# Exploring Relationships Between Country-Wide Health Markers and World Cup Performance


Cat Johnson
Computer Science
CU Boulder
Boulder, CO USA
johnsocf@colorado.edu

Sabine Hollatz
Computer Science
CU Boulder
Boulder, CO USA
sabine.hollatz@colorado.edu

Greg Giordano
Computer Science
CU Boulder
Boulder, CO USA
grgi1374@colorado.edu



## ABSTRACT

Over 200 different countries entered the qualification process for the 2018 Féderation Internationale de Football Association (FIFA) World Cup. Thirty-two teams qualified for the tournament, with France winning the ultimate prize for the second time since the first World Cup in 1930. FIFA has previously required that players have citizenship in the country that they wish to represent, but now requires that players demonstrate a “clear connection” to said country. The purpose of this project was to examine the relationships between a very wide variety of country-level markers including those of physical player skill, national purchasing power, generosity of the people, social support, and additional measures which in the fabric of life itself could be connecting with and influencing qualification and performance at the FIFA World Cup.

A few select yet varied datasets proved to provide statistical insight into the details of nationally categorized characteristics which included strong markers in emotional and physical health. Specifically, we acquired statistical data on national factors of health in a dataset called, ‘The Happiness Report’ which gave data points on markers including, ‘happiness rank’, ‘life expectancy’, and ‘government trust’. An adjunct dataset which included stats on player skill level and nationality allowed us to derive a transformed dataset showing a current average national skill level with normalized scores for ability markers including physical abilities such as, ‘acceleration’, ‘ball control’, and ‘crossing’. Health Data from the World Bank included ‘infant mortality,’ ‘life expectancy,’ and ‘death rate.’

Ultimately we sought to find insight, connection, and analysis with three important result containing datasets that could provide multiple perspectives on connection based on dimensional subsets. One of these key target datasets was derived from a dataset which included every FIFA game played in the men’s leagues. We aggregated this to map total World Cup game wins for each country. Another we derived gave us data on men’s World Cup results for each country in 2018. Our third result dataset allowed insight into the perspectives of Women’s soccer, containing FIFA World Cup wins per country; giving a look into gender based distinctions on this topic. With three result sets to cross reference with national markers, the plan was to gain insight into historical, recent, and gender based connection to country markers. We were expecting to find strong

similarities in connection with markers, yet also anticipating subtle distinctions to surface through our assessments.

Our process was evolutionary and discovery based, giving us room to step through exploratory data analysis, identify correlations at many levels, assess the meaning of our result. Using the Apriori algorithm to determine frequency sets from high values, we used principles of Active Learning to fine tune datasets to contain only attributes that looked to dynamically connect with results. We explored multiple Machine Learning approaches using these resulting determinations to understand possible predictability of future WC results based on Nationality from the data.

Classification was used to assess the relationship between qualification and markers of health and happiness with moderate success showing in best case scenarios an accuracy of 72%, precision of 39%, and recall of 61%. Results ranged in accuracy with granularity of the size of percentile into which we made the prediction. For instance finding strong accuracy at the 50th percentile (two labels) could sometimes mean linearly less accuracy to trying to predict teams that performed in the 95th percentile (20 labels). With some data, the results proved to be more consistent. Regression analysis of these markers with goals and wins yielded low $R^2$, values, although significant attributes ($p<0.05$) were found (life expectancy, positive affect).

## INTRODUCTION

The Féderation Internationale de Football Association (FIFA) has held 21 World Cup tournaments since its inception in 1930. Seventy-nine different countries have qualified to participate in the World Cup in this same timeframe, with eight different countries becoming champions. The purpose of this project is to explore relationships between a wide range of markers aggregated by country with performance at the FIFA World Cup (e.g. goals scored, wins, qualifications).

Physical, emotional, and cultural climate markers of country environment including life expectancy, happiness, anger, generosity, social support, and freedom to make life choices define the keys to the 'Happiness Report' data; Just one dataset amongst a number of dimensions we investigated. Current FIFA player data was aggregated and averaged based on nationality, massaged further to provide a transformed set for insights into national strengths based on average skills of players including those in categories of, 'ball control', 'short passing', and 'corner kicks'.

The FIFA Women's World Cup has held eight competitions since 1991 that have been won by four different countries. This project also seeks to explore relationships between country-level markers in conjunction with qualification and performance at the Women's World Cup. In addition, these relationships can be compared with the relationships observed in the men's World Cup.

The World Cup is an international competition with players required to demonstrate a "clear connection" to the country that they wish to represent. This connection between players and the country they represent leads to the question of how markers of the health of the country relate to its success in competition.

## RELATED WORK

Since 1966, FIFA has conducted a technical study after each world cup. These reports focus on physical skills, psychological attitude, team performance, and on-site conditions. Included are descriptive statistics of each athlete and the official results[1].

Goldman Sachs Global Investment Research, as one example of the big investment banks, constructs and trains predictive models for the 2018 World Cup to predict the tournament's outcome. They mined data on team characteristics, individual players, and recent team performances to analyze the number of goals scored in each match[2].

Statista, a market analysis portal, provides statistical analysis for the 2018 world cup with diverse focuses such as average player age, compensation, stadium capacities, and media and fan interest to name only a few[3]. However, neither of those resources considers country-wide markers of physical and emotional health or anything in the realm of country lifestyle climate, which may be more abstractly connected.

Houston and Wilson found positive correlations between country-level income and World Cup appearances, hypothesizing that increased income allows for increased leisure time to increase proficiency in sport[4]. Population was also found to determine World Cup appearances, possibly due to its increasing the likelihood that a star player is born[4].

Hoffman et al. found differences in the economic factors that influence men's vs. women's international soccer performance. In particular, gender inequality as measured by the ratio of women's to men's earnings explained soccer performance[5].

Previous research has examined economic and political factors that are related to performance in the World Cup. While related, this project is unique in that it specifically looks at how measures of country-level emotional health and well-being are related to performance.

## DATASETS

Six datasets out of two different domains are used to answer the question if relationships exist between country-wide markers of physical, emotional health, physical based strengths, and cultural climate and performance at the FIFA World Cup. The following four datasets provide information about the state of the sport.

Measures of World Cup performance will include home goals, away goals, wins, and qualifications, among others. This dataset includes 855 data objects with 12 attributes each. Results from all of the World Cup matches by country and year for all years 1930-2014 will be accessed from Sports Viz Sunday:
https://data.world/sportsvizsunday/sports-viz-sundays-2018/workspace/file?filename=World%20Cup%20Results.xlsx.

Measures of Women's World Cup performance from 1991-2015 will include year, team, score, and round. This dataset includes 232 data objects with 6 attributes each. This dataset can be accessed:
https://data.world/sportsvizsunday/womens-world-cup-data

Data scraped from sofifa.com on 18,000 individual FIFA players with 66 attributes including player rating, club rating, sprint speed, and shot power will be accessed from:
https://data.world/raghav333/fifa-players
And:
https://data.world/sawya/football-world-cup-2018-dataset

The next two datasets provide country-wide markers of physical and emotional health as well as information about populations and GDPs per country..

Country level health statistics from 258 countries from 1960-2015 with 34 attributes including life expectancy, malnutrition, and health expenditure will be accessed from the World Bank:
https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics

Measures of emotional health in 130 countries from 2005 - 2016 by country and year from the Gallup World Poll with 27 attributes including positive affect, negative affect, and social support will be accessed from the World Happiness Report: https://data.world/laurel/world-happiness-report-data. Many variables are subjective such as well-being or 'freedom to make life choices'. Survey participants were asked to rate their assessment on a scale from 0 to 10 or to answer binary yes/no questions. The answers were mostly averaged per country and year. Data from other studies are included such as the World Bank's Global Economic Prospects from 2016, and healthy life expectancy statistics from the World

Health Organization (WHO) and the World Development Index (WDI).

## MAIN TECHNIQUES APPLIED

Our general approach took the form of an evolutionary sequential analysis based on exploration using popular data mining techniques. We began a discovery based process to get familiar with the data by visualizing feature distributions as boxplots, which also helped to detect outliers. We determined insight on what attributes were most strongly correlated to results in each set by looking at heat maps drawn on marker datasets with result columns joined in on each item by matching on country. We were able to log correlation using the, 'Pearson Correlation Coefficient' for each marker/result pair to find the strongest connections between attributes and results, and to look to understand the trends and patterns we saw in our findings. We built significance sets for Apriori analysis by turning numeric values into binary points based on a determined level of significance (e.g. above 70th percentile value would mapped to a true value in a transformed set, otherwise mapped to false and not included). We mined frequency sets from this transformed binary dataset to show larger antecedent sets that lead to selections of singular consequent results (historical results, 2018 results, or women's FIFA WC results). We logged important factors that showed significance of these results, including confidence, support, and lift for each 'country markers {w, x, y} lead to result type z' pattern analysis. Classification and Regression models were built as a last round of analysis. Fine tuning and exploration were initially based on results from correlation and Apriori analysis to investigate possible predictive strength of markers that were found to be insightful. They were then further tuned, experimented with, and analyzed to look closely at how they changed based on the level of detail for which the prediction was modeled to determine, and assessed based on R^2 (explanatory power) and P value (marginal significance).

**Tools**

We preprocessed and analyzed the data using the Python programming language and relevant statistical libraries in a Jupyter environment. The Pandas library was selected to help aggregate, clean, and organize data in preprocessing and integrating datasets. Analysis libraries including scikit learn, mlxtend (machine learning extensions) and numpy were used to perform data mining techniques.

**Preprocessing**

The multiple datasets included in this project required significant preprocessing before integration. Missing values for continuous variables were replaced with the mean for that attribute. Missing values for categorical variables were filled with the mode for that attribute. Missing values for a team's nationality in the individual FIFA players dataset were filled with the player's nationality, which seems to be a reasonable

estimate even though player's are not required to play for their country of origin.

Outliers were detected by visual techniques such as boxplots and computation of the interquartile range for each feature. Since the number of outliers were relatively small, the data sets were trimmed down to valid data points within the interquartile range.

Min/max normalization was conducted on all numeric attributes to prevent differences in the scale of the values of different attributes to influence analysis.

Country names also needed to be matched between datasets. The Women's World Cup data used abbreviations for its country names, and some of the data dating further back in time included countries that no longer exist. Functions were created to rename countries in datasets as appropriately as possible using an intermediary map set which connected country codes with full names. Full country names proved to be a consistency we relied on across all data sets to allow us to join any result set with attribute set based on country value.

A dictionary was created to track countries that qualified for the World Cup each year. Total goals and wins for each country each year were calculated using the scores attributes and carefully aggregated from FIFA match data to arrive at sets that included total WC match wins based on countries.

**Integration**

Following preprocessing, datasets were merged using country values as the joining attribute. In some instances first an aggregation was made to average over a series of years. Consistency in these processes was key. If we used years, we used a set of years for which we had an ample amount of data for most countries. The rows in the resulting dataset represented a country for a specific year or span of years. Consistency was maintained for the set, but differences existed between approach on individual data sets (Happiness Data compared to World Bank Data), since the initial sets were so distinct.

**Dimensionality Reduction**

Dimensionality reduction was achieved by aggregating wins, goals, and qualificational data. Or to create new rows with which to test trends. We also used forward stepwise selection as an association rule generating technique to select for sets meeting specifications for strong rules which qualify as interesting based on support and confidence levels and a determined threshold.

The dimensionality is reduced from 66 attributes to 22 in preparation for future prediction mining tasks. These 22 attributes include the player's skills such as stamina, agility, and ball control, as well as their name, nationality and overall rating.
The world cups results dataset is reduced in multiple ways. One version reduced the original 12 attributes to 4 since the final teams in a tournament can be seen as most successful. Attributes such as the date, time, and stadium seem to be less useful currently. Another table generated from this dataset provides the 4 previous attributes plus 4 frequency attributes for overall goals scored during a World Cup, the number of qualified teams, played matches and the average attendance in the stadium.

The world happiness dataset is reduced from 27 attributes down to 13. The original dataset included GINI Index values and descriptive statistics that will be self calculated when needed. Missing values are replaced by the countries mean for that given attribute over the years.

**Correlation**

Correlation was determined by mining single dimensional and multidimensional association rules

between data attributes including country related factors and world cup results. Strength of correlation was determined using Pearson's Correlation Coefficient. We were able to utilize program generated heat maps to look for visual trends and anomalies in the data overall, providing benefit of observation from a high level.

**Frequent Sets**

The Apriori Algorithm was used to approach finding these sets using a complementary technique to help us mine out the most important markers using the threshold level for pruning out uninteresting attributes while sorting significant attributes for each country into a set before Apriori analysis was run. the Apriori showed frequent sets of items in a subset of high performing countries. We kept in mind that correlation doesn't imply causation and worked to understand our findings from as centered a source of truth as possible.

**Classification**

Methods including Decision Trees, K-nearest neighbors (KNN), Bayesian, and Support Vector Machines will be used evaluate classification questions. Classifiers were used to predict qualification for the World Cup. In addition, performance at the World Cup were evaluated by dividing countries into four quartiles based on the number of World Cup wins that the country has achieved. The classifier tried to predict which quartile the country fell into based on the supplied attributes. Metrics, such as accuracy, precision, and recall were computed and used to analyze the quality of classifiers. Confusion matrices were also used to visualize classifier quality.

A programmatic approach to finding the 'sweet spot' where accuracy was high and prediction was in the highest percentile possible was configured to run a series of sequential approaches for model building using 1/x where x went from 2 to 20, resulting in the possibility of predicting into the (1 - (1/x) * 100)the percentile at best. If x was 2, this was the 50th percentile, if x was 10 this was the 90th percentile, etc. We looked for trends in terms of how well accuracy stayed at a peak as granularity increased in how close we attempted to predict the very top team of the WC from the data, as well as generally how well the models performed. We fine tuned the active learning in what attribute selections were chosen from for the minimized, 'noise free' datasets for classification analysis, and sought to find how correlation level and frequency set mining led to accuracy, precision, and recall in classification.

**Regression**

An ordinary least squares (OLS) regression was used to examine relationships between attributes of interest and performance at the World Cup, as quantified by number of wins and the number of goals for each country each year. Regression was evaluated by examining $R^2$ values and p-values with statistical significance set at $p<0.05$. Forward stepwise selection was used to find significant attributes.

## KEY RESULTS

Cross referencing datasets containing markers aggregated by country with those containing World Cup results allowed us to look at statistical connections and draw insights.

### Exploratory Data Analysis

After dimensionality reduction we used the following numeric country attributes from the World Bank Dataset (most split in total, male, female ):

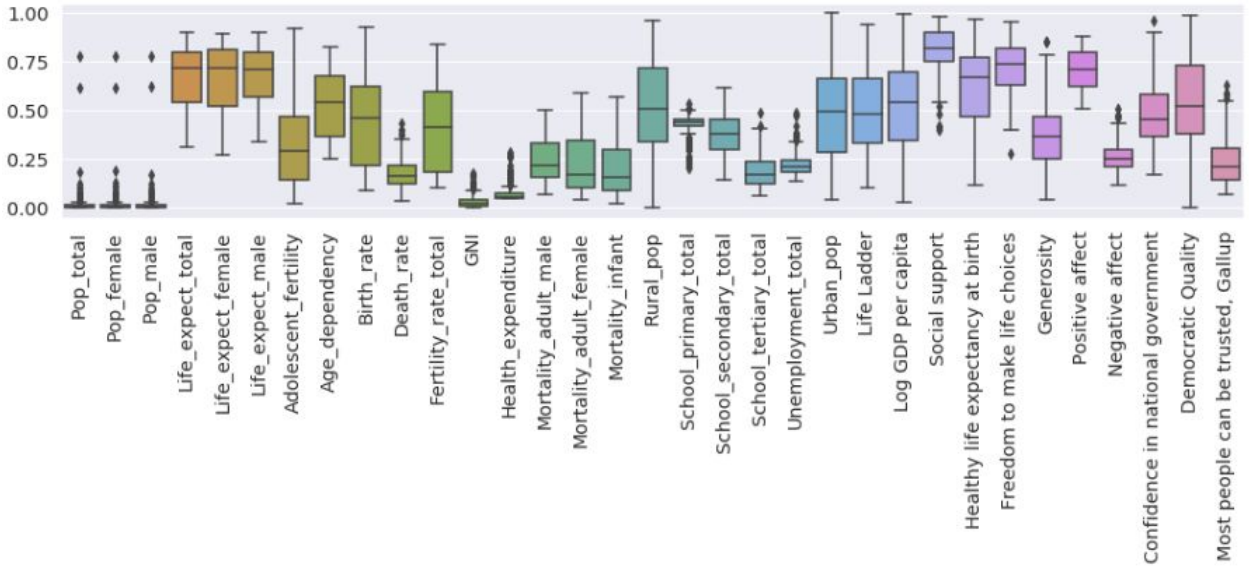



**Fig. 1.** Country attributes from World Bank data and world happiness report.

population (urban and rural), life expectancy at birth, fertility rate (also for adolescents), age dependency ratio, birth and death rates, school enrollment and unemployment rates. These World Bank features are combined with Gallup Poll surveyed factors from the World Happiness Report. All of the following attributes are continuous numeric data types, because they are national averages of individual survey answers. The World Happiness Report features of an emotional and country climate nature are: feeling happy and well, having social support in times of trouble, being satisfied with the freedom to make life choices, having confidence in the national government, democratic quality, and thinking most people can be trusted. Furthermore, positive effects are included, which combine the experience of happiness, laughter and enjoyment, as well as negative affects, that involve worry, sadness and anger. The distribution of each variable is shown as boxplots in Figure 1. Except the attributes populations, GNI, and health expenditure, which show exponential behavior, the variables appear to be normally distributed. Healthy life expectancy and freedom to make life choices are left skewed. That means that in more than fifty percent of the included countries people rate these features to be high. Mortality rates are right skewed, which

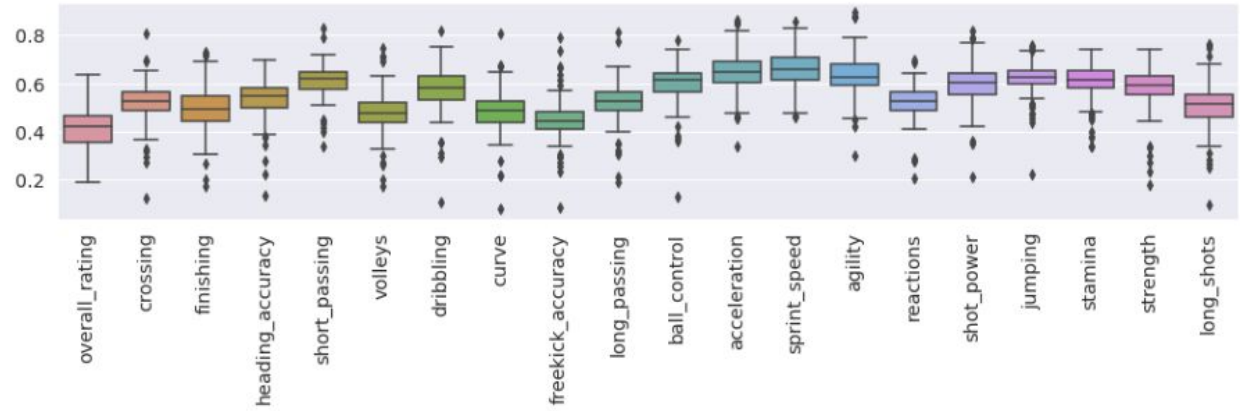



**Fig. 2.** Player skills summarized by skill over all participating countries.

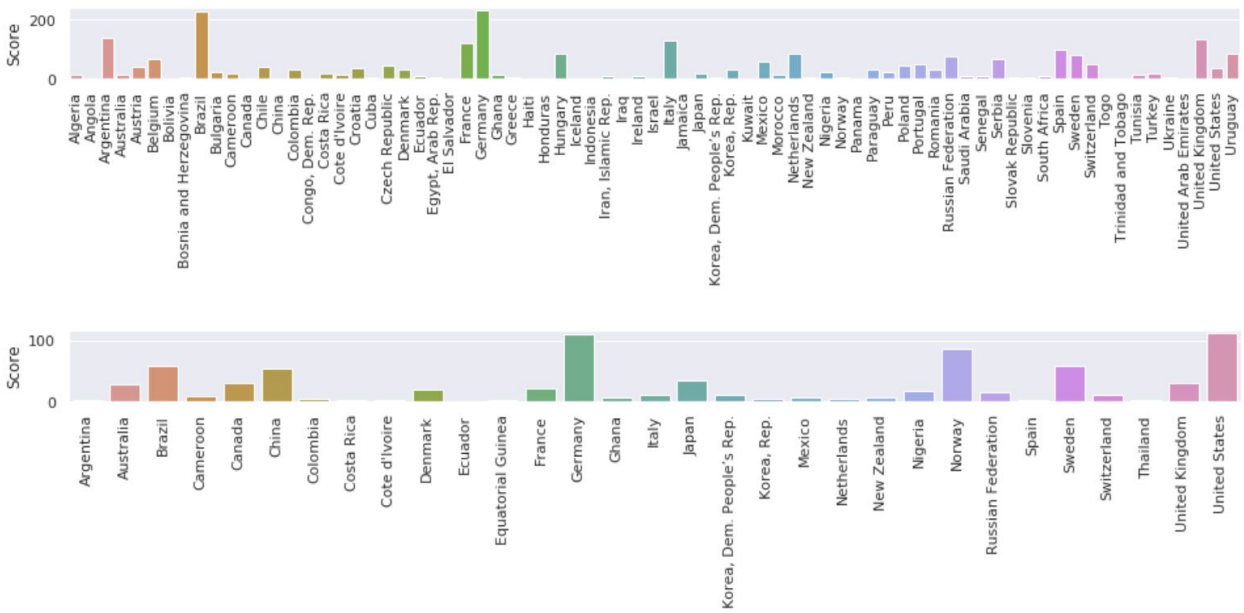



**Fig. 3.** Comparison of the number of overall goals by country in FIFA men's (top) and women's (bottom) world cups.

indicates that the median trends towards the lower end. This data was normalized based on existing min and max, however it shows that mining creates it's own scales of significance in what might be areas that overall are significant across the spectrum.

Player skills are integer ratings between 0 and 100. Unfortunately, this dataset is limited to male players. Figure 2 shows the selected features measured over all qualified teams and hosts in a world cup. All attributes seem to be normally distributed with none to small skewness. The attribute overall_rating shows a wider range than others, but all of the features have a small standard deviation between 0.06 and 0.14 of player skills averaged by country. This might be due to the players being already world class and reaching humanly possible physical limits.

Comparing the number of total goals from men's world cups with women's world cups by country shows a very different distribution. Besides the fact that more countries participate in the men's world cup (75) than in the women's (29), different countries are among the top scorers. Figure 3 shows that Brazil, Germany, and Argentina have made the most goals in all of men's world cups. USA, Germany, and Norway have made the most goals among the women's teams. Interestingly, Germany is the only country in both top three scoring national teams. Future analysis based on further training details will be sought to provide insight into this interesting space.

## Correlation Analysis

For Correlation results we used a common scale for reference which predetermined results in the range of .5-.7 to be moderate, and those between .3-.5 to be low. This left any values outside of these ranges by default as either very high or very low. This scale aside, trends within the results proved to be significant even though their score itself may not have been deemed assertively centric in numeric significance. Negative correlation can be assessed in this range as well after taking the absolute value of the correlation, and very relevant for evaluation in impact.

Exploring the Happiness Data and total women's matches won in the World Cup, led to

finding that moderate correlation existed in the general context with connection with total men's World Cup wins historically for the country. The highest correlation with a marker of happiness was technically low at .344 but relatively very high amongst positive correlations on all Happiness Data with results. Showing Democratic Quality as contextually an important marker. Negative correlation was found with Positive Affect and Negative Effect. Both levels of emotion could be seen as healthy connections with sports performance. Expectation for stronger correlation with this data, especially in regards to assumptive connection to the attribute, 'Freedom to make life choices', was not met. It is possible that the dataset size and collection method which differs for the Women's WC results, impacts this.

The set of Happiness Data and the total number of men's matches won in the World Cup showed moderate correlation with both women's results at .73 and overall player rating at .7. Most of the correlation in this set was quite low below .24. Noticeably low levels of negative correlation was found with, 'Most people can be trusted' at -.29, and 'Positive Affect' at -.24. It is merely speculative what this connection signifies however one might imagine that opportunities in less comfortable environments are more scarce, and negative correlation could be a motivating factor to pursue sport, or negative emotions a fire to fuel sports performance as an outlet and escape.

The set of Happiness Data and Total Men's matches won in 2018 showed some strong correlation to result with Overall Player Rating at .42, and the strongest absolutely with, 'Confidence in National Government'. This observation is relatively easily reasoned as existing since the player rating data point was determined from current FIFA players. Surprisingly correlation with historic results was low, and no strong correlation could be found, here.

Overall Happiness Attributes could be seen as most strongly correlated (excluding negative or positive differentiation) with confidence in national government, and there was relatively strong though lower correlation with Democratic Quality, and negative relatively strong correlation with Trust of the People and, 'Freedom to make Life Choices'. The more swayed strength of the negative correlation makes one wonder if football can flourish in otherwise unhappy nations. Perhaps negative correlation shows incentive for teamwork and less competition amongst the individual, and an opportunity for beating the odds.

National player skill strengths and total Men's WC wins showed medium correlation in freekick accuracy(.68), long passing(.55), reactions(.77), short passing(.57), crossing(.53), and curve(.57). Negative correlation was found in acceleration, sprint speed, agility, jumping, and stamina at a very low level. The stronger traits found to be correlated with results are midfield strengths which tie the team together. Showing that teams with strong consistent skills that contribute towards teamwork are likely to be the most successful. Correlation doesn't show the whole picture and importance of specialty skills, but it shows the importance of teamwork.

National player skill strengths and recent Men's 2018 WC wins showed medium correlation with, short passing(.68), long passing(.52), and overall lower correlations on attributes compared with total wins in history. Consistent negative correlation was seen on acceleration, sprint speed, agility jumping, and stamina. Correlation was relatively consistently low with a slight trend towards a more consistent trend, showing possible effects of globalization and truth behind rumors that sport

selects best for generalists over specialists. Overall we found some similarities here with the historic win correlations.

**Frequency above average values in 70th Percentile**

The Happiness Data and Total Women's matches won in the World Cup surfaced a high value frequency set which included social support.
The set of Happiness Data with the total number of men's matches won in the WC showed (Log GDP per capita, Democratic Quality, Healthy Life Expectancy at Birth, overall player rating), and (Negative Effect, WC wins in 2018, overall player rating). The set of Happiness Data and Total Men's matches won in 2018 showed a frequency set of (negative affect, total wc wins for the country, and overall player rating).
In the Happiness Data 'Social Support' and 'Negative Affect' were common in frequency sets overall.

Frequency sets on the player skillset levels were too numerous to run. The number of possible combinations of the 22 attributes created a programmatic overload which led me to the assumption that frequent sets in this group were almost too numerous for insightful analysis.

## Classification

Prediction for Total Women's World Cup matches won based on Country Happiness markers was built on a very small selected set to model from, since correlation seemed low on this connection to number of wins. 'Democratic Quality', and 'Social Support' were selected as the two most connected attributes from the dimension. Due to lower correlation than other data groups, prediction was assumably lower. Naive Bayes achieved the highest predictor in this model at a relatively very high 60% for the 50th percentile as a peak. As the prediction percentile became granular and the number of labels decreased, accuracy of classification dropped very quickly.

Prediction for the Total number of FIFA Men's WC matches based on Happiness Data we based on the attributes 'Negative affect', 'Most people can be trusted, Gallup', and 'Freedom to make life choices', as derived from the determination of relatively significant correlation and frequency sets. This served as a better prediction of results with Naive Bayes, especially. Naive Bayes models out performed KNN, here, with the highest level of accuracy at 65% for two labels or the 50th percentile. As prediction became more granular the results quickly dropped off in accuracy again, and the model's predictions became irrelevant. Tests were run regarding the accuracy of Naive Bayes and KNN for up to 20 labels, showing merely 7% accuracy in prediction of the 95th percentile.

Prediction for the Happiness data and Total Men's matches won in 2018 specifically was shown by a much stronger model in using both Naive Bayes and KNN which predicted up to the 95th percentile at 20 labels at an accuracy of 61.5% for KNN at the highest and 53% for Naive Bayes at the same level. This level of prediction could be seen as attained at lower levels of granularity as well. An explanation for this likely is associated to the selected attributes Negative Affect, Most People can be Trusted, and Freedom to make Life Choices.

In general Classifiers on the Happiness Data performed better with lower level of expectation as to the level at which the prediction was made. If you were looking for good teams versus bad teams, the models predicted relatively accurately. But if you were looking for the very top team, you were left wondering. With a rough set of a binary condition, however, they faired surprisingly well at above 50%.

Player skill set data Classification for total World Cup wins showed promise based on highly

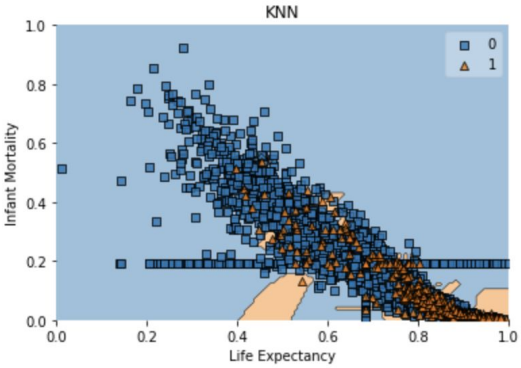
correlated data, however accuracy with Naive Bayes at the highest at determining between two labels was lower than anticipated at 54% in the analysis for up to 20 label items. Accuracy decreased relatively linearly as granularity of prediction increased.

Player skill set data Classification with Women's WC wins seemed uninformative since the player data was collected from male players. Here bias is seen at the collection phase. Player skill set data Classification for total World Cup wins in 2018 resulted in a more successful model, which could predict up to 20 labels with a 61% accuracy using KNN. Accuracy for both KNN and Naive Bayes were slightly more successful at lower granularity, however very consistent.

Overall player skills served as strong predictive markers for recent WC results and the models build would likely be effective for rough predictions in future World Cups, and granular predictions given a current Player dataset of similar quality.

Classification methods were also used to predict whether a country qualified to participate in the World Cup in a given year based on health data from the World Bank and happiness data. Only ~12% of countries qualify for the World Cup in a given year, which resulted in an imbalanced dataset for classification. This imbalance meant that the classifiers could score high in accuracy without having to classify anything as a World Cup qualifier. Early models that did not account for this had accuracies above 90%, while recall scored less than 20%. This imbalance was adjusted for by naive oversampling and use of the Synthetic Oversampling Technique (SMOTE). Naive oversampling generally outperformed SMOTE for these classifiers, and was used throughout. KNN also generally tended to perform best for these classification tasks.

Using health data from the World Bank alone, the KNN classifier had an accuracy of 83%, precision of 35%, and recall of 52%, suggesting some proficiency. A 2 item subset (Life expectancy, Infant Mortality) of the attributes used in the main classifier is shown below for visualization purposes. It can be seen that decision regions cluster toward the maximum of life expectancy and minimum of infant mortality.



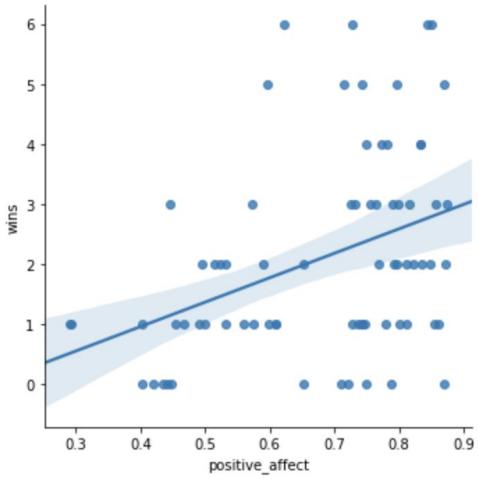

Adding happiness data to the model tended to increase to precision (39%) and recall (61%), while reducing accuracy (72%). Some of this accuracy loss may be due to a decreased sample size, as the World Bank data includes data from 1960-2015, while the happiness data is limited to 2005-2016. The datasets were linked with an inner merge, thereby dropping years that were not present in both.

These classifiers for qualification performed similarly with the Women's World Cup data.

**Regression**

OLS regression was used to examine relationships between total goals and total wins for a country in a given year. This regression analysis generally yielded poor results, with very low $R^2$ values (<10%). However, a few attributes did achieve a significant p value ($p<0.05$), including life expectancy and positive affect, both of which had positive coefficients. These attributes with low $R^2$ and low

p-values may be viewed as having an effect on the dependent variable, although they explain little of the variance.




Interestingly, the calculated difference in life expectancy between males and females achieved significance and had the greatest $R^2$ value in a number of models. Multiple linear regression with this difference in life expectancy and positive affect explained 23% of the variance in the model.

## APPLICATIONS:

Applications of insights gained from process and analysis are related directly to the questions asked of the data and our understanding of the importance of the data. The questions of which we looked for pertaining insight weren't prequalified as obvious connections. They asked for a level of abstract analysis, an unsuspective association into life, while realizing that what surfaced might imply correlation yet not causation. Observing patterns and connections in this way creates a creative inquisition which can illuminate new insights into what might seem to be an overly explored problem space, riddled with existing assumptions and repeated problems from these. In World Cup soccer, the fans are so intense and numerous that not only does a country perhaps connect through health and environmental climate to results, but so do games and scores connect to the people that live there. Applications in sports psychology, coaching, and performance at the national level are very numerous. Looking at how attributes are connected with resultant targets in data as we have done in this project results in information that can be further explored in software applications built for providing leaders in these fields with data available for progressive approaches.

Sports psychologists might use correlation between country markers and success to provide insight into how a player might form a unique mindset that allows them to succeed on the field outside of the box of 'positivity' and 'obedience to 'hierarchy' that can be assumed to be part of a demanding physical occupation on the field. By learning to challenge the norm from data driven perspectives, unexpected and unique results can emerge.

Coaches can use national player skill data to learn what individual skills can contribute the most to team success. A platform can be made to assess players on a smaller scale on these same levels and used as a training tool to connect with and assess improvements in players. Coaches could analyze data like this and come to their own conclusions as to the source of the observations, drawing unique inferences from which to guide their teams.

Data exploration like this can yield fresh ideas and insights into sports psychology, coaching, and performance. It can allow the beginnings of ideas in performance relationships to form where they weren't initially apparent. If nothing else it makes us ask questions about our data, how we gather it, how consistent we are with this, and how we select our points of data collection. Even asking questions of data has a resultant effect of our own consciousness and awareness in our interactions in everyday life. To

begin to look at detail within what might otherwise be assumptive answers to these questions, we look for specific, scientific based, data driven associations, perspectives, and insights. Even if we are at risk of coming up dry or in a place that leaves us asking even more questions to explore the narrative, further, we have made progress in cognitively understanding the problem space, and formulated steps to continue to do so.

## REFERENCES


[1] FIFA Technical Study Group (TSG), Technical Reports, https://www.fifa.com/about-fifa/who-we-are/official-documents/development/technical-study-group-reports

[2] The Goldman Sachs Group Inc, 2018 The World Cup and Economics, https://www.goldmansachs.com/insights/pages/world-cup-2018/multimedia/report.pdf

[3] Christina Gough, Soccer - Statistics & Facts, https://www-statista-com.stanford.idm.oclc.org/topics/1595/soccer/

[4] R. G. Houston and D. P. Wilson, Income, leisure and proficiency: an economic study of football performance, https://www.tandfonline.com/doi/pdf/10.1080/13504850210140150?casa_token=7X0eEqYC3n0AAAAA:_FJr1GP4nVoSQNnmGQ26Ak5Kkv4eB_meUrserhl19sURJ7zN6lP6pFetmNNJ3O_jDI7VcZ8GVwYJkw

[5] R. Hoffmann, L C. Ging, V. Matheson, B. Ramasamy, International women's football and gender inequality, https://www.tandfonline.com/doi/full/10.1080/13504850500425774