Group 4 Capstone Project - World Cup 2022

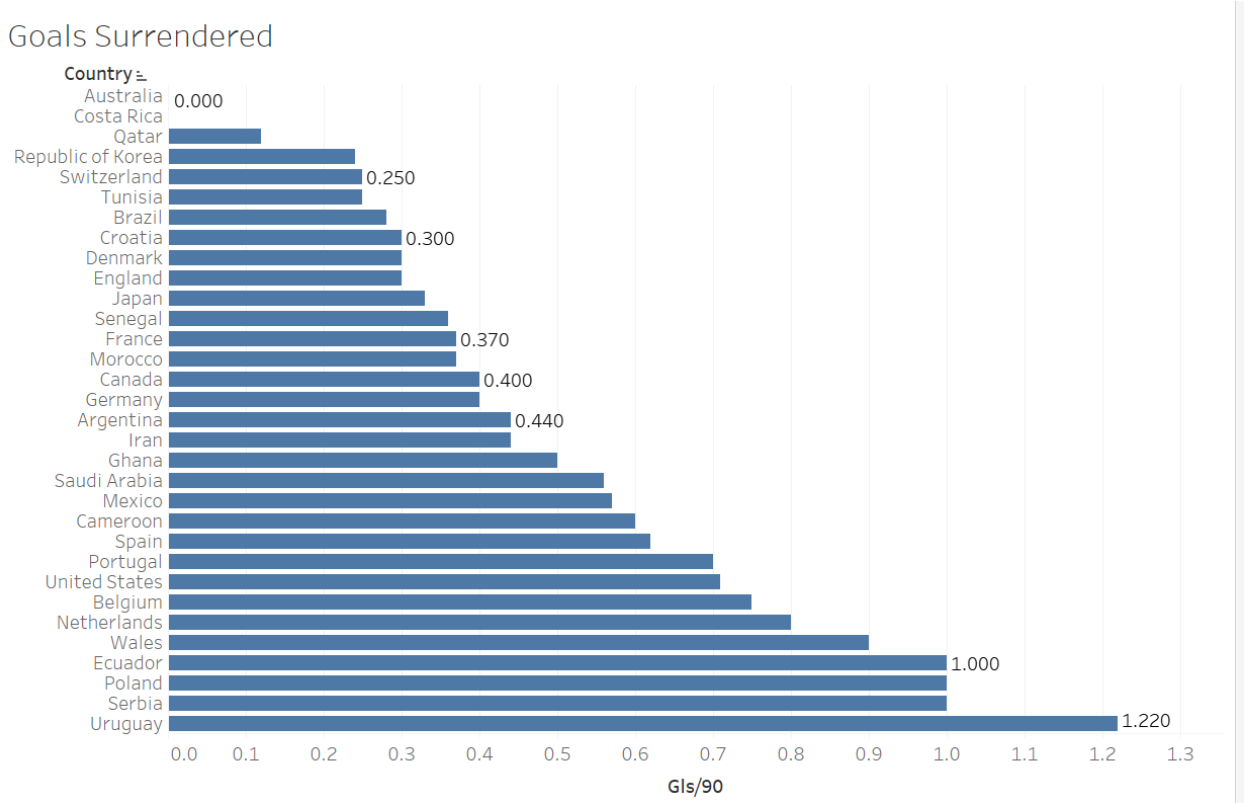
We want to use previous World Cup data along with our own machine learning model to run 100,000 simulations in order to predict the most likely outcomes at various stages for every team included. We will use machine learning to train the models and predict what the results of the group stage will be, then the elimination rounds.The 2022 FIFA World Cup Logo (Qatar) Colors with Hex & RGB Codes has 4 colors which are Ocean Boat Blue (#1077C3), Picton Blue (#49BCE3), Mikado Yellow (#FEC310) and Dark Scarlet (#56042C).

Most of us are sports fans and if not then we are all interested in current events happening around the world. The FIFA World Cup begins this Fall in Qatar and many fans will be tuning in. In 2018 it was estimated that a combined 3.57 billion viewers watched the World Cup which is almost half the world's population. Throughout the bootcamp we have seen examples of how data science and machine learning can be applied to sports and we thought we would use this as inspiration. There is a ton of historical data that we can use to build our machine learning model in hopes to predict certain outcomes within the World Cup.

Given the availability of public odds for the World Cup, we put an emphasis on building a machine learning model that would produce results that make sense and ultimately give us a sense of how teams / groups will perform. Our group expected to build a model and in the end we succeeded in doing so, however there were definitely some hiccups and difficulties. Throughout this bootcamp we have run into many hiccups and working through these alone or with a team prepared us for some of what we would run into while working on the project. Below

is a side by side comparison of the implied odds that our model calculated vs what the sharps in Vegas are saying. For the most part the shape of the graphs are mostly similar, however in just looking at these results I'm curious as to why our model does not favor the South American teams as much as teams from Europe.

With a topic like ours, we found it easy to get lost in drawing up ideas of what to create and what we could possibly create. First, what data would we gather, how would we gather it, and then at what level would we want to analyze it? The possibilities are fairly endless and we quickly learned that in most instances simplicity and consistency would be key, especially when it came to choosing what data to use in creating databases. A couple of the metrics that we decided to focus on were goals and goals against as these dictate the outcomes we wanted to predict. We decided to look at this on a team level because of the simplicity and of course that also lends itself to an area of criticism when looking at our model's accuracy. See below for average goals per game in World Cup Qualifying:



Immediately from looking just at the goals per game you can see that the teams who our model likes the most, are also near the top of goals scored per game in qualifying. Another limiting factor analysis in this piece is that not every team plays the same number of games and some (host country) automatically qualify. Qualification is done versus the countries in your region so if a region in general is not as quality as say Europe, it's difficult to account for this. Understanding this it's easy to see that comparing these metrics across teams as if they are all equal is a necessary assumption that we had to make. Given more time I think it would be really interesting to try and do this at an even more granular level, as the final rosters have not been released yet. I would expect that if we factored in more data on an individual basis we would get more accurate results.

We decided to make a few databases; one focuses on the US potential opponents in each round of elimination and their most likely opponent at each stage. The other database we created contains all knockout games played and a sample query is provided below:

dbcreation.ipynb ● .gitignore

DB > dbcreation.ipynb > con = engine.connect()

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Interrupt | Variables ··· PythonData (Python 3.7.11)

```
con = engine.connect()

query = """
    SELECT *
    FROM knockoutdb
    WHERE "Group"="Finals"
    AND "Match Winner"="Netherlands"
    ;
    """

df = pd.read_sql(query, con)
df
```

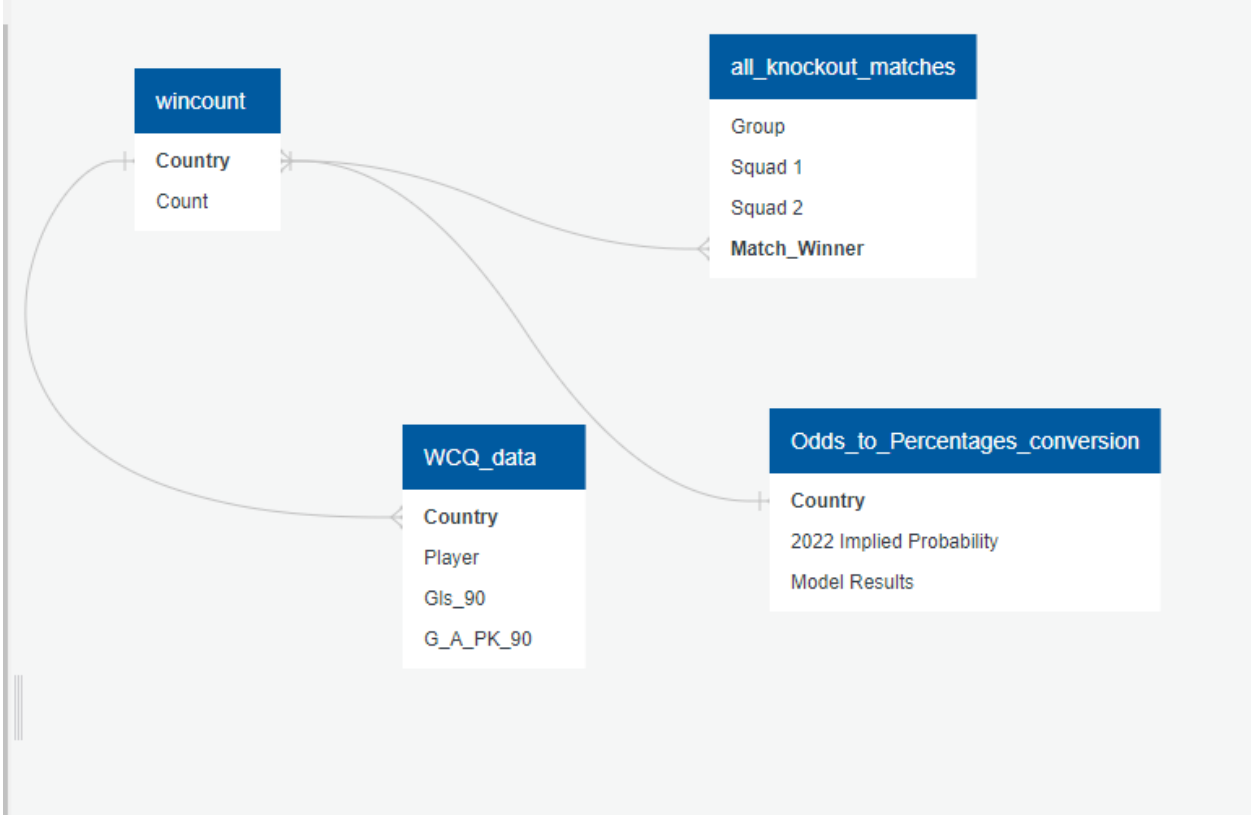
[9] ✓ 4.9s Python

	index	Match	Squad 1 Key	Squad 2 Key	Group	Stage	Squad 1	S1%	Squad 1 Seed	Squad 2	S2%	Squad 2 Seed
0	1400001	63	Knockout Round	Knockout Round	Finals	Knockout	Brazil	0.815674	M61	Netherlands	0.847293	M62
1	1400015	63	Knockout Round	Knockout Round	Finals	Knockout	Netherlands	0.847293	M61	Germany	0.819723	M62
2	1400024	63	Knockout Round	Knockout Round	Finals	Knockout	Netherlands	0.847293	M61	Argentina	0.713177	M62
3	1400027	63	Knockout Round	Knockout Round	Finals	Knockout	Netherlands	0.847293	M61	Brazil	0.815674	M62
4	1400030	63	Knockout Round	Knockout Round	Finals	Knockout	Netherlands	0.847293	M61	Denmark	0.778587	M62
...	...	...	...	...	...	...	...	...	...	...	...	...
13915	1499960	63	Knockout Round	Knockout Round	Finals	Knockout	Netherlands	0.847293	M61	France	0.780360	M62

This query gives us a dataframe containing all the finals matches in our simulation that Netherlands wins. As you can see, the index of 13,915 is equal to the implied 13.92% chance of winning that our model predicts.

Name	Type	Schema
Tables (4)		
Odds to percentages ...		CREATE TABLE "Odds to percentages Conversion" ("Country" TEXT, "2022Odds" INTEGER, "2022ImpliedProbability" TEXT, "ModelResults" TEXT)
Country	TEXT	"Country" TEXT
2022Odds	INTEGER	"2022Odds" INTEGER
2022ImpliedProba...	TEXT	"2022ImpliedProbability" TEXT
ModelResults	TEXT	"ModelResults" TEXT
all_knockout_matches		CREATE TABLE "all_knockout_matches" ("Match" INTEGER, "Squad1Key" TEXT, "Squad2Key" TEXT, "Group" TEXT, "Stage" TEXT, "Squad1" TEXT, "S1%" REAL
Match	INTEGER	"Match" INTEGER
Squad1Key	TEXT	"Squad1Key" TEXT
Squad2Key	TEXT	"Squad2Key" TEXT
Group	TEXT	"Group" TEXT
Stage	TEXT	"Stage" TEXT
Squad1	TEXT	"Squad1" TEXT
S1%	REAL	"S1%" REAL
Squad1Seed	TEXT	"Squad1Seed" TEXT
Squad2	TEXT	"Squad2" TEXT
S2%	REAL	"S2%" REAL
Squad2Seed	TEXT	"Squad2Seed" TEXT
S1_Prob	REAL	"S1_Prob" REAL
S1_wins	INTEGER	"S1_wins" INTEGER
Simulation	INTEGER	"Simulation" INTEGER
MatchWinner	TEXT	"MatchWinner" TEXT

wc_rosters		CREATE TABLE "wc_rosters" ("Country" TEXT, "Player" TEXT, "Age" INTEGER, "MP" INTEGER, "Starts" INTEGER, "Minutes" INTEGER, "fullgmsplayed" REAL, "Gls
Country	TEXT	"Country" TEXT
Player	TEXT	"Player" TEXT
Age	INTEGER	"Age" INTEGER
MP	INTEGER	"MP" INTEGER
Starts	INTEGER	"Starts" INTEGER
Minutes	INTEGER	"Minutes" INTEGER
fullgmsplayed	REAL	"fullgmsplayed" REAL
Gls	INTEGER	"Gls" INTEGER
Ast	INTEGER	"Ast" INTEGER
G_PK	INTEGER	"G_PK" INTEGER
PK	INTEGER	"PK" INTEGER
PKatt	INTEGER	"PKatt" INTEGER
CrdY	INTEGER	"CrdY" INTEGER
CrdR	INTEGER	"CrdR" INTEGER
Gls_90	REAL	"Gls_90" REAL
Ast_90	REAL	"Ast_90" REAL
G_A_90	REAL	"G_A_90" REAL
G_PK_90	REAL	"G_PK_90" REAL
G_A_PK_90	REAL	"G_A_PK_90" REAL
wincount		CREATE TABLE "wincount" ("Country" TEXT, "WinCount" INTEGER, "Win%" TEXT)
Country	TEXT	"Country" TEXT
WinCount	INTEGER	"WinCount" INTEGER
Win%	TEXT	"Win%" TEXT



Machine Learning / Predictive Analytics

The machine learning portion of the project had two seperate parts. First, two machine learning models were created to predict match outcomes. Once we were able to build out a way to predict match outcomes we designed a simulation of the entire World Cup and ran the simulation 100,000 times and observed the results.

The two machine learning models we created used Ridge Regression to predict squad offensive strength (xGS) and squad defensive strength (xGA). In order to create our machine learning models, we used 2019-2022 football match data from 147 different countries across nine different competitions gathered from FBref.com to train our models. The data used was at the competition level - so each data point was a team's average / 90 across the entire competition. The xGS used per 90 minute rates of team statistics like shots on target, possession percentage, and interceptions won to predict offensive strength. The xGA model used opponents' per 90 rates of many of the same statistics used in the xGS model. Shots on Target / 90 was the feature with the highest correlation to the target (goals scored and goals allowed / 90) for both models. Both models also featured each team's Confederation encoded as a one-hot variable.

A number of different machine learning models were tested, but we ended up using a Ridge Regression for both models. The xGS model almost used a Gradient Booster Regressor model because the RMSE was lower, but that led to more lopsided simulation results i.e. the Netherlands won ~20% of all World Cups instead of the 14% that they won in our final 100,000 World Cup Simulation. We know that the real World Cup is difficult to predict, so we would rather have more random results - especially if our predicted winner, the Netherlands, has roughly an 8% chance according to Vegas Odds. Our final xGS Model had an R^2 value of 60% when tested with the World Cup Squad data used to make our predictions. The final xGA model had a lower R^2 value - 45%, but the root mean standard error (RMSE) was lower than the xGS model's RMSE.

Once we had xGS and xGA values for each world cup squad, we plugged those numbers into a "Percent of Points Taken" formula that is similar to the Pythagorean expected wins formula used in baseball. This number gives us the "percentage of points" that a squad could expect to get from a matchup in a vacuum i.e. no opponent information needed.

$$Percent\ Points\ Taken\ = \frac{xGS^{1.2}}{xGS^{1.2} + xGA^{1.2}}$$

These values were then plugged into the Log5 equation to get the probability that one squad would beat another. According to Wikipedia, the Log5 equation is a method of estimating the probability that team A will win against Team B based on the "odds ratio" between the two teams against a larger set of teams. Instead of an odds ratio, we used the % of Points Taken Calculation. Below is the formula (PPT = Percent of Points Taken):

$$Probability\ Squad\ 1\ Wins\ =\ \frac{(Squad\ 1\ PPT) - (Squad\ 1\ PPT * Squad\ 2\ PPT)}{(Squad\ 1\ PPT + Squad\ 2\ PPT) - (2 * Squad\ 1\ PPT * Squad\ 2\ PPT)}$$

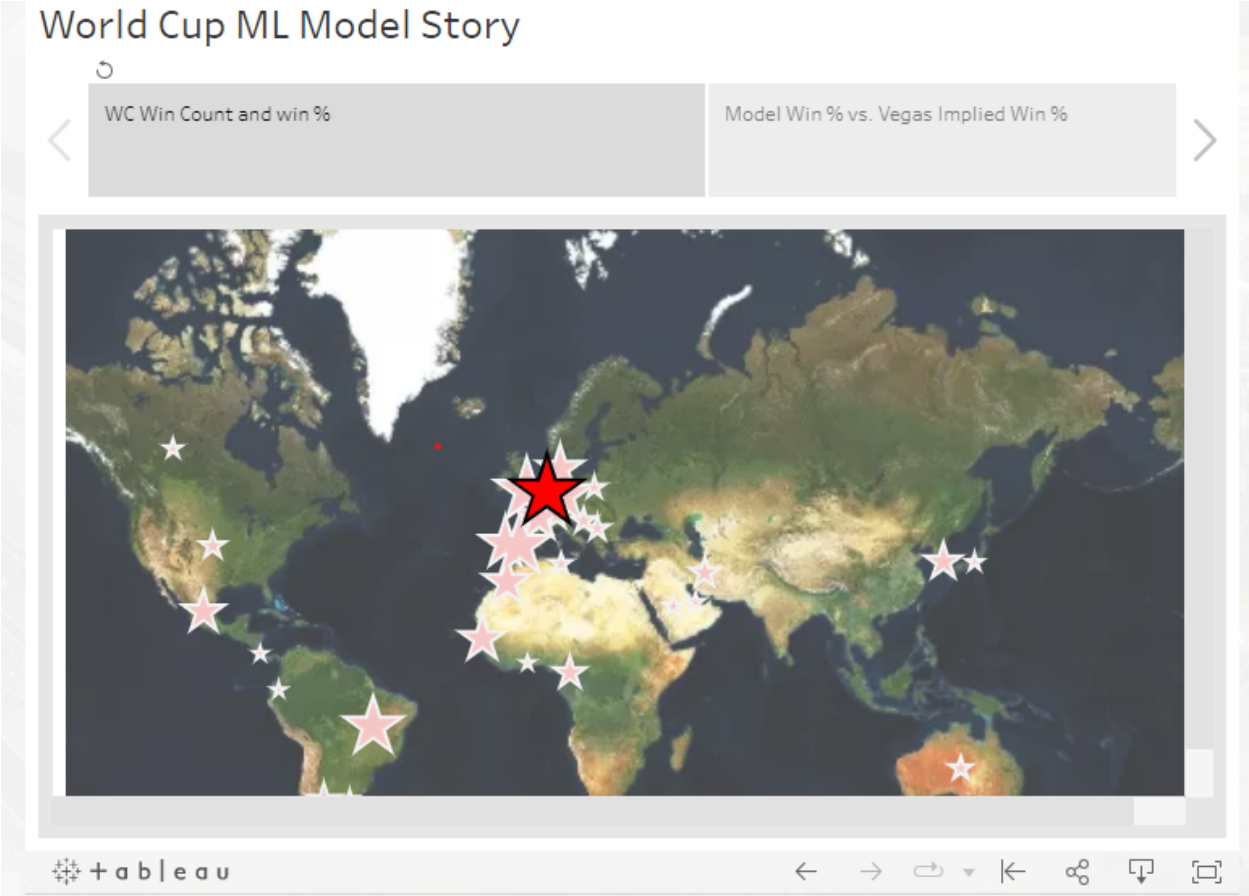
We then used these probabilities as “weighted coin flips” and utilized the binomial distribution to simulate a single matchup. This is all the information needed for the first aspect of our Web App - the World Cup Matchup Predictor. The user can select two teams, and the app will return the % probability that team 1 beats team 2 according to our models and formulas across 10,000 simulations. Netherlands vs. Ghana is the least competitive match according to our predictions - Netherlands wins 86% of the time.

Once we were able to calculate a single match, we were able to then setup a series of matches to simulate the entire World Cup. Group Stage matchups are already known, so we simulated each of those matchups 100,000 times. We then took the teams that advanced out of the group stage and placed them in the knockout rounds. The knockout round matchups already have planned placements based on group stage finishing positions - for example, if the United States makes it out of the group stage they will play either the winner or runner up of Group A depending on how the United States finishes. This is bad news for the United States, as they ended up having to play the Netherlands (the best team according to our predictions) 55% of the times they made the round of 16. All of the other knockout round placements are predetermined as well - i.e. the winner of the quarter finals match will play the winner of the second quarter finals match in the semi-finals. We accounted for all of these predetermined matchups when building out our simulation.

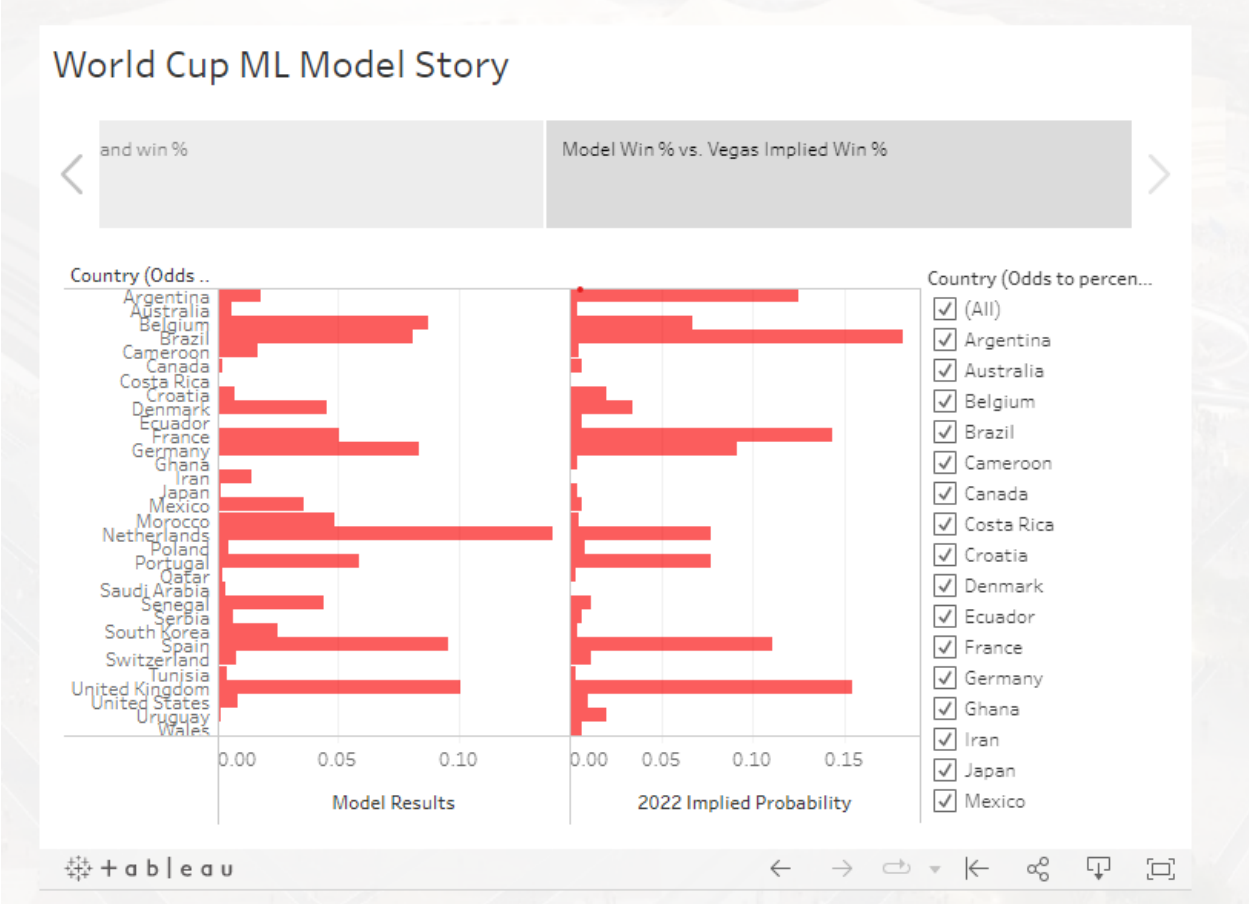
Each round of the World Cup was essentially its own simulation, with data from the prior round being used to determine what teams were in the current round. We then ran the overall simulation 100,000 times and analyzed the results.

Tableau Visualization(s)

As a team we decided to create two separate groups of visuals, one to display the findings of the Machine Learning section and the other to explore some historical data regarding the main International Tournaments (AFCON, Asian Cup, Copa America, and the Euros). It is important to note that not all data was used and some statistics are lacking significant Players/Country data. Starting with the Machine Learning visuals, Jackson created the following visual stories for -

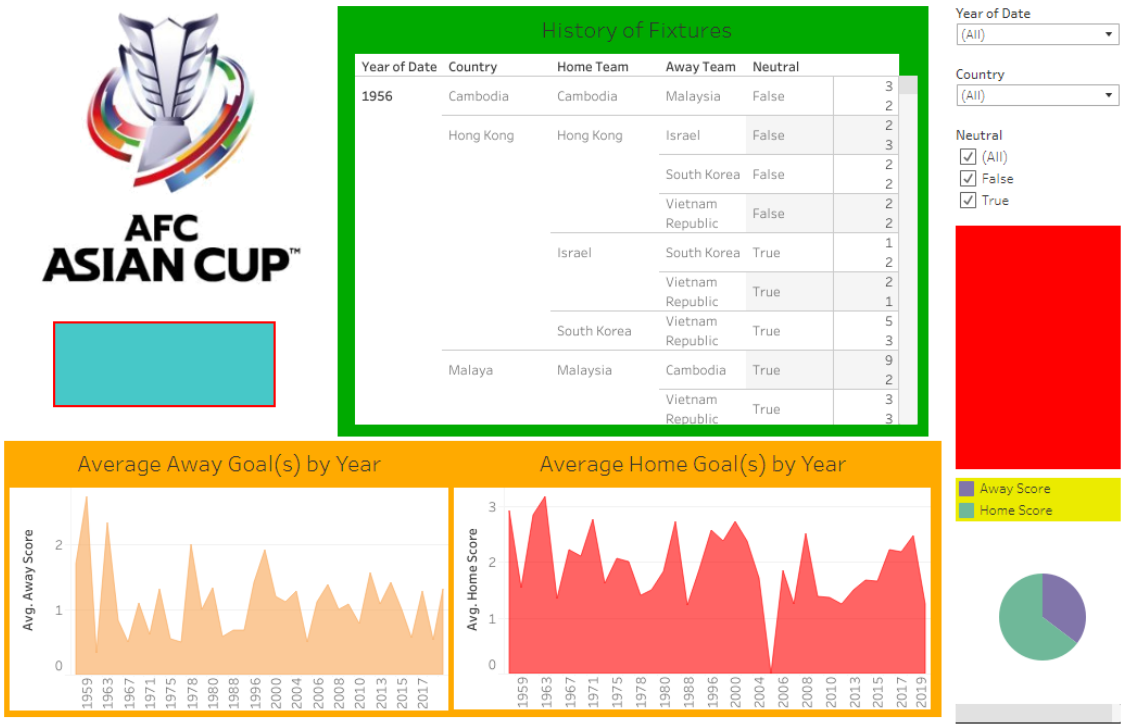


This first visual showcases all of the countries that qualified to the World Cup this fall, 2022. The data presented in this visual is each countries' total wins (National Team wins) and win percentage for the upcoming World Cup. From the previous Machine Learning section, we learned that the Netherlands have the best odds to win the World Cup. The size of the star is in reference to their win count.

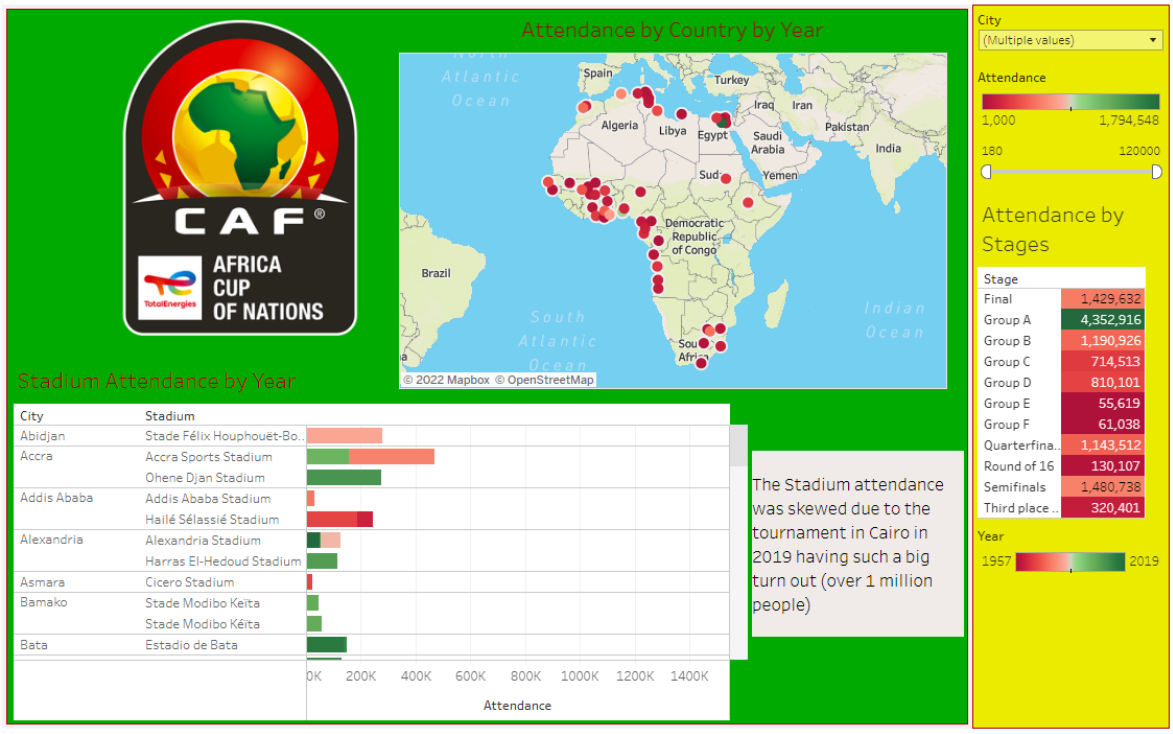


This second visual of the Machine Learning visuals shows how our Machine Learning model differed from the Vegas Implied Probability. Our model was tighter together besides the Netherlands where the Vegas odds were almost half of the odds we got from our Machine Learning. Argentina, Brazil, France, and the United Kingdom's odds were all higher in the Vegas odds compared to ours. On this visual, you can also sort by specific Country and see just their odds.

Our second Tableau tab consists of 9 dashboards in 6 stories about historical tournaments. To get them to all appear in one tab, we utilized dynamic loading, a feature in Tableau that allows us to host multiple dashboards. By using this, we were able to put the following dashboards together for AFCON, AFC Asian Cup, Copa America, and the Euros. For each competition, we tried to use the colors from the most recent competition.

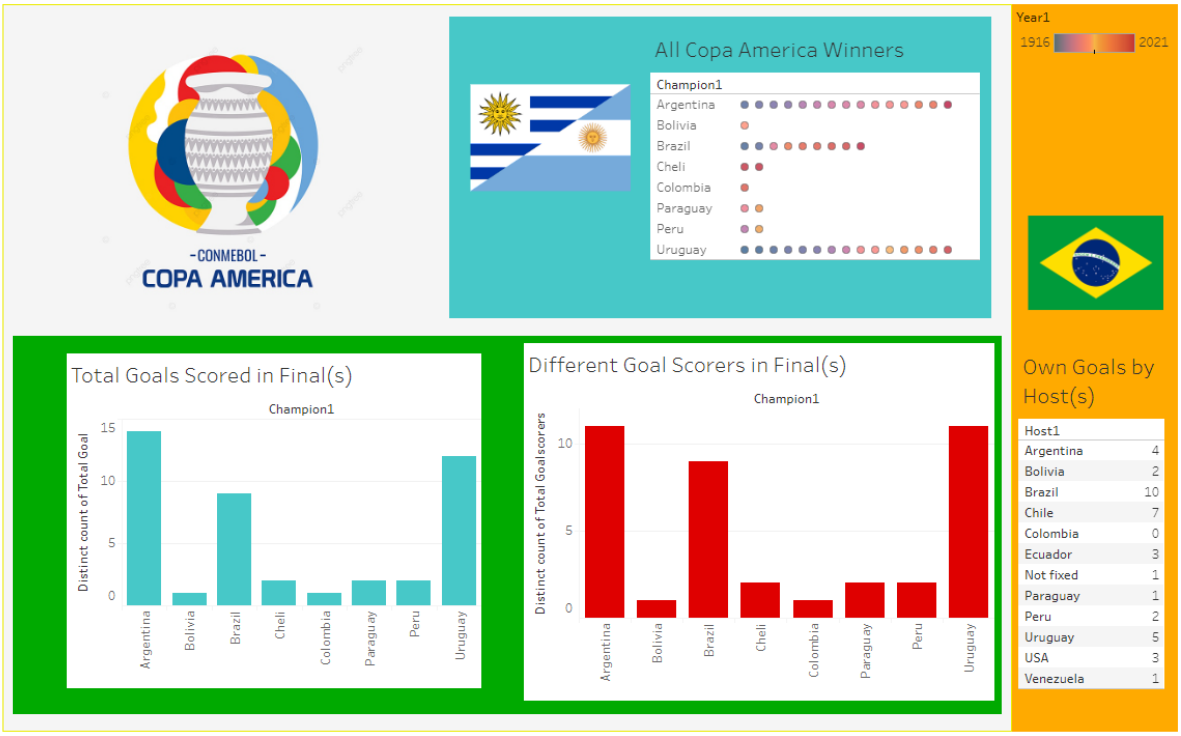


The AFC Asian Cup had the least public information out of the historical tournaments chosen. Utilizing the data we had, we created a visual that displays a match log from 1956 to 2019 that is filterable by Year and Country. We also had two graphs to display how a Home venue is better as your average goals is higher than if you are the away team.



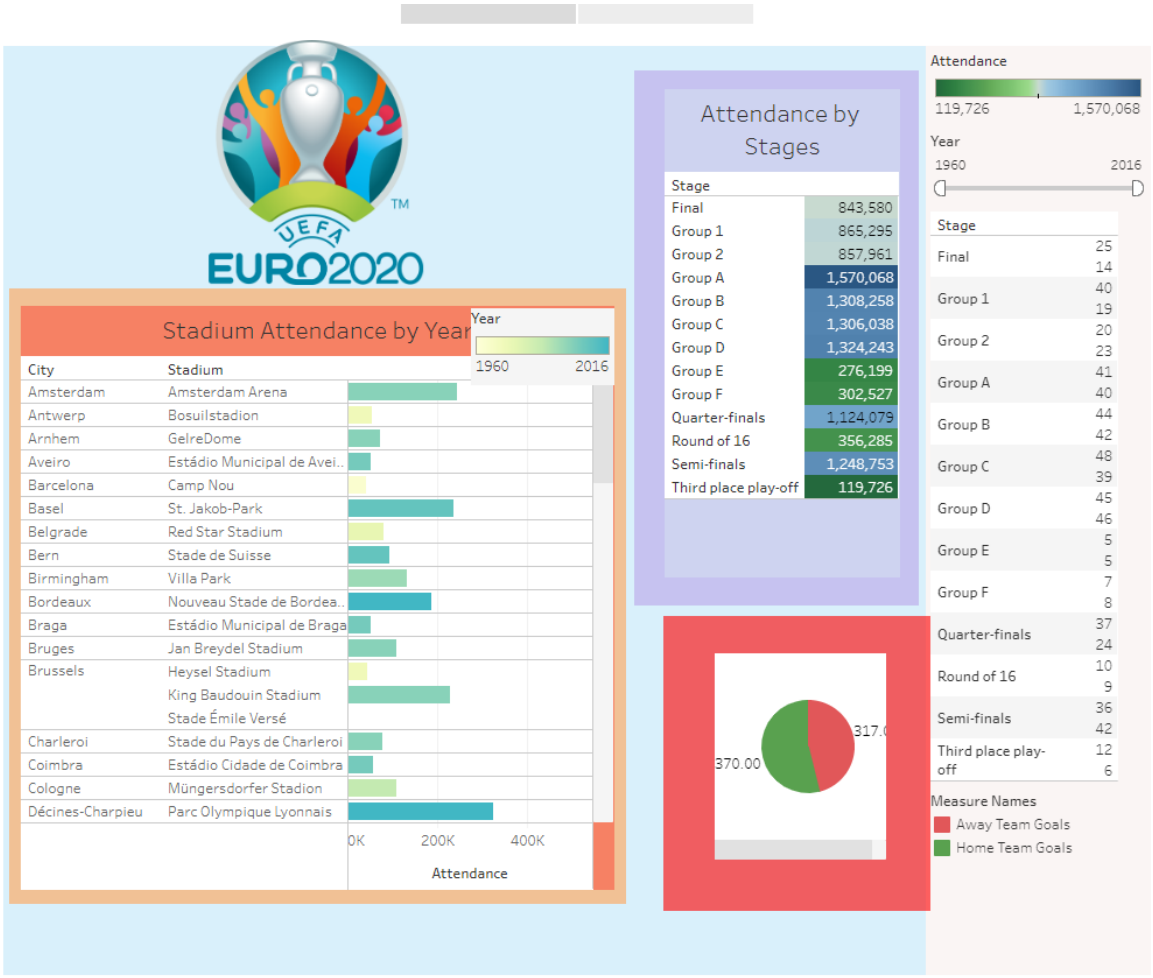
Our first AFCON page showcases the attendance statistics for the tournament and is filterable by City. There is also a map of Africa that shows each City and Stadium and their attendance statistics. There is a slider filter for the attendance as well.

Our second page of our AFCON visualization is more player orientated and focuses more on showcasing each player (that had recorded statistics) and their success. There is a table that lists each player by Country. That can be sorted by the Players' position or Country. There is also a Top Scorers table to showcase the most lethal strikers in Africa. We also added a note about Samuel Eto'o and how his stats weren't included and showed how he fared against Drogba, another African legend.

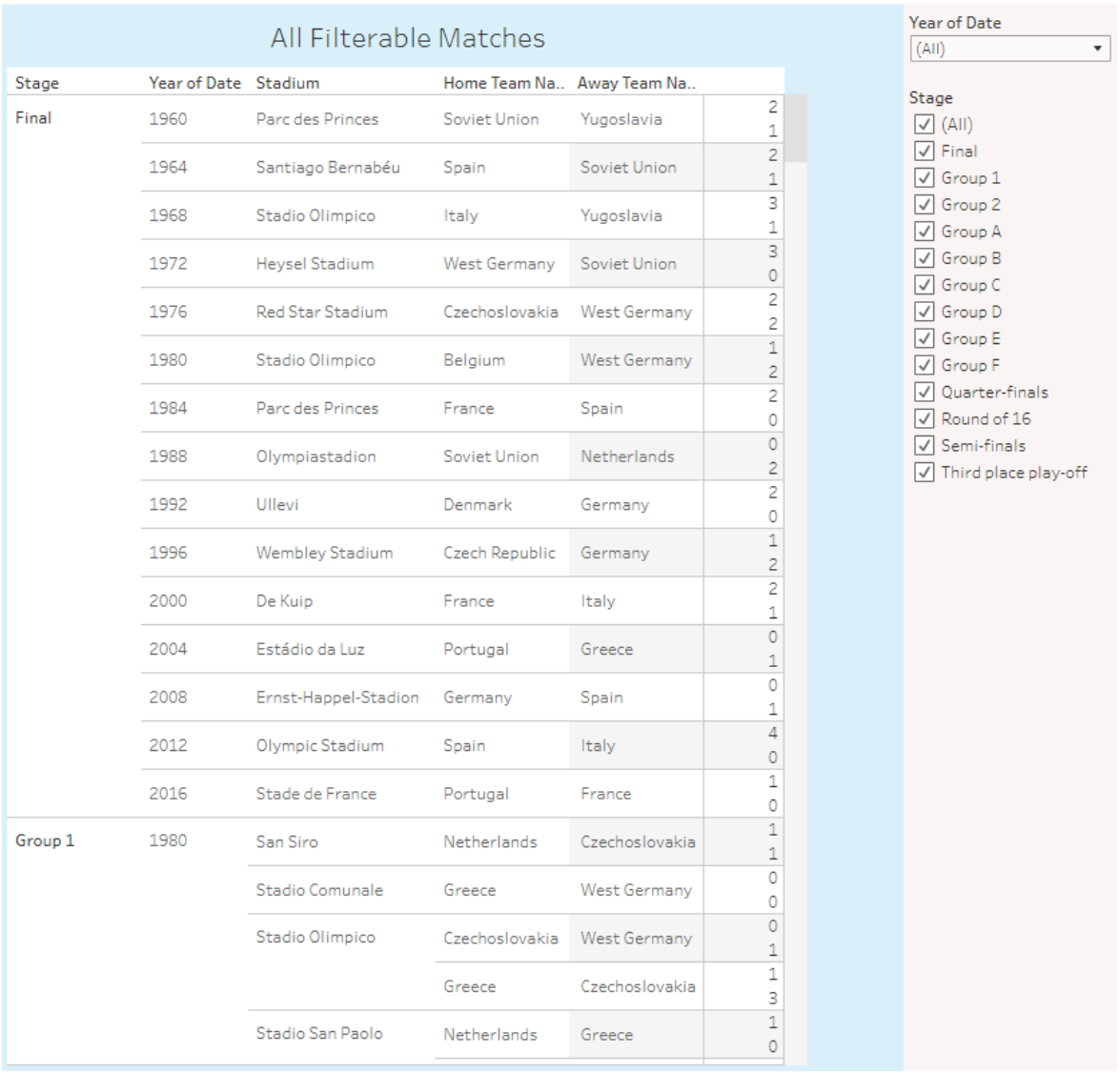


Our first Copa America visualization was to showcase the Tournaments top winners and what year they won. Two graphs were also here to show how many different players had scored for the Tournament winners and the total amount of goals scored by the Tournament winners. Another interesting statistic was how many own goals had occurred in a tournament based off of who was the host country.

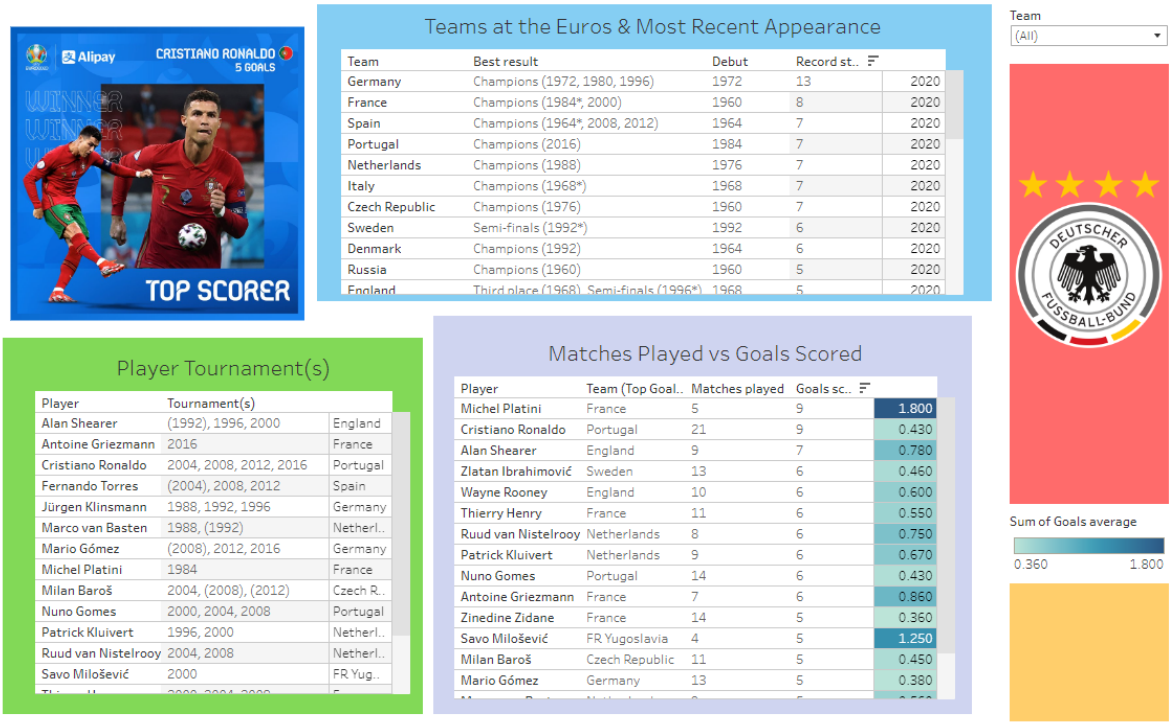
The second visualization was the same as our AFC Asian Cup dashboard, showcasing the match log history and goal scoring graphs.



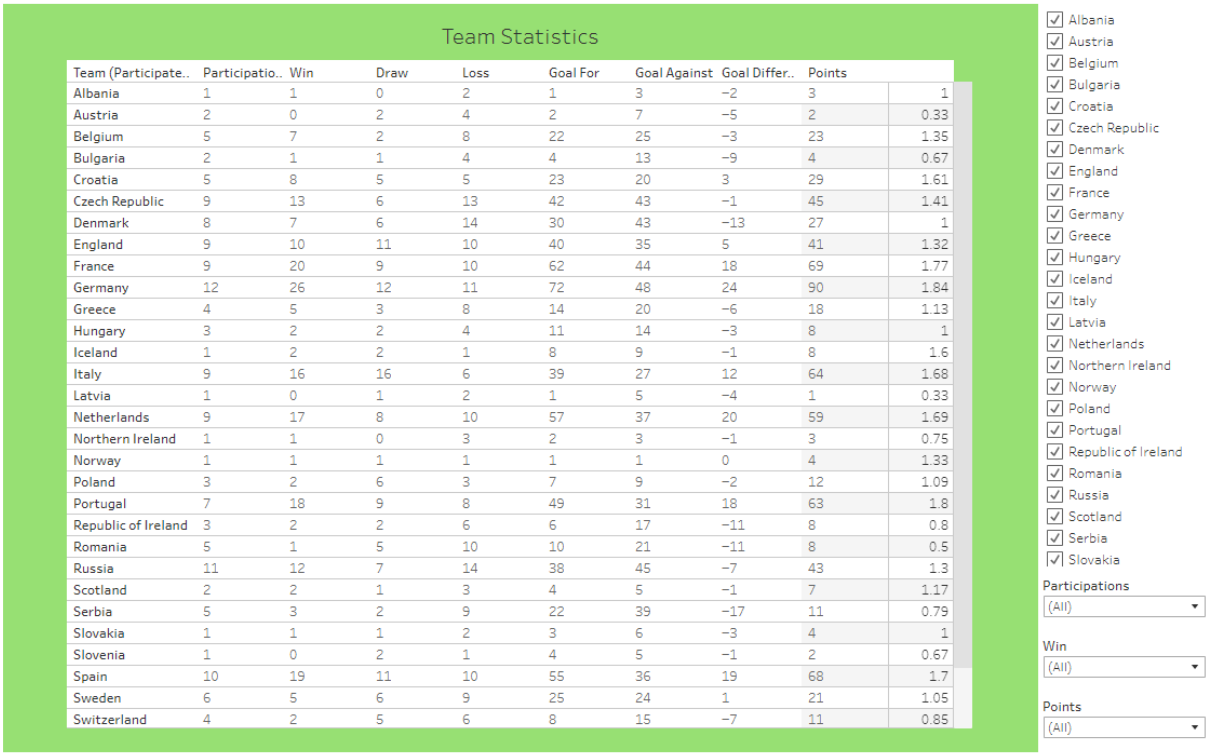
Our first Euro visualization was Attendance based, like the AFCON one. A table was made to display which stage of the tournament had the highest attendance and a table with the stadium and attendance by year was also added. A slider to sort by year(s) was added.



The second visualization was a big table of all recorded matches in our dataset. It is filterable by year and stage of the tournament.



This visualization is also like the second AFCON one, player orientated. There is a table that discusses each countries success in the Tournament historically and their record appearances in a row and their most recent.



The last visualization was a big table of each Countries statistics in the Tournament and is sortable by Country, appearances, Wins, and Points.

Some closing thoughts regarding the Historical Data were that AFCON and the Euros had the most data for team and player statistics. The most shocking find from the data was that Uruguay has the same amount of Copa America wins as Argentina and is way smaller. They also had 6 more Copa America trophies than Brazil which is insane considering the amount of talent in history that came from Brazil. The most frustrating part of the Historical Tableaus' was the lack of equal data, the Asian Cup had nowhere near as much data as the rest of the competitions and that makes it stick out much more.

Web App design.

The theme of the website was taken from the free Phantom bootstrap template provided by ThemeFisher. All of the colors were changed in the CSS files to reflect the four World Cup theme colors (Ocean Boat Blue #1077C3, Picton Blue #49BCE3, Mikado Yellow #FEC310 and Dark Scarlet #56042C.) The hexagon theme was chosen as it resembles a soccer ball. A image featuring one of the new World Cup statdiums with overlay was selected for a background.

The website has eight different pages starting with the Index/Home page which consists of seven hexagon links to each page. Each subsequent page contains a Previous/Next page button on the bottom of the page to move on to the next section. The X located in the top right corner allows the viewer to close and return back to the index.

World Cup Predictions: a queryable database allows the visitor to select two teams which are then run through 10,000 simulations to find a result.

About Us: Page containing bios and links to each of our group members.
Historical Data: Multiple Tableau pages are hosted dynamically together hosted by connecting to tableau via api. Additional buttons are added at the button to scroll between each themed Tableau viz.
Qualifying Teams: Additional Tableau Visualization
Database: Contains an image of the ERD and link to SQL database created. Additional table was added containing each teams chances as determined by simulations.
Project Details: a scrollable PDF of this written paper.
Data Sources: A hexagon themed page displaying info on the types of programs used with links added.

The web app is deployed and hosted on Heroku.
https://uofm-worldcup-capstone-group4.herokuapp.com/

Conclusions / Call to Action

There are a lot of interesting takeaways from our World Cup predictions. If our group was to place a wager on one team to over perform in the World Cup based on our predictions, it would definitely be the Netherlands! According to our predictions, the Netherlands is much more likely to win the World Cup than according to what Odds Makers think. Our World Cup simulation does not think that England, France, Brazil, and Argentina are as likely to win the World Cup as Odds Makers do. Brazil is the current World Cup favorite, with the implied odds giving Brazil an 18% chance to win the World Cup. Our predictions have them winning the world cup 8% of the time. Our models have Brazil's defense as one of the best, but their offensive abilities are not as strong as some of the UEFA teams that ended up winning the World Cup more times in our simulation.

Another intriguing conclusion from our World Cup simulation is the success of Morocco. Our models really like Morocco's defense - they have the lowest xGA of any team in the World Cup. We have Morocco winning the World Cup 5% of the time - much higher than the 1 in 200 chance according to the current odds. The United States' chance to win the World Cup is 1% according to our simulations and odds makers - we hope that the United States is more successful than that!

We don't think we can conclude that our predictions will be more accurate than what the odds makers expect - they are likely using much more advanced models to predict team success than what we are using. The World Cup is the most popular sporting event in the world, and there will likely be millions of dollars bet on teams to win the World Cup. We would expect that the odds makers put a lot of effort into making their own odds, with models that are likely more complex than ours!

Limitations and Future Work

There are a few aspects of the machine learning model that we would tweak given the opportunity. The data we used to build the model was a good start, but we believe that adding more data to the model would have potentially led to more accurate models. For example, We used tournament and competition level data that probably should have been match level data. In our current model, the data coming from a competition where each team played 15 matches is weighted the same as the data coming from a competition where a team might have only played 3 matches in the group stages and then was eliminated. Match level data would give us more data points, while also giving us more variation in the data since the competition level data is just an average of all the matches a team played.

All of the competition data was acquired from Fbref.com via manual copying and pasting of CSVs. If we were to do any additional football analysis, the immediate next step would be to come up with an efficient system for getting data from Fbref - especially match level data. We could then experiment with different statistics to see what stats are best at helping predict xGS and xGA. We did not include predictive statistics like expected goals (xG) or expected assists

(xA) - these statistics are predictive themselves, and reflect how many goals or assists a team could *expect* to have based on the quality of shots taken and passes made. Fbref did not have xG for all competitions, so we decided to only use match statistics like shots on target / 90 and possession %. Other potential football match data sources include Statsbomb and Opta Sports. These sources likely have the predictive data available for the matches we need, but for a fee. We also thought about including professional team data in the model, but decided against it. It would be interesting to see if including that data increases the predictive abilities of our model.

The models would also benefit from a feature that involves some sort of team power ranking or rating. We believe that our models likely underestimate South American teams. We thought about incorporating FIFA World Rankings into the model, but that would have required historical data separate from Fbref. Using team offensive and defensive skill ratings from FIFA video games is also another interesting potential datapoint to potentially include in future models. However, due to licensing issues that data is incomplete. For example, Qatar wasn't featured in the most recent FIFA video game - but it was included in Pro Evolution Soccer (a FIFA competitor). We would need to do additional research to see if these ratings can be used at the same time to predict squad ability.

Our model also doesn't account for draws. That doesn't matter in the knockout rounds, but in the group stages it does. For example, a team may try to lock down defensively against a superior opponent and play for a draw in the group stages of the actual World Cup. We would have to redo our calculations and system of predicting matches, but it would be useful to incorporate draws in future simulation builds.