Andrew Tremante
STAT 231 - Data Science
Final Project Proposal

Mapping Value: Nationality and Its Economic Implications in European Football

Introduction

In August 2021, the transfer of English footballer Jack Grealish to Manchester City for $138 million reignited passionate conversations about the valuation of English players in the Premier League. Before this transfer, Grealish was valued at only $82 million, raising questions about why Manchester City needed to overpay by $56 million to acquire him. This deal reinforced the notion of an English Tax, which suggests that players' nationalities, particularly English, unjustly inflate their market value. This project seeks to broaden this inquiry to Europe's most prominent football leagues, utilizing data from FBref and FootStats to explore the impact of nationality on players' market values and team affiliations. Through its analysis, the project aims to offer a more nuanced understanding of the international football market, enhancing discussions on player valuation and recruitment strategies.

Data

This project will source data through HTML scraping from FBref and from the *Most Expensive Footballers 2021* from Kaggle. The Kaggle dataset contains the market values for the five hundred most expensive players in 2021, and this data will be joined with performance data scraped from FBref for the 2021 season. Also, some data will be provided by the GitHub dataset *countries_and_coordinates.csv*. This dataset will be joined with the player data to provide geographic locations for the nationalities and leagues of the players. By combining these sources, the final dataset will include performance metrics, demographic details, and financial valuations.

Analysis and Product

The main focus of the project will be on spatial data analysis and clustering, aiming to map and interpret the geographical distribution of players' nationalities and their market values. By implementing interactive Leaflet visualizations in a Shiny app and combining new clustering techniques, the project will allow users to explore different relationships between player's nationalities, leagues, and individual statistics. The web app will include three pages for different visualizations. The first will be an interactive Leaflet map that displays the nationalities of different players and allows for the filtering of players based on several performance metrics. It will help to display the relationship between value and nationality. The second visualization will be another interactive map that will display the relationship between player's nationalities and the countries where they play. This visualization should display new properties of Leaflet and

geospatial data that have not been previously covered. The final visualization will implement hierarchical clustering to show the relationship between goals, assists, and player value. The results of the new clustering algorithm will be displayed on this page of the Shiny app.

Conclusion

By combining advanced spatial analysis and clustering, this project hopes to uncover how nationality influences the careers and market perceptions of professional football players in Europe. It hopes to shed light on common debates in modern football, such as the English Tax, and also provide insight into the recruitment methodologies and approaches of the top clubs and leagues around the world.

Plan

The following is an outline of the steps needed to complete the project:

1. Scrape and wrangle data from FBref, import data from Kaggle and GitHub (April 15)
2. Research Leaflet and possibilities for Shiny App (April 18)
3. Implement visualizations (pages in-app)
    a. Visualization 1: Interactive Map 1 (April 22)
    b. Visualization 2: Interactive Map 2 (April 25)
    c. Visualization 3: Clustering (April 28)
4. Present in class and complete peer feedback (May 1, May 6)
5. Finalize app and finish report (May 7)
6. Complete project with reflection (May 15)