# Data Exploration

To make a Neural Network model that will select the best candidates, the first step is to explore the data from the used dataset. The dataset contains information from prior candidates and the hiring decisions.
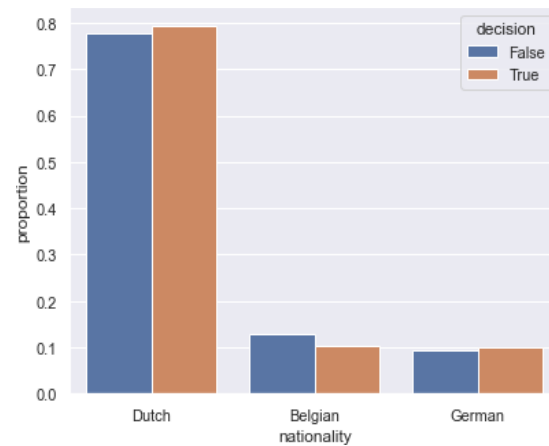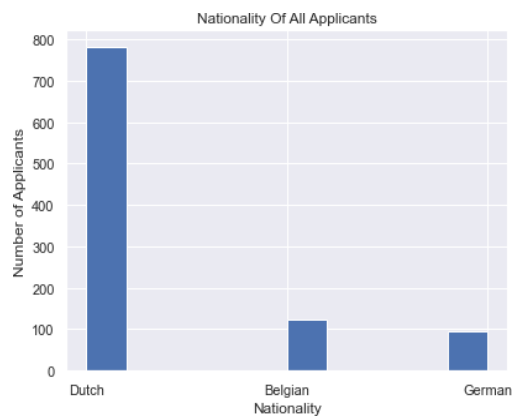
The following information has been shown out of the data exploration:

From the 1000 applicants of the company, 742 (74,2%) applicants have been rejected and 258 (25.8%) have been accepted. Zooming in each descriptors the following has been detected:
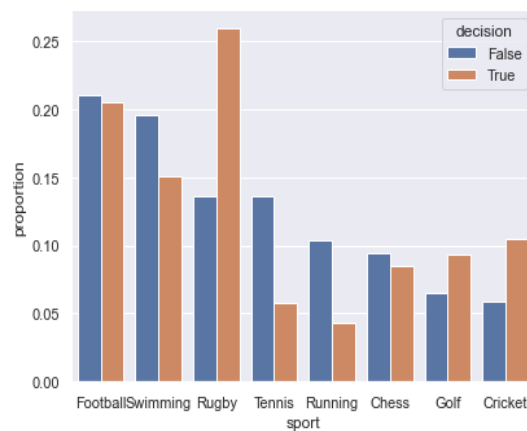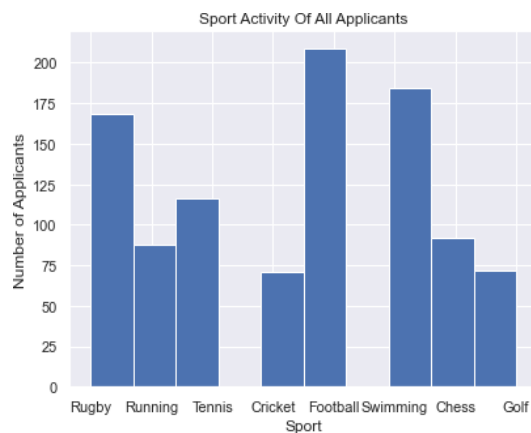
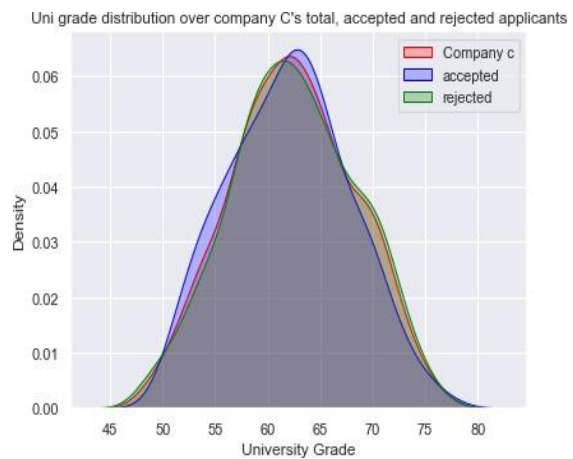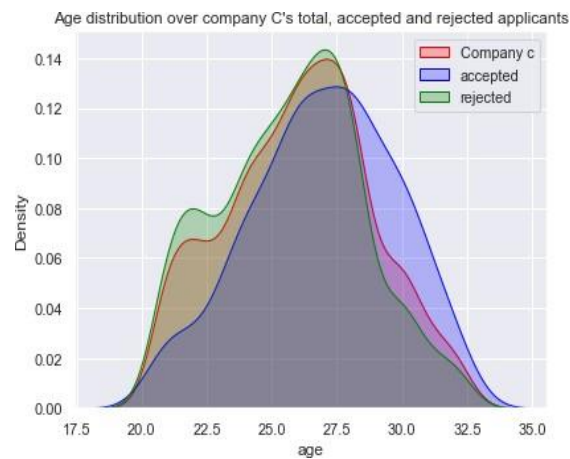### - **Gender**
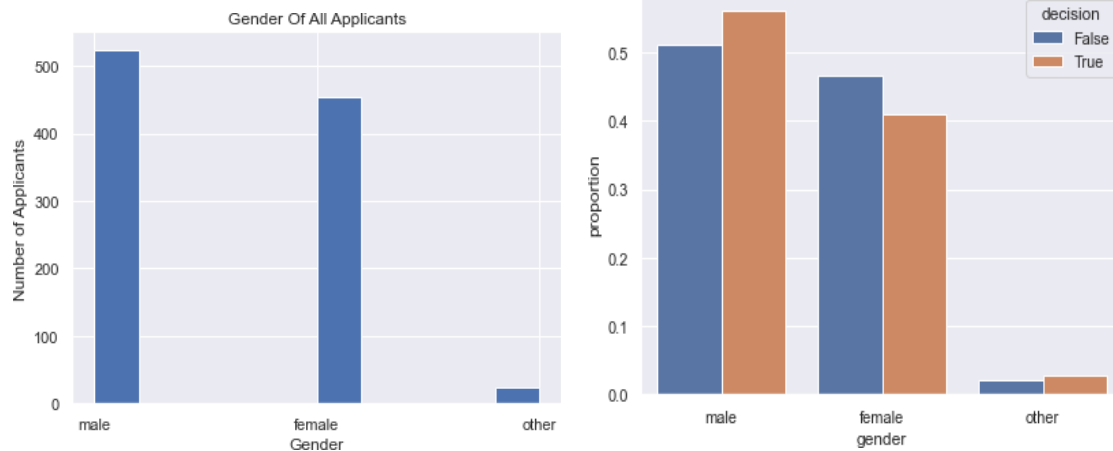
- Nationality

- Sport

Interesting observation: Despite having vastly different acceptance percentages. Rugby and Tennis have the same rejection percentages

- Age and University Grade:

Distribution of Age of total applicants and rejected applicants are very similar. Increase of accepted applicants on the older side, between age 27.5 and 32.5. Slight spike in rejected applicants on the younger side, between age 20 and 22.5.

Distributions of University Grade between all 3 are similar. Some slight increase of accepted applicants with a university grade between 50 and 57, and decrease between 67.5 and 75.
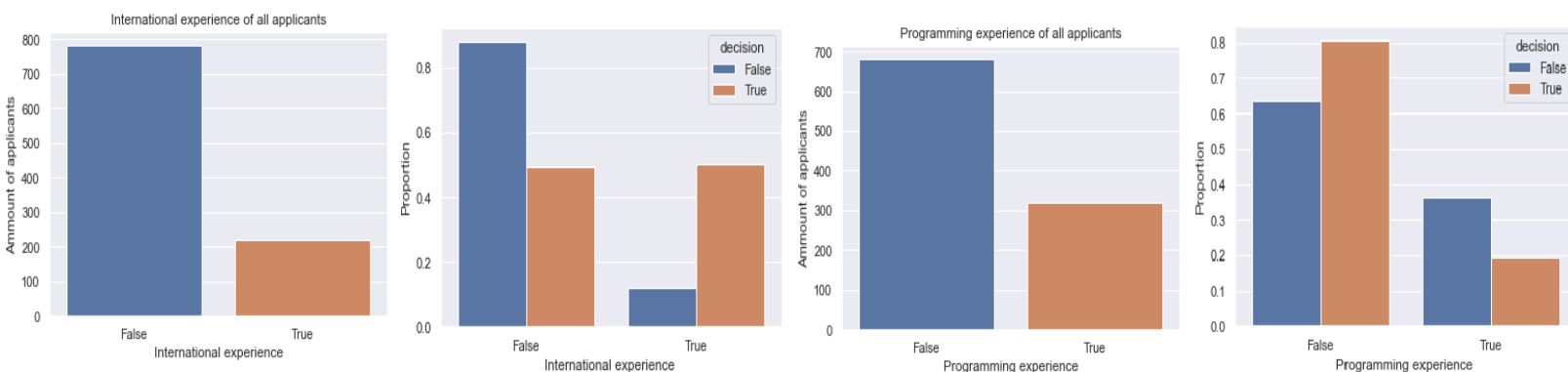
Interesting observation: The distributions for Age and University grade among applicants for Company C and all 4 companies are surprisingly similar.
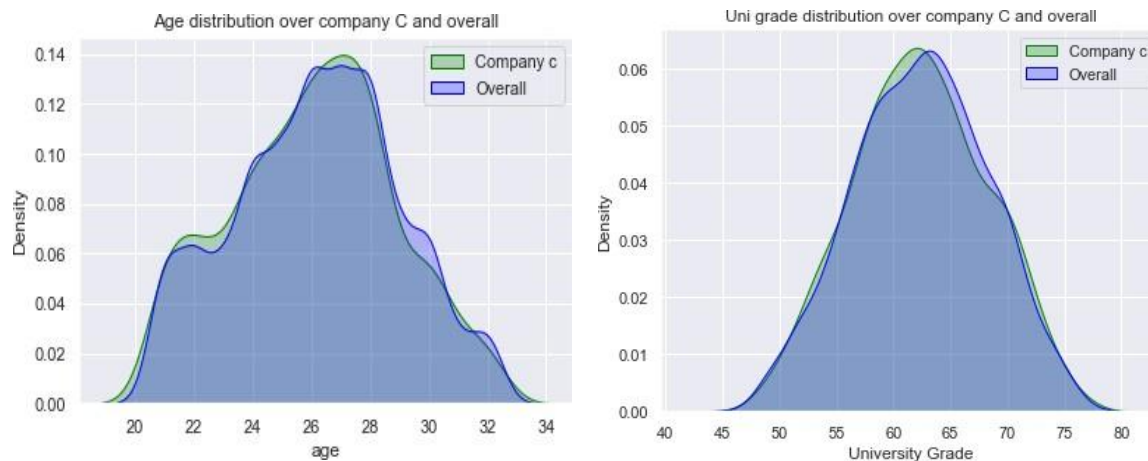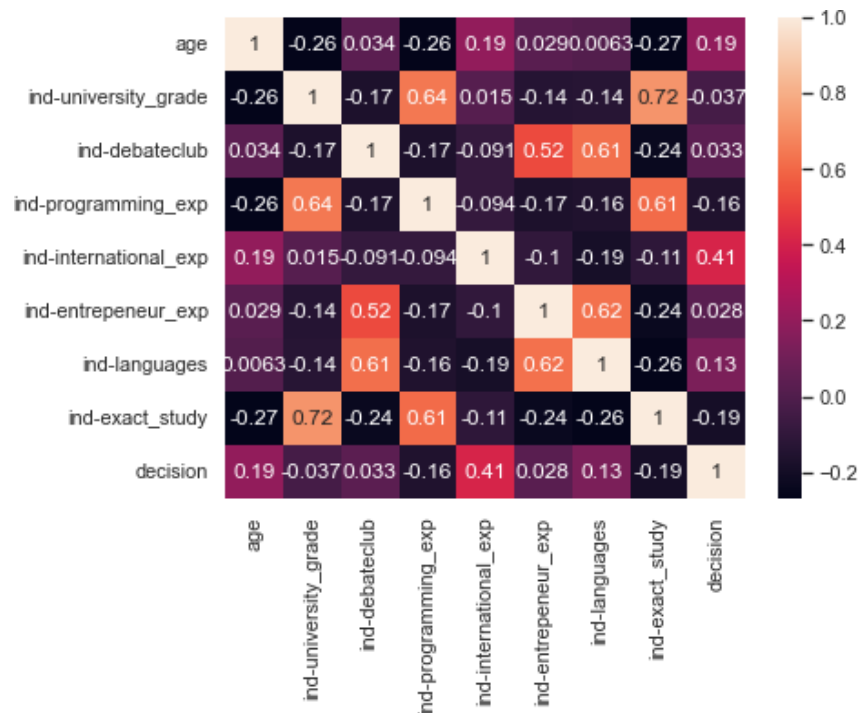

- Entrepreneurship + Scientific Study
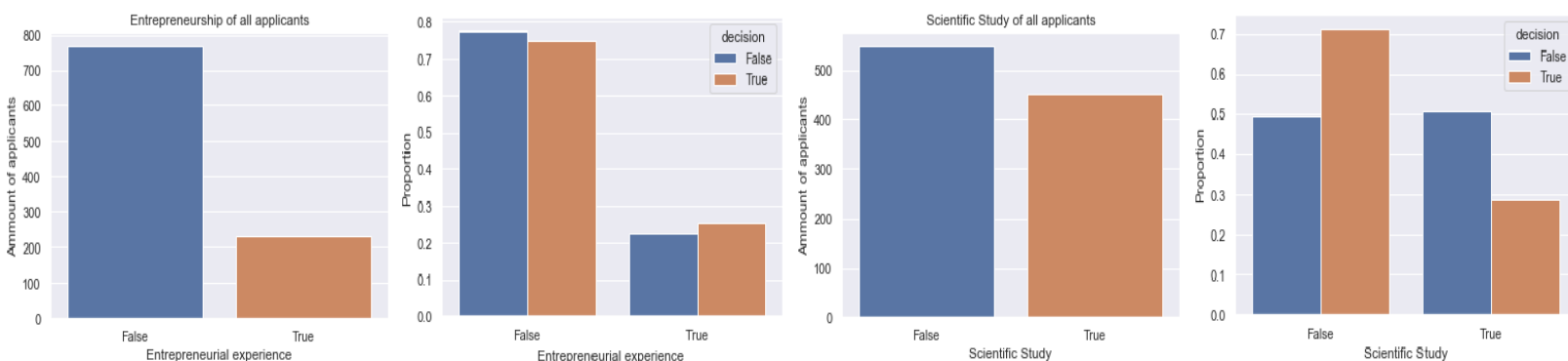
## - Programming + International experience

- Correlation

Heatmap of correlation matrix, using all possible data (i.e, numerical data) from all applicants. According to the chart, the most correlated value to decision is international experience, followed by age and languages. Other correlations are also visible, the most prominent ones being University grade correlating with programming experience and whether they did any scientific study, and being in the debate club correlating with languages and entrepreneurial experience.

- Bias:

Looking at the descriptors, there is a slight possibility of bias in Gender and Age. 41.1% of all accepted applicants were female, so there might be a slight bias against females. As for those with gender "other", there's a small number of applicants, so it is difficult to say with the metric used (what percentage of all accepted/rejected applicants are male/female/other). As for age, there is a large increase of acceptance of applicants aged between 27.5 and 32.5, indicating possible bias. For nationality, since there is an overwhelming amount of Dutch nationalities compared to German and Belgian, it is difficult to say whether there is any bias with this metric. For Sport, there is a high chance of bias. Despite Rugby being the third most popular sport, it has the highest acceptance rate of all sports. Furthermore, despite Cricket being the least popular sport, it has the 4th highest acceptance rate.

Using instead the metric of "what percentage of e.g female/male/other are accepted/rejected may give a clearer picture:
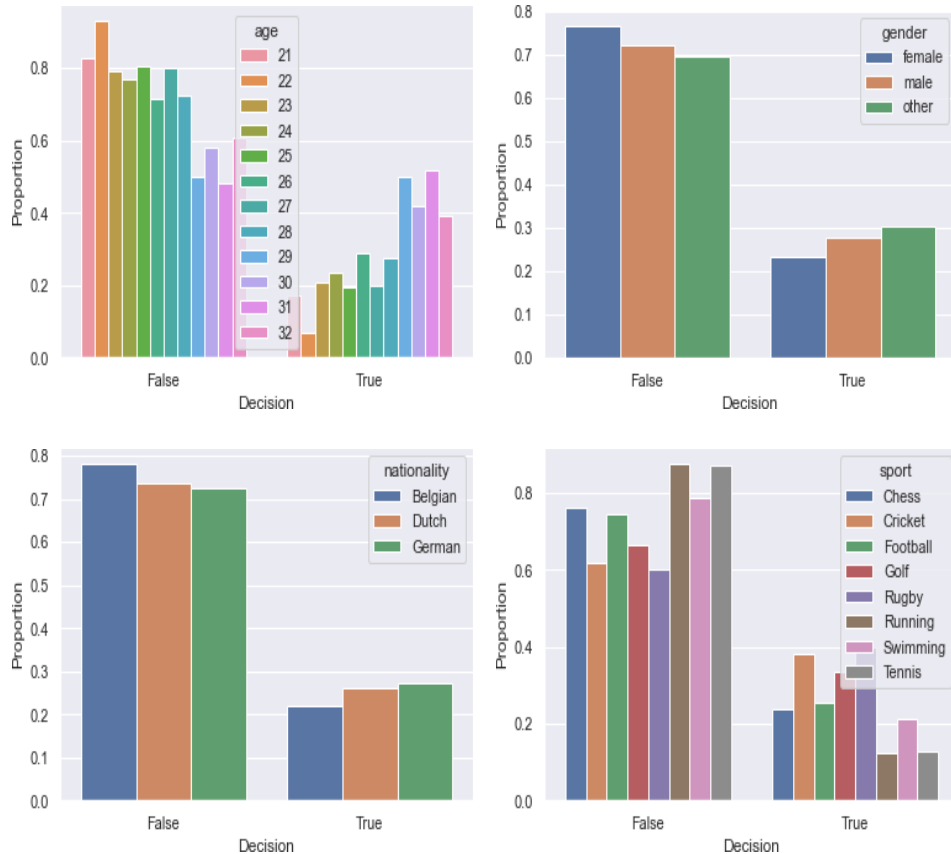
Using this metric instead, it can be seen that there is indeed bias for all descriptors. In case of age, the proportion of applicants between 29 and 32 have a higher acceptance compared to the rest. For sport, Rugby, Cricket and Golf players have a higher acceptance compared to the rest. For gender and nationality, although it seems there is slight bias towards "other" and Germans, the percentages are very close to each other (less then 10% difference) and may be therefore considered acceptable.

Indicators were not tested for bias, since that is less considered "bias" and more the capabilities and skills the company is looking for.