Project on Performance Analysis of Football (Soccer) players in FIFA 2022 World Cup

By

Pretom Ghosh

Problem Statement

- I run a football Analytics company where I have many big football clubs as my clients who often looks for recommendations which players are they need to buy to build better teams for the club and thus to ensure best performance in their respective football leagues.
- As a football fan, I also love to watch football and want to understand how are the players performing from different perspectives. I love to respond to different discussions in social media and in my YouTube channel and discuss them from the point of view of Analytics.
- In last year, FIFA World Cup, which is called the Greatest Show on Earth took place from 20th November 2022 to 18th December 2022 which was hosted by Qatar where 32 teams from all over the world took part in it to win the World Cup and finally Argentina won this competition by beating France football team.
- My clients want to me to find out the top players who performed very good in the world cup so that they can recruit them in their respective teams to bolster their team performance.
- I have found out Top 10 players from each category like Strikers, Midfield players, Defensive players and Goalkeepers for my clients which will give them a wide variety of players to pursue.
- I have also answered some of the debated and most discussed topics about this World cup to the football fans from Data Analytics point of views.

AFC (6)
- Australia (38)
- Iran (20)
- Japan (24)
- Qatar (50) (hosts)
- Saudi Arabia (51)
- South Korea (28)

CAF (5)
- Cameroon (43)
- Ghana (61)
- Morocco (22)
- Senegal (18)
- Tunisia (30)

CONCACAF (4)
- Canada (41)
- Costa Rica (31)
- Mexico (13)
- United States (16)

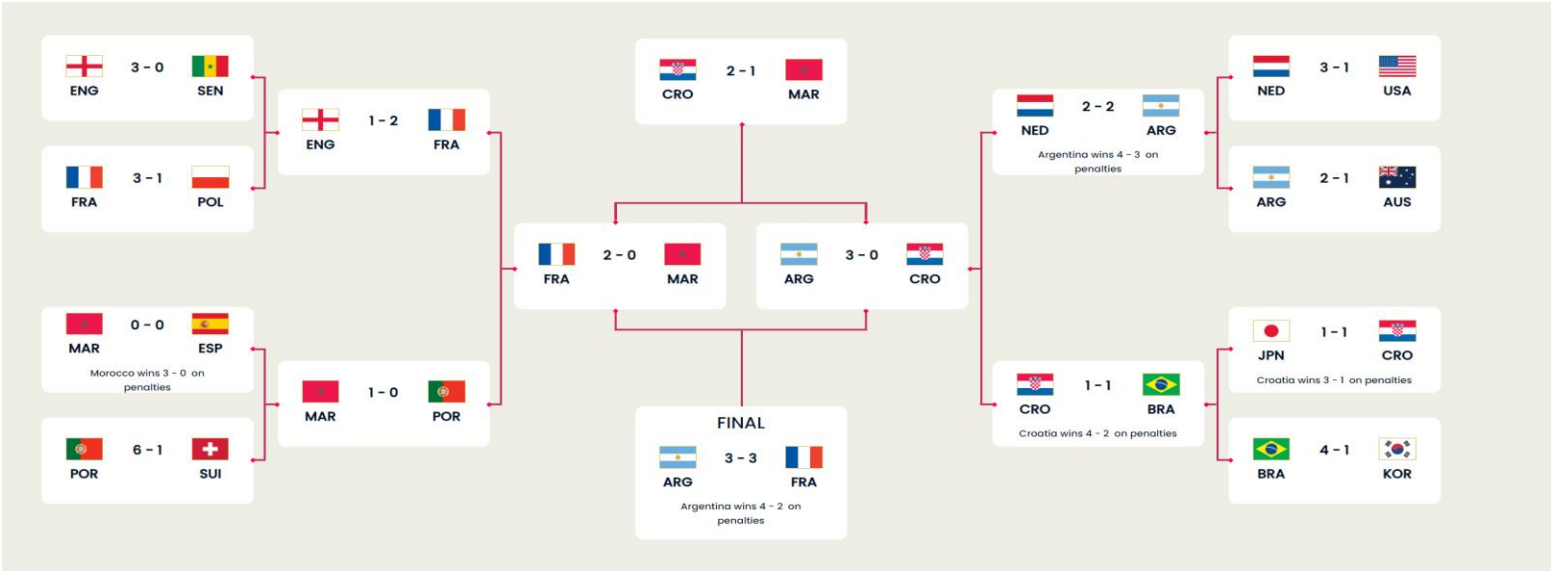
CONMEBOL (4)
- Argentina (3)
- Brazil (1)
- Ecuador (44)
- Uruguay (14)

OFC (0)
- *None qualified*

UEFA (13)
- Belgium (2)
- Croatia (12)
- Denmark (10)
- England (5)
- France (4)
- Germany (11)
- Netherlands (8)
- Poland (26)
- Portugal (9)
- Serbia (21)
- Spain (7)
- Switzerland (15)
- Wales (19)

Background Information :

- Total 32 countries competed
- 16 of them made it to the Round of 16
- 8 of them played the Quarter finals
- 4 to the Semi finals
- Final is played between Argentina and France where Argentina won the World Cup.



Dataset: The dataset is named as players_list.csv which have been collected from

https://www.kaggle.com/datasets/tittobobby/fifa-world-cup-2022-player-stats

Features of the dataset:

```
df.shape
```

```
(680, 42)
```

- The dataset contains 680 rows and 42 columns in total.
- The rows contain the specific information of a player based on different attributes in the columns. We have taken into consideration different attribute columns for meaningful analysis.

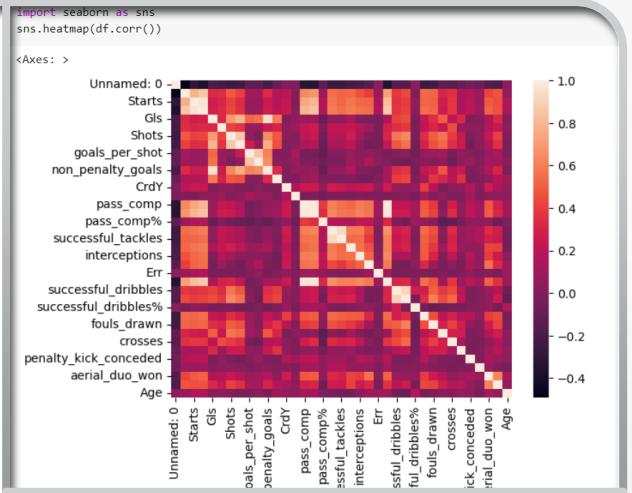
	Unnamed: 0	Player	Match_played	Starts	Minutes_played	Gls	Ast	Shots	shots_on_target	goals_per_shot	...	penalty_kick_won	penalty_kick_conceded	own_goal	aerial_duo_won	aerial_duo_lost	Pos	Age	Club_country	Club	Team
0	0	Emiliano Martínez	7	7	690	0	0	0	0	NaN	...	0.0	0.0	0	2.0	0.0	GK	29	eng	Aston Villa	Argentina
1	1	Lionel Messi	7	7	690	7	3	27	13	0.11	...	1.0	0.0	0	2.0	7.0	FW	34	fr	Paris S-G	Argentina
2	2	Nicolás Otamendi	7	7	690	0	1	1	0	0.00	...	0.0	1.0	0	21.0	13.0	DF	34	pt	Benfica	Argentina
3	3	Rodrigo De Paul	7	7	599	0	0	7	3	0.00	...	0.0	0.0	0	2.0	3.0	MF	28	es	Atlético Madrid	Argentina
4	4	Nahuel Molina	7	6	567	1	1	2	1	0.50	...	0.0	0.0	0	0.0	6.0	DF	24	es	Atlético Madrid	Argentina

5 rows × 42 columns

Exploratory analysis (EDA)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 680 entries, 0 to 679
Data columns (total 42 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                680 non-null    int64
 1   Player                    680 non-null    object
 2   Match_played              680 non-null    int64
 3   Starts                    680 non-null    int64
 4   Minutes_played            680 non-null    int64
 5   Gls                       680 non-null    int64
 6   Ast                       680 non-null    int64
 7   Shots                     680 non-null    int64
 8   shots_on_target           680 non-null    int64
 9   goals_per_shot            431 non-null    float64
 10  goals_per_shot_on_target  247 non-null    float64
 11  non_penalty_goals         680 non-null    int64
 12  penalty_goal              680 non-null    int64
 13  CrdY                      680 non-null    int64
 14  CrdR                      680 non-null    int64
 15  pass_comp                 677 non-null    float64
 16  pass_attempt              677 non-null    float64
 17  pass_comp%                674 non-null    float64
 18  tackles_made              677 non-null    float64
 19  successful_tackles        680 non-null    int64
 20  Blocks                    677 non-null    float64
 21  interceptions             680 non-null    int64
 22  clearence                 677 non-null    float64
 23  Err                       677 non-null    float64
 24  Touches                   677 non-null    float64
 25  successful_dribbles       677 non-null    float64
 26  dribbles_attempted        677 non-null    float64
 27  successful_dribbles%      435 non-null    float64
 28  fouls_commited            680 non-null    int64
 29  fouls_drawn               680 non-null    int64
 30  offside                   680 non-null    int64
 31  crosses                   680 non-null    int64
 32  penalty_kick_won          677 non-null    float64
 33  penalty_kick_conceded     677 non-null    float64
 34  own_goal                  680 non-null    int64
 35  aerial_duo_won            677 non-null    float64
 36  aerial_duo_lost           677 non-null    float64
 37  Pos                       680 non-null    object
 38  Age                       680 non-null    int64
 39  Club_country              679 non-null    object
 40  Club                      679 non-null    object
 41  Team                      680 non-null    object
dtypes: float64(17), int64(20), object(5)
```

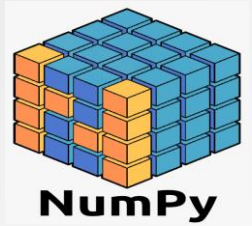
```
import seaborn as sns
sns.heatmap(df.corr())
```



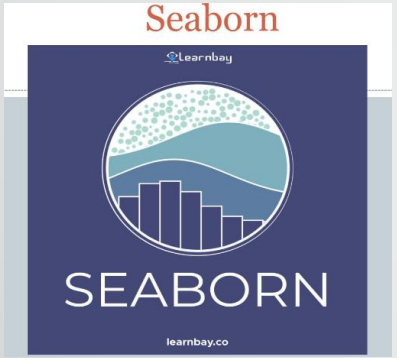

```
CrdR                      0
pass_comp                 3
pass_attempt              3
pass_comp%                6
tackles_made              3
successful_tackles        0
Blocks                    3
interceptions             0
clearence                 3
Err                       3
Touches                   3
successful_dribbles       3
dribbles_attempted        3
successful_dribbles%    245
fouls_commited            0
```

- So we saw that there are some null values in various columns but it is common because in football players play as per their positions and roles in the field, it is common t hat a player from a specific position will not show a ll attributes as for Example player "Nicolás Otamen di" is a player in the defense (DF) position, so his go als (Gls) are showing o and it is normal for a player not to score goals but to prevent goals and that's w hy his blocks, clearances are showing values.
- Again, We also plotted the heatmap of the dataset to initially see relationship between the attributes and found some strong correlation exist between some attributes.

Libraries used for visualization and their introduction



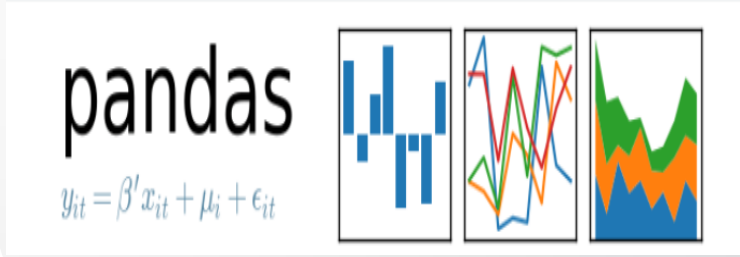
NumPy, which is a popular Python library used for numerical computations. It provides support for arrays and matrices, as well as a large collection of mathematical functions to perform operations on them. NumPy is widely used for scientific computing, data analysis, and machine learning, and is an essential tool in the Python data science ecosystem.



Seaborn is a Python library for data visualization that is built on top of Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is designed to work well with Pandas dataframes, making it a popular choice for data visualization in the data science community.

My Project

Pandas is a popular open-source Python library for data manipulation, analysis, and cleaning. It provides highly optimized data structures for working with structured and semi-structured data, making it an essential tool in the data science and machine learning communities.

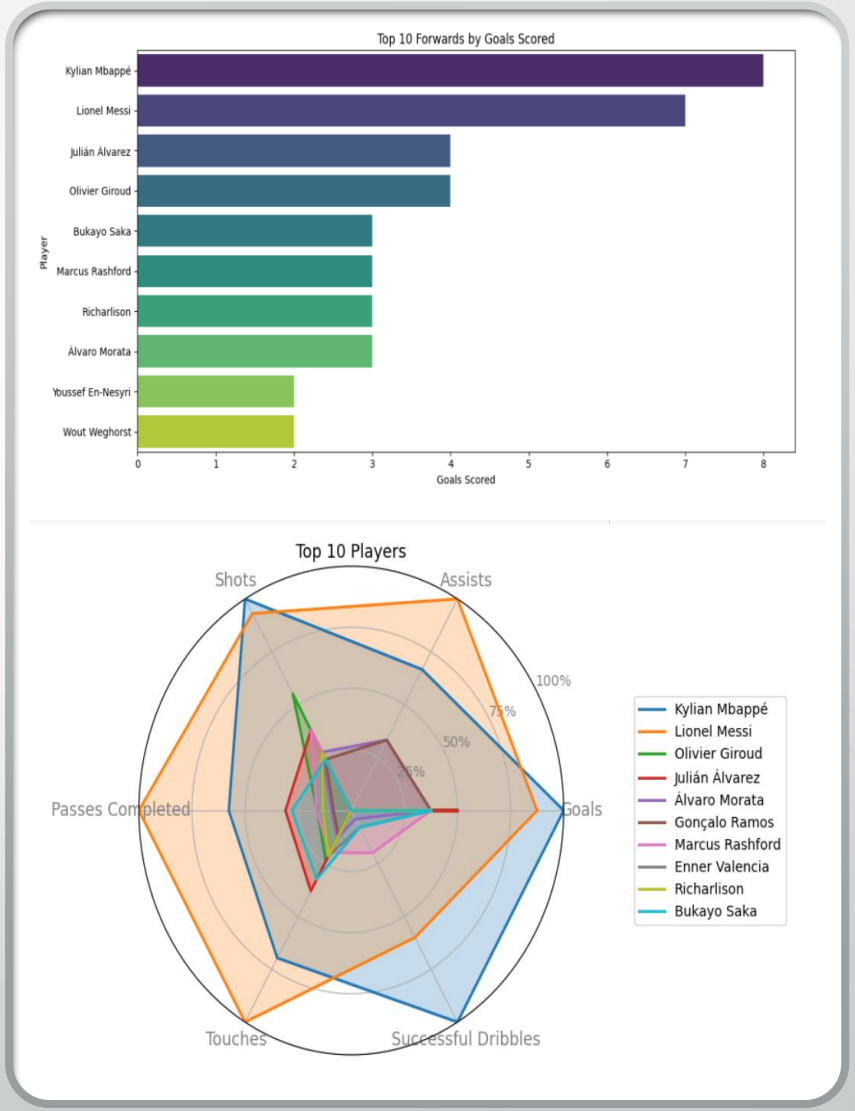


Plotly Express is a high-level Python library built on top of the Plotly graphing library that simplifies the creation of interactive visualizations. It provides a wide range of easy-to-use functions for creating many types of plots, including scatter plots, line plots, bar plots, pie charts, histograms, and more.

Matplotlib is a popular open-source Python library for creating static, animated, and interactive visualizations in Python. It provides a wide range of 2D and 3D plotting options, including scatter plots, line plots, bar plots, histograms, and more. Matplotlib is highly customizable, making it a powerful tool for creating publication-quality graphics and visualizations.

In the Hunt of Top players

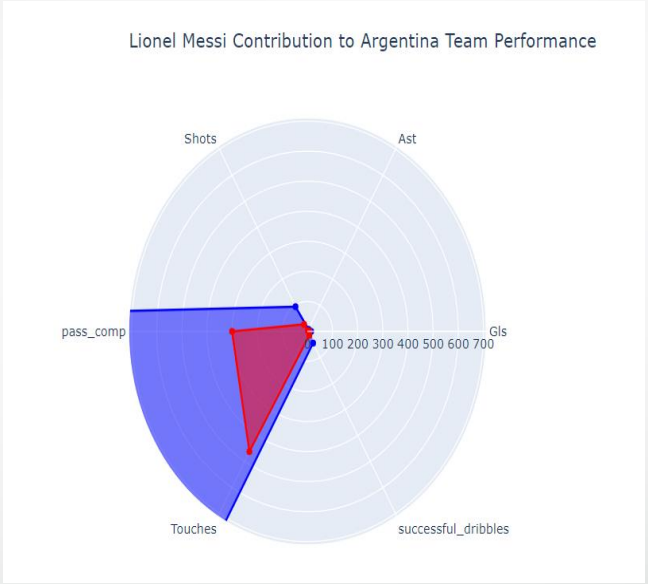
- First, we tried to find out the top 10 players who scored highest goals individually
- Kylian Mbappe from France have scored the highest 8 goals.
- Then Lionel Messi of Argentina scored 7 goals and he is in the 2nd position.
- Julian Alvarez and Oliver Giroud scored 4 goals each and are in 3rd and 4th positions, respectively.
- Then I deep dived and took Goals, Assists, Shots, Passes completed, Touches and Successful Dribbles as attributes to find out who were the most "Complete players" and plotted a Radar chart. I plotted Radar chart because in sports analytics it is often used to compare multiples players based on multiple attributes.
- Kylian Mbappe is the player who had most goals, completed most dribbles successfully and had most shots amongst the players.
- Lionel Messi had the 2nd highest goals, most assists, had the most passes c ompleted and most touches on the ball amongst the players.
- No other forward players went nearly to the performance of Kylian Mbapp e or Lionel Messi in this case when considered the above attributes.

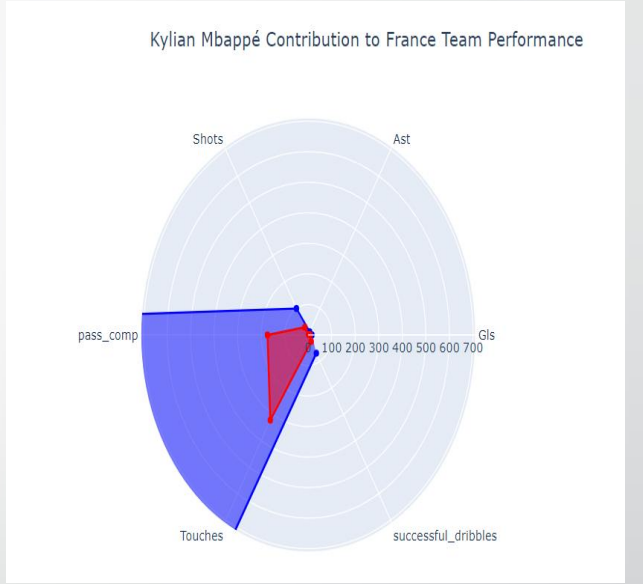


Who deserves the “Golden Ball”???

VS

FIFA awarded Lionel Messi the ‘Golden Ball’ Award for the FIFA 2022 Football World Cup which is given to the best player of the tournament. Many football fanatics raised questions against it because the competition of the performance between Lionel Messi and Kylian Mbappe was awesome as they went toe-to-toe and very tough. Still, we can see that Mbappe had most goals, successful dribbles and most shots but Lionel Messi had most assists, most passes completed, most touches, he is only 1 goal short then Mbappe, and finally he had nearly as many shorts as Mbappe. Moreover, Lionel Messi became the FIFA World Cup winner this time. If we compare the contribution of Lionel Messi to Argentina team's overall performance, we can see that he also outperformed Kylian Mbappe because Messi contributed more to the success of his team but compared to that Mbappe contributed less which was a crucial factor behind the secret of Argentina's World Cup win. So, I think FIFA awarded the Golden Ball to Lionel Messi when considering this facts and was a fair decision.

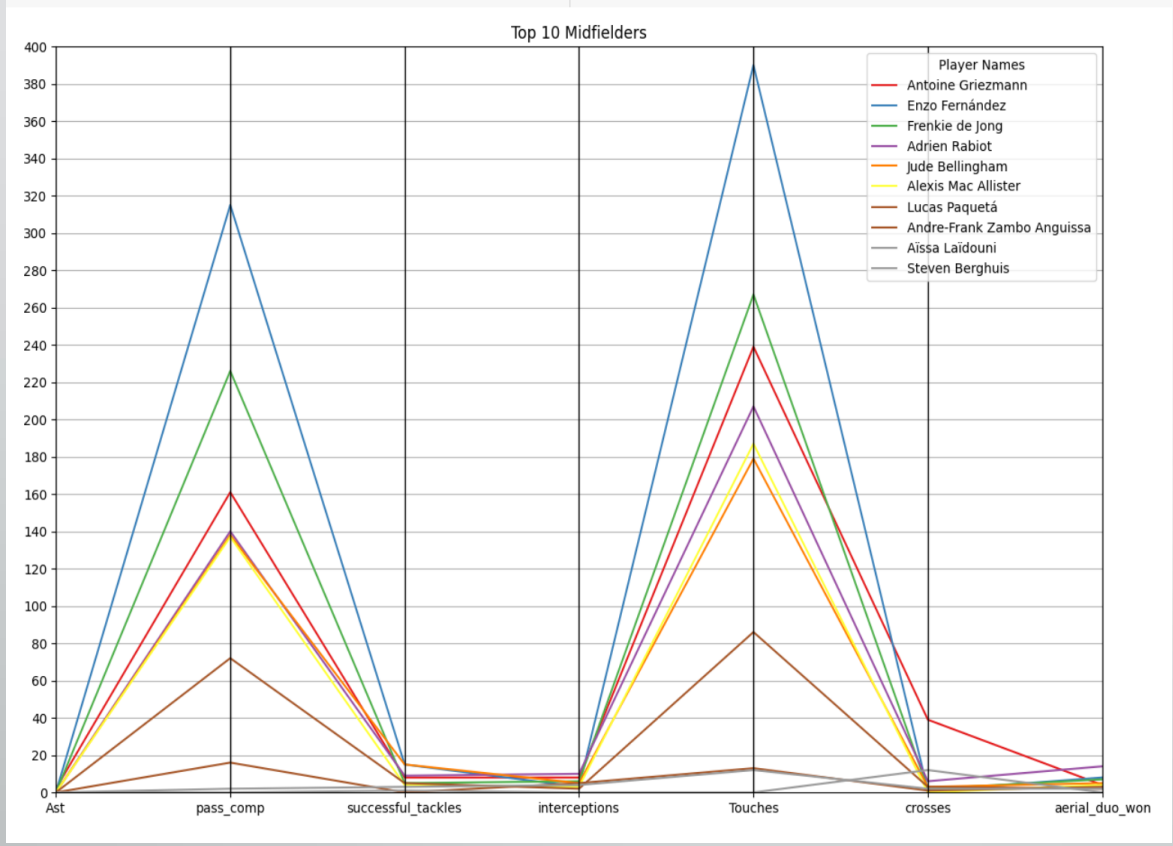




top_players

	Player	Gls	Ast	Shots	pass_comp	Touches	successful_dribbles
0	Kylian Mbappé	8	2	29	173.0	322.0	25.0
1	Lionel Messi	7	3	27	300.0	462.0	15.0
2	Olivier Giroud	4	0	16	48.0	111.0	0.0
3	Julián Álvarez	4	0	11	93.0	176.0	0.0
4	Álvaro Morata	3	1	8	28.0	60.0	1.0
5	Gonçalo Ramos	3	1	7	25.0	56.0	0.0
6	Marcus Rashford	3	0	11	49.0	92.0	5.0
7	Enner Valencia	3	0	8	40.0	99.0	2.0
8	Richarlison	3	0	8	40.0	99.0	0.0
9	Bukayo Saka	3	0	7	84.0	150.0	2.0

The Game 'Actually' controlled by the Midfielders



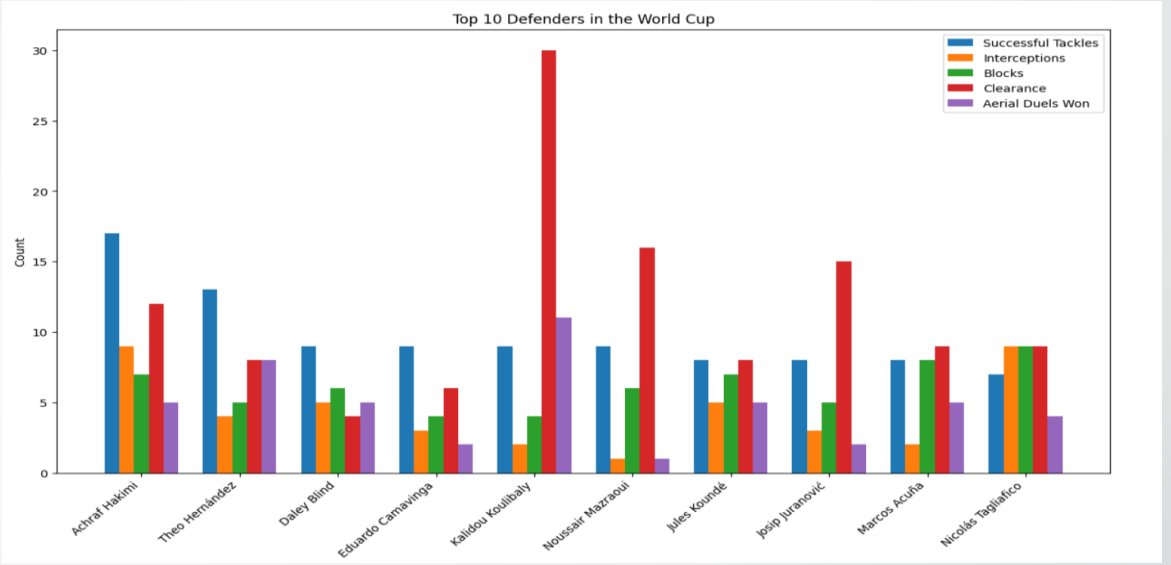
- Though the forwards win the best player awards and most other awards historically, midfield players are the unsung heroes of the game because without midfield, a team cannot win.
- I took Assists, passes completed, successful tackles, interceptions, touches to the ball, crosses and aerial dual won which are the features of a midfield player and plotted a parallel coordinate plot because parallel coordinate plots are used to compare multiple attributes in football analytics.
- Antonio Griezmann from France topped the chart whereas Enzo Fernandez from Argentina was 2nd, the Dutch midfielder Frankie De Jong is in the 3rd and so on.

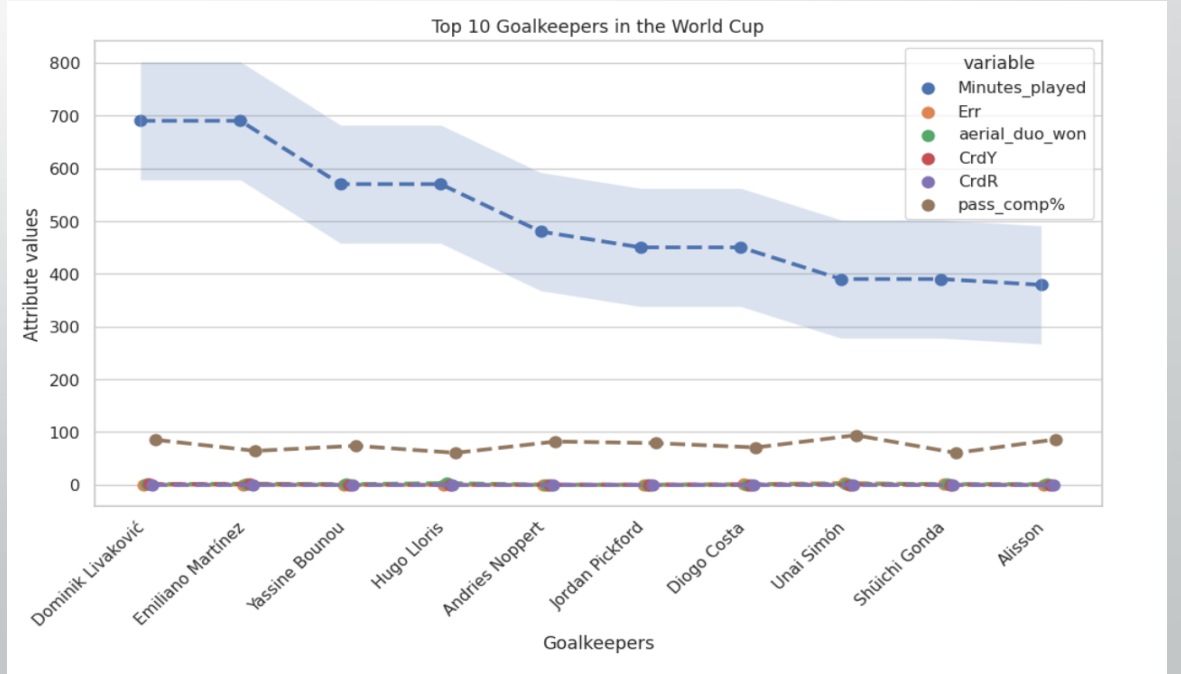
My clients also want to recruit some superb defenders and goalkeepers in their teams

- Defense is very important for a team's success

As a great football manager Sir Alex Ferguson said, "Attack win you matches, and defense win you the tournament. Defenders and Goalkeepers are the foundations of a team's defense.

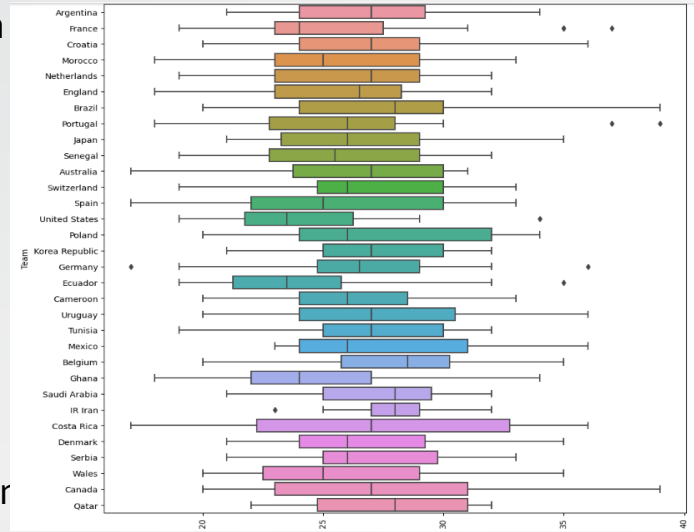
- Successful tackles, Interceptions, Blocks, Clearance and Aerial duo won are taken as defenders attribute. I choose stacked bar plots here because it is simple and easy to interpret and yet powerful visualization.
- Achraf Hakimi, Theo Hernandez and Daley Blind were some fantastic defenders in the world cup.
- Now while finding out the top goalkeepers, I took Minute played, Error, aerial duo won, number of red and yellow cards and passes completed as attributes as those indicated the longer playing capabilities, less error prone and discipline of a goalkeeper. I choose line plots here because it is simple and easy to interpret.
- Dominik Livakovic and Emiliano Martinez dominated the competition of being the best goalkeepers in the world cup.

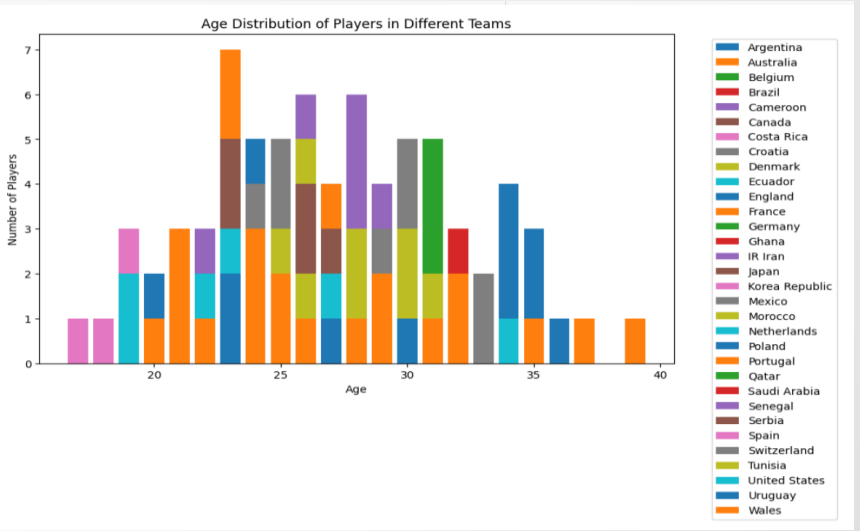




Does Age matters in Football?

- This is a long-discussed question amongst the football fans.
- Football is a physically demanding game where Age must have a meaningful contribution.
- But we saw many players have played for longer in their career even in their late 30s also.
- I first plotted a box plot and a bar chart to see the Age distribution in all the teams.
- Then I found out the top 10 youngest and 10 most old teams, then plotted line plots to see total goals scored and total attempted shots by the youngest and oldest teams by mean age to compare them against the bar plots of ages to see which portion performs better. Found out the younger teams like France outperformed the oldest and experienced teams like Brazil or Belgium which indicates that Age matters in Football.





Conclusion and Discussion

- At the end, I was able to find out top players from different field positions and recommend them to my clients to recruit.
- Able to answer some questions of the football fans after analyzing the and scrutinizing the facts.
- As a football fan, it was satisfying for me.

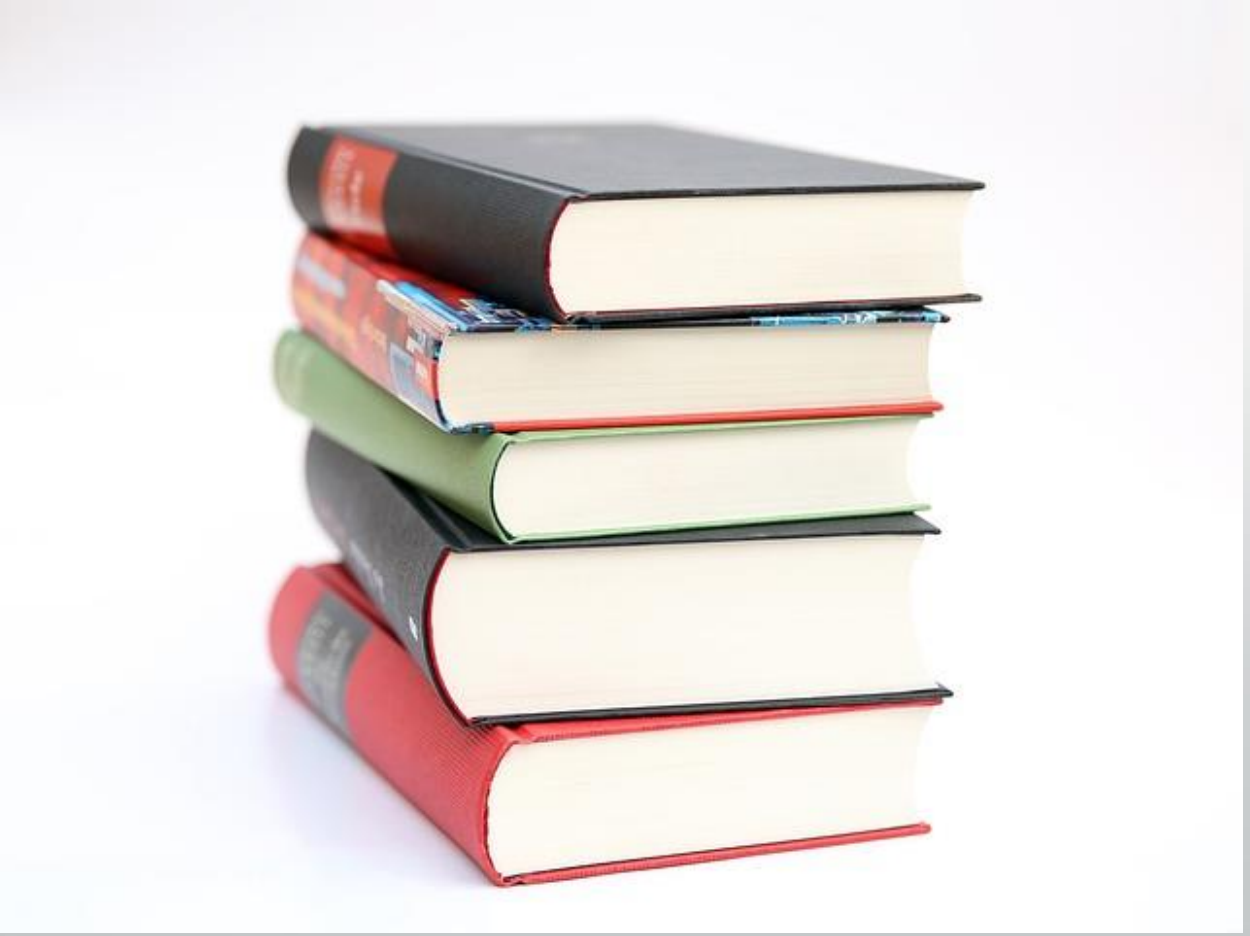
Key Learnings and Future Work

- Key learnings:
- Application of visualizations in football analytics.
- Story telling.
- Insight on performance of different players in the FIFA World Cup 2022.

- Future Work:
- A lot of different insights can be found out and variety of visualizations can be tried accordingly to more detailed dataset and this can be used to analyze different player's performance in different matches as well if more detailed match by match data can be found.

References :

https://www.kaggle.com/datasets/tittobobby/fifa-world-cup-2022-player-stats



Thankyou