Problem Statement

The prevailing image of a migrant is that of a low-skilled refugee or perhaps an asylum seeker fleeing war. Less attention is paid to high-skilled migrants who are highly mobile and even fewer attention is paid to football players who play for countries other than those of their birth.

As an example, the 2018 Men's World Cup in Russia was won by the French who fielded a team that had two foreign-born players while the runners up team Croatia had four foreign-born players. I therefore intend to evaluate the hypothesis that having foreign-born players on a team leads to better FIFA rankings in men's football. I hope to provide more clarity on the boon to a host nation of having more open migration policies, with an emphasis on the social cohesion and cultural benefits of having successful men's national football teams. The target audience for this analysis is policy makers evaluating and making decisions on migration, national football governing bodies and fantasy football enthusiasts.

Dataset Wrangling:

DataFrame 1

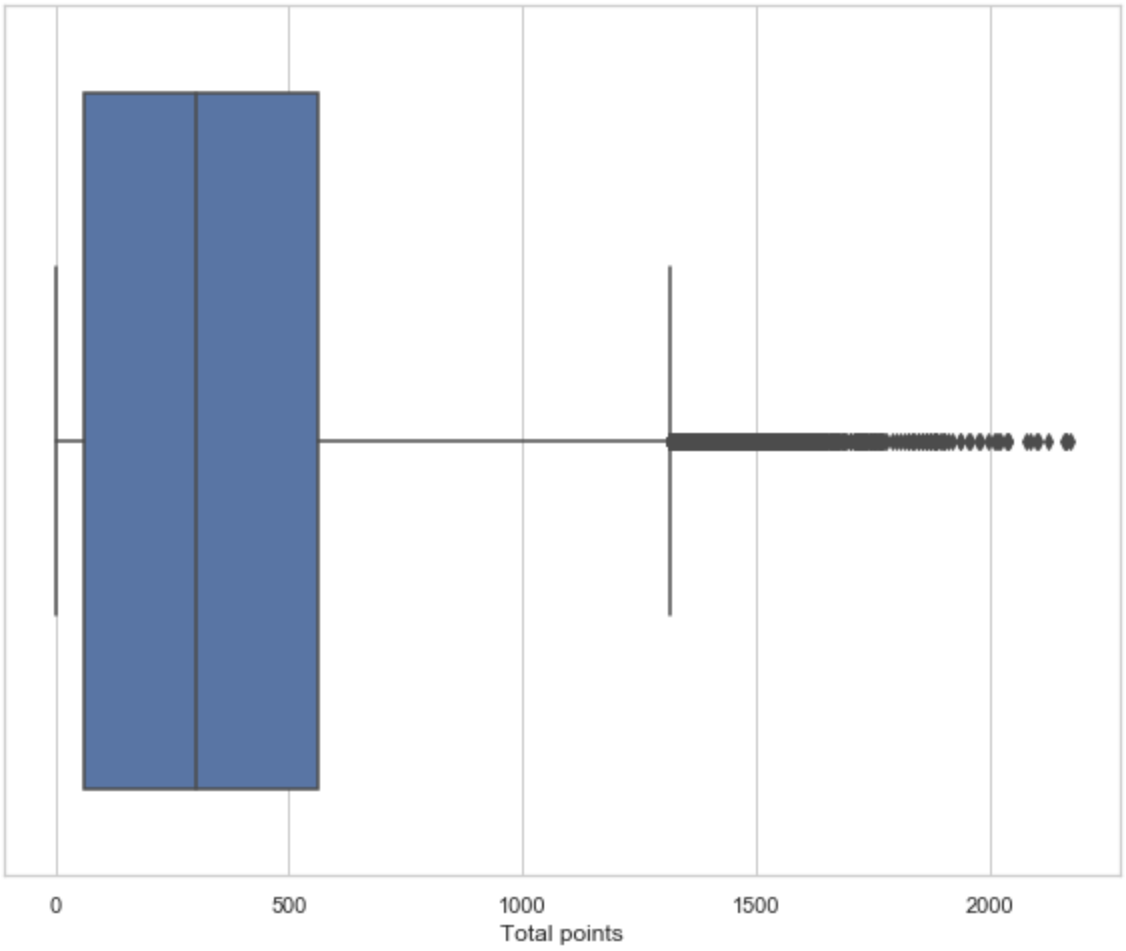
The first dataset was a .csv file compiling FIFA rankings and was found on the Kaggle website. The source material was obtained from the official FIFA website tracking countries which have participated in any FIFA tournament including confederation matches and World Cup matches.

I loaded the file as a DataFrame named [df1]. The resulting DataFrame had 9 columns, all of which had the date range 2019-12-19 to 1992-12-31. There were about 60,000 rows and nine columns. Inspection of the DataFrame revealed no missing values. I then noticed that the DataFrame had a column named "previous_points" with similar values to the column named "total_points" the only difference being the presence of extra zero digits so I decided to eliminate the former.

For my analysis I decided to focus on the columns named “country_full”, “rank_date”, and “total_points”. I then renamed the “country_full” column to “country” and ‘rank_date’ to ‘date’ in order to prepare the DataFrame for an inner merge with the second DataFrame. I also dropped duplicate values.

I created a box-plot of the “total_points” column and found it to have an outlier value at around 2000 points. After inspecting the “total_points” column, I realized that the rankings in 2018 were higher than in all other years (including 2019) which is a bit counterintuitive as shown in the next plot.

a. Box plot of the ‘total-points’ column



I came to the conclusion that this might be because that year was when the world cup was held. It’s likely that there’s a points bump for countries that participate in the tournament. I decided to use the “rank” column instead to track a country’s progress. I eventually decided to focus only on three [df1] columns for the final statistical analysis: ‘rank’, ‘date’ and ‘country’. The whittled down DataFrame still had about 60,000 rows.

DataFrame 2

The second set of data was obtained after a search through Google’s dataset website. The .xlsx file was created by three researchers at the University of Rotterdam. They pulled data from player Wikipedia profiles and other sources to create tabular data of players who have participated in each FIFA world cup tournament from 1930 until 2018. Special distinction was made between those playing for countries other than their country of birth and those playing for their countries of birth and effort was made to distinguish first-generation migrants and second-generation migrants.

I loaded the file as [df2] . The resulting DataFrame had 12 columns and about 10,000 rows. Close inspection revealed no outliers. Some column names had whitespace in between and in front of names which made it difficult to slice so I trimmed the whitespace. The columns “NationalityFather”, “NationalityMother”, “NationalityGrandmother”, “NationalityGrandfather” had significant chunks of missing values, but I decided these were not relevant to the analysis so I discarded them.

The columns to be included in statistical analysis are “NameFootballPlayer”, “'International”, “FIFAWorldCup” and “Foreign-born”. I renamed the “International” column to “country” and the “FIFAWorldCup” column (which contains the year the player participated in the tournament) to “date” in preparation for merging.

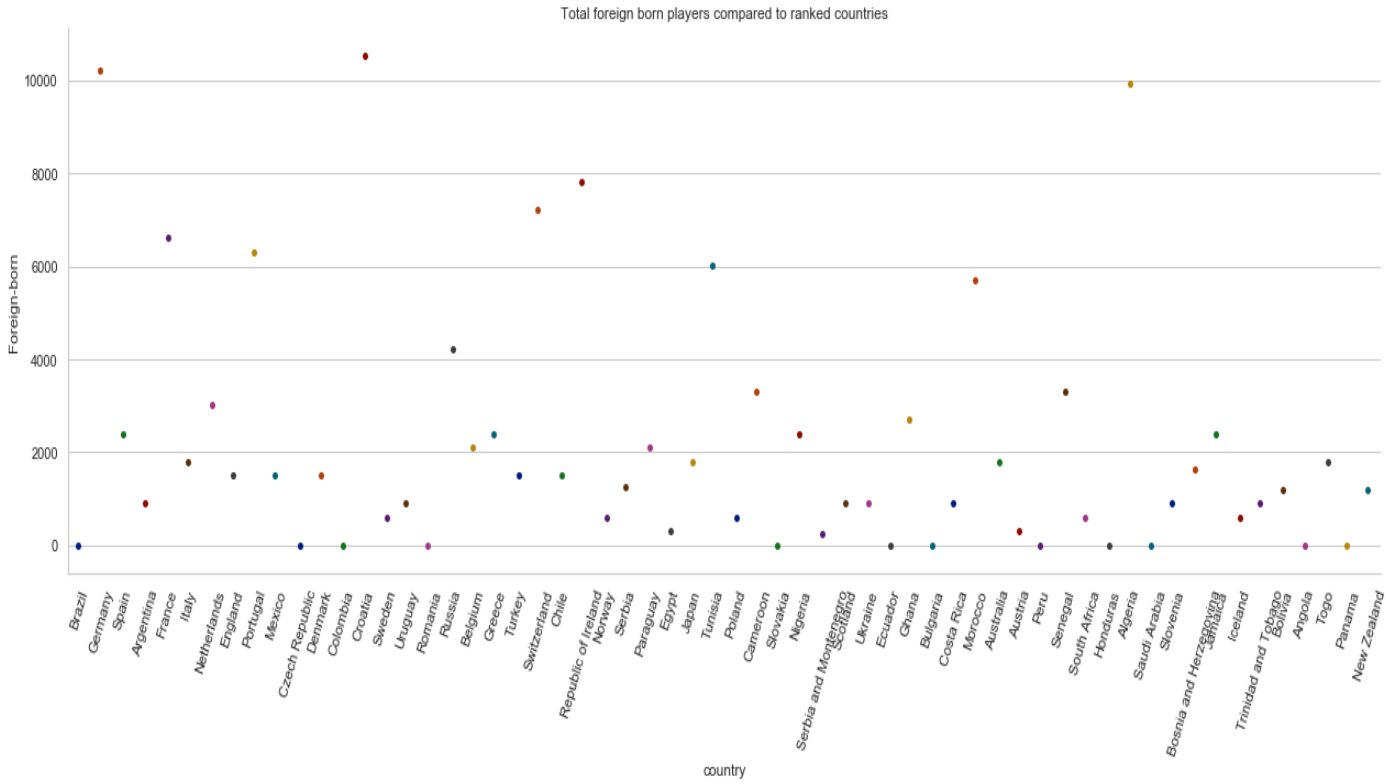
I did an inner merge on [df1] and [df2] on the “country” column in each respective DataFrame. I then dropped the extra date column from [df2]. After merging, I realized the new date column had the digits 1970 added onto the beginning of each date entry. After consulting with my mentor, we decided to do the merge with both original date columns as numeric type. I changed the date column in df1 from object to integer type then continued with the merge with df2 date column which was already in numeric form. This resulted in all the rows in the date column reverting back to the four digit year form.

The merged DataFrame [new_df] now had four columns: ‘date_y’ and ‘rank’, which were numeric, ‘country’, ‘NameFootballPlayer’ which were strings and ‘Foreign-born’ which was Boolean. In preparation for statistical analysis I converted the rank column into category type to minimize the memory requirement. After performing preliminary statistical analyses (described in detail below), I realized the new DataFrame consisted of inflated data most likely from the merging process.

In an effort to minimize the number of rows (at this point numbering at about 2 million) I sliced the rows to only focus on dates from January 1994 onwards. This is because the date range on [df2] stretched back until 1930 and the new FIFA ranking system wasn’t operational until 1993. This cut [new_df] rows by about half. I also went back to the original DataFrames [df1] and [df2] and dropped duplicates which resulted in about 1000 fewer rows in the final merged DataFrame.

I grouped [new_df] according to the ‘country’ column and aggregated the ‘rank’ column with the mean function and the ‘Foreign-born’ column with the sum function. I then made the plot shown next.

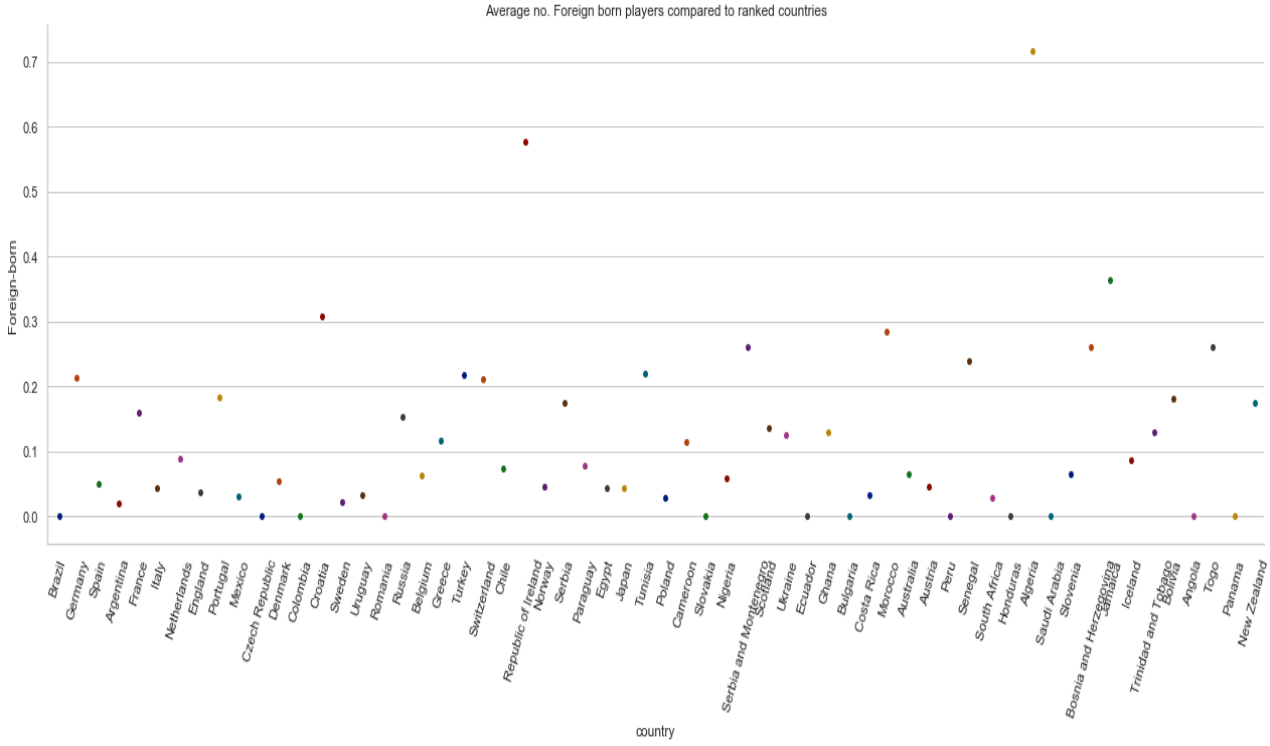
b. Categorical plot of the ‘Foreign-born’ column totals vs ranked countries



The image shows the number of foreign-born players playing for countries sorted by FIFA rank; highest average rank being Brazil and lowest average rank being New Zealand. This image shows no clear relationship on the rank of countries having no foreign-born players vs countries that do. The image shows the aggregation on the ‘Foreign-born’ column has some atypical values for the sum of foreign born players. It indicates that Germany and Croatia for example have each had a total of over 10,000 players foreign-born players between 1993 and 2018.

I decided to change the aggregation function on the ‘Foreign-born’ column to the mean instead of the sum and created the next plot.

c. Categorical plot of the ‘Foreign-born’ column mean vs. ranked countries



It is now evident that most countries who have competed in the World Cup have had at least 1 player who is foreign-born but there’s no clear relationship between that and the rank of a country.

Hypotheses

- Null hypothesis: there is no difference in ranking between countries that have more foreign-born players vs countries that have no foreign-born players
- Alternate hypothesis: countries which have more foreign-born players have better FIFA rankings

In the new DataFrame the output variable is a country’s FIFA ranking (in the ‘rank’ column) and the input variable of interest is in the column ‘Foreign-born’. Since both these columns are categorical, I chose a Chi-squared test to determine whether or not countries that have a higher number of foreign-born players have higher FIFA rankings. The test compares observed frequencies to expected frequencies and from there determines whether an input is independent of the output variable or not. This is determined by comparing the chi-square

values: small chi-square values indicate negligible dependence (independence), large chi-square values indicate a gulf between observed and expected frequencies and therefore we can be sure that values are dependent. From there, any independent inputs can be discarded from our analysis.

The resulting chi-squared value was very high: 5,042,748 (the observed frequency is very high as opposed to the expected frequency under the null hypothesis) and the p-value was 0. The large chi-squared value is unusual but the original data set was also very large with about 2,450,000 rows. This inflated data was most likely as a result of the merging process where duplicate entries in the country columns of both original DataFrames resulted in multiple combinations in the merged DataFrame.

At this point, I sliced out and decided to focus on rankings from November 1993 as described above to further minimize the DataFrame, and dropped duplicates from the original DataFrames which halved the original number of rows. After running the test again, the new value was 2,632,748 with a p-value of 0. In tandem with acknowledging the limitations of our result above, we can then reject the null hypothesis that there is no difference in ranking for countries that have more foreign-born players vs countries that don't. This is because if the assumption of the null is true, we would get the result above only 0% of the time. While there is a regrettable limitation, the structure of our original DataFrames doesn’t allow any other wrangling that doesn’t drastically whittle down our data.