EDA for a Soccer Data

Francis Addae

Created on May25th, 2021 Last Edited: June 16, 2021

Contents

1 Summary

This is an Exploratory Data Analysis report which focuses on exploration and gaining insights.

2 Libraries

Below are the libraries used in this report:

- **dpylr** : For data query, wrangling, manipulation and transformation.
- **ggplot2**: For data visualization and info messaging
- **forcats**: For categorical variable manipulation and wrangling
- **lubridates**: For date-time variable analysis and reporting

3 Data

This dataset is a soccer data obtained from kaggle.It is the collection of all intenational games played from the inception of association football or soccer. Association football's history and understanding can be found on Wikipedia

3.1 Summary

This data is the collection of games played between nations from 1800's to 2020. It data has a dimension of 41865, 10; such records are attributed to 10. Each record is match between two teams. Any recorded game has a unique home and away team. The attributes are:

- **date**: Date when a game was played. This isnt a date-time variable.
- **home_team**: Name of the home team
- **away_team**: Name of the away team
- **home_score**: Score made by the home team in that particular match
- **away_score**: Score made by the away team in that particular match
- **tournament**: The type of tournamnet in whcih the game was held.
- **city**: The city where the game was officiated.
- **country**: The country in which the city is located.
- **neutral**: The advantageous venue of each match.Think of it as home_court advantage
- **result**: The final result of the game.

Although these attributes are great, the dataset could be served with attributes like; temperature, time, attendance, stadium name etc, which can give great insght as to how the overall game look like data-wise.

3.2 Cleaning

The dataset has 0 null(s) values. Despite this, the dataset could use further cleaning if need be. For example, instead of have both the **city** and **country** as a seperate entity, they can be combined as a unique value. This can shorten the dimension of the data. A lot of cleaning can come in place after the formatting of the dataset.

3.3 Data Formatting

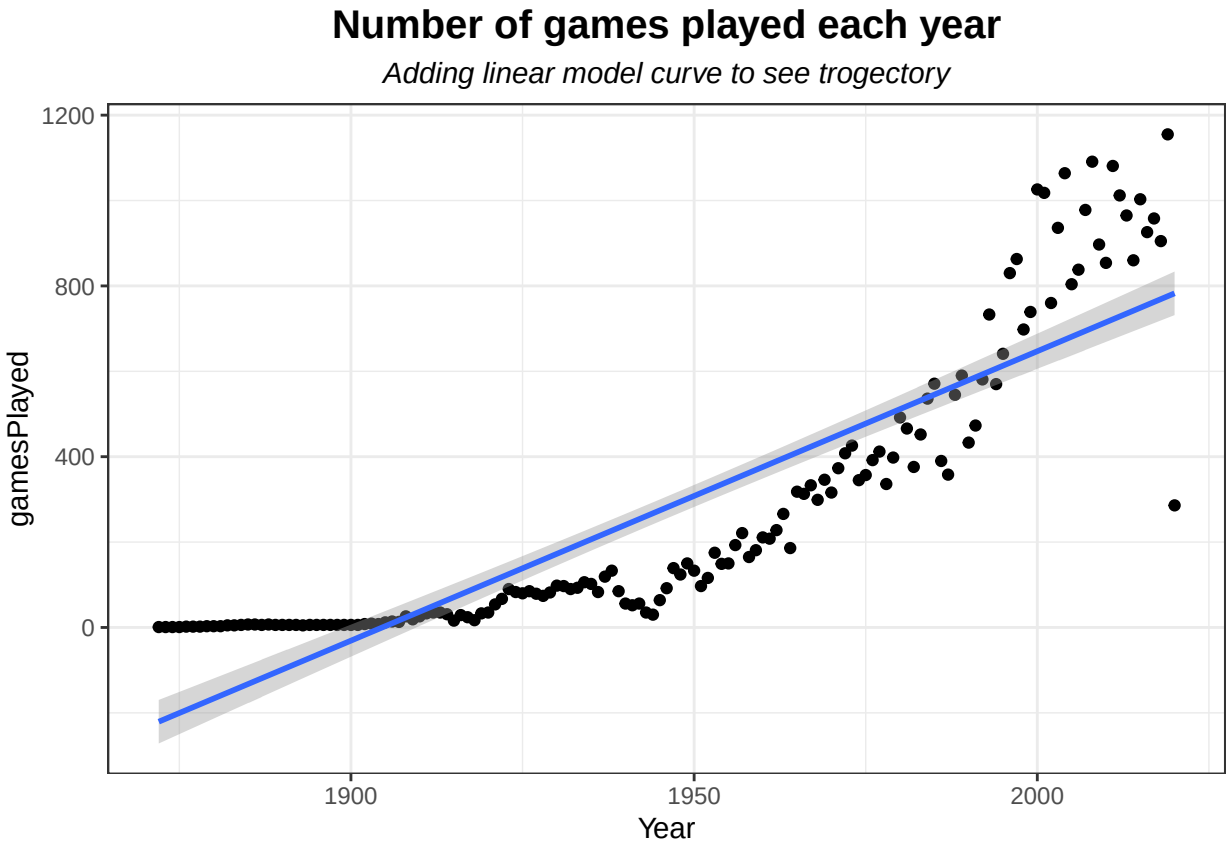
As stated in the dataset, the columns needs to be stated in their correct data structures. The qualitative(categorical) varaibles are home & away teams, city, country,neutral and result. The quantitative(numerical) variables are home and away scores. Despite most qualitatives been strings, neutral needs to be an ordered level category(Win, draw, loss) in that order.This ordering can always change, therefore in this report it will be left as is.Date is formated in a Year-Month-Day format.

4 Objective

4.1 First Objective

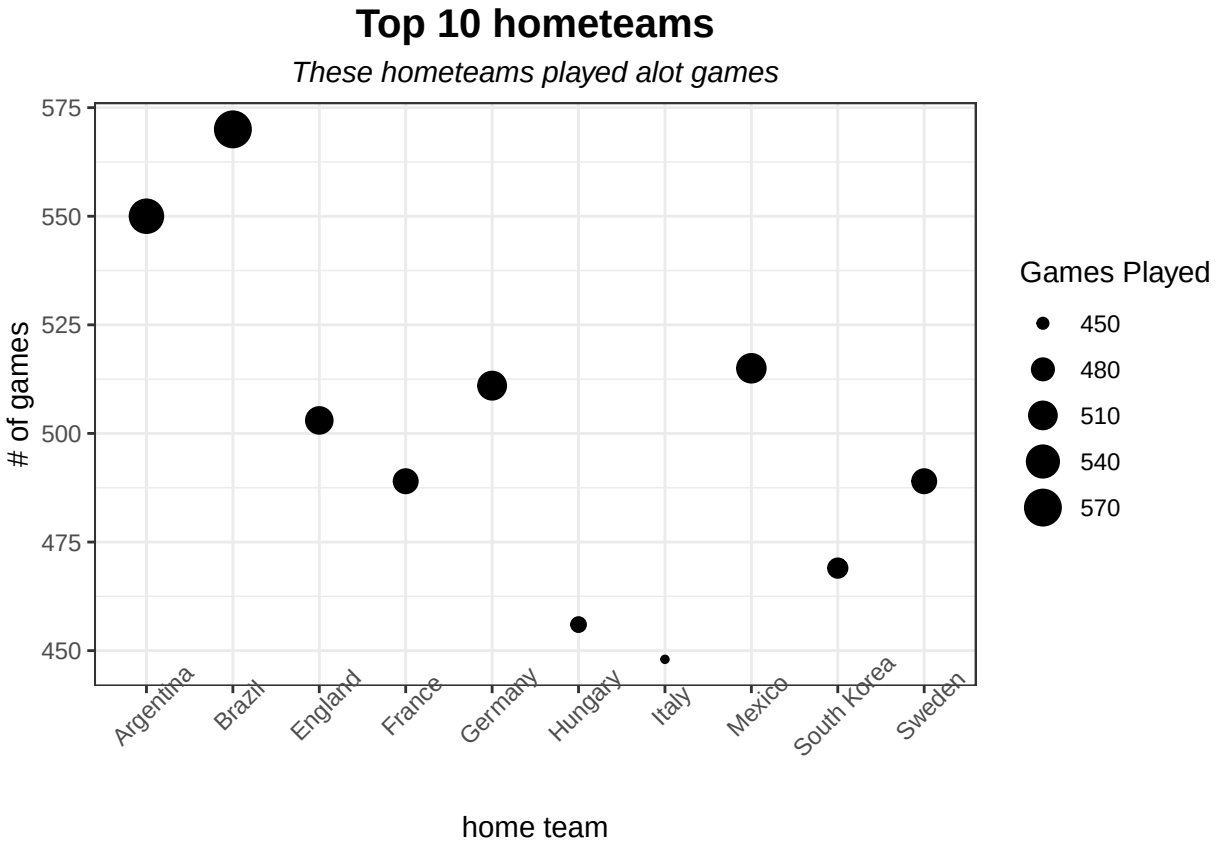
Using exploratory data analysis to gain insight into dataset and see meaningful questions to ask. This step is cruicial because it helps in viewing the data from an general point of view.

1.How many games were played each year?



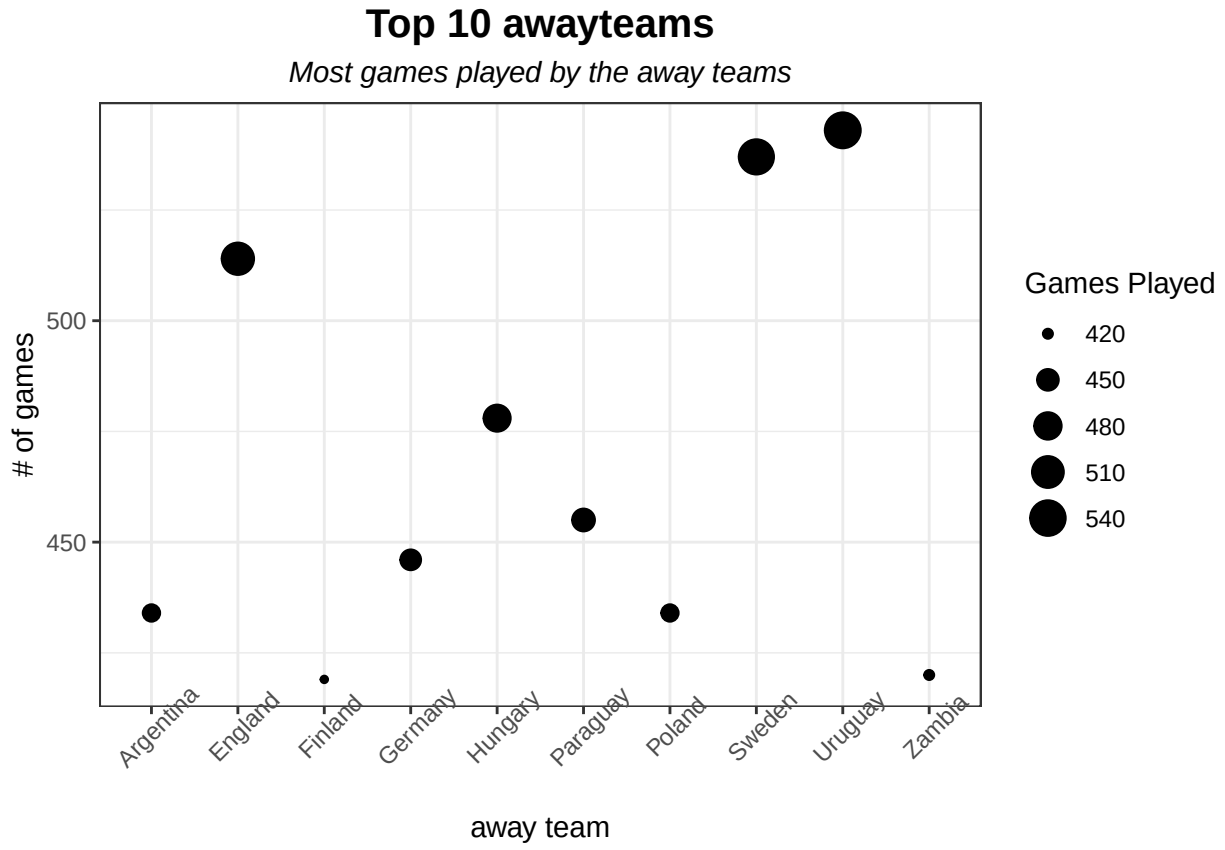
The number of games rose as year increases and this can be an indication of numerous teams joining alot of tournament and also hosting alot of games. Between 1920's and 1940's, it seems there was drastic decrese in games played. This could be due to World War 1 and the beginning of Wolrd War 2.

2.What are the top 10 home teams in soccer?



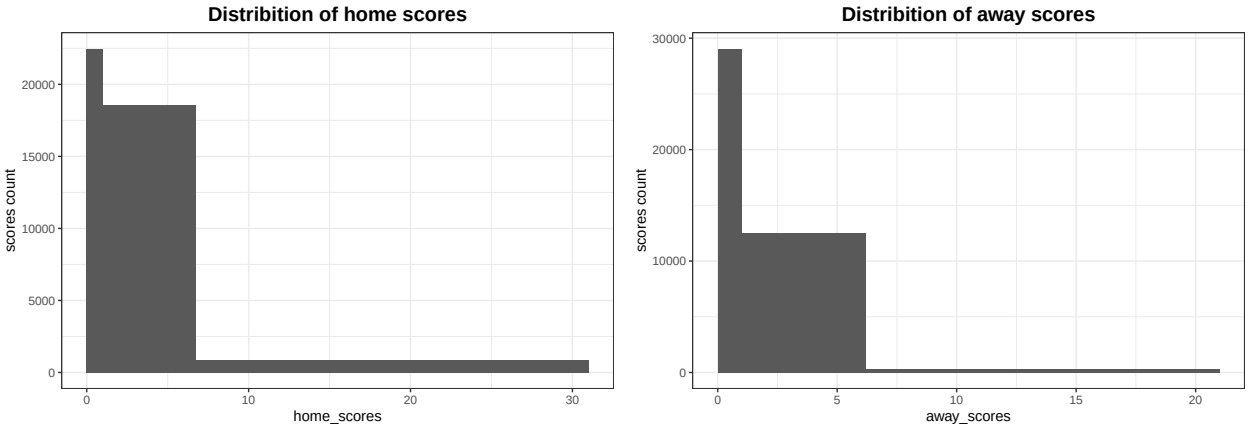
It seems Brazil played the most games out of everyone in the entire dataset. Judging by this, it seems that European teams tend to play the most games as compared to South American teams despite Brazil and Argentina having more games than anyone.

3.What are the top 10 away teams in soccer?



Argentina, England, Germany, Hungary & Sweden are some teams that appeared to be in the top 10 away teams as well as in the home teams. It seems we have Zambia, an African team which played the most away game the dataset. One interesetd thing to look at; can be taking a top 25 countries(both home and away) and see if how many times those two meet and the observations of each record.

4. Distribution of scores



It seems the range of the goals scored during each game changes by the minutes. Most of the scores were between 0-4, with the average goal for both teasm being around 2. The data can be cleaned, and there are multiple ways to do that: * **Use the yearly average to change outliers scores greater than 5**

```
* **Use the mean of the entire score greater than 5**

* **Find the number of games scored with each goal was socred and see if the outlier
```

All cleanups have consequences; due to the fact that the cut of point may chnage the outcome of result. And this can affect any other analysis further down.

The graph below shows the total number of games scored each year and that tells the same story as the number of games played.

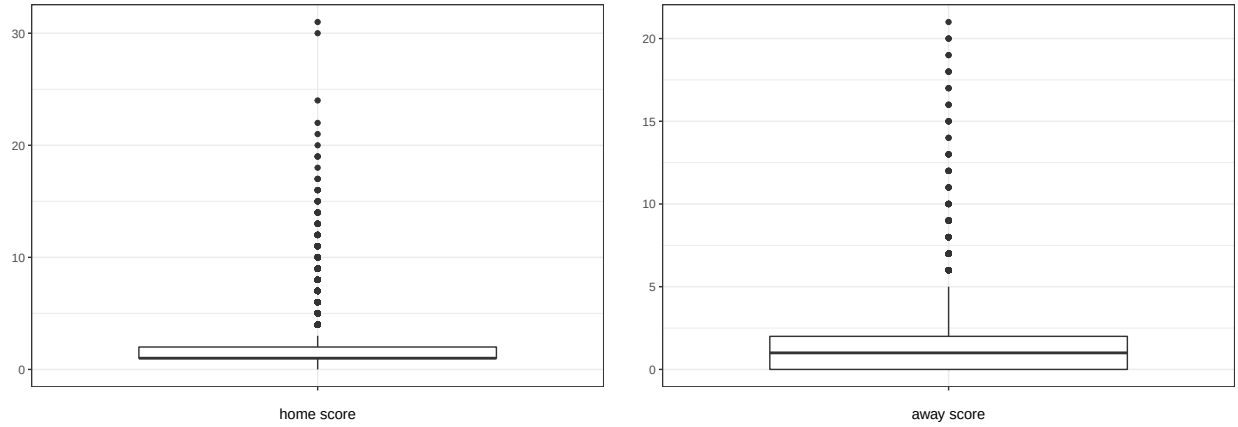
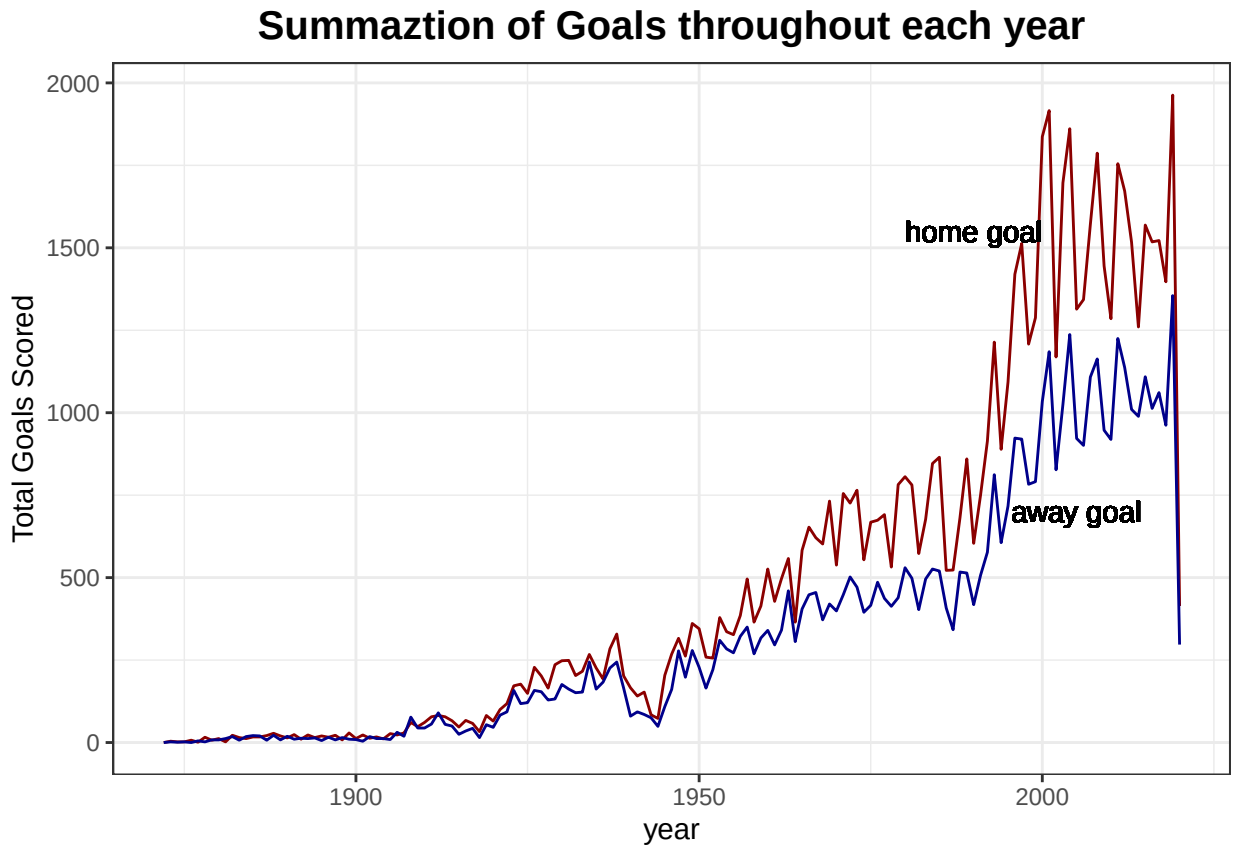
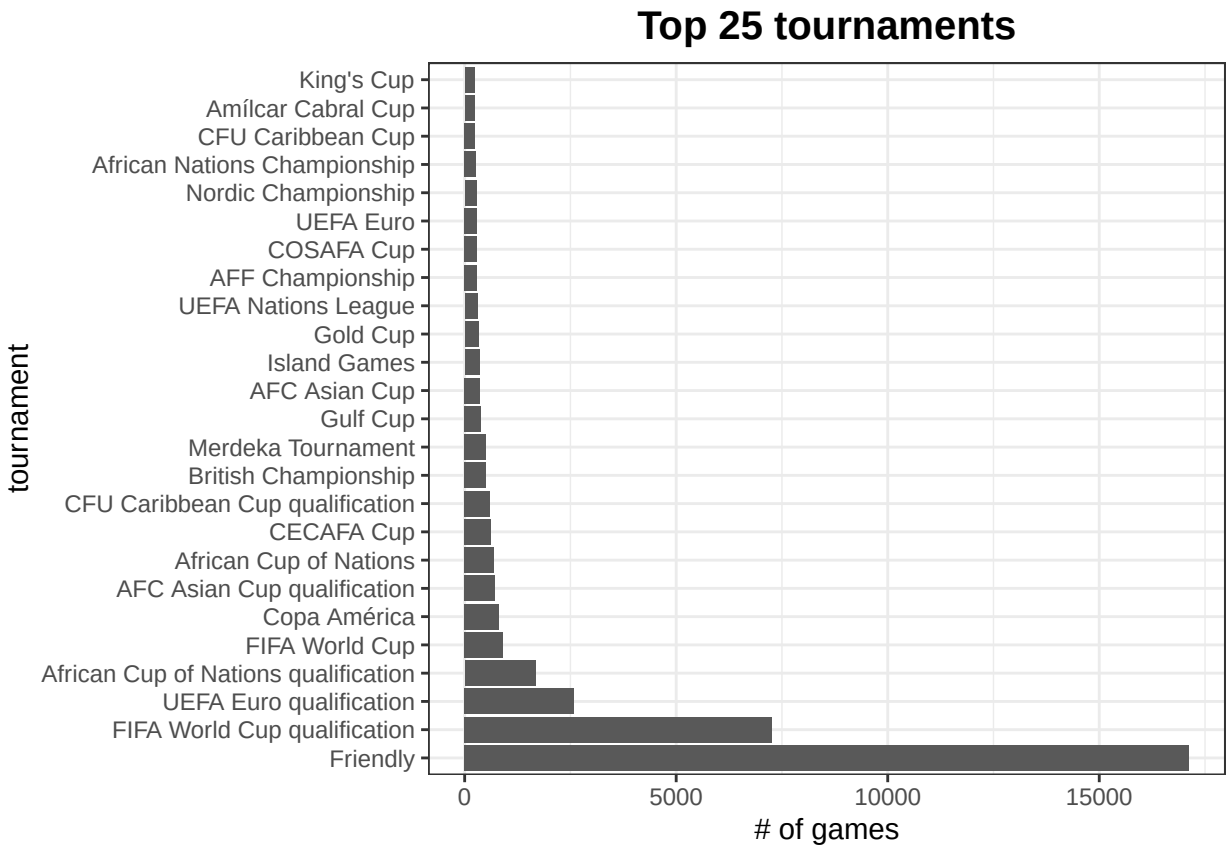


Figure 1: Distributive summary statistics



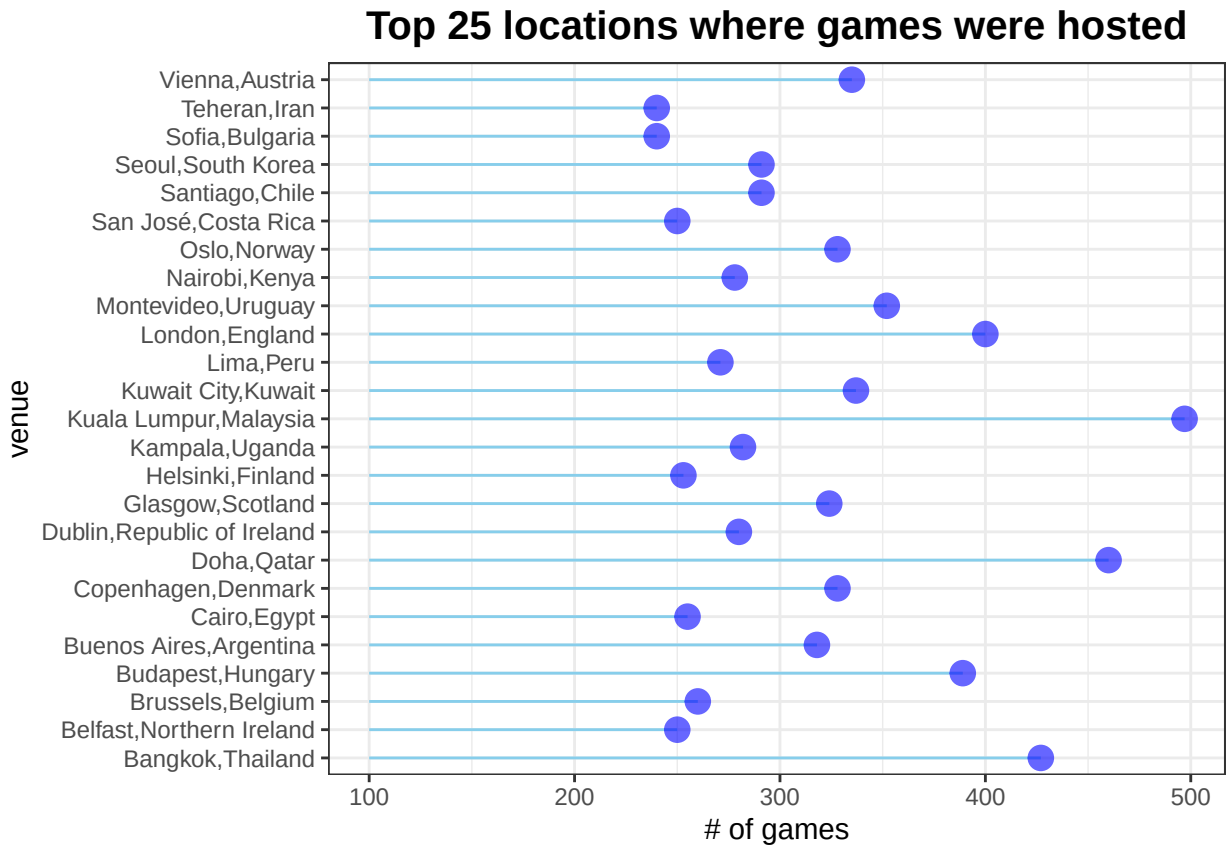
Observation: As the number of games played increase so does the total goal scored. This was to be expected since the linear model shows a positive trend. It most most year, home team goals are mostly higher than away scores.

5.What are the top 25 tournaments in played?



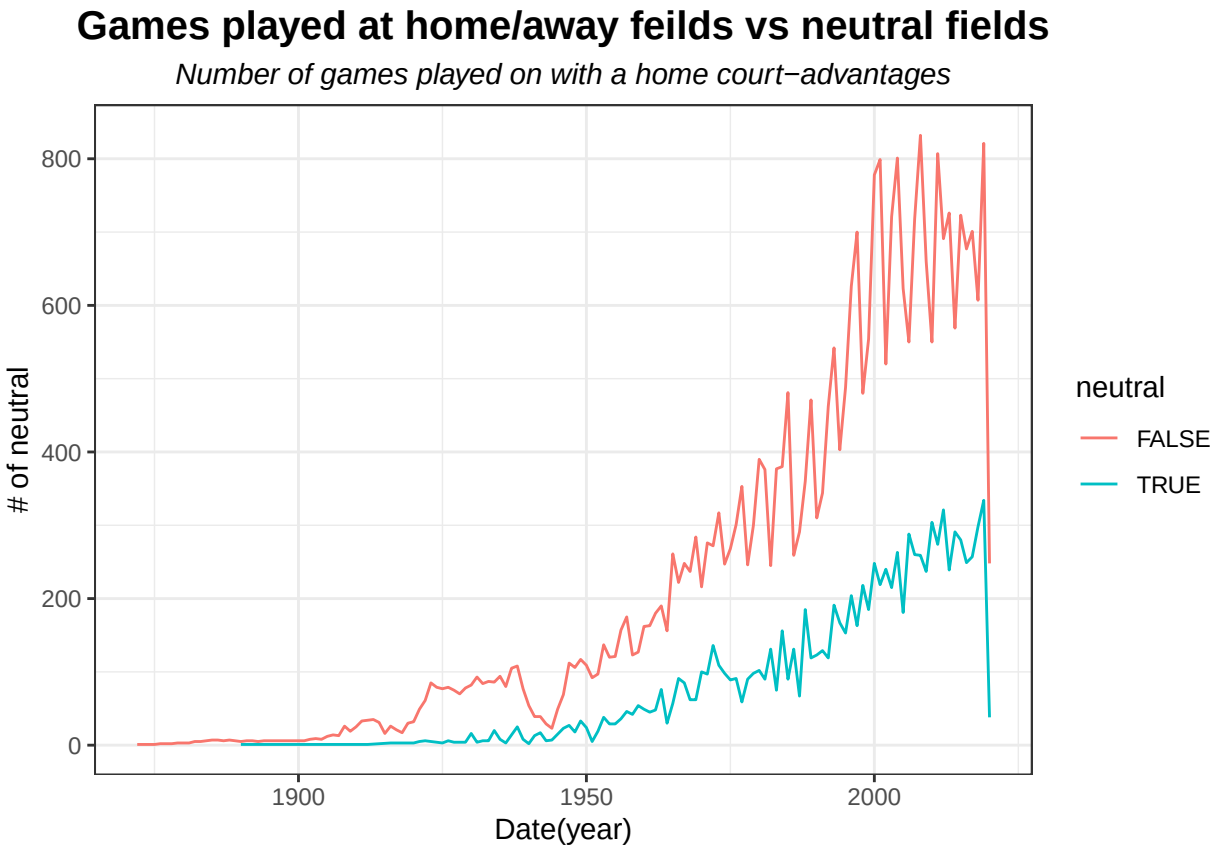
Friendly games are games where a person teams play to asset the abilities and needs of teams in terms of talents and improvements. Mainly the reason why we have alot of friendlies in the dataset.

6. What are the top 25 locations where a game was hosted?



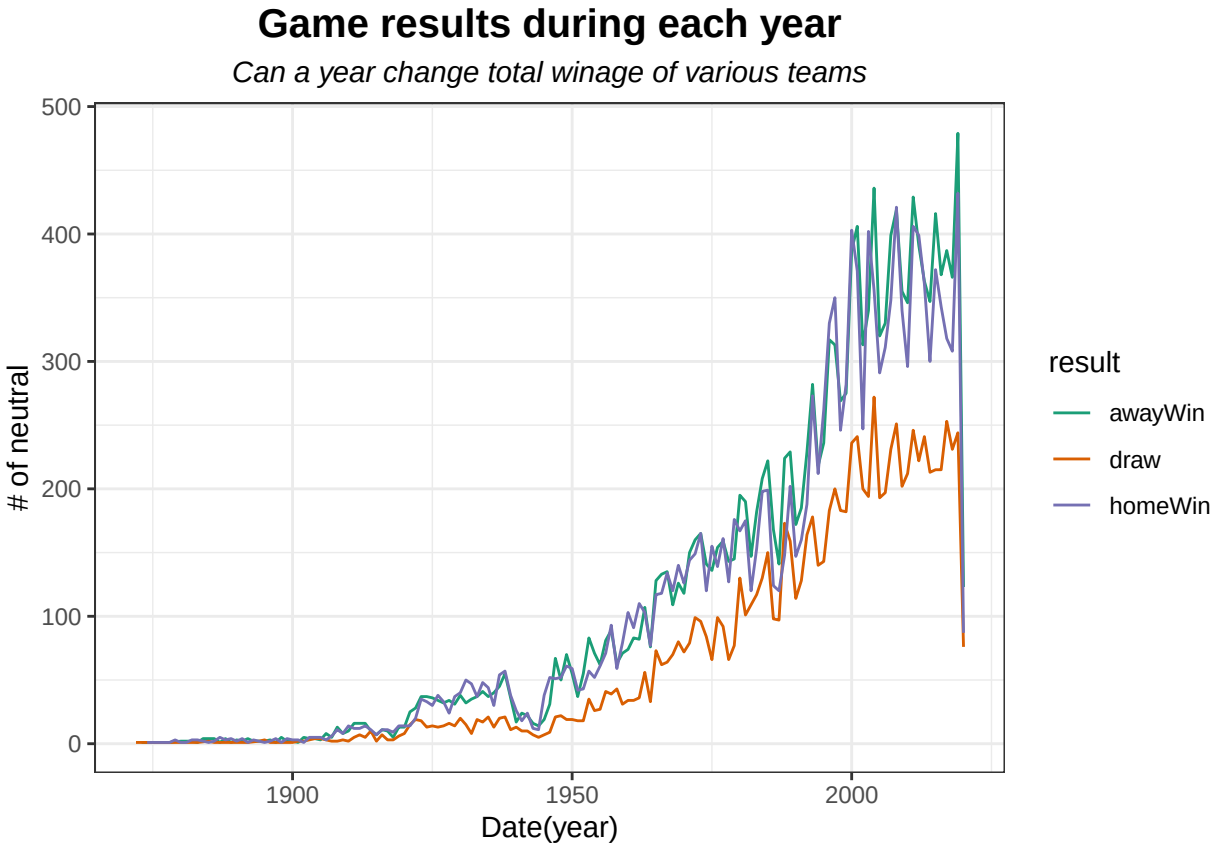
The locations are a combinations of both the city and country. This makes the naming conversion a bit easier. The top 25 locations are scattered all over the world. This shows the randomization of locations where games were played.

7. Distribution of Neutral Zones



Neutral refers to the home-court advantagement of the home_team playing. Its highly unlikely for a home team to have alot of games played on thier tuff especially if it's not a Friendly.

9. Results



The results are quite fascinating. It seems draw are quite irrelevant as we see alot of home and away wins. Most years seem to have away wins as compared to home wins. Away wins almost all the time expect between 1940-1965; and then picks up again in the late 1990's to elary 2000's. This trend is worth investigating.

Before we begin further investigating any interesting facts; the dataset has a location data which can't be pinpoint which continent that game was played. Due to this a second dataset containing names of countries need to join with the original dataset.

This dataset consists of 283 countries with its appropraite continent. A lot of countries change thier official names as things progress so this might make things abit difficult if we want to calculate for overall trends when it comes to. For example, if we want to see the win average of counties throughout the entire dataset, that will be a huge problem since a couple of countries change thier name alot and a couple of counties were colonies when those games were played.

10. What are the top 10 places to a game was hosted on each continent?

This will help in knowing which countries were more favored and comparison can be down relative to the top 25 locations where a game was hosted. Before this, let's check if there are any abbrreviated country names like US or UK. Most abbreavitions are less or equal to 3 characters.

There are abbreviations in the dataset. Therefore, we dont need to account for those issues later on.

Number of total games played on each continent

Are some continents more prone to hosting games than orhers?

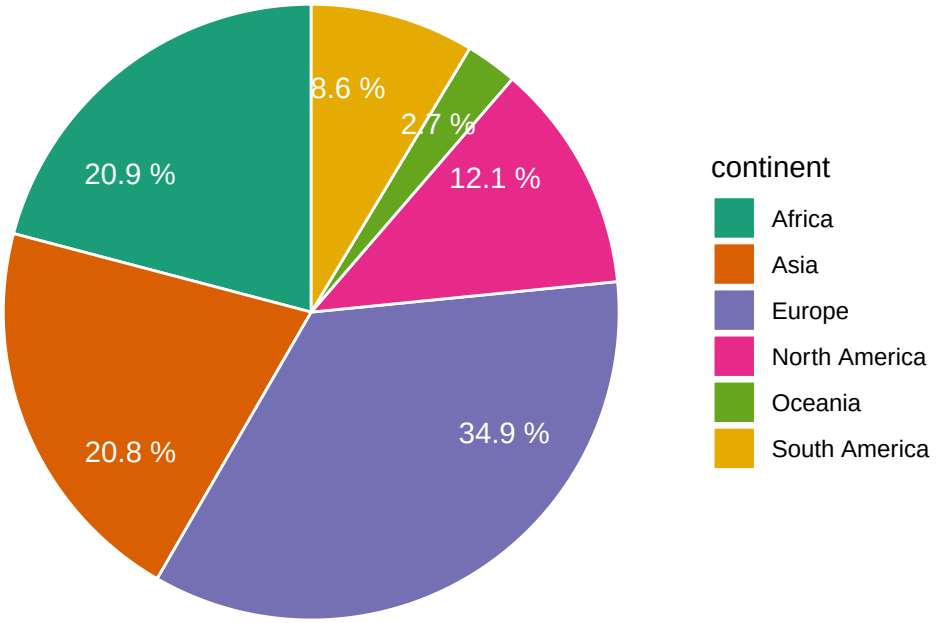


Table 1: Top 10 hosted countries on each continent

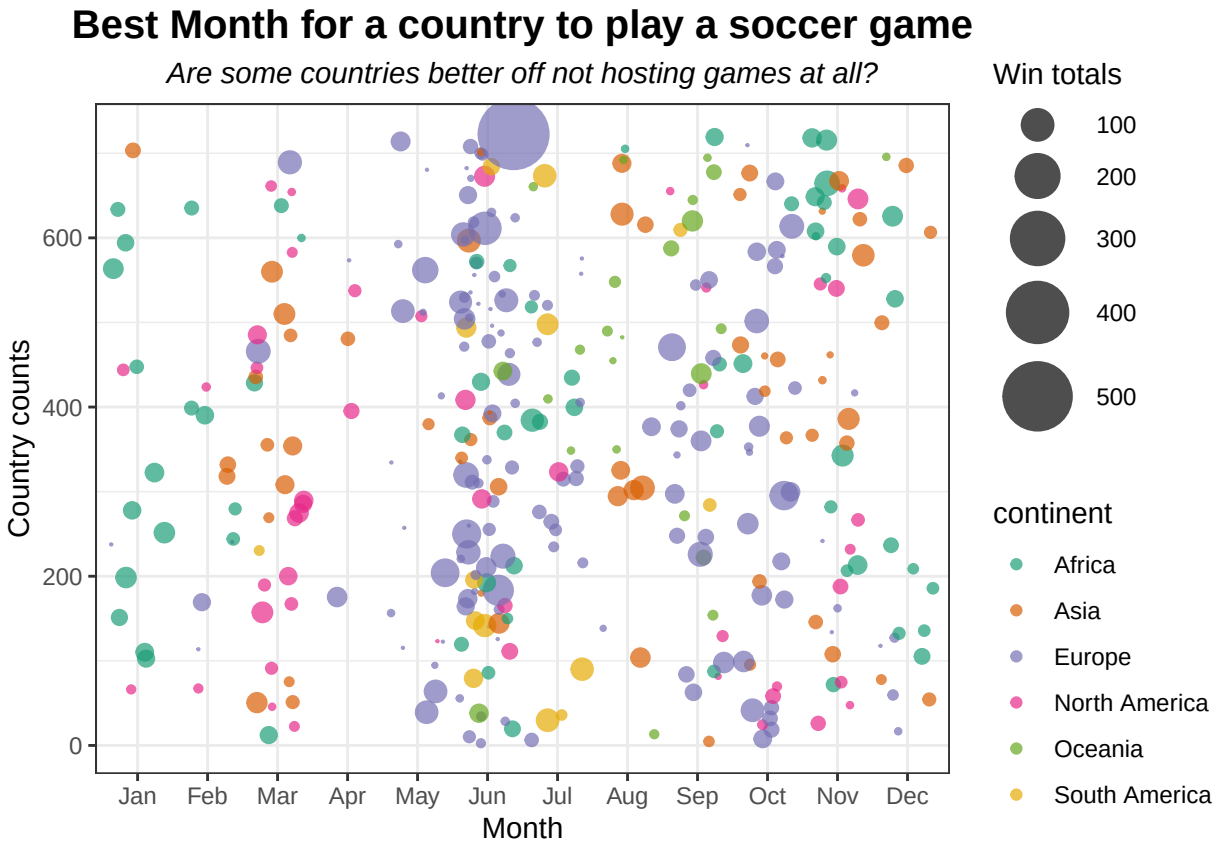
n	Africa	Asia	Europe	North America	Oceania	South America
1	Tanzania	Turkey	Slovenia	Puerto Rico	Northern Mariana Islands	Ecuador
2	Togo	Turkmenistan	Soviet Union	Saint Kitts and Nevis	Palau	French Guiana
3	Tunisia	United Arab Emirates	Spain	Saint Lucia	Papua New Guinea	Guyana
4	Uganda	Uzbekistan	Sweden	Saint Martin	Samoa	Netherlands Antilles
5	United Arab Republic	Vietnam	Switzerland	Saint Vincent and the Grenadines	Solomon Islands	Netherlands Guyana
6	Upper Volta	Vietnam DR	Ukraine	Sint Maarten	Tahiti	Paraguay
7	Zaïre	Vietnam Republic	United Kingdom	Trinidad and Tobago	Tonga	Peru

n	Africa	Asia	Europe	North America	Oceania	South America
8	Zambia	Yemen	Vatican City	Turks and Caicos Islands	Tuvalu	Suriname
9	Zanzibar	Yemen AR	Wales	U.S. Virgin Islands	Vanuatu	Uruguay
10	Zimbabwe	Yemen DPR	Yugoslavia	United States	Western Samoa	Venezuela

4.2 Second Objectives

This section will focus on finding insghts of the data based the summary's found in the dataset.

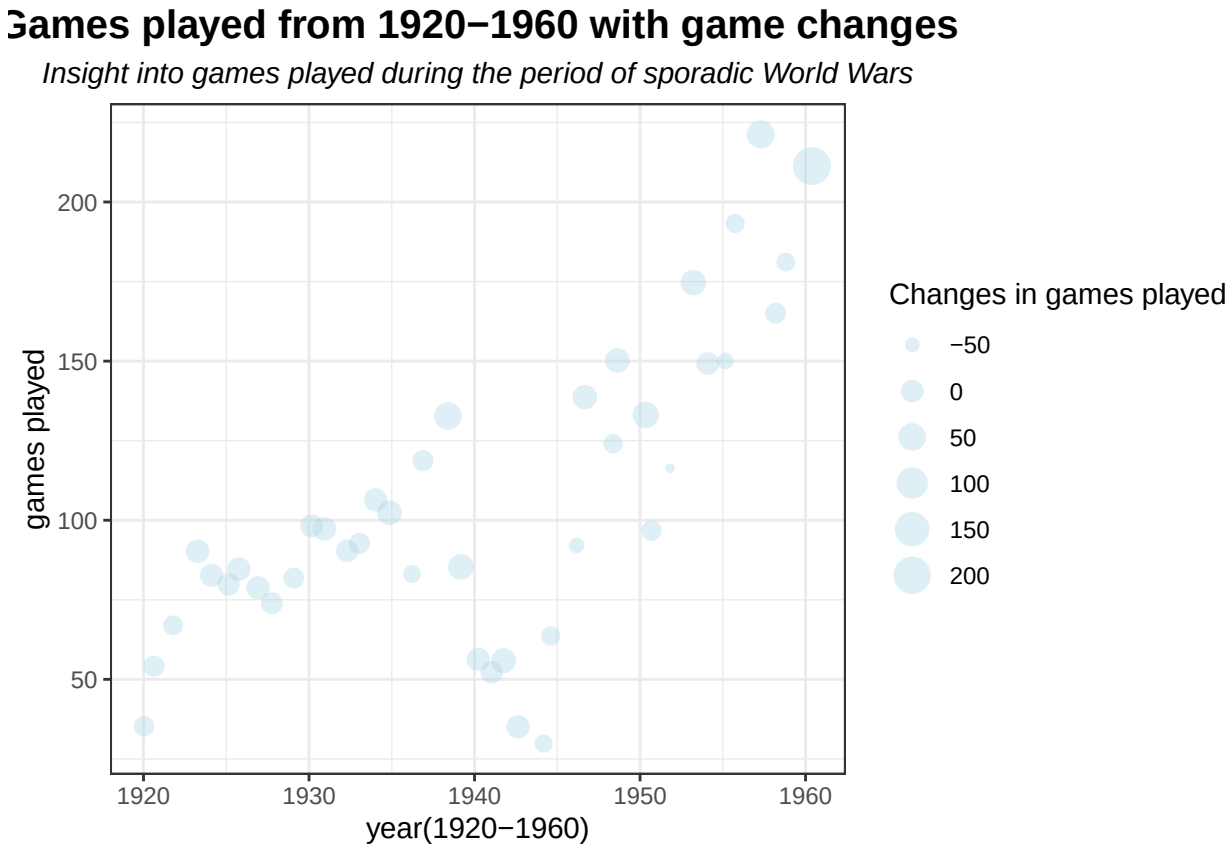
1. What was the best month for a team to play a soccer game? Solution: We need to do a group by using two fields. First by month then by team. The team is a little tricky since there are home and away teams, but this can be remeded using a case when statement.



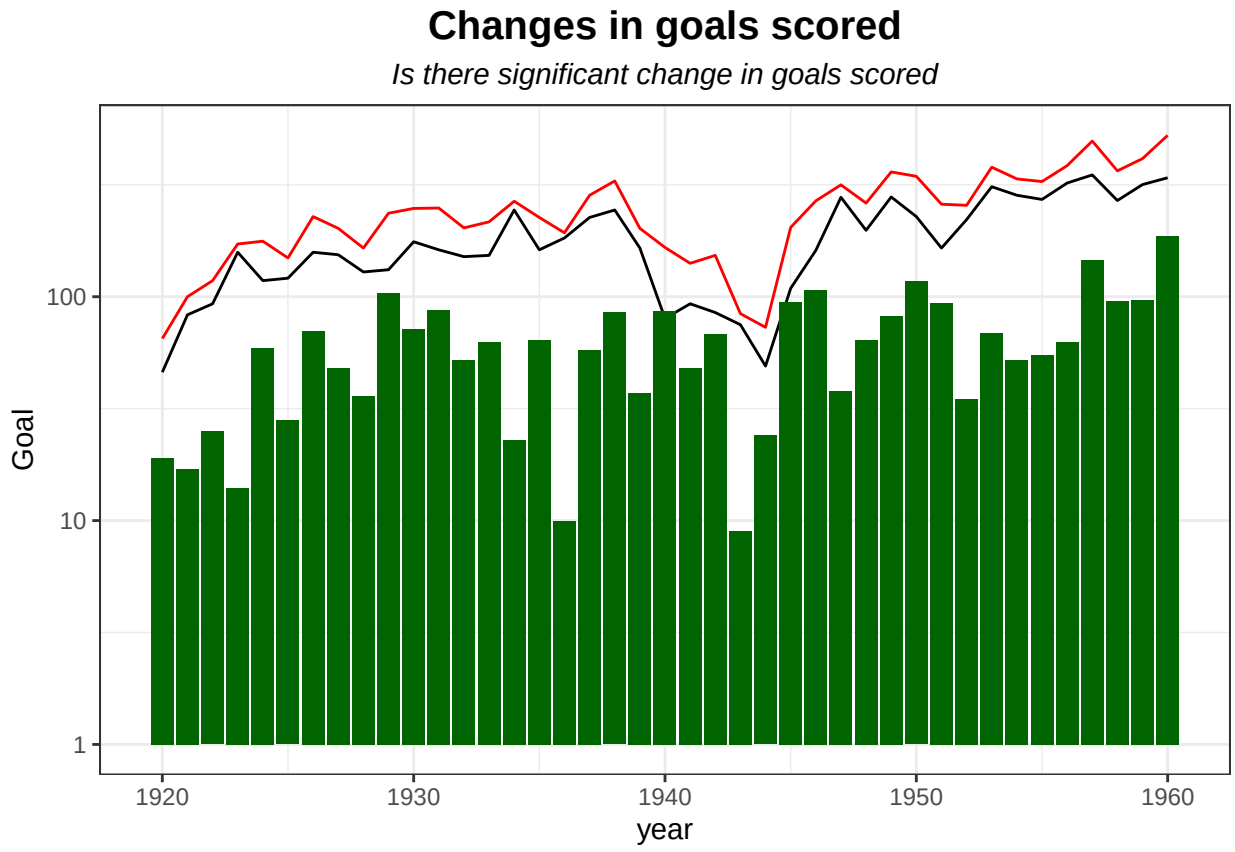
2.From 1920's -1960's: *i. What are the changes in games played ii. Highest home/away goal scored in each year iii. Rolling Average of total goals scored iv. Changes in game played on each continent*

Solution: Filter the dataset to games played between January,1920 - December,1960.

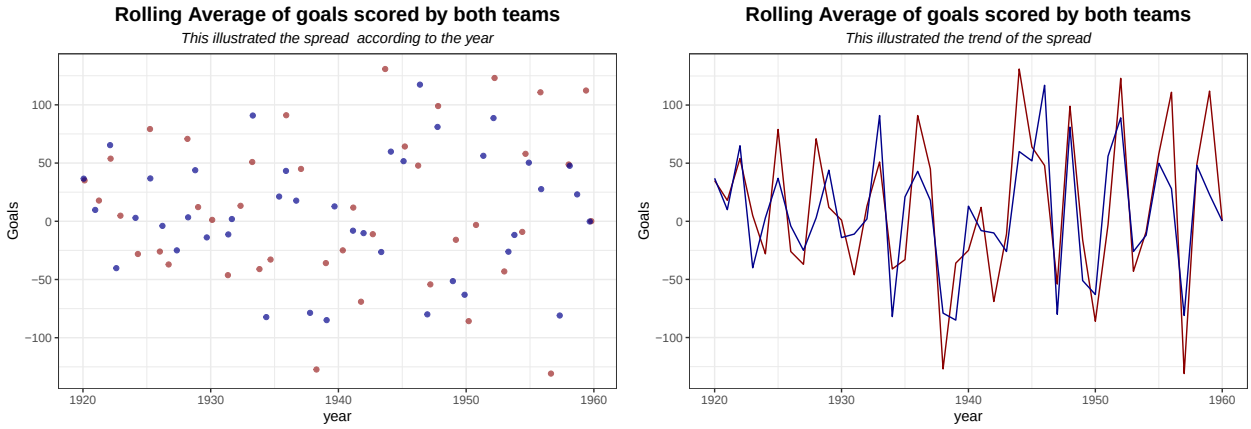
i. Count the number of games played per year, then use the lag function to calculate pr



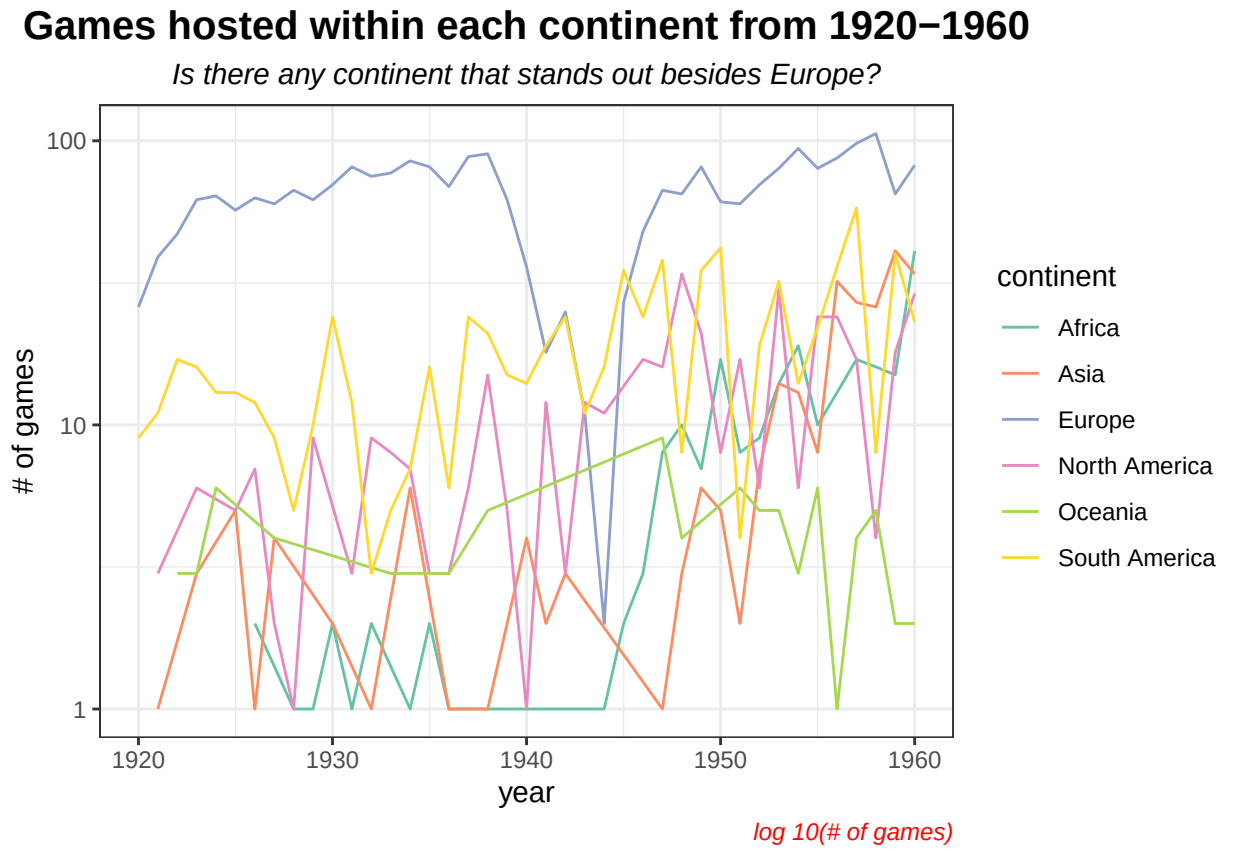
ii. Add both home and away goals the subseting it by year.



iii. Use the Answer from (ii) as base for the rolling average calculation. We can also calculate for the difference in goal scored in year. We will use the lead method to determine the changes before comparing them.



iv.Same concept as the (i) focusing only on the continent per year.



3. Excluding Friendly and qualification games, Based on the top 10 tournamenet, what are the total goals scored during each decade?