Table Of Contents

Report Information

This report is intended to discuss the upcoming assignment and give an overview of the Dataset being reviewed and the statistical techniques to be used in Python. The Dataset is based on the FIFA World Cup 2022 and is sporting/entertainment based.

The main data to analyse, is the Player/Squad information from each of the Football Associations attending the World Cup.

Why this data? Soccer in general is a field that is of great interest Worldwide and Data Analysis within Football has become a large field in recent years. The World Cup is the showpiece of International Football and comes around only once every 4 years, so typically gains mass appeal.

Football commentators, Newspapers, Betting Sites will all use the Player data to make predictions on the World Cup and will mainly favour National Teams who have the more experience players, who are currently playing their club football with the Top Teams in Europe. It will be interesting to get a breakdown of Players by Age, Caps (international games played to date) and where they are currently playing their Club Football (not always in their home league).

Source Data and References

The source data being used has collected from the following websites:

https://en.wikipedia.org/wiki/2022_FIFA_World_Cup_squads

https://en.wikipedia.org/wiki/Confederation_of_African_Football

https://en.wikipedia.org/wiki/Asian_Football_Confederation

https://en.wikipedia.org/wiki/UEFA

https://en.wikipedia.org/wiki/CONCACAF

https://en.wikipedia.org/wiki/Oceania_Football_Confederation

https://en.wikipedia.org/wiki/CONMEBOL

Date all above websites accessed 19th November 2022.

A List of Players attending the World Cup can also be viewed on the FIFA Website at Teams https://www.fifa.com/fifaplus/en/tournaments/mens/worldcup/qatar2022/teams.

The Game data was manually updated daily, using the following website:

https://www.flashscore.com/football/world/world-cup/#/823QwKIu/draw

Licencing Restrictions

I understand that there will be no ethical or licencing restrictions to the data being used, as this information is readily and freely available over numerous websites. According to the Source pages on Wikipedia, the data is available to view and copy. The data from the above Wikipedia pages was copied to excel. The data from Flashscore and FIFA websites were merely viewed and updated manually, if required.

Other Data Sources

There are other interesting World Cup Data source on Kaggle, however this data tends to be a historical overview of previous World Cups, see link below. Wikipedia also gives Player data on the listed pages, however all my data presented, will be arrived at using excel and Python formulas.

https://www.kaggle.com/code/shivan118/fifa-World-Cup-data-analysis

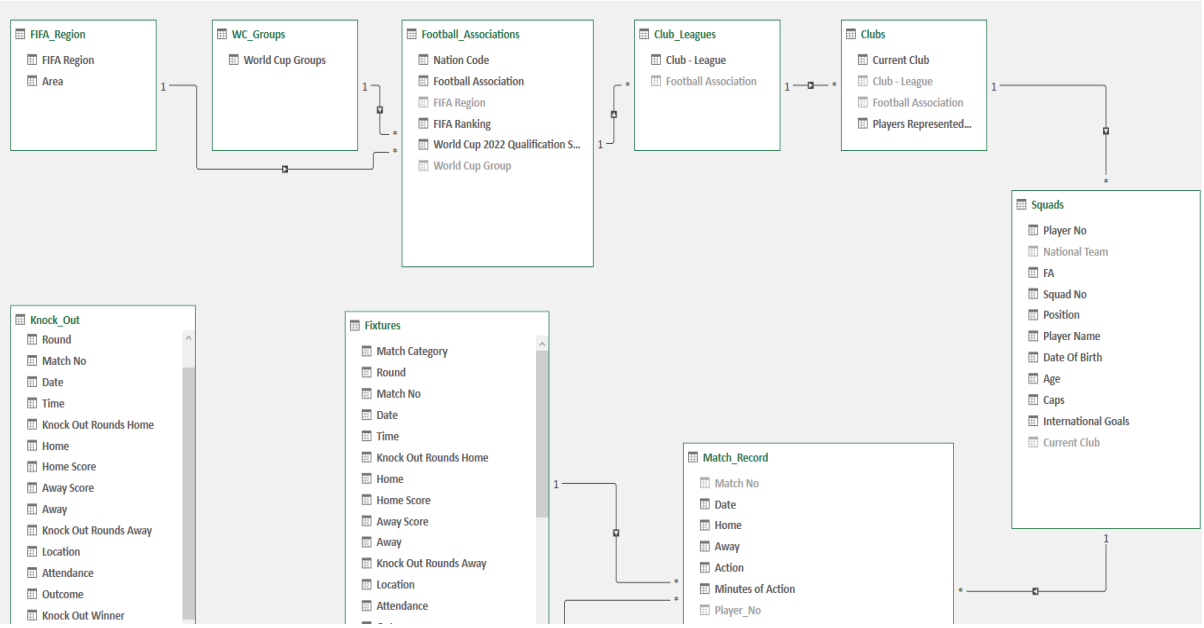
Data Cleaning

The data mentioned above, from the Wikipedia pages, regarding the FIFA Organisation and the Squad Players involved, have been copied to Excel and cleaned by creating separate tables and deleting any images and formatting the columns. Separate excel tables have been created on:

- FIFA Region
- Stadiums
- World Cup Groups
- Football Associations
- Club Leagues
- Clubs
- World Cup 2022 Matches
- Squads
- Match Report

As the focus of the report will be on the Players, a new field for **Player No** has been created in excel, starting at 1001, intended be the primary key. All the tables have been imported into Excel's Data Model, which allowed me to create relationships within the Tables and merge required data.

The data will be cleaned and exported as a csv file, called "Squads.csv". In excel, Statistical data has also been created on Player Representation allowing for presentation of pivot tables, Statistical data and charts. As a new user to Python, I wanted to have an excel version of the data for comparison purposes.



To run Python and to read the Squads.csv file, the Pandas library will be required. From this a Data Frame will be created, called "df_World_Cup". The Matplotlib library will be required to run the charts, for visualization. The datetime library will be imported to clean the "Date of Birth" column, from object to datetime. The numPy library will also be required to run the mathematical formulas. The Seaborn library will be used to create a scatterplot chart.

```
uploaded = files.upload()
df_World_Cup = pd.read_csv(io.BytesIO(uploaded['Squads.csv']),
header=None)
```

A second csv file has also been created to list the Final Outcome of each Team in the Tournament. It will be called "Outcome" and imported to Python also:

```
uploaded = files.upload()
df_Outcome = pd.read_csv(io.BytesIO(uploaded['Outcome.csv']), header=None)
```

Data Cleaning in Python

Once the data is uploaded, it will not require a refresh within Python. The csv file will be loaded without a header and to show the proper presentation in Python, a Dictionary will be created to assign headers to the Data Frame as:

- National_Team_ID
- National_Team
- Player_No
- Squad_No
- Position
- Player_Name
- Date_Of_Birth
- Age
- Caps
- Goals
- Current_Club
- Football_League
- FA
- FIFA_Region

When imported to Python, Player_No will need to be set as the index (Primary Key) and it will be important to check the index. There should be 830 rows in total. Any rows with missing or incomplete data will be dropped.
The National Team data will import with Country name and Code in the 1 column (eg..argentina ARG), so this needs to be split into National Team and Code and then drop the Code Field, as it already exists under National_Team_ID. The National Team name will need to be capitalised and check that the Date of Birth column is set to European date time

Please note, the term “National Team” used for player information is interchangeable with “Football Association” and “FA”.

Aims

The main aim of the project is to assess the players in terms of age ranges and international match experience (Caps). As the World Cup is a Tournament played once every 4 years, this presents a lot of opportunities for young Footballers to experience their 1st World Cup.

Show Player experience. The statistical techniques used in Python, will be:

- Create a Function and IF statement to categorize player by number of Caps won to date
- Use this to create a new column, called “Experience” to define the Player’s Cap Experience
- Create a Mini Date Frame to show only:
- Player Name, Caps, National Team, Experience and Age
- Group and count this data by Experience and visualise on a chart.
- Show a list of the top 20 most experienced Players
- Run a Max query to show the Player with the most caps and list that players information.
- Create a scatterplot to visualize the Age of players in relation to match experience (Caps)
- Show a Distribution curve on Player’s Experience

Show Player Age Data. The statistical techniques used in Python, will be:

- Run a describe query to find the Mean and max of Players Ages
- Create a Function and IF statement to categorize Players by Age
- Use this to create a new column, called “Age Range” to define the Player’s Age grouping
- Group and count by Age Range and visualise the data
- Find the Oldest Player and show Player’s information
- Plot a bar chart and distribution curve to show visualization of Player Age Ranges

Show the Top Players based on Goals scored to date. The statistical techniques used in Python, will be:

- Create a Mini Data Frame to breakdown information
- Give a calculation to show Goals / Caps and add this as a new column called “Goals Per Game”
- Create a new variable to show only, where players have 20 or more Caps (any less may skew the data)
- Sort to show the Top 10 Players only by Goals Per Game

Show Player Data on the last 4 Teams in the Tournament:

- Load the Outcome csv file and create a Data Frame called “result”
- Clean and Merge the Data Frame with the df_World_Cup
- Create a Mini Frame called “df_Last_Four” to Show the Last 4 Teams, ie Winner, Runner Up, Third and Fourth
- There should be 104 Players
- Compare the Mean, Median and Mode between the original List and the Last 4 List

Show Data on Football Clubs. The statistical techniques used in Python, will be:

- Create a new Mini Data Frame showing: Current_Club, Football_League, FA and FIFA_Region
- Group and Sort by highest Value, showing the Top 20 Football Clubs most affected by the World Cup
- Show the mean, median and Mode of the data
- Run similar queries to show data by Football League

Typically, a Football Season in Europe runs from August to May, with the World Cup usually played in June. However, this year the World Cup is being played from November to December and the Football Leagues of Europe have been suspended, until the tournament is completed. This could result in huge disruption for many clubs, as this affects Team Training, possible injuries, revenue, etc. For example, the German Bundesliga has a winter break each year in January and will now also have a break from 20th November to 31st December, for the World Cup, so no German Bundesliga games will take place between 14th November 2022 and 19th January 2023. Hopefully the above data will give a breakdown of where the players at the World Cup are currently playing their club Football and what leagues / Clubs will be affected the most.

Objectives

The objective regarding the Players is to show data on experience, age and Goals per game, to give an overview of the Central Tendency and show visualization of the data. By showing the experience of the players in terms of Caps won, within a particular range, it can visualize a better breakdown of the information and it will be based on the following:

Caps <= 20, "Newcomer",

Caps >= 20 and Caps < 40,"Novice",

Caps >= 40 and Caps < 80,"Regular",

Caps >= 80 and Caps < 101,"Master",

Caps >= 101 = "Legend"

By creating the above, we can get an idea if this is a Tournament of Newcomers, or more experienced players.

Using a similar analysis for the Age Range of Players, we get a breakdown of Player age brackets, to get an idea if this is a Tournament of Youth or more senior players. The age breakdown will be shown, as follows:

Age < 18: = 'Under 18'

Age >= 18 and Age <= 21: = '18-21'

Age > 21 and Age <= 25: = '22-25'

Age > 25 and Age <= 29: = '26-29'

Age > 29 and Age <= 33: = '30-33'

Age > 33 and Age <= 37: = '34-37'

Age > 37 and Age <= 42: = '38-42'

Age > 42: = 'Over 42'

The final objective is to show the Football Clubs with the most players assigned to international duty. From this data, we can see which Club Teams and their respective Football Leagues will be most affected by this World Cup, which is unusually being played this year during the European Football season.

By using the “Outcome” csv file, we can compare the data between the original list of players and only those players still involved by the Semi-Final.

Methodology

I intend to run the same data within excel, using the Squad Table in the Data Model and create formulas to run those IF statements and use Pivot Tables and charts to visualise the data, again for comparison purposes.

Once the Data has been cleaned in excel and uploaded and cleaned in Python, it will be interesting to run similar data requests using various libraries within Python and to transform the data for Analysis purposes.

Result Expectations

Prior to looking at the data, in reference to the Football clubs, I would expect the Top Clubs in Europe to mainly have most players drafted to the World Cup. With the Premier League in England, La Liga in Spain and the Bundesliga in Germany, I would expect those leagues to be impacted the most. Those are the leagues with the biggest financial Budgets and tend to scout players from around the Globe.

On a Club level, I would expect Bayern Munich, Paris Saint-Germain, Barcelona, Real Madrid, Manchester United, Manchester City and Liverpool to be the most impacted clubs, again based on their success, financial budgets and global reach.

On a Player lever, I expect most Managers to bring the more experienced Players as the majority of their 26-man squad. As it generally takes 2 years to Qualify for the World Cup, managers tend to bring the player who have played in a Tournament before and who already helped the National Team to qualify. Based on this I would think that most players will have more that 40-50 caps and be around 25-30 years old. Most players are only breaking into the Club scene in their early 20's, it usually takes a few years to become an established international.

The expected outcomes with Python, is to:

1. Clean the Data
2. Format the data
3. Group and sort data
4. Merge data
5. Provide statistical analysis
6. Show the results of the data in terms of visualization

As a new user of the Python software, I am hoping to get a basic understanding of the powerful data analysis tools available within Python, which will lead to further research into its full capabilities.