Do foreign-born players increase a country’s FIFA ranking?

Teresiah Kahura: Capstone 1

Problem Statement

The prevailing image of a migrant is that of a low-skilled refugee or perhaps an asylum seeker fleeing war.[1]

Less attention is paid to high-skilled migrants including football players who play for countries other than those of their birth.[2, 3]

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 runner's up team Croatia had four foreign-born players.[4]

Objectives

I intend to evaluate the hypothesis that having foreign-born players on a team leads to better FIFA rankings in men's football.

This in turn may lead to greater social cohesion as a result of having a successful men's football team.[5]

Target Audience

National football governing bodies

Policy makers evaluating and making decisions on migration

Fantasy football enthusiasts

Data Sources

List of FIFA world rankings of men's national football teams from 1992 to 2019[6]:

.csv file obtained from Kaggle: 9 columns, about 60,000 rows.

-Named df1.

List of foreign born players playing for men's national football teams in the FIFA World Cup from 1930 until 2018[7]:

.xlsx file obtained from Google Datasets: 12 columns, about 10,000 rows.

-Named df2.

Wrangling DataFrame 1

1

2

3

9 Columns, ~60k rows

Date range: 1992 - 2019

Includes 4 dtype: object columns, 5 dtype: int64 columns

Columns to focus on:

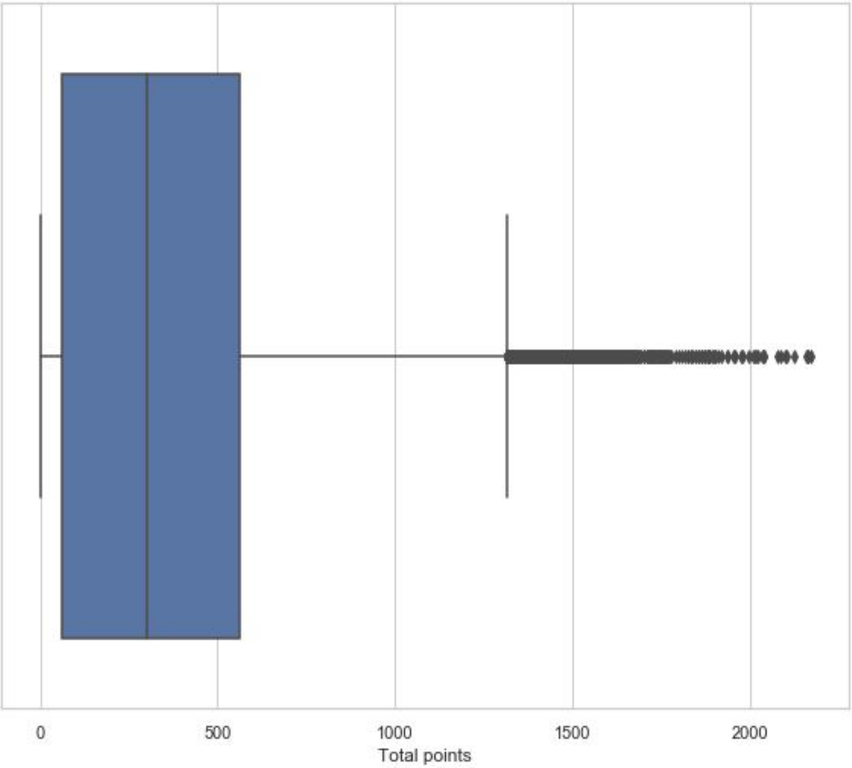
‘Country-full’

‘Rank-date’

‘total-points’

Outlier Detection

A box-plot of the ‘total-points column revealed outlier values at around 1400 points.



Wrangling Dataframe 1

- After inspecting the “total_points” column, I realized that the total points in 2018 were higher than in all other years (including 2019) which is a bit counterintuitive.
- That year was when the world cup was held so maybe there’s a points bump for countries that participate in the tournament.
- I decided to keep the 'rank' column (no outliers) to show a country's improvement instead of the 'total_points column.
- I also renamed the ‘country-full’ column to ‘country’ and the ‘rank-date’ column to ‘date.

Wrangling: DataFrame 2

1

12 columns, about 10,000 rows:

- Date range: 1930 -2018
- Included 9 dtype: object columns, 3 dtype: int64

2

Outliers and missing values

- No outliers
- Significant missing values in 4 columns which were excluded from analysis

3

Columns to focus on:

- ‘NameFootballPlayer’
- ‘International’
- ‘FIFAWorldCup
- ‘Foreign-born’

4

Renaming columns:

‘International’ renamed to ‘country’.
‘FIFAWorldCup’ column renamed to ‘date’

Merging DataFrame 1 and DataFrame 2

1

Inner merge df1 and df2 on ‘country’ column

2

Dropped ‘date’ column from df1

3

New DataFrame (new_df)

Five columns:

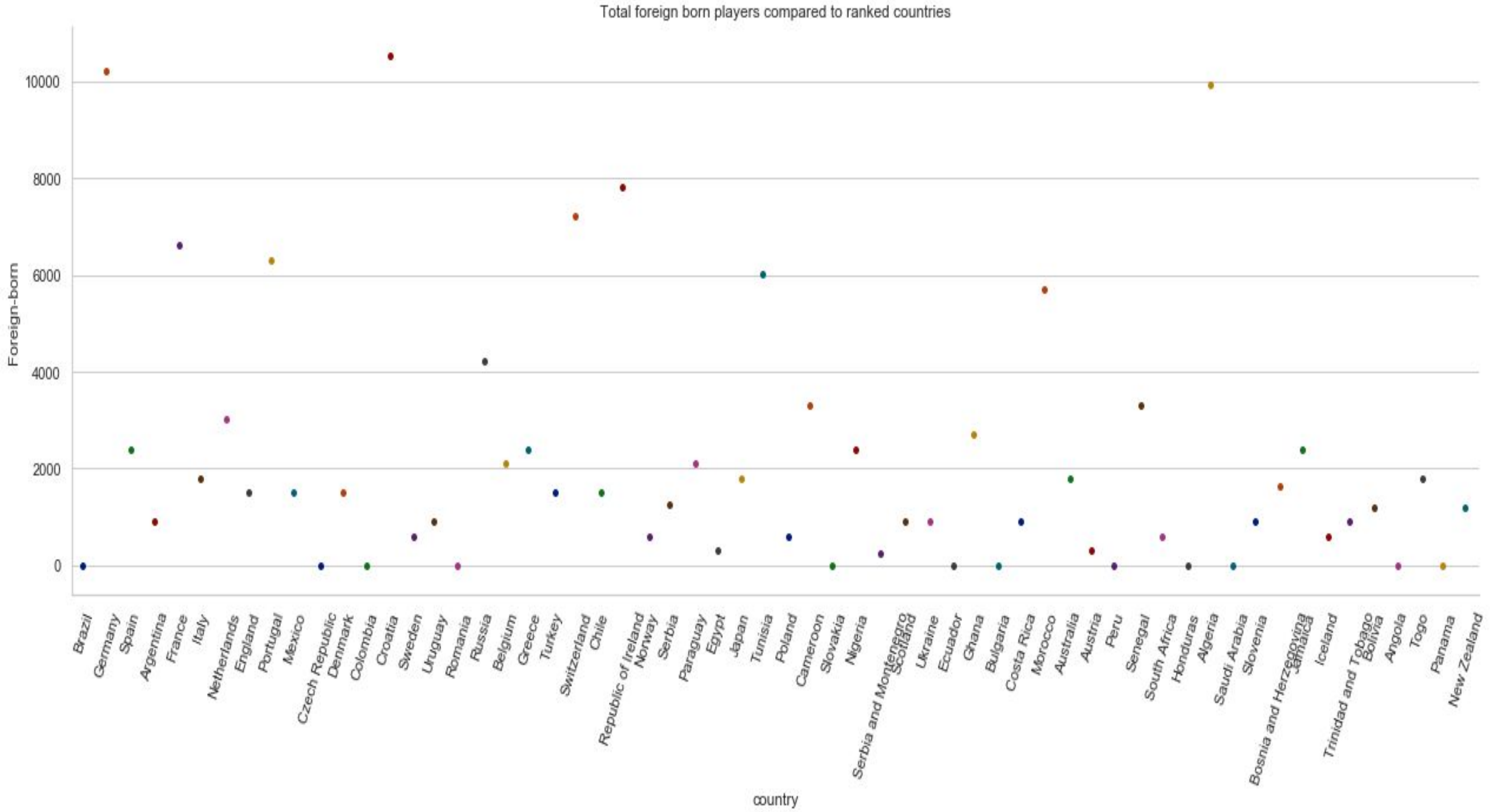
- Integer columns: ‘date_y’, ‘rank’
- Object columns: ‘country’, ‘NameFootballPlayer’
- Boolean column: ‘Foreign-born’

About 2 million rows:

- multiple instances of same country in each dataframe likely resulted in inflation.

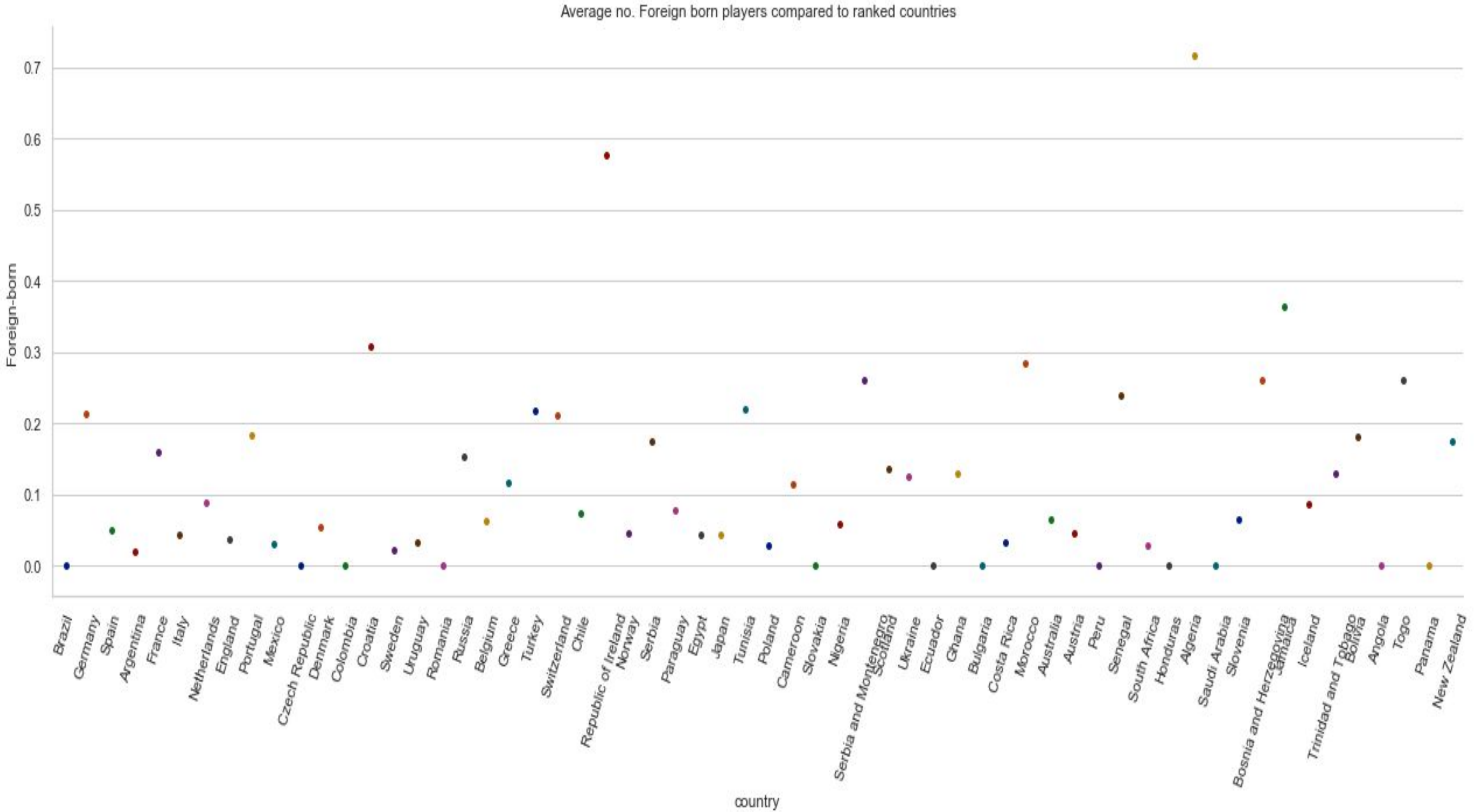
Exploratory Data Analysis

- I sliced out dates prior to December 1993 after finding out FIFA instituted new ranking system in 1994.[8]
- The resulting dataset from 1994 onwards had about 1 million rows.
- 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.
- Created plot shown on next slide.


Total foreign born players compared to ranked countries
Foreign-born
0
2000
4000
6000
8000
10000
country
Brazil
Germany
Spain
Argentina
France
Italy
Netherlands
England
Portugal
Mexico
Czech Republic
Denmark
Colombia
Croatia
Sweden
Uruguay
Romania
Russia
Belgium
Greece
Turkey
Switzerland
Chile
Republic of Ireland
Norway
Serbia
Paraguay
Egypt
Japan
Tunisia
Poland
Cameroon
Slovakia
Nigeria
Serbia and Montenegro
Scotland
Ukraine
Ecuador
Ghana
Bulgaria
Costa Rica
Morocco
Australia
Austria
Peru
Senegal
South Africa
Honduras
Algeria
Saudi Arabia
Slovenia
Bosnia and Herzegovina
Jamaica
Iceland
Trinidad and Tobago
Bolivia
Angola
Togo
Panama
New Zealand

Analysis

- The image shows the aggregation on the ‘Foreign-born’ column has some atypical values for the sum of foreign born players.
- Germany and Croatia for example have each had a total of over 10,000 players foreign-born players between 1993 and 2018.
- This is most likely as a result of the inflated dataset obtained after merging.
- I decided to change the aggregation function on the ‘Foreign-born’ column to the mean instead of the sum and created the next plot.



Average no. Foreign born players compared to ranked countries
Foreign-born
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Brazil
Germany
Spain
Argentina
France
Italy
Netherlands
England
Portugal
Mexico
Czech Republic
Denmark
Colombia
Croatia
Sweden
Uruguay
Romania
Russia
Belgium
Greece
Turkey
Switzerland
Chile
Republic of Ireland
Norway
Serbia
Paraguay
Egypt
Japan
Tunisia
Poland
Cameroon
Slovakia
Nigeria
Serbia and Montenegro
Scotland
Ukraine
Ecuador
Ghana
Bulgaria
Costa Rica
Morocco
Australia
Austria
Peru
Senegal
South Africa
Honduras
Algeria
Saudi Arabia
Slovenia
Bosnia and Herzegovina
Jamaica
Iceland
Trinidad and Tobago
Bolivia
Angola
Togo
Panama
New Zealand
country

Statistical Analysis

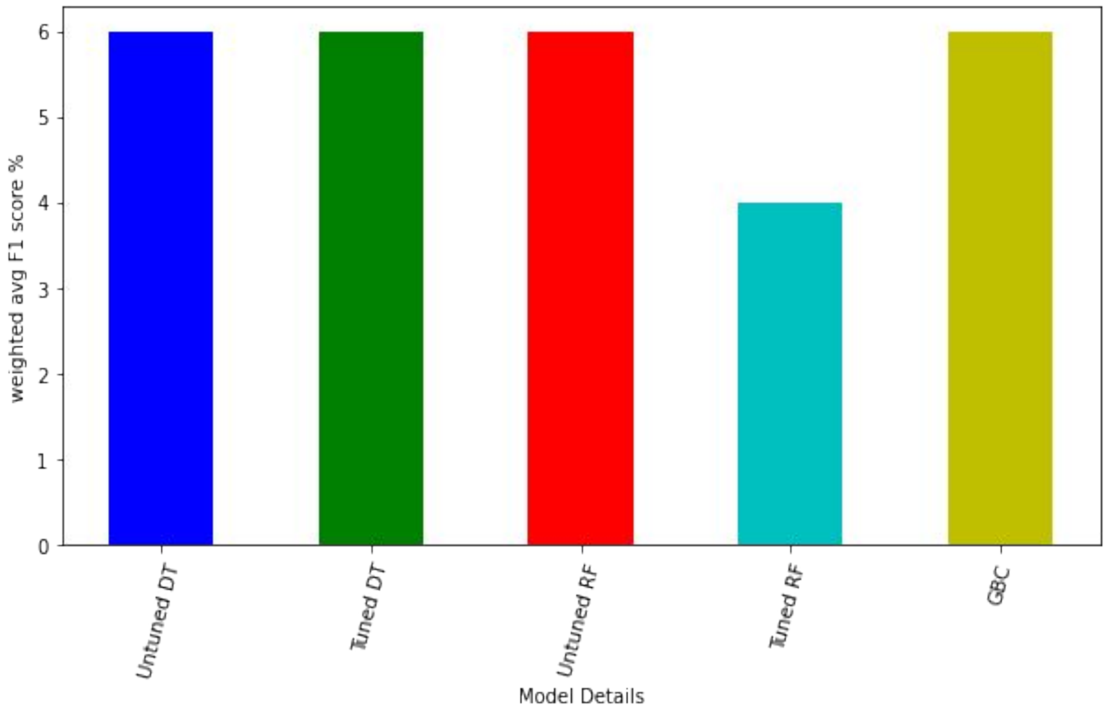
- Null hypothesis: there is no difference in ranking between countries that have more foreign-born players vs countries that have no foreign-born players.

- In the merged DataFrame the output variable is a country’s FIFA ranking (in the ‘rank’ column) and the input variable of interest is in the column named ‘Foreign-born’ (both categorical).

- I chose a Chi-squared test of independence and the resulting p value was 0 therefore we can reject the null hypothesis.

Modeling

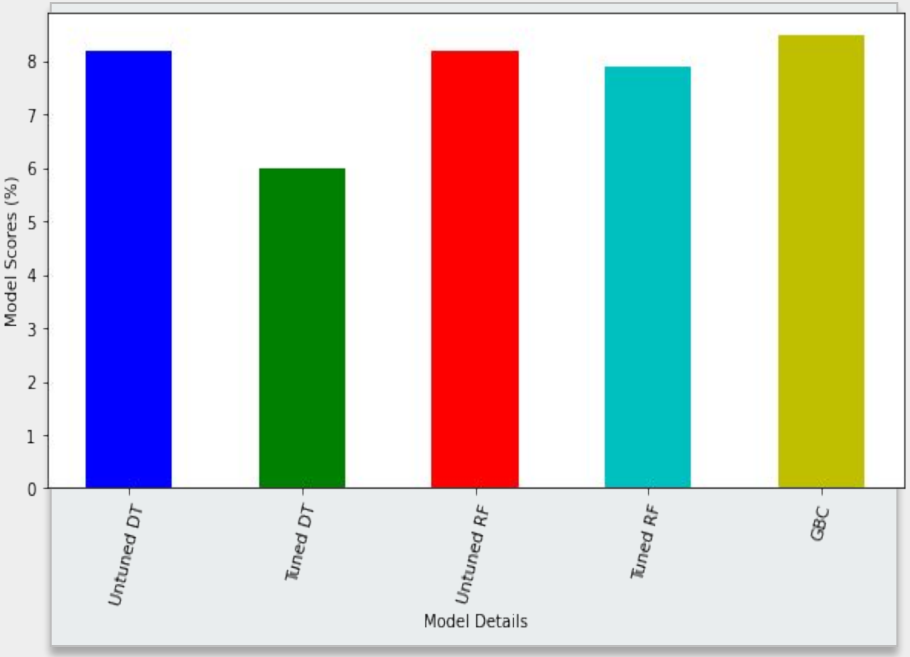
- Approached the problem as multiclass classification problem.
- Chose three models to train on an imbalanced dataset:
    - Decision Tree
    - Random Forest
    - Gradient Boosting Classifier



★ All three classifiers performed poorly with a weighted average F1 score of around 6%.

Modeling

- The classifier scores were similarly modest.
- The majority of the classifiers had a score hovering at around 8%.
- Despite the poor performance across the board, the Gradient Boosting Classifier seems to have better capability with this particular dataset.
    - Limitation: long training time



Conclusion

- All classifiers performed poorly, due to training on flawed data.
  - The matrix of features was very minimal.
  - The merging process created an inflated dataframe:
    - unavoidable instances where duplicate country entries in each original dataframe resulted in multiple combinations in final dataframe.

Future Scope

- Future scope:
  - Bolster the number of dependent features.
  - Bolster the datetime column to enable time-series analysis.