King Hussein School Of Computing Sciences

Data Visualization Technical Report

**Project: Main factors associated with hiring at UN agencies**

Group 5

**Prepared by:**

 Sondos Ali (20200438)

Alksandra Aljabery (20200129)

AbdelRahman Yaghi (20200355)

Heba Najdawi (20190959)

Raghad Joudeh (20200777)

Rakan Armoush (20190009)

Omar Attawneh (20200191)

**Tutored by:**

Dr. Osama Abdel Hay

**Contributions**

Sondos Ali: Designing and writing the reports.

Alksandra Aljabery: Designing and writing the reports.

Omar Attawneh: Data cleaning(roaster), finding statistical insights from dataset.

Abdel Rahman Yaghi: Data cleaning(recruitment),data visualization, finding statistical insights from dataset.

Rakan Armoush: Data visualization, finding statistical insights from dataset.

Raghad Joudeh: Presentation

Heba Najdawi: Presentation

**Content**

[1. Abstract](#_heading=h.xoozbx2r7kpr) 3

[2. Introduction](#_heading=h.kobsa6if7adi) 3

[3. Methodology](#_heading=h.fhmiu49hpk6o) 4

[3.1. Understanding and Preprocessing data](#_heading=h.ayhfpt1muegh) 4

[3.2. Statistical Features and Characteristics](#_heading=h.993op36xx1am) 4

[4. Results](#_heading=h.i4kc1vr1kvzw) 6

[5. Appendix](#_heading=h.4d97fpg20hr6) **7**

**Table of Figures**

[Figure 1: Age to level employment rate 6](#_Toc124762102)

[Figure 2: language effect on recruitment rate 7](#_Toc124762103)

[Figure 3: The region effect on recruitment rate 8](#_Toc124762104)

# Abstract

The purpose of this project is to extract the main three observations from two given datasets, roster and recruitment, retrieved from The United Nations Agencies. By this report, we will be able to demonstrate the technical work, beginning at preprocessing the datasets and finalizing at providing a proposal for further analysis. This report will go through understanding the features and extracting information from the data. As well as providing a step by step procedure to the visualizations presented.

The major findings and conclusions depend on records from both datasets, the people who got accepted to the UN whether they were in the roster or not.

# Introduction

Preprocessing is usually a long and boring concept, where changing the raw data into a clean data set takes place. The complete cycle of data-preprocessing which will turn data into beneficial insights is called ETL[i]. In this project extracting the data was made by merging the roster and the recruitment datasets, taking into consideration the people that got recruited but not included from the roster.

Moving on to the transformation process, where checking for anomalies, finding patterns, missing values, or noisy data is the key to find the best visualizations and to meet the objectives asked. Lastly, when data is being fully transformed and ready, loading the data is the next step to take recommended decisions.

The general-purpose programming language used throughout this project is Python, and the libraries generated are the following; pandas for dataframe manipulation; seaborn and matplotlib usually used for the visualizations; and numpy, for more mathematical operations.

[i] ETL: stands for Extract – Transform – Load.

# Methodology

## 3.1. Understanding and Preprocessing data

As briefly stated before, the transformation contains the technical work. We saw that to understand the data given, then we proceeded by looking over the glossary provided and going along with each variable one by one, making sure to understand each one perfectly.

In both datasets, multiple functions have been used in preprocessing, such as info() to get the overall information about the data. Detecting missing data or null values can be handled by replacing them with the mean in continuous data or mode in categorical data. Any dependent continuous features has been recalculated to check the correctness of the dependent values. For example, the Recruitment Timeline in Days attribute, depends on subtracting the SelectionMonth from the PostingMonth, but in some instances, we found in the Recruitment Timeline in Days column, rows where the value was way above the logical amount. So we've assumed that the SelectionMonth minus the PostingMonth was what made sense.

Most frequent issue faced during the transformation was finding, and dealing with inconsistent data. It is mostly based on the person’s understanding of the data provided. A built-in function, unique(), has been used to determine any wrong or misspelled value.

## 3.2. Statistical Features and Characteristics

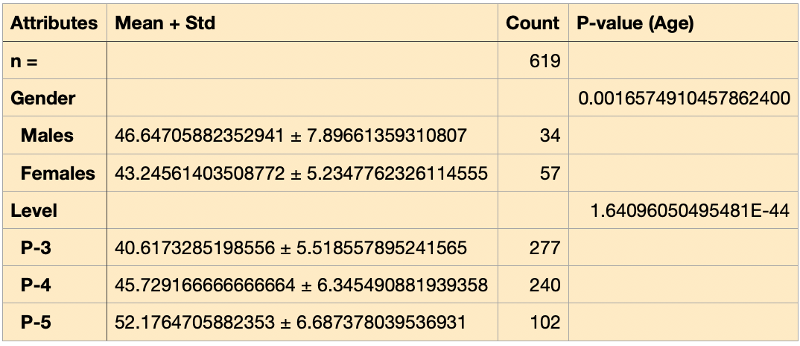
Once satisfied by the transformations done, statistical features can be derived from box plots or from p-value to find out the significance of that feature; and to find correlations, first it is recommended to look for the state of bivariate. The following tables show some statistical features found.

Table 1 in appendix shows the number of chosen Females, Males, and number of people who got chosen from every Nationality Region; in addition to the count, P-value has been calculated from chi2\_contingency.

Between Gender and being chosen the p-value with value 0.00002 is less than the significance level (0.05), we can reject the null hypothesis that there is no association, and conclude that there is a statistically significant association.

Between the Nationality Region and being chosen the p-value with value 0.3351 is greater than the significance level, we fail to reject the null hypothesis, which means we don't have enough evidence to suggest that there is an association, we can conclude that there is no statistically significant association.

Table 2 shows the mean, standard deviation, count, and p-value of ages when grouping the gender and the levels of professionalism.

*Table 2: Statistical Characteristics (Mean, Sd****[ii]****, Count, and p-value)*

Performing t-test to both genders, p-value is less than the significance level, we can reject the null hypothesis that there is no association and conclude that there is a statistically significant association.

Performing non-parametric test like kruskal to the levels, because data doesn't show normality, p-value is less than the significance level, we can reject the null hypothesis that there is no association and conclude that there is a statistically significant association.

# Results

As a result, the objectives have been met by concluding that the three main features that affected the recruitment of people who applied, are mentioned below:

1. The language fluency

Figure 2, found in the appendix, illustrates that being fluent in Chinese gives the candidate a higher possibility to get recruited.

1. The region

In figure 3, found in the appendix, it has been found that Africa got the highest percentage of recruitment, although employees from the Western Europe and Others region are the highest in number.

1. Age

Figure 1, shows the Age to level, with the addition to the number of employees in each level.

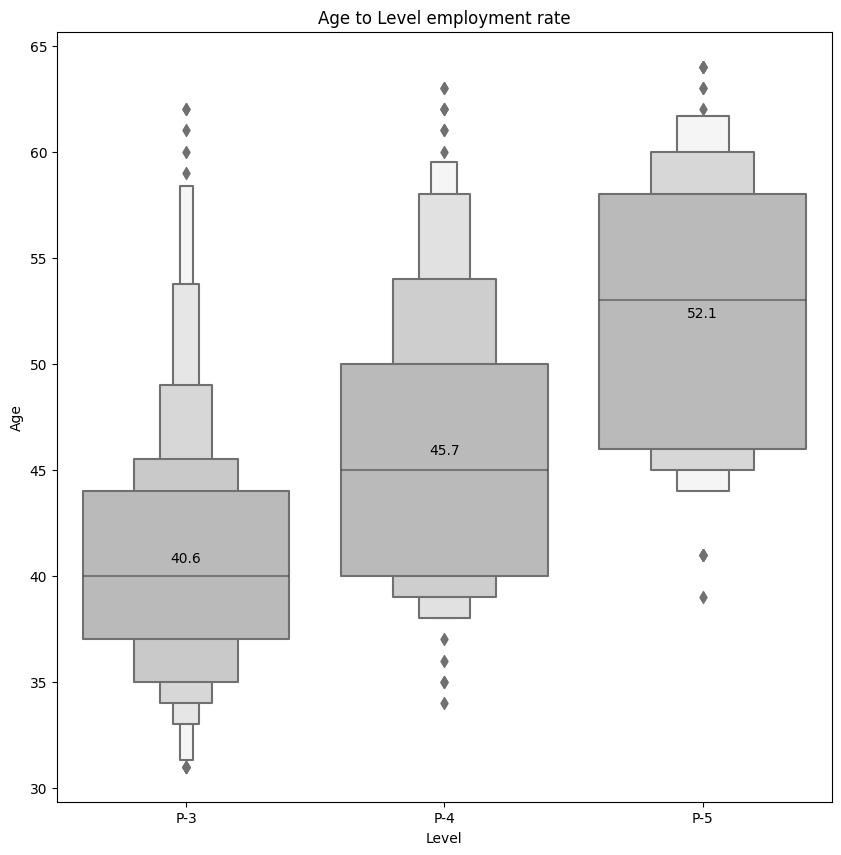
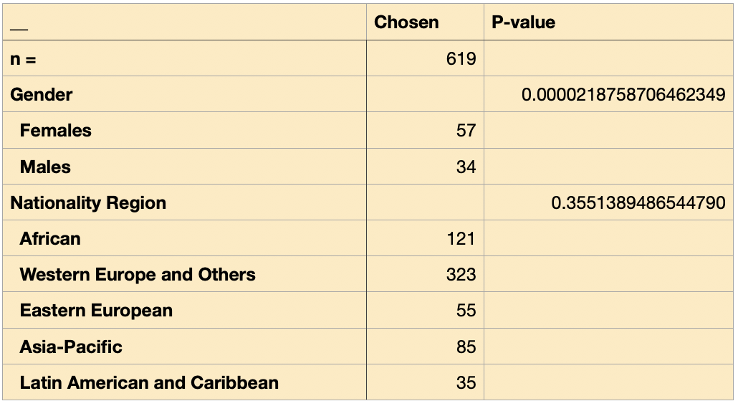


Figure 1: Age to level employment rate

[ii] **Sd:** Standard Deviation

# Appendix

Contains supplementary figures and tables which may be helpful in providing a more comprehensive understanding of the research problem.

Table 2: Statistical Characteristics (Count and p-value[i])

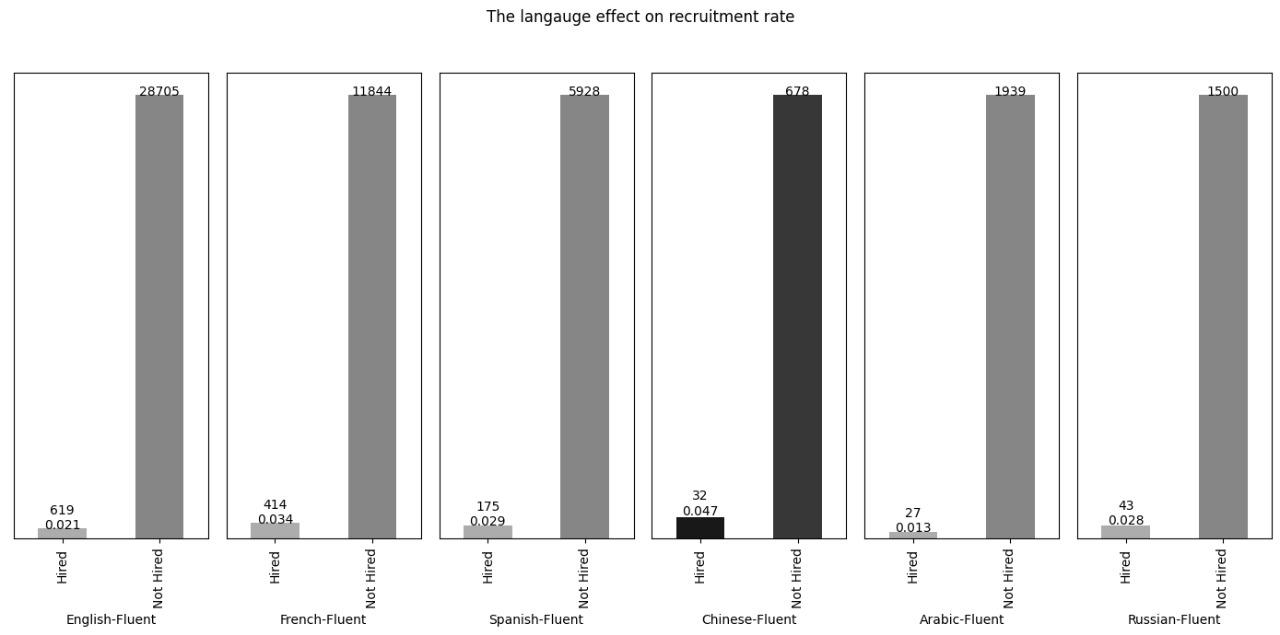


Figure 2: language effect on recruitment rate

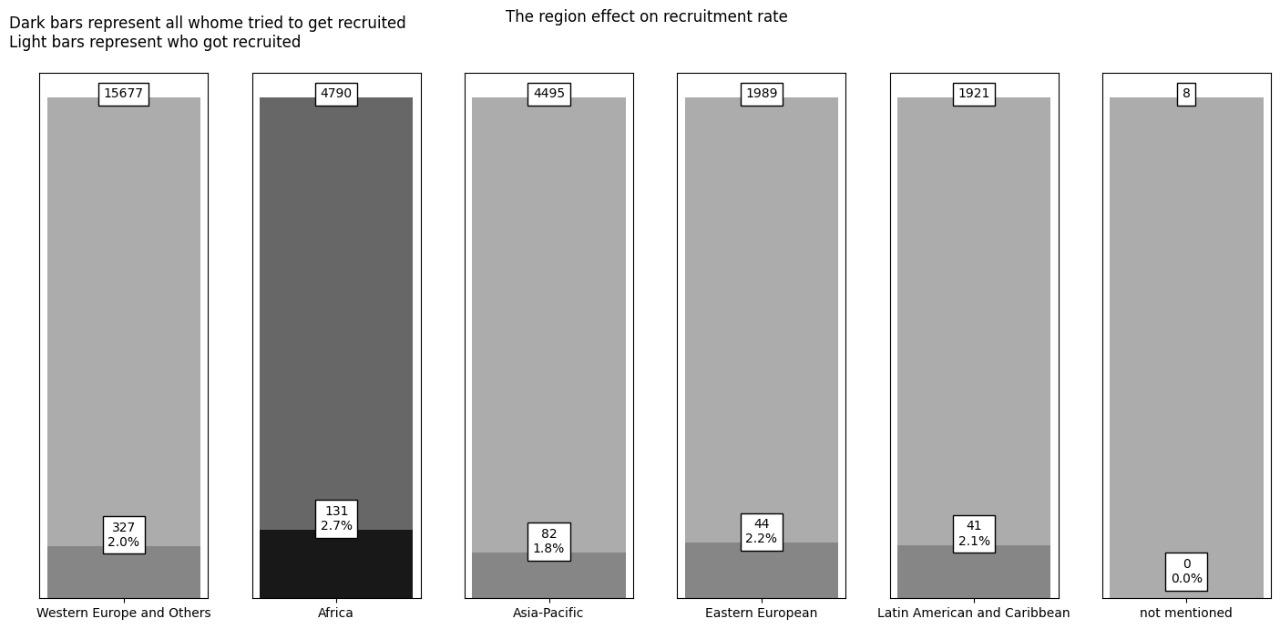


Figure 3: The region effect on recruitment rate