Course Number: DATA-301 *Instructor: Osei Tweneboah*

Data Analysis on the 2022 Qatar World Cup

Chris Valente[1], Tariq Alagha[2], and Nikita Gerzhgorin[3]

[1] Ramapo College of New Jersey; cvalent1@ramapo.edu
[2] Ramapo College of New Jersey; talagha@ramapo.edu
[3] Ramapo College of New Jersey; ngerzhgo@ramapo.edu

Abstract: The World Cup is a major international sporting event that is watched by billions every 4 years; the best national teams in world football qualify to reach the pinnacle of sport competition. Because the tournament doesn't occur annually, many different factors can change between tournaments; different star players, different successful teams, etc. Our interest was in observing the tournament from multiple different aspects; crowd attendance, player performance, team performance, and matches individually to try and understand the big factors that went into this particular iteration of the World Cup; the 2022 Qatar World Cup. Using PANDAS to read and clean multiple different `csv` files consisting of different types of tournament data, we were able to effectively visualize our findings and results by utilizing libraries such as Plotnine, Streamlit, Seaborn, and Matplotlib. Our research found our answers to our unique research questions:

1. We were able to ascertain which country had the most loyal fans in terms of attendance. We were also able to generate statistics that showed average attendance, and found correlation between stadium venue and attendance as well.

2. We implemented several advanced statistics - expected goals (xG), expected assists (xA), and shot creating actions (SCA) to calculate and identify which players were at the top of these respective categories in the tournament, thus narrowing down the list of best offensive players in the tournament.

3. By taking a look at several different attributes involved in each match (goals scored, shots attempted, total passes, fouls committed), we judged the intensity of matches, and aimed to observe if we were able to notice a substantial difference in these attributes as the rounds of the tournaments went on.

4. We analyzed the results of teams that had majority possession in their matches to attempt to understand if the concept of controlling the game via possession was a tactic that had a particularly strong or weak effect throughout the tournament.

Keywords: World Cup, FIFA, sports statistics, advanced football statistics, shot creating actions, player efficiency, tournament attendance, Qatar

1. Introduction

The World Cup is a tournament consisting of some of the best qualifying nations in international football. Countries are divided into groups, where the top two teams from each group move on to the 'knockout rounds', effectively dwindling down the competition until one team is left standing. The World Cup is a global event that transcends borders, cultures, and languages, captivating audiences with the sheer talent, passion, and dedication of the world's top footballers. The tournament also serves as a platform for countries to showcase their national identity and pride, and to unite their citizens around a common goal. Whether you are a casual fan or a die-hard supporter,

the World Cup is an unforgettable experience that captures the imagination and inspires a sense of unity that extends far beyond the pitch. And for the players, it is considered the highest honor in the sport; one that demands a lot of passion and excellence in order to persevere and bring home glory to their country. The 2022 World Cup in Qatar has recently concluded with an Argentinian extra-time victory in the final over France; a dramatic victory that perfectly captures the essence of the tournament. The dataset that we will be utilizing in our project focuses on each individual match from the tournament, and contains a large amount of data about the match itself, specifically for each team; possession, shots attempted, shots on goal, total passes, etc. We are hoping to showcase some trends, and insights that can summarize and visualize the tournament effectively through data. Our goal is to provide some background on the tournament; the teams, the players, the matches, and the results. We will be using the dataset to answer some questions that we have about the tournament, and to provide some insights that can help us understand the tournament better. Outside our background analysis, some questions we hope to answer are:

1. Which matches were the most/least attended? We are interested to see if there are trends between attendance and the teams playing, or the stage of the tournament.
2. Who was the most efficient player throughout the tournament, in regard to the attack? (we can calculate some statistics for this, such as G/A, etc.)
3. Does the difference in rounds have an effect on the way that teams play? (this can be measured by shots attempted, possession, etc.)
4. Having possession is an emphasis in modern football, but does having more possession actually lead to more success?

2. Materials and Methods

We used multiple datasets in order to capture our needed information on the 2022 World Cup. The first dataset is titled "Fifa World Cup 2022: Complete Dataset" [3], the second is titled "FIFA World Cup 2022 Player Data" [2], and both datasets were sourced from Kaggle. "Fifa World Cup 2022: Complete Dataset" consists of a single `csv` file containing information about every match during the cup such as which teams were playing, how much possession each team had, and how many goals each team scored. "FIFA World Cup 2022 Player Data" is composed of multiple `csv` files, each containing different statistics about individual players such as their age, position, country, and so on. The data was further split by category; offensive statistics, defensive statistics, advanced statistics, etc. We didn't utilize all of this data due to the focus of our research questions, but there was a large amount of data available that we could use if we wanted to expand our research in the future.

In total, we ended up with a `matches` dataset (which contained information about each individual match of the tournament), a `players` dataset (which contained information about each player in the tournament), and a `countries` dataset (which was created based on information computed about each country in the *matches* dataset.

Using PANDAS, we were able to read and clean our multiple datasets consisting of different types of tournament data. There were many instances where columns had to be imputed, or where data types needed to be modified. Once all of our data was cleaned, and appropriately typed and manipulated to our liking, we were able to effectively visualize our findings and results by utilizing libraries such as Plotnine, Streamlit, Seaborn, and Matplotlib.

3. Results

This section may be divided into subheadings. It should provide a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

3.1. Are there common factors between the most attended matches and the least attended matches in the tournament?

Based on the time of year (midseason instead of in the summer like usual), location for the tournament (far from many participating countries, worried about fan attendance), and newsworthy alcohol ban, the attendance was a big question for this particular tournament. We found that the average attendance for all matches in the tournament is: 53191.44 people per match. The following tables are the results of sorting our match data by the most and least attended matches.

Table 1. The table below shows the matches with the top five highest attendance numbers.

team1	team2	category	attendance	venue
Argentina	France	Final	88966	Lusail Iconic Stadium
Argentina	Croatia	Semi-final	88966	Lusail Iconic Stadium
Argentina	Mexico	Group	88966	Lusail Iconic Stadium
Portugal	Uruguay	Group	88668	Lusail Iconic Stadium
Netherlands	Argentina	Quarter-final	88235	Lusail Iconic Stadium

[1] It can be inferred that 88966 is the maximum capacity of Lusail Iconic Stadium

For the most attended matches, the Lusail Iconic Stadium ended up being the venue for all of these matches, and an interesting insight was that Argentina was a team participating in 4 of the top 5 matches in respect to attendance.

Table 2. The table below shows the matches with the lowest five attendance numbers

team1	team2	category	attendance	venue
Switzerland	Cameroon	Group	39089	Al Janoub Stadium
Mexico	Poland	Group	39369	Stadium 974
Cameroon	Serbia	Group	39789	Al Janoub Stadium
Spain	Costa Rica	Group	40013	Al Thumama Stadium
Belgium	Canada	Group	40432	Ahmed bin Ali Stadium

For the least attended matches, there is a larger distribution of venues which can be explained by a further visual later, but we do notice that Cameroon appears twice in our low attendance matches, so there is at least a commonality here.

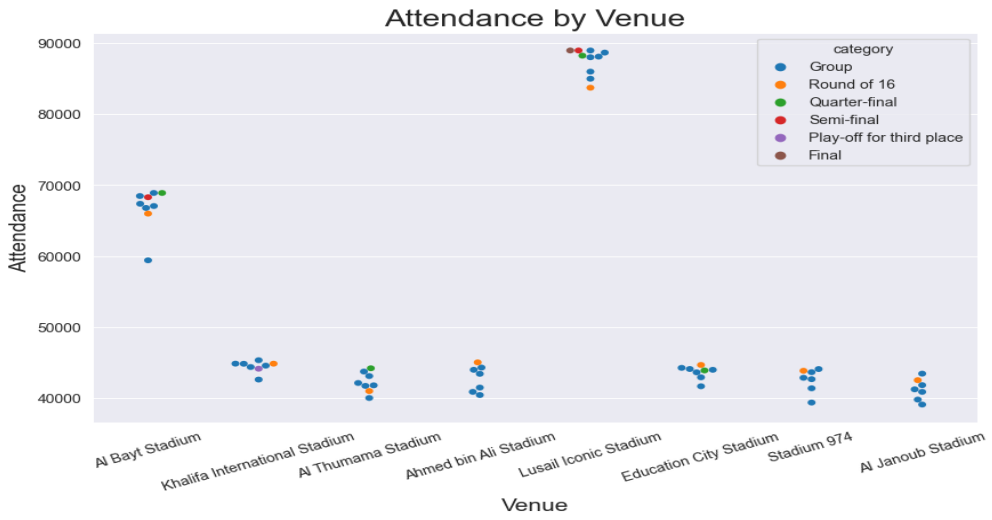

Figure 1. Swarm plot that shows the distribution of attendance by venue, faceted by round of the tournament.

Interesting things to note here is that all of the most attended matches were at Lusail Iconic Stadium, and 4 out of the top 5 matches had Argentina playing. So while we can see that Argentina is a very popular team, we can also see that the Lusail Iconic Stadium is a very big stadium, which has also hosted the most matches in the tournament. So, it is not surprising that the most attended matches were at this stadium if you keep that in mind, and consider that two of the most attended matches at this stadium were the World Cup Final, and Argentina's semifinal match against Croatia.

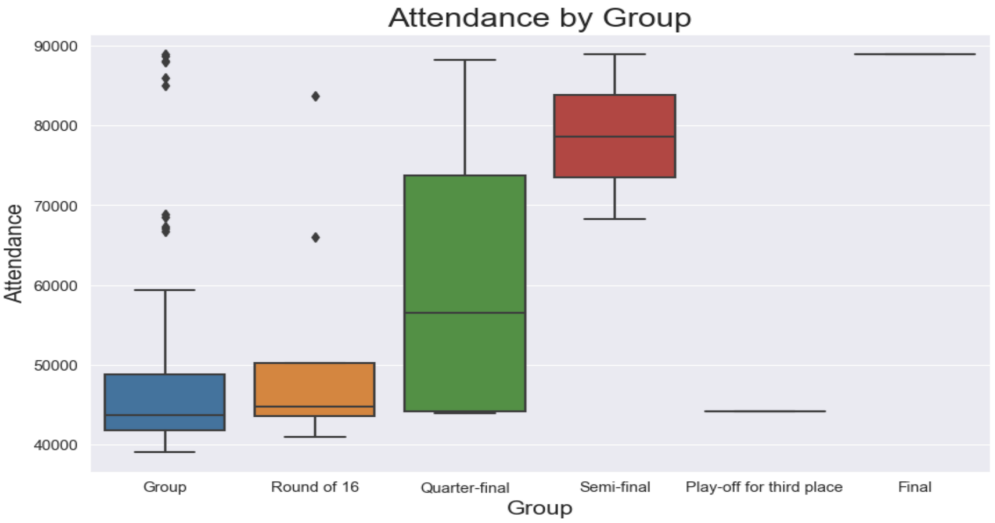

Figure 2. Box plot that shows the distribution of attendance faceted by each round of the tournament

The above table shows that quarter-final games had much higher mean attendance than group games, yet almost every stadium that hosted quarter-final games also had group games that saw more attendance except for Al Thumama. Figure 1 also shows

that there were a lot more group games, so their mean is more representative than quarter-finals which had a very wide range of attendance without many observations.

Additionally, Table 1 tells us that Argentina is a very popular team, playing in the top 3 most attended matches. We also see that the Lusail Iconic Stadium is a very big stadium, which has also hosted the most matches in the tournament. So, it is not surprising that the most attended matches were at this stadium if you consider that two of the most attended matches at this stadium were the World Cup Final, and Argentina's semifinal match against Croatia.

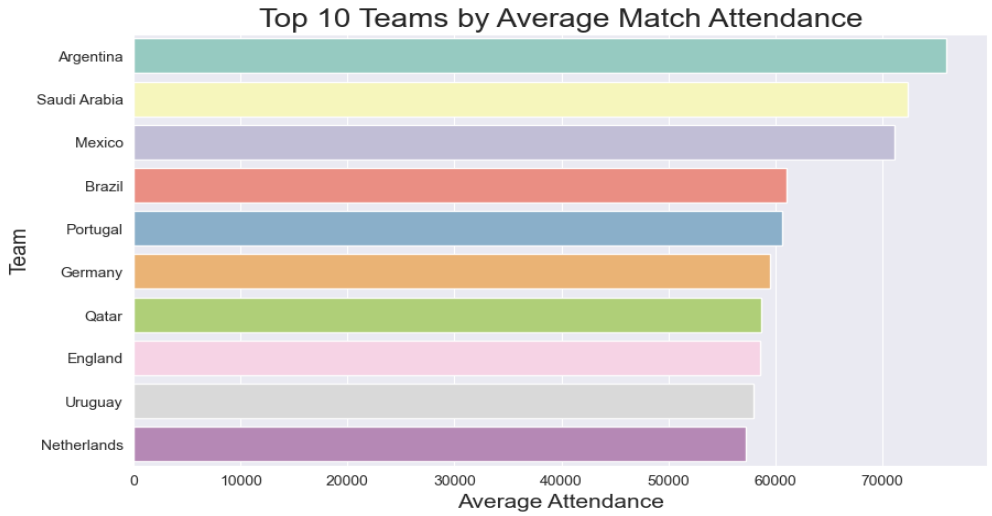

Figure 3. Bar plot that shows the top teams in terms of average match attendance

Our previous suggestions have been proven true; throughout the World Cup, Argentina has tended to carry the largest crowds to their matches. Other relevant top 10 national teams would include Saudi Arabia, who is closer to all of the other participating nations to the actual location of the tournament in Qatar, as well as Qatar itself. Fans from countries such as Brazil and Mexico would likely have had to travel a very long distance to get to the tournament.

3.2 Who was the most efficient player throughout the tournament, in regard to the attack?

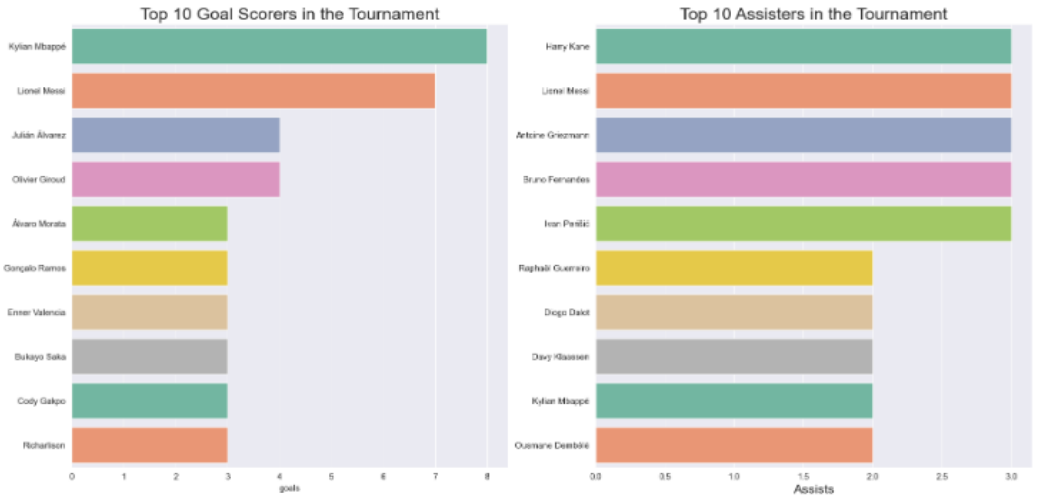

Figure 4. Two bar plots that show the top goal scorers and assisters respectively throughout the World Cup

From these plots, we can see that Kylian Mbappe and Lionel Messi were clear at the top of goals scored with 8 and 7 respectively, with the rest of the top 10 being rounded out by multiple 4 and 3 goal scorers. There were no sole assist leaders, but Harry Kane, Lionel Messi, and a few others all shared the top spot with 3 assists throughout the tournament. We will be building off these basic measurements of football success to determine how efficient some of these players ended up.

xG, or Expected Goals, is a metric that is used to measure the quality of a shot taken by a player. The higher the xG, the more likely the shot is to go in. xA, or Expected Assists, is a metric that is used to measure the quality of a pass that leads to a shot. The higher the xA, the more likely the pass is to lead to a goal. These are key metrics in modern football, and are frequently used to measure the quality of a player's performance. We will be exploring these metrics to determine which players are deemed most efficient in the tournament based on their criteria.

Figure 5. Scatter plot comparing expected goals against goals scored by players in the World Cup

Looking at the visual, we can see right away that two prominent outliers are Kylian Mbappe and Lionel Messi. They both have scored more goals than their xG output, meaning that they've scored more goals than they were expected to realistically finish.

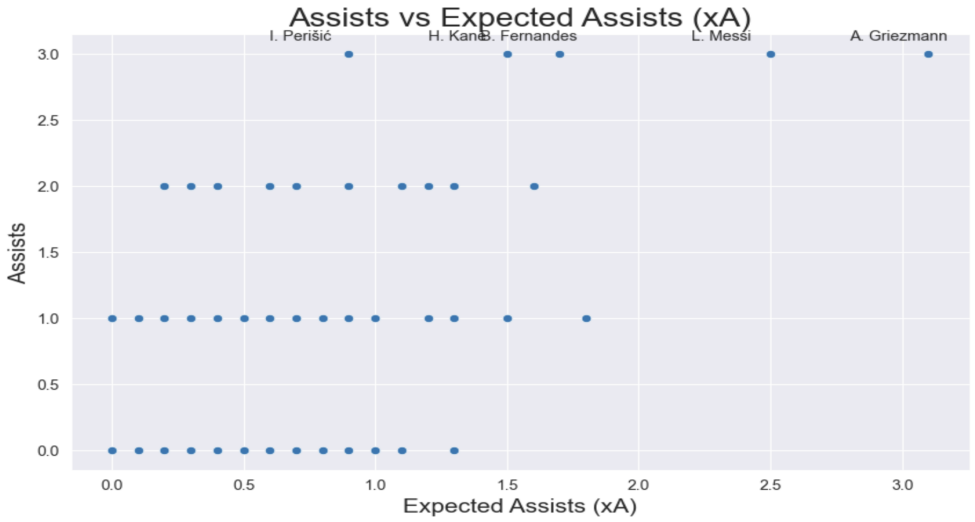

Figure 6. Scatter plot comparing expected assists against assists by players in the World Cup

It is worth noting that while Kylian Mbappe had the most goals this World Cup, Antoine Griezmann, who plays on the same team, had the most assists. This suggests that Mbappe may have relied on Griezmann to create opportunities for a goal. On the other hand, we don't see any players from Argentina on the top of Figure 6 apart from Messi, suggesting that as well being a top goal scorer for his country, he is also a large part of the build-up, and creating opportunities for his teammates.

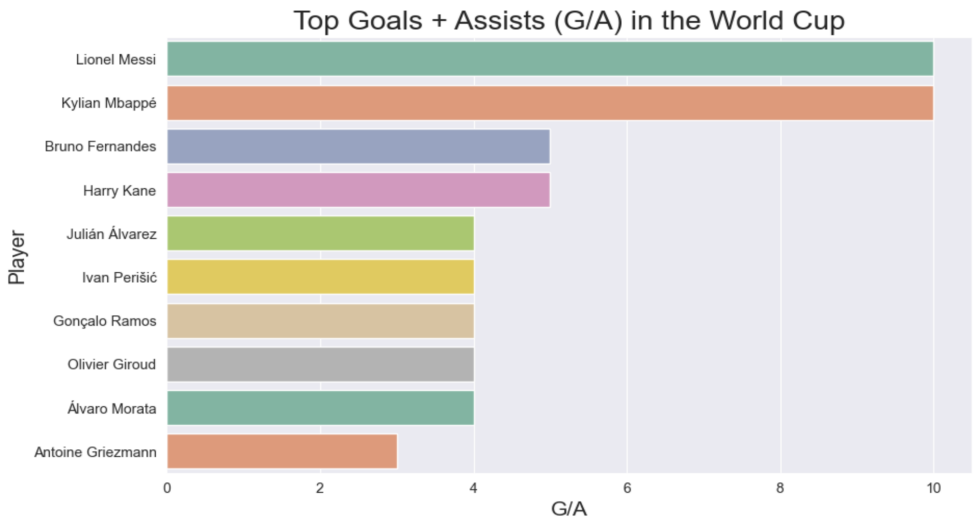

Figure 7. Bar plot that shows the players with the top combined goals + assists throughout the World Cup

G/A is a metric that is used to measure the total amount of goals and assists that a player has, by simply adding together the two. Ultimately, an easy indicator of a player's performance is their goals + assists (G/A), and we can see that tied at the top, is none other than Lionel Messi and Kylian Mbappe, who both had a total of 10 G/A throughout the tournament. Again, it's important to note that both Messi and Mbappe played in every possible match in the tournament, which can serve as a reason why their G/A output is so high compared to others. If we weren't aiming to find the most productive players, and instead were looking for the most efficient player, we might be able to use this metric as a way to help determine that.

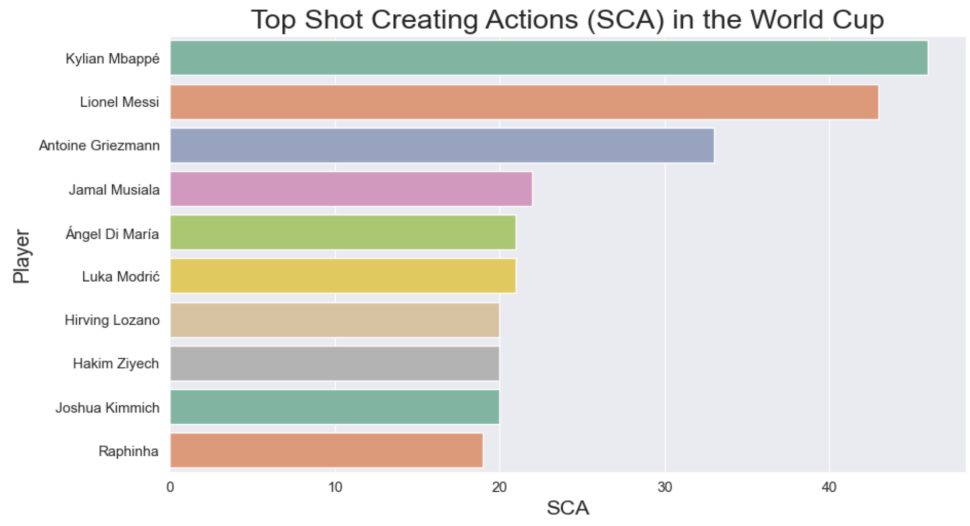

Figure 7. Bar plot that shows the players who have accumulated the most shot creating actions throughout the World Cup

A shot-creating action(SCA) is an advanced statistic that is defined as a pass, dribble, or drawing of a foul that leads to a shot. This is an interesting statistic because it can show us who is creating the most chances for their team, and who is the most involved in the attack. Here, we are continuing to see familiar faces at the forefront; Lionel Messi and Kylian Mbappe. Antoine Griezmann, our assist king, and other new names have also come forward as creative players throughout the tournament, like Jamal Musiala of Germany, and Angel Di Maria of Argentina.

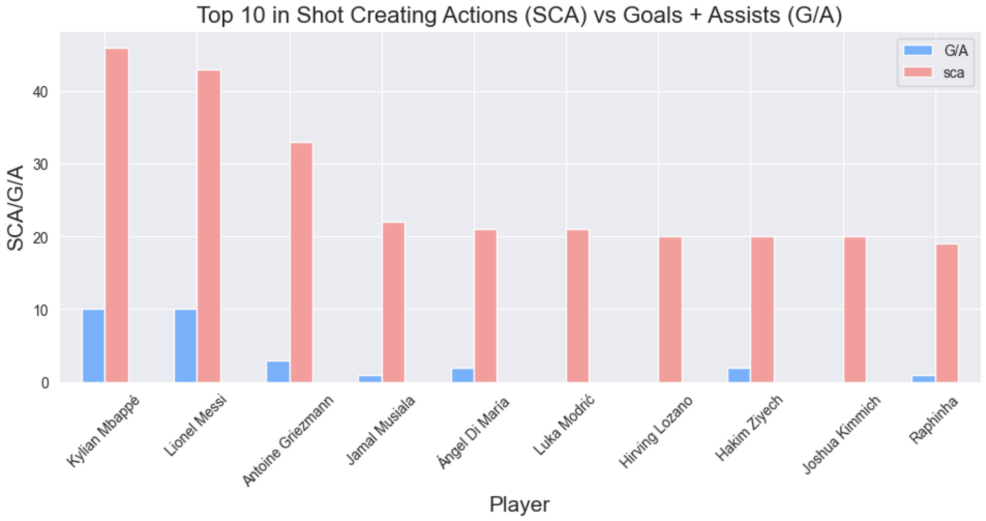

Figure 8. Side by side barplot that compares the leaders in shot creating actions and their combined goals + assists

In total, there *does* seem to be a correlation between SCA and G/A, but it is not necessarily a very strong one. While we do see some of the same names, like Mbappe, Messi, and the king of advanced assist statistics in Antoine Griezmann, we also see some new names, like Luka Modric, Hirving Lozano, and Joshua Kimmich, who have 0 G/A combined. This is an interesting observation, because it shows that while SCA is a good indicator of goal scoring opportunities (which in turn leads to goals and/or assists), it is not necessarily a strong indicator of G/A. However, it is still a good indicator of who is creating the most chances for their team, and who is the most involved in the attack.

Overall, it is clear that Lionel Messi, and Kylian Mbappe have consistently been the most productive players in the tournament. This is not surprising, as they are both world class players, and are both considered to be among the best players in the world. Having produced identical G+A outputs, and leading every advanced statistic together; with Messi showing a slight edge in terms of xG and xA, and Mbappe showing a slight edge in terms of SCA, it is clear that they are two players with different styles of play, but with a similar level of production.

3.3 Does the difference in rounds have an effect on the way that teams play?

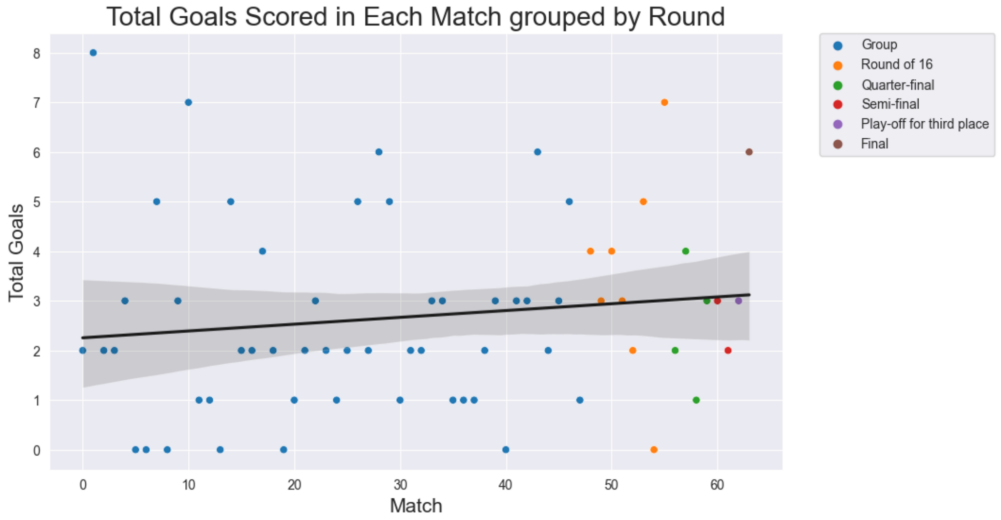

Figure 9. Scatter plot that shows the distribution of goals throughout each match of the tournament, grouped in color by the particular round that the match occurred in.

We can see that most matches resulted in a game with a little under 3 or less goals per game, but 11 of the 16 knockout matches also resulted in 3 or more goals scored, which shows us that knockout matches were more likely to perform above the mean in terms of goals scored. This is an interesting observation, because it shows that teams might be more likely to play more aggressively in knockout matches, which makes sense because there is more at stake in a knockout match than in a group stage match.

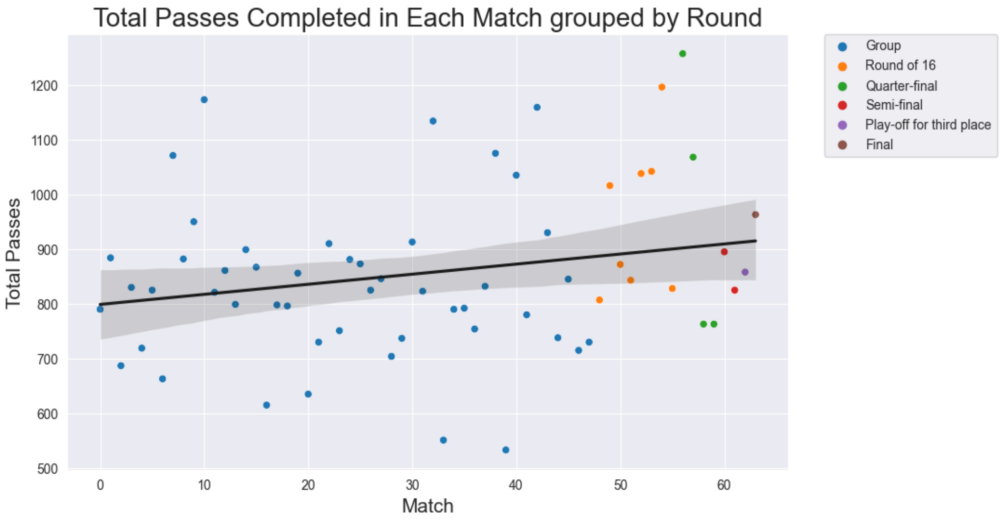

Figure 10. Scatter plot that shows the distribution of total passes throughout each match of the tournament, grouped in color by the particular round that the match occurred in.

A trend we're able to see here is that a majority of matches throughout the tournament end up with around 800-900 total passes completed by both teams combined. It's also interesting to see that there is a positive correlation between match

number and total passes, indicating that the players play more conservatively in the later matches

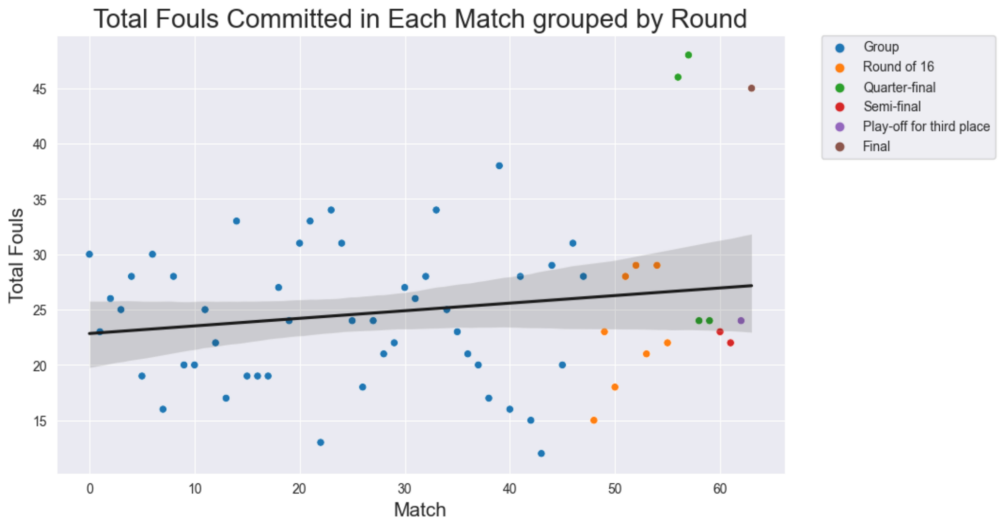

Figure 11. Scatter plot that shows the distribution of total fouls committed throughout each match of the tournament, grouped in color by the particular round that the match occurred in.

We are actually observing a tendency for less fouls than the mean to occur in knockout matches, with the exception of two quarter final matches, and the final, which were all outliers in the sense that 45+ fouls were committed in all three of these matches, which is a lot more than the mean of around 24.

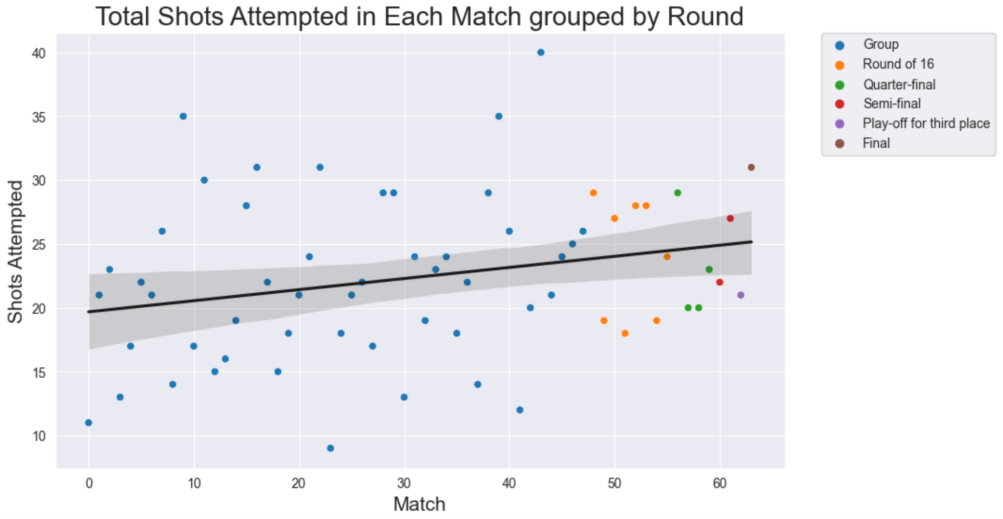

Figure 12. Scatter plot that shows the distribution of total shot attempts throughout each match of the tournament, grouped in color by the particular round that the match occurred in.

To conclude, by looking at all of these statistics, we don't particularly see a substantial change in the way teams are playing in knockout matches compared to group stage matches. However, we do see a slight increase in goals scored, and a slight decrease in fouls committed, which could be an indicator that teams are adjusting their play in knockout matches, but it is not a strong enough indicator to make a definitive conclusion.

3.4 Having possession is an emphasis in modern football, but does having more possession actually lead to more success?

We wrote code that went through each match, calculating the result for the team that had majority possession; if the team won, drew, or lost, the result would be noted. After calculating the results for every match, the team with majority possession ended up with a record of 22 wins, 15 draws, and 27 losses. The figure below shows the distribution of results, faceted by the round of the match.

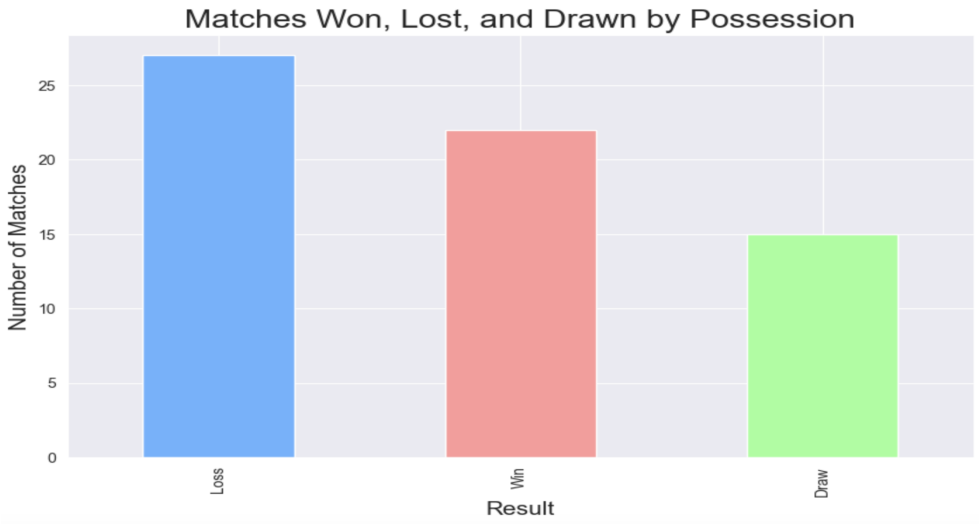

Figure 13. Bar plot showing the distribution of wins, losses, and draws for teams with majority possession in the World Cup

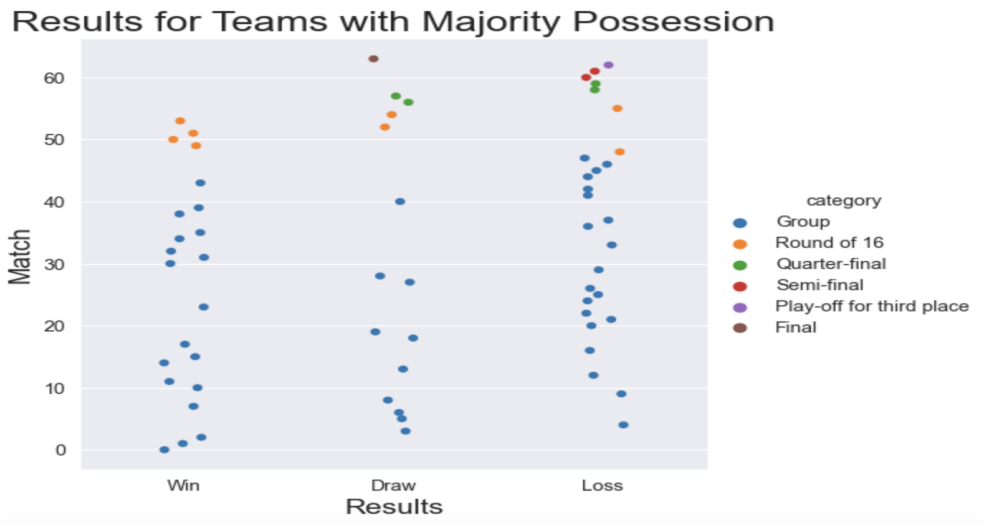

Figure 14. Catplot that shows the distribution of match results based on the team with majority possession

We can see that broken down by round:

- More teams lost with majority possession in group play, this can be attributed to the fact that there are more matches in group play, and that there are more teams in group play

- The round of 16 was the only example of a round where the team with more possession ended up with a positive record
- No team with majority possession won in the quarterfinals or in the semifinals
- The third place match resulted in a loss for the team that held majority possession
- The final resulted in a draw after extra time, and went to penalty kicks

So, we can see that possession is not a very good indicator of success in the World Cup, and that it is not a very good indicator of success in the later rounds of the tournament. This is interesting because it is a common belief that the team that holds majority possession is the team that is more likely to win the match, and that in high pressure tournament situations, there is room for many different strategies to win. For example, in the knockout match between the United States and the Netherlands, although the United States had the majority of possession, the Netherlands played to their tactics and were in complete control of the game, as they were able to win 3-1. This is a good example of how possession is not a good indicator of success, and how teams can play to their strengths and win matches, as reflected in the data.

4. Discussion

After going through all of our data visualizations, we feel confident that we have answers to all of our research questions. We are able to definitively say that the commonality between the most attended matches is the Lusail Iconic Stadium and Argentina playing. This is because the Lusail Iconic Stadium seems to be the largest stadium, Argentina is a very popular team with a devout fanbase, and Argentina played the top two viewed matches, the semi-final and final, in the Lusail Iconic Stadium.

As for who we found to be the most efficient player, it depends. Mbappe scored the most goals with 8 while Messi scored 7. However, one of Mbappe's teammates, Griezmann, topped our charts for assists while none of Messi's teammates appeared near the top for assists. This suggests that while Mbappe may have received the ball more often in positions to score, our research also suggests that many times Messi was working from a playmaking position, which makes it impressive that he was able to be among the leaders in both goals and assists. Keeping in mind that Mbappe only had one more goal with the top assister at his side, it becomes debatable as to who the most efficient player was.

We also found that as the rounds go on and tensions become higher, teams play more conservatively with an increase in passes during the later games.

Finally, we challenged the intuition that most have while spectating a game that the team with more possession is playing better and more likely to win. We found the opposite to be true. In general and especially in the later games, the team with the majority of possession is actually less likely to win based on the last world cups data.

5. Conclusions

Overall, many of our conclusions were expected, but some of them were not. For example, we would have assumed a more significant difference in the playstyle of teams between rounds, rather than a more similar approach based on the stats we compared (except for total passes, which showed a difference worth discussing previously.) Also, we were expecting to see the opposite of our results on attendance. There were many

factors that were mentioned in section 3.1 that led us to believe that there might be a decline in attendance, but we were able to see the opposite.

In the future, we would want to attempt to solve some of the following: Create an algorithm that determines a 'Starting XI' of the best players in the tournament; we would need access to more statistics that we didn't have, and more research on those statistics/what is valued more or less from a statistical point of view. Additionally, we would have wanted to go more in depth on the differences between rounds; we chose to not do this due to fear of repetition (potentially not interesting to continuously observe the same similarities/differences multiple times)

Author Contributions: Conceptualization, N.G.; methodology, C.V.; software, C.V., T.A; formal analysis, C.V.; investigation, N.G, C.V, T.A.; resources, T.A.; data curation, C.V.; writing—original draft preparation, N.G, C.V, T.A.; writing—review and editing, T.A.; visualization, C.V., T.A., N.G.; All authors have read and agreed to the published version of the manuscript.

References

1. Statista Research Department. "FIFA World Cup Average & Total Attendance." *Statista*, 8 Dec. 2022, https://www.statista.com/statistics/264441/number-of-spectators-at-football-world-cups-since-1930/.

2. Swaptr. "FIFA World Cup 2022 Player Data." *Kaggle*, 19 Dec. 2022, https://www.kaggle.com/datasets/swaptr/fifa-world-cup-2022-player-data/.

3. Iron486. "FIFA World Cup 2022: Complete Dataset." *Kaggle*, 18 Dec. 2022, https://www.kaggle.com/datasets/die9origephit/fifa-world-cup-2022-complete-dataset/.

4. "Men's Ranking." FIFA, https://www.fifa.com/fifa-world-ranking/men/.