FIFA World Cup 1930 - 2018

Abstract:

An analysis of the FIFA World Cup dataset from 1930 to 2014, starting with an examination of proper data preparation techniques. This report includes a statistical and graphical study of the match attendance over the years, and correlations of note, if any. An analysis of the match performance of the teams participating in the FIFA World Cup games was also done, and used as a basis to predict qualifying teams, and the outcomes of FIFA 2018 match-ups.

Team Members: Jevonne Peters, Muhammad Rizwan Kalim, Xiaming Gu

Course / Group: University of Toronto - SCS 3250 - Group Assignment 2018 (Group 2)

Date: August 10 2018

Data Context

The FIFA World Cup is a global football competition contested by the senior men's national teams from the 208 Member Associations of FIFA. The competition is held every four years since the inaugural tournament in 1930, with two exceptions. It is the most prestigious and important trophy in the sport of football.

Kaggle 'FIFA World Cup Data' [1] was used as the data source, with the origin being courtesy of the FIFA World Cup Archive website. It belongs to the 'Event' category, and comprised of three .csv files: World Cup, World Matches, World Cup Players. An additional, well-prepared data file featuring World Cup Hosts was compiled using the FIFA World Cup website, to supplement this dataset. (Fig. 1) shows the data dictionary of the raw data.

Element/Column	Description
RoundID	Unique ID of the round
MatchID	Unique ID of the match
Team Initials	Player's team initials
Coach Name	Name and country of the team coach
Line-up	S=Line-up, N=Substitute
Shirt Number	Shirt number if available
Player Name	Name of the player
Position	C=Captain, GK=Goalkeeper
Event	G=Goal, OG=Own Goal, Y=Yellow Card, R=Red Card, SY = Red Card by second yellow, P=Penalty, MP=Missed Penalty, I = Substitution In, O=Substitute Out

Fig. 1

Element/Column	Description
City	The city name, where the match was played
Home Team Name	Home team country name
Home Team Goals	Total goals scored by the home team by the end of the match
Away Team Goals	Total goals scored by the away team by the end of the match
Away Team Name	Away team country name
Win conditions	Special win condition (if any)
Attendance	Total crowd present at the stadium
Half-time Home Goals	Goals scored by the home team until half time
Half-time Away Goals	Goals scored by the away team until half time
Referee	Name of the first referee
Assistant 1	Name of the first assistant referee (linesman)
Assistant 2	Name of the second assistant referee (linesman)
RoundID	Unique ID of the Round
MatchID	Unique ID of the match
Home Team Initials	Home team country's three letter initials
Away Team Initials	Away team country's three letter initials

Element/Column	Description
Year	Year of the world cup
Country	Country of the world cup
Winner	Team who won the world cup
Runners-Up	Team who was the second place
Third	Team who was the third place
Fourth	Team who was the fourth place
GoalsScored	Total goals scored in the world cup
QualifiedTeams	Total participating teams
MatchesPlayed	Total matches played in the cup
Attendance	Total attendance of the world cup

Data Preparation

It is said that **80% of data analysis is spent on the process of cleaning and preparing the data** [2]. With this in consideration, the techniques outlined in (Fig. 2) were employed as part of the data preparation process.

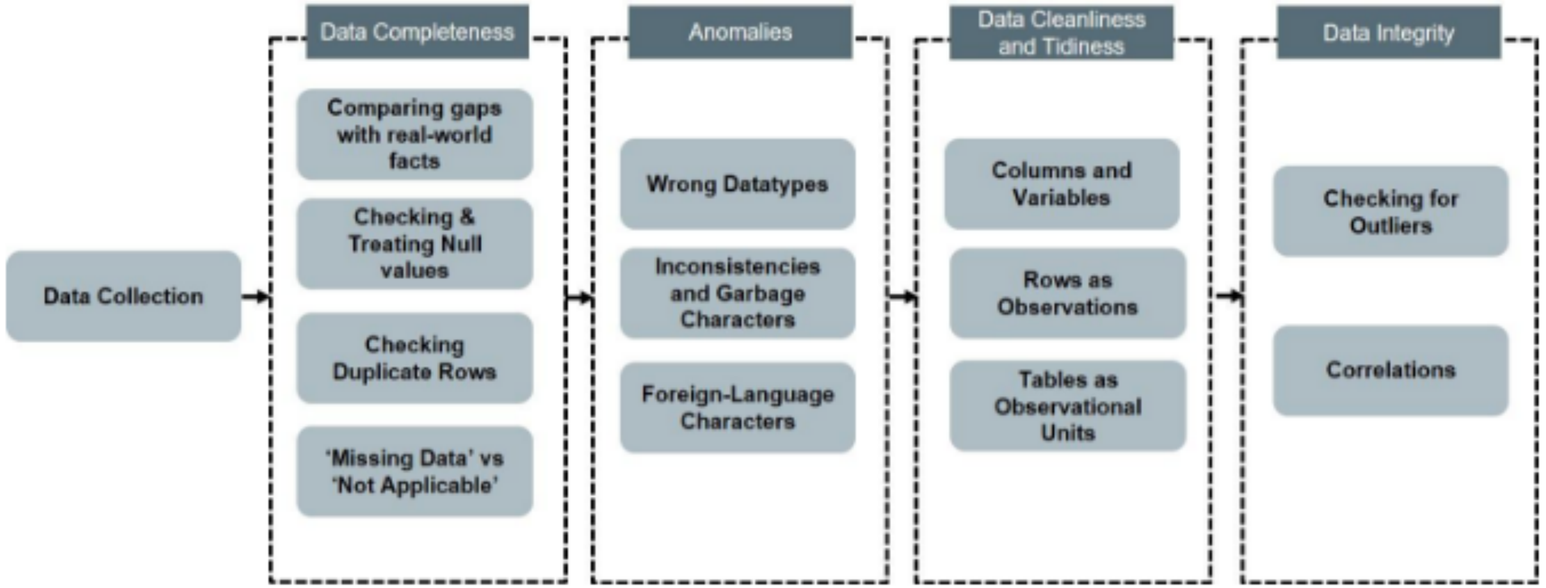


Fig. 2

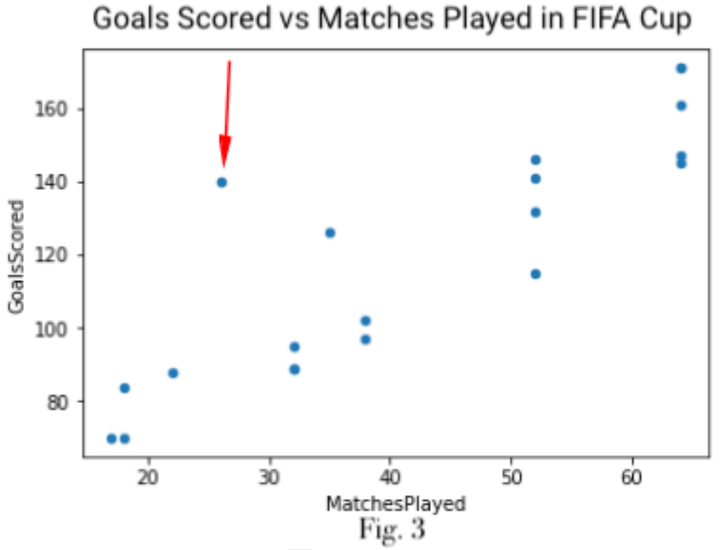


Fig. 3

FIFA World Cup Data Analysis

Jevonne Peters - Attendance Analysis

Xiaming Gu - Match Analysis

Muhammad Kalim - Cup Analysis

Attendance

- Statistics
- Graphs
- Correlations / Prediction Ability

Matches

- Statistics
- Analysis
- Predicting Qualifying Teams

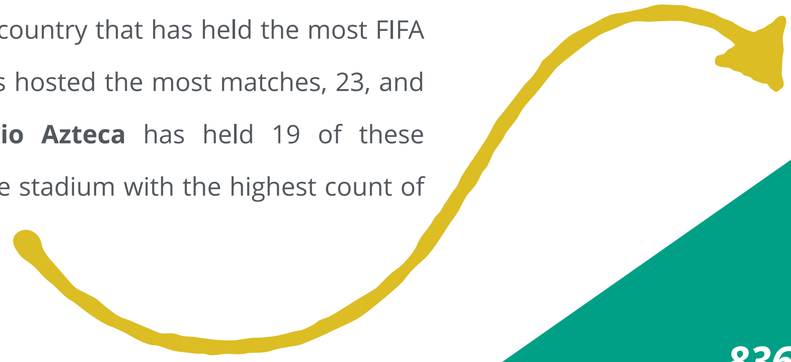
Cups

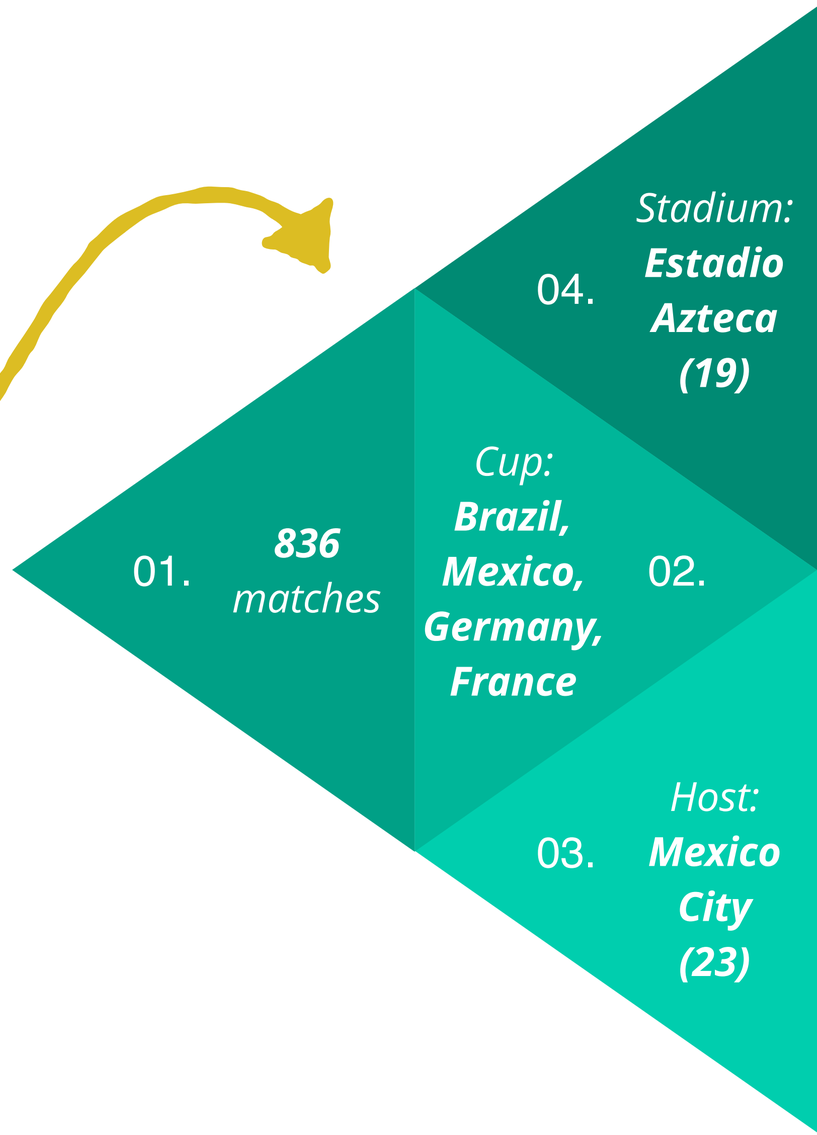
- Statistics
- Predicting the Outcome of 2018 FIFA Matches

Attendance Analysis

01. Statistics

A statistical analysis of the data showed that there were a total of **836 matches played between 1930 and 2014**, with **Brazil** being the country that has held the most FIFA cups. **Mexico City** has hosted the most matches, 23, and their stadium, **Estadio Azteca** has held 19 of these matches, making it the stadium with the highest count of FIFA matched.





03. Correlations / Prediction Ability (1)

An analysis of interest was to determine whether **attendance in earlier matches can predict a winner in later games.** That is, do fans have an intuition, which translates into attending the matches of teams that will win. This analysis was done for seven games over the years, and was determined to have **no significant prediction** relation with the overall winner of the FIFA cup for that year.

For attendance during the Round of 16 stage, only one of the seven randomly chosen dates (1990) being somewhat promising as a predictor.

03. Correlations / Prediction Ability (2)

The total attendance of the FIFA Cup Attendance that year leading up to the Final was also considered. Of the sample years taken into consideration, only 1998 seems promising, however, this match was held in France. **There seemed to a higher correlation between host country and attendance of matches for the domestic team,** that attendance of matches and FIFA winner.

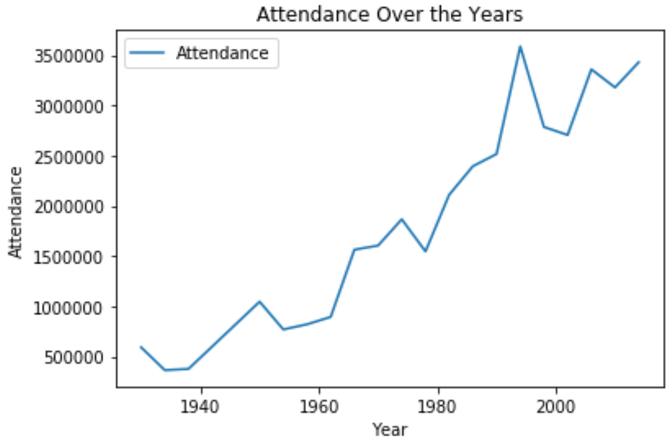


Fig. 5

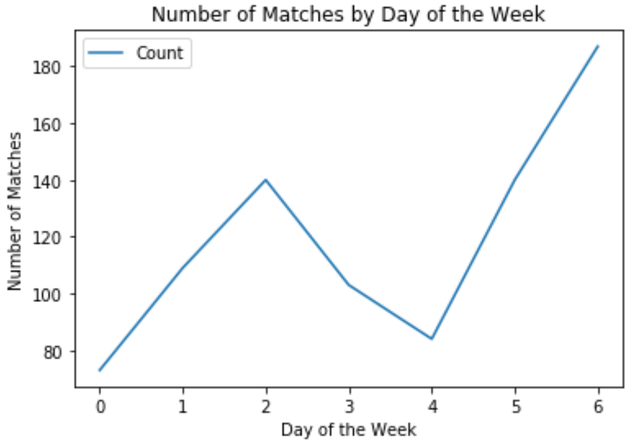


Fig. 6

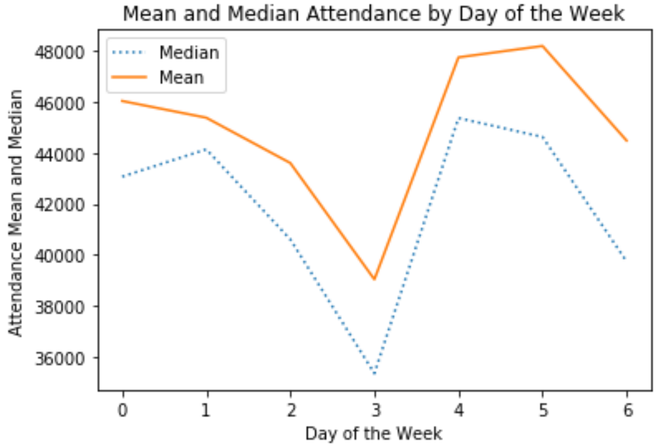


Fig. 7

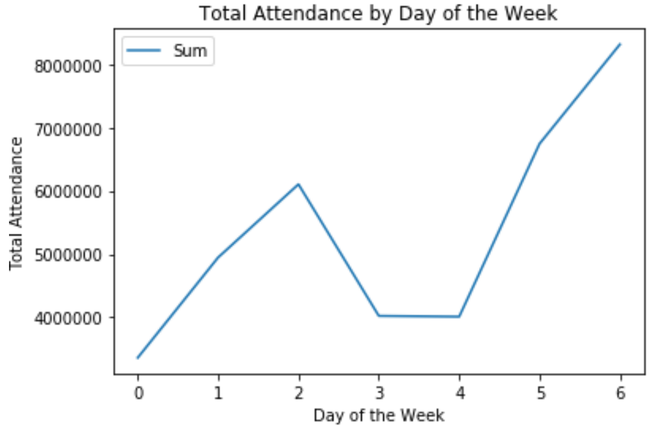


Fig. 8

Attendance Analysis

02. Graphs

(Fig. 5), drawn using matplotlib, shows the attendance over the years. It demonstrates an upward trend between 1930 and 2014. **Most matches were held on Sundays** with the lowest count being on Mondays, followed by Friday (Fig. 6 and 7).

However, despite the **low number of matches being held on Fridays** historically, the matches on **Fridays are very well attended** as the second highest. Saturday matches have the highest overall attendance (Fig. 8). **Finals are also highly attended matches,** and over the years, matches with **Brazil or Germany** playing are well attended.

Match Analysis

01. Statistics

Brazil, Germany, Italy, and Argentina were found to have played in FIFA World Cup matches the highest number of times.

01. *Most matches played:* ***Brazil***

02. ***38*** *reverser matches*

Top: ***Brazil***

03. *High placement (1930 ->):* ***Brazil***

04. *High placement (1998 ->):* ***Germany***

02. Teams that can turnaround a game

An analysis, shows teams with the most potential to turn a game around (called **reversers** in this report). These are teams who may start off seeming as the losing team, but ultimately are victors. There are **38 matches** of this type in the history of FIFA that was analysed. Amongst them, **Brazil** was noted to appear 6 times on the list. (Fig. 14)

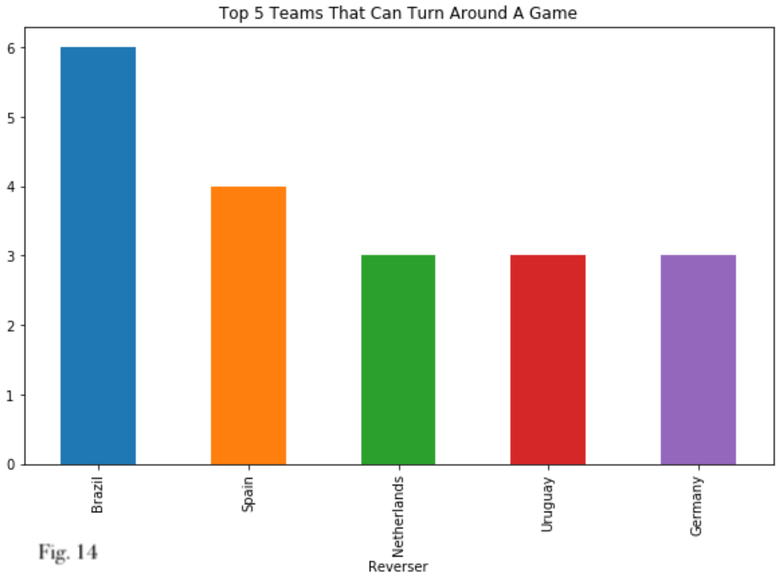


Fig. 14

Match Analysis

02. Which Team is the Best

(1) **Brazil, Germany, Italy, and Argentina** were found to place highly most frequently, as shown

(2) In 1998 the number of qualifying teams allowed was changed (it was increased to 32). If the ranks post-change are considered, **Germany** tops the list as shown in (Fig 11).

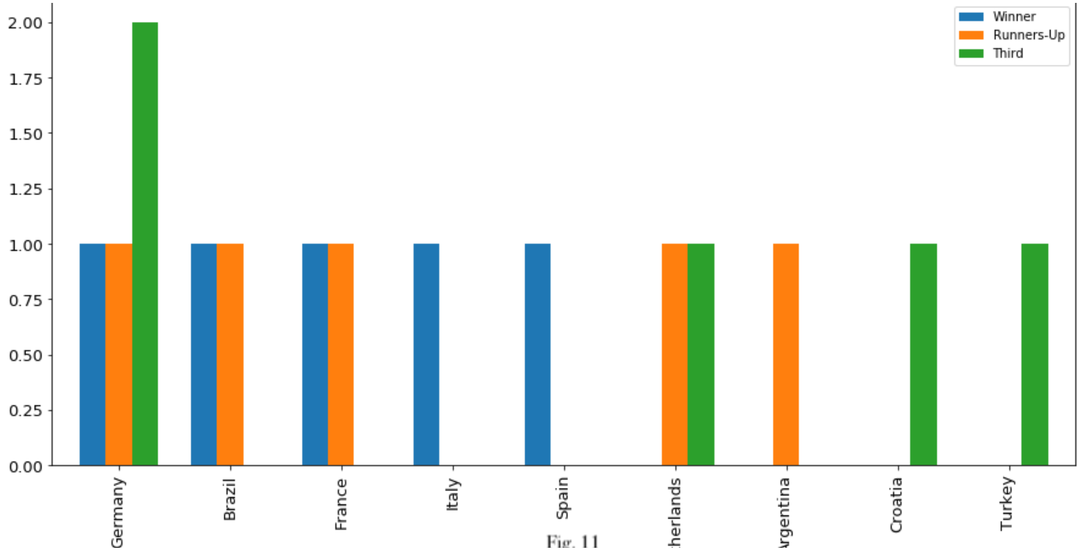


Fig. 11

(3) The goals scored were combined, and the finding was that **Germany, Brazil, Argentina and Italy,** were amongst the top in terms of overall goals.

(4) **Hungry, Brazil, Germany and Turkey** lead in terms of highest average goals per match (Fig. 13).

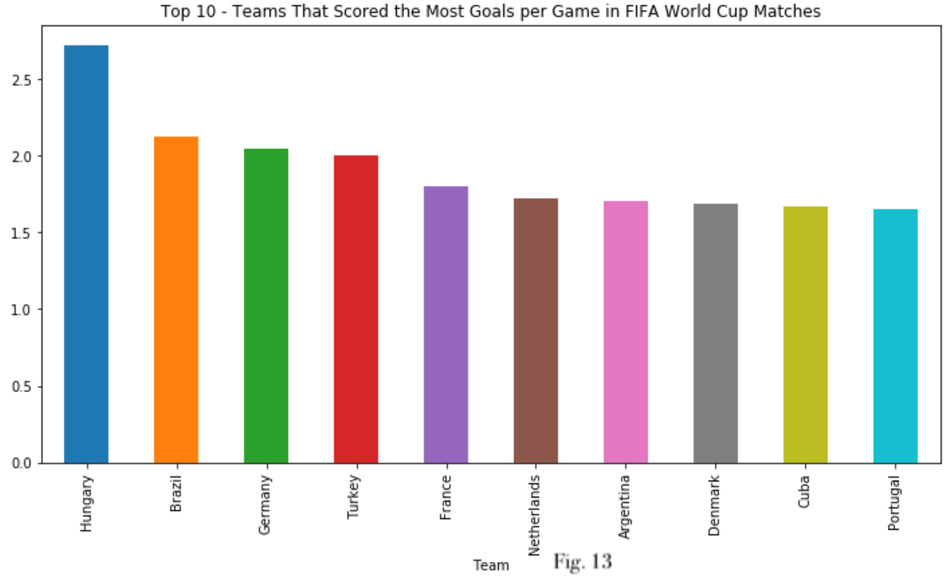


Fig. 13

Match Analysis

03. Predicting Qualifying Teams

Three methods of determining the best FIFA teams were demonstrated above:

(1) the top three teams in terms of placement in historical FIFA World Cups (1930 - 2014),

(2) the top three teams in terms of placement in historical FIFA World Cups (1998 - 2014),

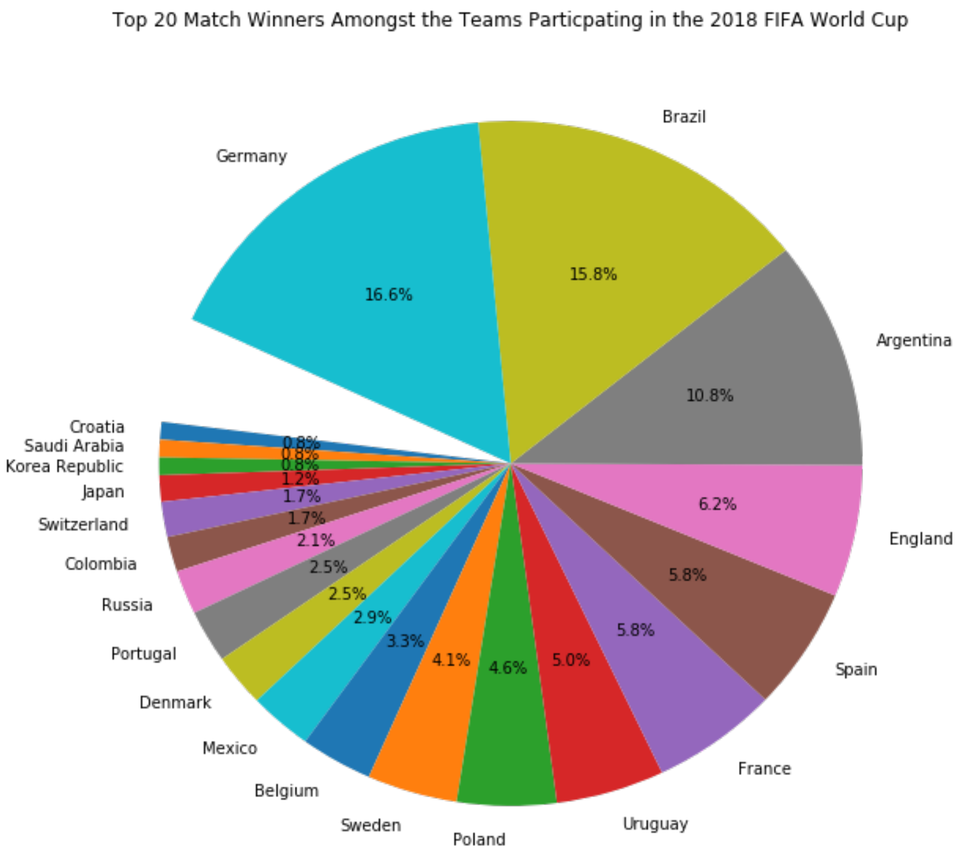
(3) the top 10 teams that scored the most goals per game in FIFA World Cup matched.

A combination of these can be used to determine the best teams, and would-be qualifiers. A predictive list of 14 of the 32 could be: **Argentina, Brazil, Cuba, Denmark, England, France, Germany, Hungary, Italy, Netherlands, Portugal, Spain, Turkey, Uruguay**. (Table 2) shows the 2018 qualifiers, with highlighted teams being the predictions. The full-list prediction was **56.25% on target (18 of 32 correct).**

Group	Qualifying Teams in 2018 FIFA Games
Group A	Russia, Egypt, Saudi Arabia, Uruguay
Group B	Morocco, Spain, Portugal, Iran
Group C	France, Australia, Peru, Denmark
Group D	Argentina, Iceland, Croatia, Nigeria
Group E	Brazil, Costa Rica, Serbia, Switzerland
Group F	Germany, Mexico, Sweden, Korea
Group G	Belgium, Panama, Tunisia, England
Group H	Poland, Senegal, Colombia, Japan
Incorrect Predictions	*Italy, Netherlands, Hungary, Turkey, Cuba* *Bosnia and Herzegovina, Côte d'Ivoire, Yugoslavia, USA, Romania, Austria, Paraguay, Csechoslovakia, Chile*
Yellow: Best-Effort Prediction Orange: Full List Prediction	

Table 2

Cup Analysis



01. Statistics

Of the teams participating in the 2018 FIFA World Cup, **four of them have never won a match, namely, Egypt, Iran, Iceland and Panama**. Of these, Egypt and Iran have played in the FIFA World Cups before. **Iceland and Panama are newcomers** to the FIFA World Cup tournament. The **top twenty match winners include Germany, Brazil, and Argentina** who are historically known for high performance in the FIFA World Cup.

Never Won:
04. ***Egypt, Iran***

01. *Newcomers:* ***Iceland, Panama***

02.

03. *Top match winners:* ***Germany***

Cup Analysis

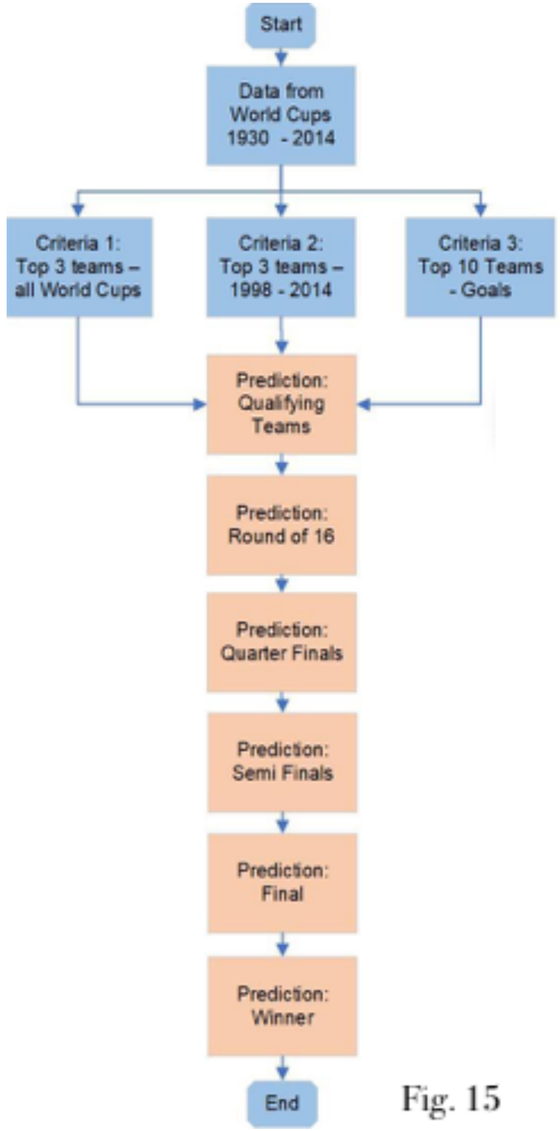


Fig. 15

Group	Round of 16 in 2018 FIFA Games
Group A	Russia, Uruguay
Group B	Spain, Portugal (Incorrect Prediction: Morocco)
Group C	France, Denmark
Group D	Argentina, Croatia (Incorrect Prediction: Iceland)
Group E	Brazil, Switzerland (Incorrect Prediction: Serbia)
Group F	Mexico, Sweden (Incorrect Prediction: Germany)
Group G	Belgium, England (Incorrect Prediction: Panama)
Group H	Colombia, Japan (Incorrect Prediction: Poland, Senegal)
Yellow: Qualifying (32) Only predictions Green: Predicted for 32 and 16 Blue: Round of 16 Only Predictions	

Table 3

02. Predicting the Outcome of Matches (1)

The remainder of this report will be based on the flow shown in (Fig. 15). The aim here is to demonstrate **the likelihood of arriving at the correct winner, now that the results at each stage of the 2018 FIFA World Cup are known**. The predictions made will compared with the actual outcome at each round.

In the previous section, the historical match performance was used to predict which teams may qualify. Now, given the full list of 32 teams that did qualify, the next step is to determine which of these are likely to move onto the next round (Round of 16) from each Group A - H. This was done by first calculating the ratio of the goals scored vs goals against for each of the teams.

This method was repeated for Groups B - H to select the top two performers who would most likely move on to the quarter-finals. (Table 3) shows the proposed round of 16 compared to the actual 2018 FIFA World Cup round of 16. The highlighted teams are the predictions.

Cup Analysis

Stage	Team Match-Ups
Quarter Finals	[France, **Argentina**], [Uruguay, **Portugal**], [**Brazil**, Mexico], [Belgium, **Japan**], [**Spain**, Russia], [Croatia, **Denmark**], [**Sweden**, Switzerland], [Colombia, England] - predicted a tie
Semi-Finals	[**France**, Uruguay], [**Brazil**, Belgium], [**Russia**, Croatia], [Sweden, **England**]
Finals	[**France**, Belgium], [Croatia, **England**]
Winner	[**France**, Croatia]
Red: Predicted in Analysis(Incorrect) Green: Predicted in Analysis (Correct)	

Table 4

02. Predicting the Outcome of Matches (2)

If the round of 16 predictions were supplemented with the predictions for the qualifying teams, meaning it was predicted in both rounds, as shown in green, the predictions would be **at best 50% on target** [3]. The prediction accuracy increases to **56.25%** if the old results are discarded, and the correct 2018 qualifying teams feed into the prediction module.

The Semi-Finals, Quarter-Finals, Finals and ultimate winner were predicted in a similar manner, each time discarding the old predictions, and using the known participants of the rounds in the calculations. (Table 4) shows the breakdown for the remaining stages of the FIFA World Cup and predicted winners of the match-ups.

Conclusion

While researching, papers that speak on methods for predicting the outcome of organised football games were consulted. A paper by Constantinou, Fenton, and Neil published in 2012, specifically looked into forecasting the outcome of the English Premier League matching during 2010/11. [4] The method of prediction used in this paper considered **both objective and subjective information**, unlike our efforts, with future efforts being on revising the **methods used for calculating the strength of the teams.**

Many other methods have been used in the past to better rank the FIFA teams. The current FIFA 2018 ranking system is closely modelled after the **Elo rating system,** which is often used to determine the relative strength of players in zero-sum games. The official FIFA system has been revised three times in the past, and criticised a few times for inaccuracy, however, the methods use richer data and more features than the crude methods used in this report. [5]

It is hypothesised that **given a better model of team strength,** combined with a feature-identification based, or otherwise **advanced method of learning what variables are important** in identifying a winner, **more accurate predictions could be performed.**

Thank You!

Xiaming Gu (Caroline)
Contribution: Match Analysis, Python Notebook Compilation.
Caroline has a B.CS in Computer Science and Technology, and several years of experience in software developing (hands-on program and mobile phone system developing). Most recently, she was a part of a team developing robotics and participated in the development of Natural Language Processing in Chinese. She has a passion for the NLP and machine learning, and is now working towards a Data Science graduate certification at the University of Toronto.

Muhammad Rizwan Kalim (Rizwan)
Contribution: Data Preparation, Cup Analysis.
Rizwan is an accomplished Information Security and IT Risk Management professional currently working for TD Bank. He holds a Masters degree in Computer Science coupled with certifications of CISA (Certified Information Security Auditor), CISSP (Certified Information Systems Security Professional), and CCSP (Certified Cloud Security Professional). His interests include the application of Artificial Intelligence and Data Sciences in the field of Information Security – a secure digital world for all – and he intends to continue upgrading his knowledge and skills in this domain after building a platform in this course

Jevonne E. Peters (Jevi)
Contribution: Attendance Analysis, Report Write-up, Presentation Design
Jevi is a Business Informatics B.ASc. McMaster University graduate, and studied Machine Learning and Artificial Intelligence (Prof. Certification) at MIT. She has worked in the IT industry over the years as a Business Intelligence Analyst, an Ethical Hacker (IT Security Consultant) and a Software Programmer, and currently works at the TRU research institute at University of Toronto. Hailed from a family of great writers, she enjoys photography, poetry and prose, and pursued the study of Art & Design at the GBC School of Design, graduating with honours. In her spare time, she works towards a Data Science graduate certification at the University of Toronto, and for a double-concentration certification in Digital Media and Social Innovation Design at OCAD U.

Endnotes

[1] "FIFA World Cup | Kaggle", Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/abecklas/fifa-world-cup. [Accessed: 08- Jul- 2018].

[2] T. Dasu and T. Johnson, Exploratory data mining and data cleaning. New York, NY: John Wiley & Sons Inc., 2003.

[3] This assumes that the outcomes of Round of 16 predictions are the same for the actual set of qualifying teams, and a set of predicted qualifying teams. As this may not be the case, the 50% quoted is an approximate likelihood of accuracy.

[4] Constantinou, Fenton, and Neil, “pi-football: A Bayesian network model for forecasting Association Football match outcomes,” QMRO Home, 01-Dec-2012. [Online]. Available: http://qmro.qmul.ac.uk/xmlui/handle/123456789/10780. [Accessed: 01-Aug-2018].

[5] S. Price, “How FIFA's New Ranking System Will Change International Soccer,” Forbes, 11-Jun-2018. [Online]. Available: https://www.forbes.com/sites/steveprice/2018/06/11/how-fifas-new-ranking-system-will-change-international-soccer/#18864e86c412. [Accessed: 02-Aug-2018].