# Final Report

## 1. Description of the Data

- **data source:** We collected our data from various online sources.
    1. The GDP data (GDP.csv) for each country (1962-2014) was obtained from the World Data Bank website. The site allows you to select the desired country and the corresponding variable (GDP in our case), along with the year range for which data is needed. Then one can download the data in a CSV or EXCEL file. Winning years of each time and details in our first picture were collected from Wikipedia and FIFA World Cup. We integrated these two datasets to make our visualization about the Year v/s GDP for each country.
    2. The Star Players data (player_ranking.csv) for each country was obtained from The Guardian. We sampled the data from two datasets: the World Cup's top 100 footballers and the World's Best Footballers in 2013. Since the World Cup's top 100 footballers dataset does not have the data of 2014, we chose The World's Best Footballers in 2013 dataset as a complementary dataset, which evaluates the overall performance of top footballers before the 2014 World Cup.
    3. The Player data for each country (2006) was obtained from FIFA. That is the only data available on this topic. To match up with the time, the population dataset was set on 2006 as well. The sources of the population data is from google, which is sourced from World Bank. This is one example: Germany. The only exception is the data of England. England does not have census every year. The only available data around this period is 2011, obtained from wikipedia. The stadium data was from wikipedia: Brazil, Germany, Italy, Argentina, Spain, England, France. Only stadiums with capacity larger than 30,000 were included because the smallest number of the Brazil dataset is 30,000.
- **data processing**
    1. (GDP.csv) The data re-formatted to its current format: switching the columns and rows and changing column names) to make the data elements accessible in D3. The file contains a column for the years the world cup has been played so far , 7 columns for each of the 7 countries containing the value of the GDP of each of them for that particular world cup, and a column to represent the winning country corresponding to the world cup year.
    2. (player_ranking.csv) We only pulled the players who played for the countries we wanted to analyze from the two data source mentioned above. We re-formated and combine two dataset so that it can be read using D3. The “World Cups” column in the original dataset was splitted into 5 columns, each representing one year in which the player appeared. For The World's Best Footballers in 2013 dataset, we checked if all the players we pulled from this dataset played for the 2014 World Cup. If they did not play, then we deleted those players from the dataset.

3. (var dataset, var datasetBarChart in the html) Number of players was divided by the population per country; Stadium capacity of a country was summed up and divided by the population of that country in order to obtain the percentage of mega-stadium capacity per person in that country. U20 world cup for each country was divided by the total U20 world cups assigned.

## 2. Data Mapping

- Graph1: (Year v/s GDP) We used linear scales to represent both the years and the GDP values on x and y axis respectively. Ordinal scale was used to assign specific colors to line graph of each country (matching our following visualization). Since the GDP values were huge numbers, we used d3 format to re-format those numbers after being read from the CSV file (when being displayed on the axis and on hover over the line) and also created a custom function to convert the values to a format like "$XY.Z T" i.e. XY.Z Trillion dollars and so on. Tick size feature was also extensively incorporated to create a grid like view so that it's easier for the user to see the exact value on the graph. We also appended "svg:image" on the line charts to display the winning years and the corresponding GDP value for that particular country to make more sense of the data and to contrast and compare different countries. We used HTML checkboxes to control what data will be drawn or removed on or from the graph based on user selection.
- Graph2 (number of star players): We used ordinal scales for both y-axis(years) and the national teams (x-axis), since they are not continuous. We also appended "svg:image" on the winning years for each nation team to show the correlation between number of star players and the championship.
- Graph3: Standard linear scale for all x and y-axis. D3.format was used to keep two decimal points for percentage if it is not .00, and eliminate decimal points if it is .00.

## 3. The story

We examined the elements ---economy, talented individual, football culture--- that accompany world cup championships. Due to the lack of GDP prior to 1960, our project focused on championships 1960-2014.

We found there is no single one element that can surely predict the success of a country; however, we found as the gap between different economies increases, GDP may function as a negative filter. Argentina's economy and its absence from the championship in recent decades seem to go side by side.

With regard to super star players, we used the number of star players as metric. Championship seems consistently associated with teams that have more star players than those without. But we cannot make too strong an assertion here because it is hard

to find an objective metric to measure the success of different players at different positions, and the top list we find may not be immune from these flaws either.

In terms of football culture in these countries, Germany really stands out in the popularity of the game. It has 20% players among the general public, as compared to 7% in the rest countries. We also notice Brazil and Argentina have won the most youth World Cups, but their performance of these two countries in recent World Cups do not seem to suggest U20 world cups can carry on to World Cup.

In sum, the relationship between the championship and the three dimensions we study in this project is complex. Chance certainly seem to have a role to play here, but especially in recent years, a decent GDP, a reasonable number of superstar players and the popularity of the game seem to be the base for a championship.