INVESTIGATE A DATASET REPORT

For the ***Investigate a Dataset*** project, I have chosen the European Soccer Database (**ESD**), compiled by Hugo Mathien, hosted on Kaggle. This report will cover all the aspects relating to the data analysis process focused on the ESD. The report consists of the following sections; an **Outline**, **Questions**, **Data Wrangling**, **Visualisations** and a **Conclusion**.

The Outline section will look at the ESD in its raw form, and include a discussion on a few preliminary SQL[1] steps that have been taken to extract the ESD data in a form which makes it more suitable for analytical purposes. The section on Questions will provide a selection of questions that have been posed about the ESD data.

The Data Wrangling section will highlight and look at all the steps, procedures and processes that have been taken in order to ensure that the data is consistent, easier to work with and that irrelevant and/or problematic data has been removed. The Visualisations section will be the part of this report where all the necessary graphs relating to the ESD can be found. Lastly, Conclusions will form the final section of this report, containing the summarised findings reached by the data analysis.

1. Outline

The ESD data was downloaded from the Kaggle website in the form of a **.sqlite** file. The data was examined using the DB Browser (SQLite) (**DBB**) software. In order to assist with the analytical process as a whole, the ESD data was manipulated within the DBB and then extracted as 2 separate **.csv** files. Please refer to the file named **final.sql** for the comments for the SQL queries that will be displayed below.

Below is the SQL query which was run in DBB in order to export the first dataset on the Big 5 Leagues (**B5L**) match data:

```
WITH big5 AS
```

[1] Structured Query Language.

```
(
  SELECT lg.name AS league,
         mt.season,
         mt.date,
         tm1.team_long_name AS home_team,
         tm2.team_long_name AS away_team,
         mt.home_team_goal,
         mt.away_team_goal,
         mt.shoton AS shots_on_goal,
         mt.possession
  FROM Match mt
  JOIN Team tm1
  ON tm1.team_api_id = mt.home_team_api_id
  JOIN Team tm2
  ON tm2.team_api_id = mt.away_team_api_id
  JOIN League lg
  ON lg.id = mt.league_id
  WHERE lg.name IN ("England Premier League", "France Ligue 1", "Germany 1.
Bundesliga",
                    "Italy Serie A", "Spain LIGA BBVA")
)
SELECT SUBSTR(league, 1, INSTR(league, " ") - 1) AS country,
       league,
       SUBSTR(league, INSTR(league, " ")) AS clean_league,
       season,
       date,
       SUBSTR(date,1, 10) AS clean_date,
       home_team,
       away_team,
       home_team_goal,
       away_team_goal,
       shots_on_goal,
       possession
FROM big5;
```

Table 1 SQL DQL for the B5L match data.

The SQL query below was run in DBB to export the second dataset on player statistics:

```
SELECT pa.player_fifa_api_id AS fifa_id,
       pl.player_name AS player,
       pl.birthday AS birth_date,
       ROUND(pl.height, 0) AS height,
       pa.preferred_foot,
       pa.overall_rating AS rating
FROM Player pl
JOIN Player_Attributes pa
ON pl.player_fifa_api_id = pa.player_fifa_api_id;
```

Table 2 SQL DQL for the player statistics data.

2. Questions

At the outset, the questions pertaining to the ESD has been split into 2 categories; namely *Match Data* and *Player Data*. Also, the *Match Data* database has been filtered to only focus on the so-called B5L, consisting of the top leagues of **England**, **France**, **Germany**, **Italy** and **Spain**. Therefore, the *Match Data* questions will only focus on these countries/leagues.

Looking at the *Match Data*, these are the questions that I have looked at:

1. What is the distribution of goals scored per match across the dataset as a whole, additionally, what is the distribution for England, Italy and Spain?

2. Which of the B5L scored the most number of goals over the time period covered by the data?

3. Has the total number of goals scored per B5L changed or remained consistent over the time period covered by the data?

With the *Player Data*, I have asked the following questions:

1. What is the relationship between a player's height and their average FIFA rating?

2. In terms of percentage, what is the distribution of left footed versus right footed players?

3. What is the relationship between a player's age and their average FIFA rating?

4. Based on the card rating system in FIFA's Ultimate Team (**FUT**) mode, what is the distribution of rankings across the players in terms of percentage?

3. Data Wrangling

To assist with the process of wrangling the data contained in both datasets sourced from the ESD, I have created a few functions which contain code which is used often and repeatedly when cleaning and wrangling the data. Details of these functions can be found in the accompanying jupyter notebook named **soccer-db-final.ipynb**. Take note that this section within the jupyter notebook is split between the *Match Data* and *Player Data*.

The first issue I looked at, with both datasets, is the conversion of the columns containing date values. Initially, these values are of the object/string datatype, which is not ideal when dealing with dates. To remedy this, I converted these values from objects/strings into a datetime datatype.

The second issue concerned the existence of redundant columns which were either carried over from the SQL extraction phase or became irrelevant at a later stage during the analysis process. Fortunately, thanks to the creation of a dedicated function to carry out the removal (or dropping) of DataFrame columns, this issue was resolved efficiently for both datasets.

Once the second issue was resolved, the third issue was the renaming of a few existing columns to make them consistent and easier to read. Again, due to the creation of specialised function, this renaming could be carried out efficiently.

The fourth issue, only found in the *Player Data*, was the occurrence of duplicated rows of data. These duplicated rows were removed in order to reduce the number of rows in the DataFrame and to provide more consistent data in order to answer the questions relating to this specific dataset.

The datasets required new columns to be created based on existing data found within the DataFrames. This was done for both datasets, with a *goals* column being

created for the *Match Data* and an *age* and *fut category* columns being created for the *Player Data*. Details regarding the creation of these new columns can be found within the jupyter notebook.

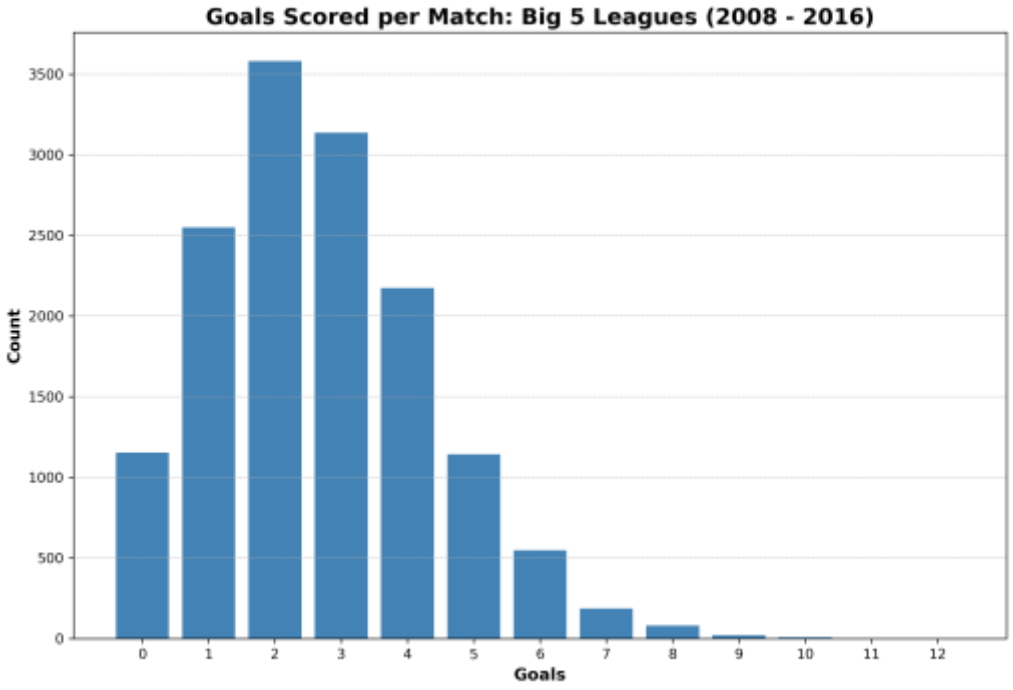
It must be noted that, in comparison with the other 4 countries/leagues in the Match data, the German Bundesliga contains 18 teams only. The other 4 B5L leagues have 20 teams each. This has the effect that Bundesliga clubs play 34 matches per season, whereas the other B5L clubs play 38 matches per season. This difference results in German clubs each playing 4 matches fewer per season.

Finally, both DataFrames were filtered and subsequently manipulated in various ways in order to obtain the necessary data that would assist in answering the questions posed above, and to provide the data required for the creation of the relevant visualisations in the next section.

4. Visualisations

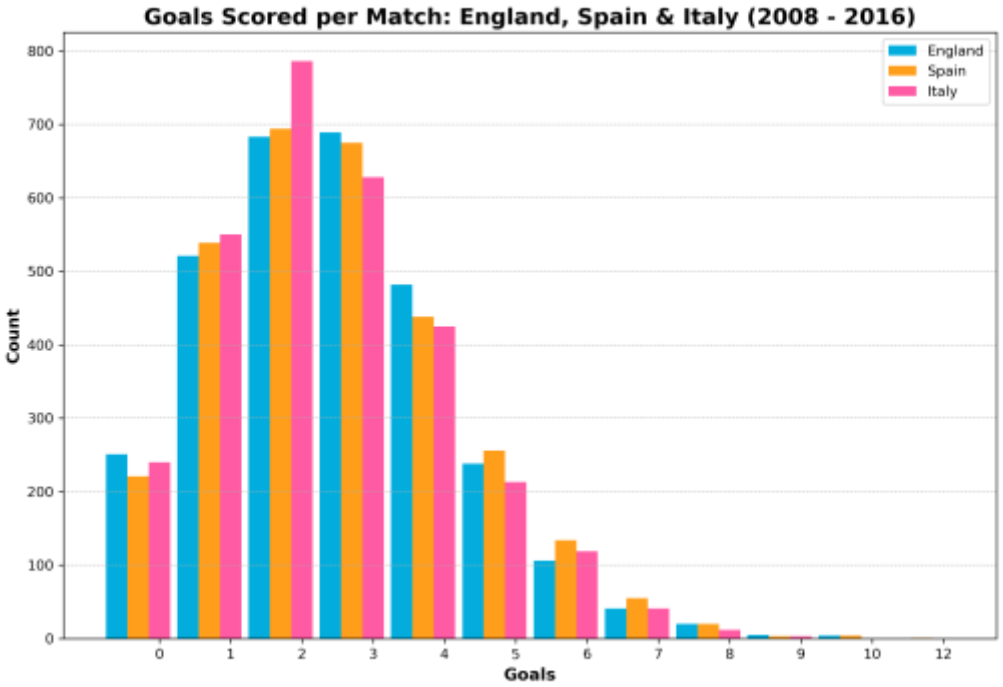
This section will be split between the *Match Data* and *Player Data* datasets. For ease of reference, the posed questions will be included for each visualisation. This section will look at the visualisations for the *Match Data* first.

1. What is the distribution of goals scored per match across the dataset as a whole, additionally, what is the distribution for England, Italy and Spain?



As shown by the graph, the data is skewed to the right. The modal goals scored per match is 2 for the Big 5 Leagues for the period covered by the dataset. There were over 1000 matches where no goals were scored by either team.

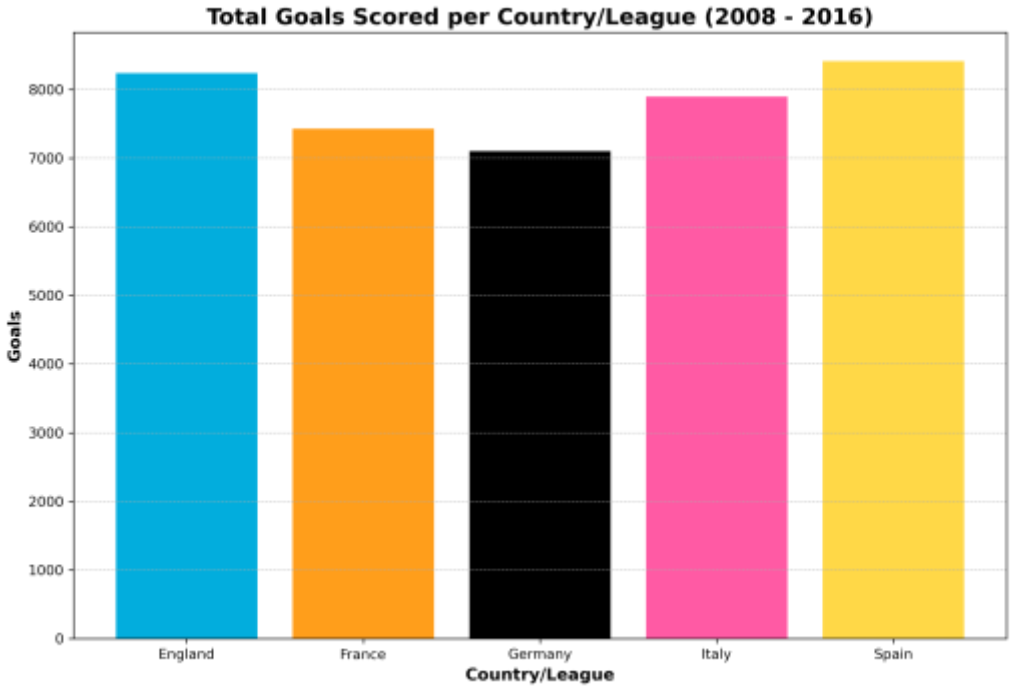
Focussing on this same question, but with the data being filtered to include only England, Italy and Spain:



As this bar graph with indicates, the 3 biggest leagues in Europe all share a similar skewness in terms of the distribution of goals per match over the period covered by the data. The data for all 3 leagues are skewed to the right.

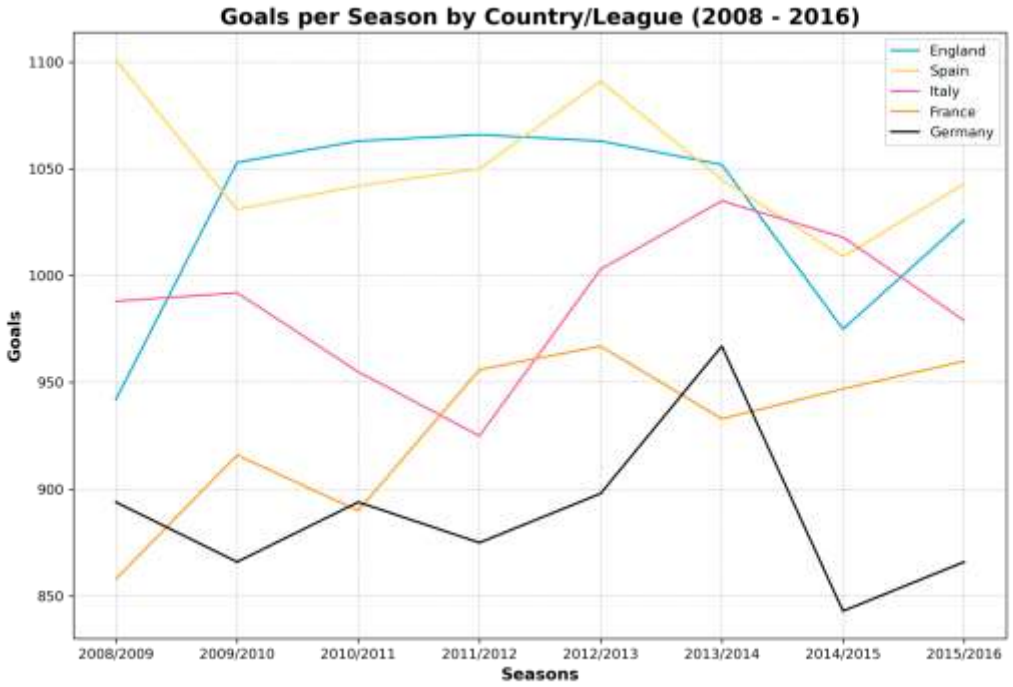
The modal goals scored per match for Italy and Spain is 2. In contrast, the modal goals scored per match for England is 3. Please refer to the country coefficients on the UEFA website for information as to why England, Italy and Spain are considered to be the top 3 leagues in Europe.

2. Which of the B5L scored the most number of goals over the time period covered by the data?



As shown by the bar graph, the Spanish LIGA BBVA scored the highest amount of goals during the period covered by the data. Note that, even though Germany's Bundesliga is the lowest scoring league, German teams play 4 fewer matches each per season compared to the teams in other leagues.

3. Has the total number of goals scored per B5L changed or remained consistent over the time period covered by the data?



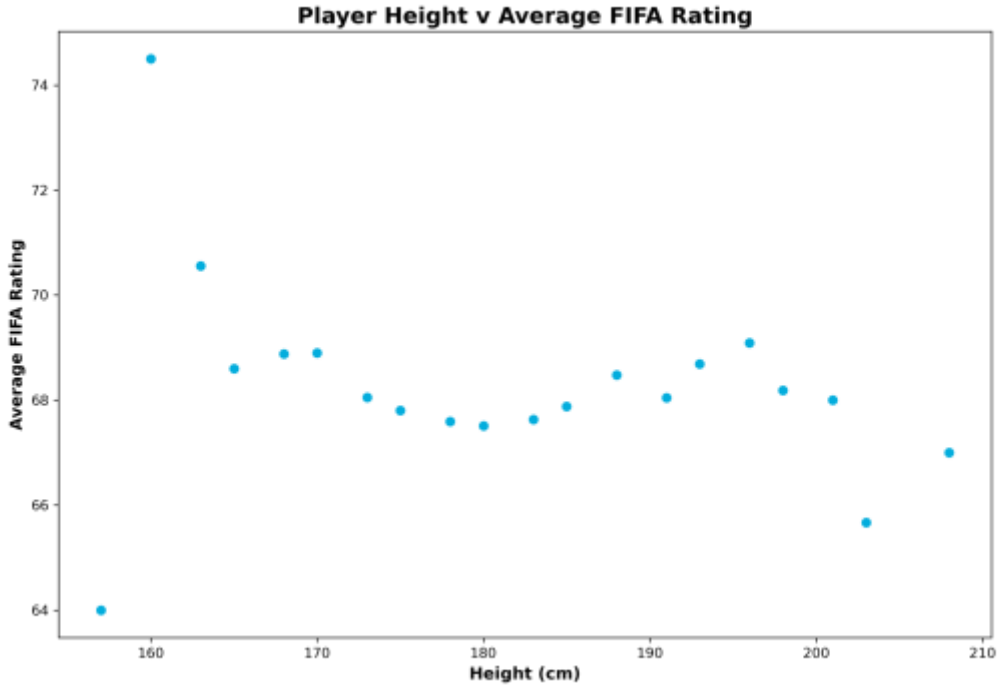
This line graph shows that Spain's LIGA BBVA was the league with the most goals scored in a season, with this occurring in the 2012/2013 season. The English Premier League showed a period of consistency from 2009/10 until 2013/14 with 5 seasons where there were more than 1050 goals scored per season. Apart from this, the general trend for each league was inconsistent, with the total goals scored varying throughout the seasons.

The English Premier League also saw a season-on-season increase of more than 100 goals from 2008/2009 to 2009/2010. The German Bundesliga had the biggest season-on-season decrease of more than 100 goals from the Bundesliga's high point in 2013/2014 to 2014/2015.

However, even with each Bundesliga club playing 4 fewer games each season compared to the other clubs in the rest of the leagues, the Bundesliga outscored the French Ligue 1 on 3 occasions, being the 2008/2009, 2010/2011 and 2013/2014 seasons.

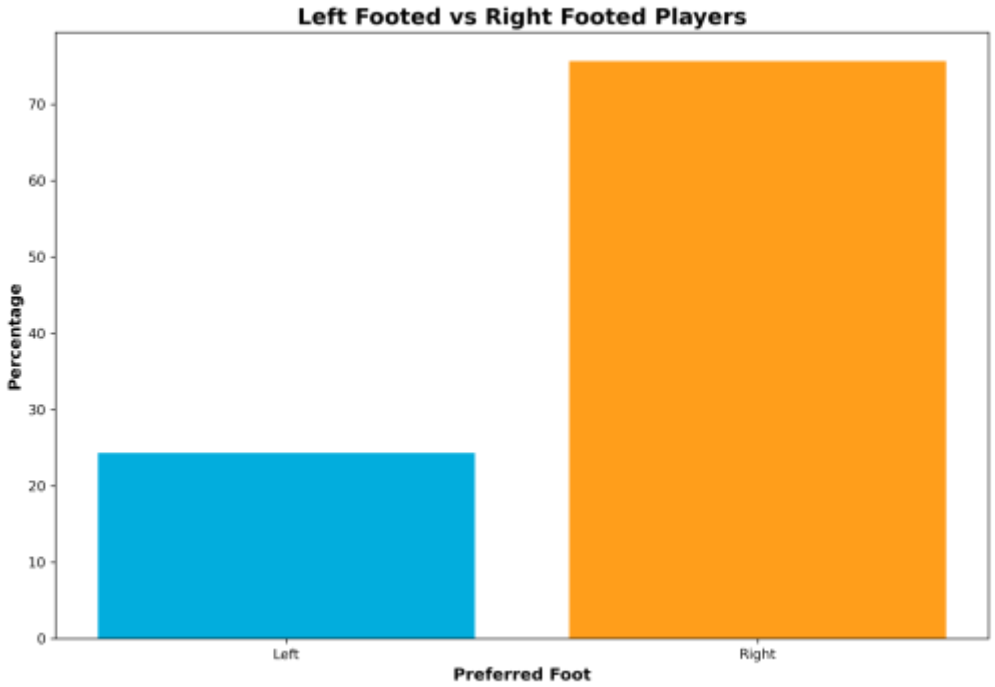
Moving onto the visualisations for the *Player Data*.

1. What is the relationship between a player's height and their average FIFA rating?



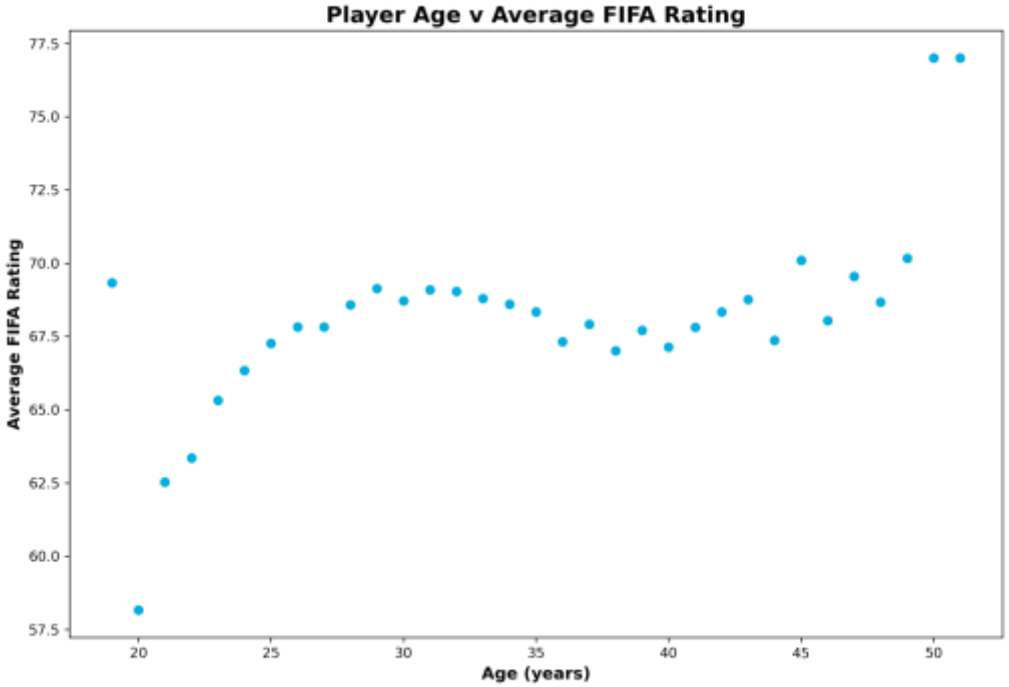
It appears that there is moderate negative relationship/correlation between height and average FIFA rating according to the scatter plot above. Apart from a few outliers, it appears that as a player's height increases, the average rating for that height decreases.

2. In terms of percentage, what is the distribution of left footed versus right footed players?



As this bar graph shows, there are much more right footed players compared to left footed players for the data contained in the *Player Data*.

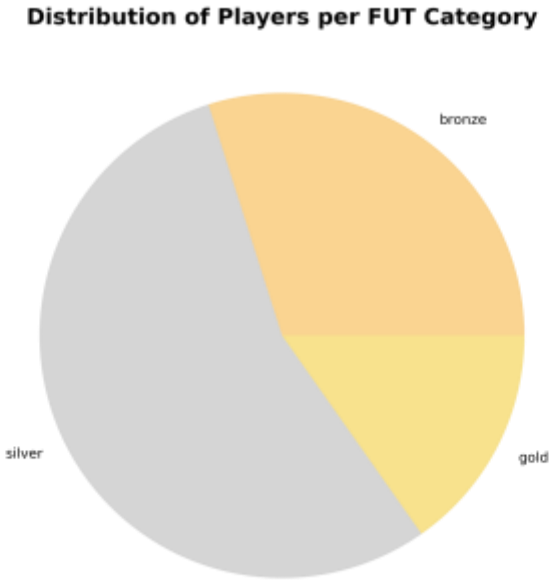
3. What is the relationship between a player's age and their average FIFA rating?



This scatter plot seems to indicate that there is a strong relationship/correlation between age and average FIFA rating. It seems that older players have higher average ratings compared to the younger players.

However, it must be noted that the dataset included data on players whose age would usually lead to the assumption that such players are no longer active professional football players. These are players aged 40 and above. A possible explanation could be that FIFA includes game modes which contain players that have actually retired in real life, but are available in the game.

4. Based on the card rating system in FIFA's Ultimate Team (**FUT**) mode, what is the distribution of rankings across the players in terms of percentage?



According to the pie chart above, it appears that FIFA's FUT mode has more than 50% of its players being categorised as Silver level players. Just over a quarter of players are classified as being Bronze level players. Lastly, the remaining percentage of players in FUT are Gold level players.

5. Conclusion

At the outset, it must be noted that one of the limitations with the datasets that were used in this data analysis was the fact that a sizeable portion of data was sourced from Electronic Art's FIFA video games series. While the FIFA data is consistent and easy to understand, the data itself, especially regarding team attributes and player attributes and ratings, are subjective and/or difficult to quantify. Thus this type of data cannot be relied upon to reach any sort of definitive findings.

To conclude, the *Match data* indicates that the modal amount of goals per match is 2. The *Match data* also shows that, for the time period covered by the dataset, Spain's LIGA BBVA was the league with highest total goals scored. The *Match data* also indicates that each of the B5L have at least 850 goals being scored per season for each league.[2]

The *Player data* seems to establish that there might be a relationship between a player's age and height and their respective average FIFA rating. However, this is merely tentative or speculative at this stage and further analysis should be done in order to determine whether or not such factors are correlated.

The *Player data* shows that there are more right footed players compared to left footed players. Finally, the *Player data* also seems to indicate that the majority of the players found in the FUT mode are Silver level players.

[2] With the exception being Germany's Bundesliga during the 2014/2015 season.

6. Bibliography

WEBSITES

Electronic Arts "FIFA 21 Ultimate Team Item Guide" available at https://www.ea.com/games/fifa/fifa-21/ultimate-team/item-guide (accessed 27 September 2021)

Future Studio "Matplotlib – Save Plots as File " available at https://futurestud.io/tutorials/matplotlib-save-plots-as-file (accessed 27 September 2021)

hilite.me "Source code beautifier / syntax highlighter" available at http://hilite.me/ (accessed 23 September 2021)

Kaggle "European Soccer Database" available at https://www.kaggle.com/hugomathien/soccer (accessed 21 September 2021)

matplotlib "matplotlib.pyplot" available at https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html (accessed 21 – 27 September 2021)

UEFA "Country Coefficients" available at https://www.uefa.com/memberassociations/uefarankings/ (accessed 23 September 2021)

w3resource "SQLite Tutorial" available at https://www.w3resource.com/sqlite/ (accessed 21 September 2021)