2025].

Giang H. & Rima T., 2018. *The Labour Market Integration in Europe: New Evidence from Micro Data.* Available: https://www.imf.org/en/Publications/WP/Issues/2018/11/01/The-Labor-Market-Integration-of-Migrants-in-Europe-New-Evidence-from-Micro-Data-46296 [10 March, 2025].

International Organization of Migration, 2020. *Chapter 3: Migration and Migrants: Regional Dimensions and Developments.*Available: https://worldmigrationreport.iom.int/what-wedo/world-migration-report-2024-chapter-3/europe [10 March, 2025].

Raijman R.& Semyonov M., 1997. *Gender, ethnicity, and immigration: double disadvantage and triple disadvantage among recent immigrant women in the Israeli labor market.* Available: https://www.jstor.org/stable/190228 [12 March, 2025].

Sprengholz M.&Schieckoff B.,2021.*The labor market integration of immigrant women in Europe: context, theory, and evidence.* Available: https://link.springer.com/article/10.1007/s43545-021-00279-3 [12 March, 2025].

Schaeffer M.&Kas J., 2023.*The Integration Paradox: A Review and Meta-Analysis of the Complex Relationship Between Integration and Reports of Discrimination* Available:https://journals.sagepub.com/doi/abs/10.1177/01979183231170809 [12 March, 2025].

United Nations, 2024. *International Migration Stock 2024.* Available: https://www.un.org/development/desa/pd/content/international-migrant-stock [12 March, 2025].

Verkuyten M., 2016. *The Integration Paradox: Empiric Evidence From the Netherlands.* Available: https://pubmed.ncbi.nlm.nih.gov/27152028/ [12 March, 2025]

Apendix of R implementation

```
knitr::opts_chunk$set(echo = TRUE)
suppressMessages(library(dplyr))
suppressMessages(library(knitr))
suppressMessages(library(kableExtra))
dataset <- read.csv("/Users/xiaojinghuang/Desktop/STA5071Z/estat_lfso_21obst01_filtered_en-2 copy.csv")
dataset <- dataset[,c(4:8,10,12)]
dataset <-dataset[complete.cases(dataset),]
dataset <- na.omit(dataset)
colnames(dataset) <- c("Country_of_birth","Educational_attainment","Discrimination_perceived","Age","Se
dataset <- dataset %>%
  mutate(across(-ncol(dataset), as.factor))

responses <- data.frame(
  Response_variable = dataset$Discrimination_perceived,
  Total_observations_in_thousands = dataset$Observed_population)

responses_table <- responses %>%
  group_by(Response_variable) %>%
  summarize(Total_observations_in_thousands = sum(Total_observations_in_thousands))

knitr::kable(responses_table, format = "latex", booktabs = TRUE,caption = "Response variables with asso
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 10) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")
Age <- factor(c("15-24 years old", "25-54 years old", "55-74 years old"))
Gender <- factor(c("Male", "Female"))
Educational_attainment <- factor(c("Less than primary, primary and lower secondary education (Level 0-2
Country_of_birth <- factor(c("EU27 except reporting country", "Non-EU27 countries (from 2020) nor repor
Hosting_country <- factor(c("Spain","Italy","Hungary","Netherlands","Austria","Belgium","Switzerland","
                            "Croatia","France","Luxembourg","Cyprus","Czechia","Germany","Portugal","Fi
                            "Estonia","Norway","Slovenia","Sweden","Denmark","Slovakia","Ireland","Lith
                            "Malta","Poland","Latvia"))
predictors <- data.frame(
  Predictor = c("Age", "Gender","Educational_attainment","Country_of_birth","Hosting_country"),
  Levels = c(paste(levels(Age), collapse = ", "), paste(levels(Gender), collapse = ", "),paste(levels(E

knitr::kable(predictors, format = "latex", booktabs = TRUE,caption = "Predictor levels") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 10) %>%
  kableExtra::column_spec(2, width = "12cm", latex_valign = "m")
```

```r
suppressMessages(library(vcd))
#Remove the response variable
dataset2 <- dataset[,c(1:2,4:6)]
cramers_v_matrix <- function(data) {
  n <- ncol(dataset2)
  cramers_matrix <- matrix(NA, nrow = n, ncol = n)
  colnames(cramers_matrix) <- colnames(dataset2)
  rownames(cramers_matrix) <- colnames(dataset2)
  for (i in 1:n) {
    for (j in 1:n) {
      if (i == j) {
        cramers_matrix[i,j] <- 1  # Diagonal=1
      } else {
        a <- table(dataset2[,i], dataset2[,j])
        cramers_matrix[i,j] <- assocstats(a)$cramer
      }
    }
  }
  return(cramers_matrix)
}
# Calculate Cramer's V for all pairs
cramers <- as.data.frame(cramers_v_matrix(dataset))
knitr::kable(cramers, format = "latex", booktabs = TRUE,
             caption = "Cramer's V scores for each pair of predictors") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"),
                            font_size = 10) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")
suppressMessages(library(nnet))
require(broom, quietly = TRUE)
suppressMessages(library(ggplot2))
#Fit the model
dataset$Discrimination_perceived <- factor(dataset$Discrimination_perceived,
                                           ordered = FALSE)
dataset$Discrimination_perceived <- relevel(dataset$Discrimination_perceived,
                                            ref="None")
#Since the dataset is in aggregated form -
#use weights as the number of observations under each aggregated group
model <- multinom(Discrimination_perceived ~ Country_of_birth +
                    Educational_attainment + Age + Sex + Hosting_country +
                    Educational_attainment*Age +
                    Hosting_country*Educational_attainment,
                  weights=dataset$Observed_population,
                  data=dataset,trace=FALSE)
suppressWarnings(tidy_results <- tidy(model))
#Hypothesis 1
H1_Sex <- tidy_results %>%
  filter(term == "SexMales")
H1_Sex <- H1_Sex %>%
  mutate(odds_ratio = exp(estimate))
knitr::kable(H1_Sex, format = "latex", booktabs = TRUE,
             caption = "Intercept terms, p-values and odds-ratio for
             Sex variable (Female as reference category)") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"),
```

```
                            font_size = 12) %>%
  kableExtra::column_spec(2, width = "5cm", latex_valign = "m")
# the effect of education predictor
H2_Education1 <- tidy_results %>%
  filter(term ==c("Educational_attainmentTertiary education (levels 5-8)"))
H2_Education1 <- H2_Education1 %>%
  mutate(odds_ratio = exp(estimate))
knitr::kable(H2_Education1, format = "latex", booktabs = TRUE,
             caption = "Intercept terms, p-values and odds-ratio for
             Educational attainment interaction terms (Less than primary,
             primary and lower secondary education (levels 0-2)
             as reference category)") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"),
                            font_size = 12) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")

suppressMessages(library(stringr))
H2_Education2 <- tidy_results %>%
  filter(str_detect(term, "levels 5-8"))
H2_Education2 <- H2_Education2 %>%
  mutate(odds_ratio = exp(estimate))
H2_Education2 <- H2_Education2 %>%
  filter(p.value <= 0.005)
H2_Education2 <- H2_Education2 %>%
  filter(estimate >= 0)
knitr::kable(H2_Education2, format = "latex", booktabs = TRUE,caption = "Intercept terms, p-values and
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 12) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")
suppressWarnings({
suppressMessages(library(tidyr))
#Fisher exact test
chi <- predict(model,type = "class")
contingency_table <- table(dataset$Discrimination_perceived, chi)
fisher_test <- fisher.test(contingency_table,simulate.p.value = TRUE,
                           B = 10000)

#Model fitting by adding one variable at a time
null_model <- multinom(Discrimination_perceived ~ 1,
                       weights=dataset$Observed_population,data=dataset,
                       trace=FALSE)
model_int1 <- multinom(Discrimination_perceived ~Country_of_birth,
                       weights=dataset$Observed_population,data=dataset,
                       trace=FALSE)
model_int2 <- multinom(Discrimination_perceived ~ Country_of_birth +
                         Educational_attainment ,
                       weights=dataset$Observed_population,data=dataset,
                       trace=FALSE)
model_int3 <- multinom(Discrimination_perceived ~ Country_of_birth +
                         Educational_attainment + Age,
                       weights=dataset$Observed_population,data=dataset,
                       trace=FALSE)
model_int4 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age + Hosting_country,
```

```
                        weights=dataset$Observed_population,data=dataset,
                        trace=FALSE)
model_int5 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age  + Hosting_country + Sex,weights=dataset$Observed_
model_int6 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age  + Hosting_country + Sex +  Hosting_country*Educat
                        data=dataset,trace=FALSE)

#BIC calculation
null <- BIC(null_model)
first <- BIC(model_int1)
second <- BIC(model_int2)
third <- BIC(model_int3)
fourth <- BIC(model_int4)
fifth <- BIC(model_int5)
sixth <- BIC(model_int6)
full <-  BIC(model)

#AIC calculation
nulla <- AIC(null_model)
firsta <- AIC(model_int1)
seconda <- AIC(model_int2)
thirda <- AIC(model_int3)
fourtha <- AIC(model_int4)
fiftha <- AIC(model_int5)
sixtha <- AIC(model_int6)
fulla <-  AIC(model)
AIC <- c(nulla,firsta,seconda,thirda,fourtha,fiftha,sixtha,fulla)
BIC <- c(null,first,second,third,fourth,fifth,sixth,full)
added_variables <- c("Null","Country_of_birth","Educational_attainment", "Age",
                     "Hosting_country", "Sex",
                     "Hosting_country*Education","Age*Education")
added_variables <- factor(added_variables)
table <- data.frame(added_variables, AIC,BIC)
table$added_variables <- factor(table$added_variables,
                                levels = table$added_variables)

#Plot
table_long <- pivot_longer(table, cols = c(AIC, BIC), names_to = "Line",
                           values_to = "Value")
ggplot(table_long, aes(x = added_variables, y = Value, group = Line,
                        color = Line)) +
  geom_line(size = 1) +
  geom_point(size = 3) +  # Add points to emphasize categories
  labs(caption = "Figure 1: Change in AIC/BIC for additional predictor
       added once at a time",
       x = "Added variable",
       y = "AIC/BIC",
       color = "Legend") +
  theme_minimal()
})
# Transform the dataset
par(mfrow = c(2, 1))
```

```r
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))
suppressMessages(library(sf))
suppressMessages(library(stats))
suppressMessages(library(rnaturalearth))
suppressMessages(library(rnaturalearthdata))
categorical_vars <- c("Age", "Sex", "Educational_attainment",
                      "Discrimination_perceived","Country_of_birth")
trans_data <- list()
# Loop through each level of variable
for (variable in categorical_vars) {
  data1 <- dataset %>%
    group_by(Hosting_country, !!sym(variable)) %>%
    summarise(count = sum(Observed_population), .groups = "drop") %>%
    group_by(Hosting_country) %>%
    mutate(total_count = sum(count)) %>%
    mutate(prop = count / total_count) %>%
    ungroup() %>%
    rename(level = !!sym(variable))
  data1 <- dplyr::select(data1, Hosting_country, level, prop)
  # Pivot the data to wide format
  data2 <- data1 %>%
    pivot_wider(names_from = level, values_from = prop,
                names_prefix = paste0(variable, "_"), values_fill = list(prop = 0))
  trans_data[[variable]] <- data2
}
trans_data <- Reduce(function(x, y) full_join(x, y, by = "Hosting_country"), trans_data)

#Perform PCA
suppressMessages(library(factoextra))
suppressMessages(library(compositions))
trans_data_noHC <- trans_data[,-1]
trans_data_noHC <- as.matrix(trans_data_noHC)
trans_data_noHC <-clr(trans_data_noHC)
trans_pca <- prcomp(trans_data_noHC,center=FALSE,scale.=FALSE)

#First 3 PCs will be chosen - explains 82% of the variation
trans_pca3 <- prcomp(trans_data_noHC,scale.=FALSE,rank=3)
results <- as.data.frame(trans_pca3$x)

#Find the optimal k - 4
within_cluster_ss <- c()
for (k in 1:25) {
  kmean_model <- kmeans(results, centers=k, nstart=10, iter.max=5000)
  within_cluster_ss[k] <- kmean_model$tot.withinss
}
table <- data.frame(1:25,within_cluster_ss)
table_long <- pivot_longer(table, cols = within_cluster_ss, names_to = "Line", values_to = "Value")
ggplot(table_long, aes(x = X1.25, y = Value, group = Line, color = Line)) +
  geom_line(size = 1) +
  geom_point(size = 3) +  # Add points to emphasize categories
  labs(caption = "Figure2: Change in total within cluster sum of squares",
       x = "k",
```

```
       y = "Total within-cluster sum of squares",
       color = "Legend") +
  theme_minimal() +
  theme(legend.position = "none")
kmean_model2 <- kmeans(results, centers =4, nstart=10,iter.max=5000)

results$cluster <- as.factor(kmean_model2$cluster)
suppressMessages(library(cluster))
suppressMessages(library(ggpubr))
sil <- silhouette(kmean_model2$cluster, dist(results))
sil_plot <- fviz_silhouette(sil,print.summary = FALSE)
ggpar(sil_plot, caption = "Silhouette Analysis for K-Means Clustering",legend="none")
# Display in a map
results2 <- cbind(trans_data[,1],results)
map <- ne_countries(scale = "medium", returnclass = "sf")
map1 <- map %>%
  filter(name %in% results2$Hosting_country)
#Malta is not in the record -> add
malta_coords <- matrix(c(
  14.5, 35.8,
  14.5, 36.1,
  14.7, 36.1,
  14.7, 35.8,
  14.5, 35.8
), ncol = 2, byrow = TRUE)

# Create a polygon and convert to an sf object
malta_polygon <- st_polygon(list(malta_coords))
malta_sf <- st_sf(
  name = "Malta",
  geometry = st_sfc(malta_polygon),
  crs = 4326
)
malta <- st_buffer(malta_sf, dist = 0.1)
missing_cols <- setdiff(names(map1), names(malta_sf))
for (col in missing_cols) {
  malta_sf[[col]] <- NA
}
malta_sf <- malta_sf[, names(map1)]
map3 <- rbind(map1, malta_sf)

# Merge the map data with PCA RESULTS
map_data3 <- left_join(map3, results2, by = c("name" = "Hosting_country"))

# Plot the map with K-means clusters
ggplot(map_data3) +
  geom_sf(aes(fill = factor(cluster)), color = "black", size = 0.2) +
  scale_fill_manual(values = c("orange","lightblue","lightgreen", "pink")) +
  labs(title = "K-means clustering result",
       fill = "k-means clusters") +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8),
```

```
        legend.key.size = unit(0.3, "cm")) +
  geom_sf_text(aes(label = name), size =1, color = "black", check_overlap = FALSE) +
  coord_sf(xlim = c(-20, 40), ylim = c(30,85), expand = FALSE) +
  theme_minimal()
# View countries in each cluster
clustered_df <- results2 %>%
  select(Hosting_country, cluster) %>%
  group_by(cluster) %>%
  summarize(countries = paste(Hosting_country, collapse = ", ")) %>%
  arrange(cluster)
clustered_df <- as.data.frame(clustered_df)
knitr::kable(clustered_df, format = "latex", booktabs = TRUE,caption="Assignment of countries to the clu
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 12) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")
```