Gregg McDonald. Greggmcd83@gmail.com

A picture containing clipart

Description automatically generated

world cup data analysis

Insights into soccer world cups

# GitHub URL

https://github.com/greggmcd/Data/tree/feature

# Abstract

Analysis of past soccer world cups with a view to extracting insights though charts and visualisation.

# Introduction

I chose this dataset as I am a big soccer fan and I found I had a decent knowledge of the data within it and I was curious to see if more meaningful insights could be extracted through the techniques we have discussed over the duration of this course.

# Dataset

The dataset I have chosen consists of three databases with data relating to FIFA world cups dating back to 1930.

The World Cups dataset show all information about all the World Cups in the history, while the World Cup Matches dataset shows all the results from the matches contested as part of the cups.

(https://www.kaggle.com/abecklas/fifa-world-cup )which is an open source for datasets and allows for easy access to the raw data. This allowed me to import the csv. Files into the Spyder IDE and work on it to produce the analysis in this report.

# Implementation Process

I was using anaconda as my IDE navigator inside anaconda I used the Spyder platform.

I began the process by importing the necessary libraries into my python workspace these libraries included :

**NumPy** is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

**Pandas** is an open source **Python** package that is most widely **used** for data science/data analysis and machine learning tasks. It is built on top of Numpy, which provides support for multi-dimensional arrays.

**Seaborn i**s a **Python** data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Matplotlib i**s a cross-platform, data visualization and graphical plotting library for **Python** and its numerical extension NumPy.

Initially I set about importing the data using the pandas function pd.read\_csv. I then assigned variables to the different datasets. The data pertaining to players was given the variable ‘Players’, the data pertaining to matches was given the variable ‘matches and finally the data pertaining to the world cups was given the variable ‘cups’.

Then i began sense checking the data using the .head() , .describe(), .shape() functions.

From there I was able to check for null values using isnull() function I then replaced null values by using the fillna() function I also dropped duplicates in all the datasets using the drop.na() function.

I then proceeded to change the index to the year on each datasets for ease of use and resaved all the datasets to the variables matches cups and players.

In the process of cleaning the data I noticed a few incorrect entries in the country name columns and proceeded to correct the incorrect entries by creating a dictionary of the old names such as the federal republic of Germany (FR Germany) I then made a dictionary of the correct country names I then enenumarated over the datasets replacing the old names with the new correct ones, using the .replace() function.

**Visualisations**

My intention was to create charts for winning countries, goals scored, attendances, matches played, qualified teams per year, and if the home or away team was more likely to have won.

I started by creating a new series by counting the values in the columns labelled winners runner up and third place and saving new variables ‘winners’ ‘runner up’ and ‘third place’. I concatenated these items to create a new data frame for ease of use. Then using seaborn I created a pairplot to identify any correlations or patterns.

Using seaborn again I created bar plots for the goals scored, attendances, matches played, and teams qualified per year. Using the in built function i was able to make colourful charts to extract insights.

For the home and away wins charts I had to create a user defined function get\_labels which iterates over the datasets and returns weather a home or away team wins based on what column has a bigger number or if they are equal it returns the result of draw. I then added the column outcome to the data frame and using matplotlib created the match outcomes by home or away pie chart.

# Results

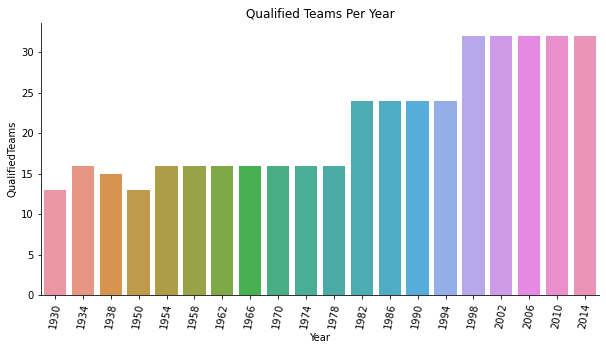
Below is my initial pairplot I used to get a grasp of the winners runners up and third place teams we can see that there a few teams that seem to be dominating here in a few areas but further investigation is needed.



We now move onto goals scored per tournament which we can observe that the later the year the higher the number of goals. This can be connected a later chart which shows the number of qualified teams has increased since the 1990 tournament as well as rule changes which made it easier to score.



As we can observe in the following bar chart we can see that the number of teams increased from 1982 to 2002 and then levelled off. Started with a very small number in 1930 increasing to an eventual 32 from 2002 on. This had an impost on the number of goals scored in each tournament.



This increase in teams qualified also had a direct impact on the following bar chart which denoted the number of matches played in each tournament which has increased from 20 in the 1930 Uruguay world cup to over 60 in Brazil 2014.

 We now move onto perhaps the most interesting charts which shows mnatch outcomes based on wether the team was home or away we can clearlly see that the home team is more twice as likley to win



# Insights

* 1954 had the most goals per games
* The average number of goals has increase with more qualified team and rule changes
* The home team has a big advantage
* The more teams qualified the number of matches increased.
* Brazil have won the most world cups .

# References

https://www.kaggle.com/abecklas/fifa-world-cup