Fifa World Cup Analysis

Rayyan Kazim

31/03/2022

Introduction: - https://www.kaggle.com/datasets/abecklas/fifa-world-cup - the data comes in a csv file from kaggle. This data is about the Fifa World Cup, which is a topic that I find really interesting as it is a sporting event that is watched almost all over the world as it incorporates a large amount of countries. The questions I will be investigating are which countries scored the most goals in world cup history? which countries participated in the most world cups? in terms of goal scoring, how well did countries perform over time? This all leads to the question, which country is the best at the sport historically? Plots will be nessasary to answer and conclude these questions.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.1.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(styler)
library(dslabs)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)
set.seed(87)

data <- "/Users/rayyankazim/Documents/HTHSCI-1M03/WorldCupMatches.csv"
data_read <- read.csv(data)
glimpse(data_read)
```

```
## Rows: 4,572
## Columns: 20
## $ Year                 <int> 1930, 1930, 1930, 1930, 1930, 1930, 1930, 1930, 1~
## $ Datetime             <chr> "13 Jul 1930 - 15:00 ", "13 Jul 1930 - 15:00 ", "~
## $ Stage                <chr> "Group 1", "Group 4", "Group 2", "Group 3", "Grou~
## $ Stadium              <chr> "Pocitos", "Parque Central", "Parque Central", "P~
## $ City                 <chr> "Montevideo ", "Montevideo ", "Montevideo ", "Mon~
## $ Home.Team.Name       <chr> "France", "USA", "Yugoslavia", "Romania", "Argent~
## $ Home.Team.Goals      <int> 4, 3, 2, 3, 1, 3, 4, 3, 1, 1, 6, 4, 1, 4, 3, 6, 6~
## $ Away.Team.Goals      <int> 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 1, 1, 1~
## $ Away.Team.Name       <chr> "Mexico", "Belgium", "Brazil", "Peru", "France", ~
## $ Win.conditions       <chr> " ", " ", " ", " ", " ", " ", " ", " ", " ", " ",~
## $ Attendance           <int> 4444, 18346, 24059, 2549, 23409, 9249, 18306, 183~
## $ Half.time.Home.Goals <int> 3, 2, 2, 1, 0, 1, 0, 2, 0, 0, 3, 1, 1, 4, 2, 1, 3~
## $ Half.time.Away.Goals <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1~
## $ Referee              <chr> "LOMBARDI Domingo (URU)", "MACIAS Jose (ARG)", "T~
## $ Assistant.1          <chr> "CRISTOPHE Henry (BEL)", "MATEUCCI Francisco (URU~
## $ Assistant.2          <chr> "REGO Gilberto (BRA)", "WARNKEN Alberto (CHI)", "~
## $ RoundID              <int> 201, 201, 201, 201, 201, 201, 201, 201, 201, 201,~
## $ MatchID              <int> 1096, 1090, 1093, 1098, 1085, 1095, 1092, 1097, 1~
## $ Home.Team.Initials   <chr> "FRA", "USA", "YUG", "ROU", "ARG", "CHI", "YUG", ~
## $ Away.Team.Initials   <chr> "MEX", "BEL", "BRA", "PER", "FRA", "MEX", "BOL", ~
```

```
# some rows have duplicates (there are 2 of the same row) at the end
data_read <- data_read[!duplicated(data_read),]

# changing names to column lowercase
data_read %<>% rename_with(tolower)

#deleting columns that are not needed
data_read <- data_read[,-c(2)]
data_read <- data_read[,-c(3:4)]
data_read <- data_read[,-c(7:21)]
```

```
colnames(data_read) <- c("year","stage", "home_team", "home_team_goals",
                         "away_team_goals",
                         "away_team")




# Changing names of values within columns
data_read$home_team[data_read$home_team == "Germany FR"] <- "Germany"
data_read$away_team[data_read$away_team == "Germany FR"] <- "Germany"


scores <- data_read %>% drop_na()

# for the next part, the tibble will be broken down into 2 tibbles, 1 of them
# reloving around home teams and one revolving around away teams, the tibbles
# will be combined to have a year column, team column and goals column, instead
# of having both away and home teams and goals, there will just be teams
# and goals

Home <- tibble(
  year = scores$year,
  team = scores$home_team,
  team_goals = scores$home_team_goals
)

Away <- tibble(
  year = scores$year,
  team = scores$away_team,
  team_goals = scores$away_team_goals
)

goal <- bind_rows(Home, Away)

# the next part will group by year and teams to figure out the total number
# of goals scored by countries in each year.

totals <- goal %>%
  group_by(year, team) %>%
  summarise(total_goals = sum(team_goals))
```

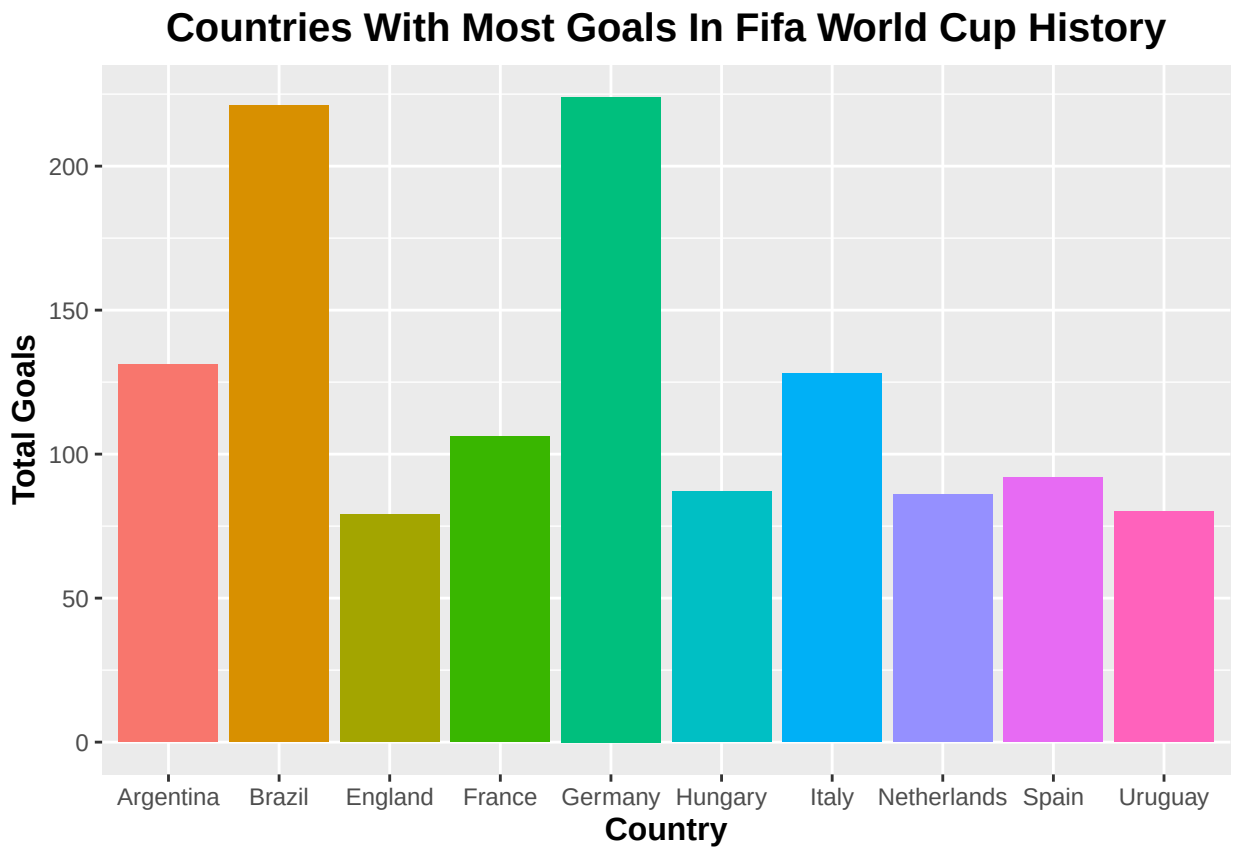
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
# the next part will be making the data wider by having each row represent
# years and each column representing a different country. The data can then
# be put back where one column represents country, one being year and one
# being goals. In the country column, the rows with the same country will be
# right on top of eachother, unlike the totals tibble where the year column
# has the same year rows right on top of eachother.

wide_data <- totals %>%
  pivot_wider(names_from = team, values_from = total_goals)
```

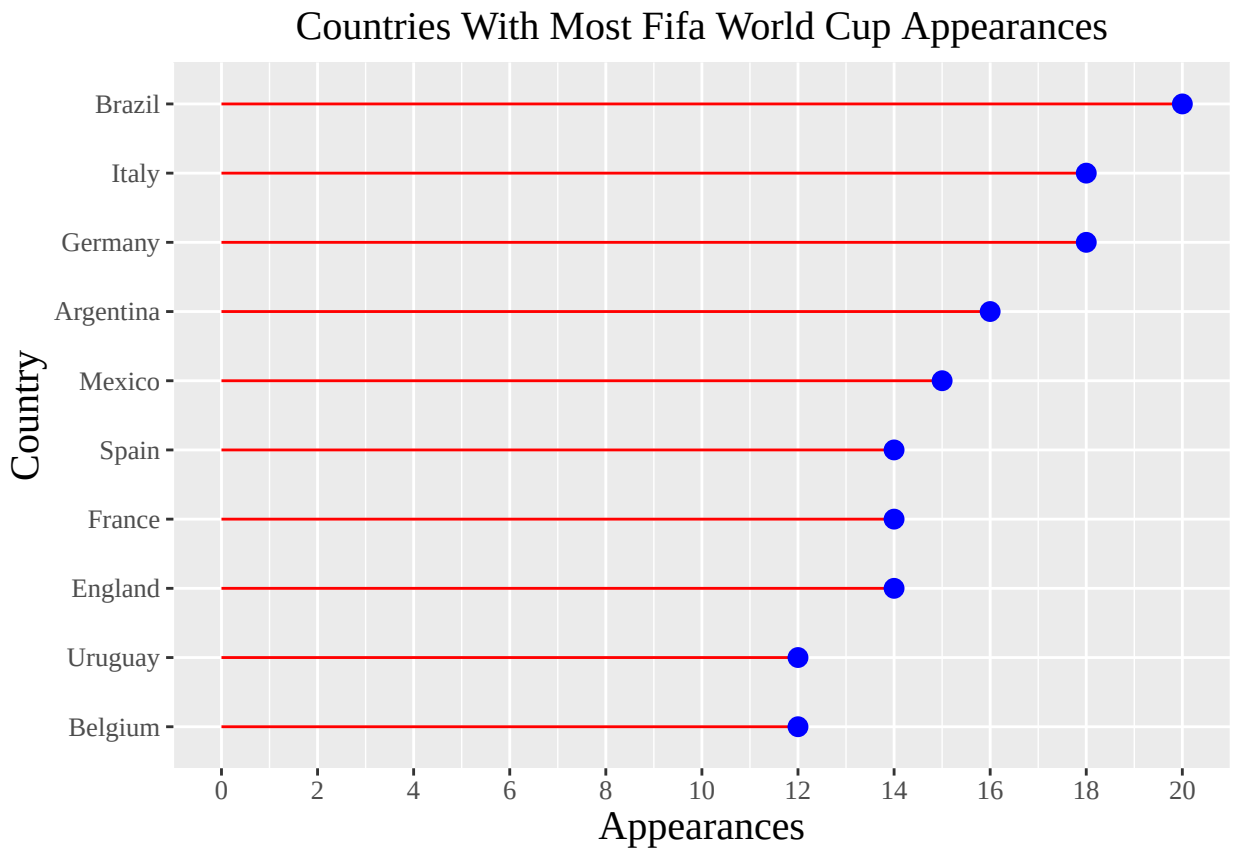
```r
allCountryData <- wide_data %>%
  gather("country", "goals", 2:83)


allCountryData <- allCountryData[c("country", "year", "goals")]

# if a country has NA in the goals column, that means that the country did not
# participate that year, in this case, those rows can be dropped.
final_tibble <- allCountryData %>% drop_na()

# final_tibble is the final clean tibble that will be worked with for
# visualizations
glimpse(final_tibble)
```

```
## Rows: 427
## Columns: 3
## Groups: year [20]
## $ country <chr> "Argentina", "Argentina", "Argentina", "Argentina", "Argentina~
## $ year    <int> 1930, 1934, 1958, 1962, 1966, 1974, 1978, 1982, 1986, 1990, 19~
## $ goals   <int> 18, 2, 5, 2, 4, 9, 15, 8, 14, 5, 8, 10, 2, 11, 10, 8, 0, 2, 1,~
```

```r
# The discussion for the plots is at the bottom with the conclusion

plot1 <- final_tibble %>%
  group_by(country) %>%
  summarise(l=sum(goals)) %>%
  arrange(desc(l)) %>%
  head(10) %>%
  ungroup()%>%
  ggplot(aes(x=country, y = l, fill = country)) + geom_bar(stat="identity") +
  guides(fill = "none") +
  labs(x="Country",y="Total Goals",
       title = "Countries With Most Goals In Fifa World Cup History") +
  theme(plot.title = element_text(size = 15,
                                  color = "black", family="Helvetica",
                                  face="bold",
                                  hjust=0.5),
        axis.title = element_text(size=12, color = "black",
                                  family="Helvetica", face="bold"))
plot1
```




```
plot2 <- final_tibble %>%
  group_by(country) %>%
  count(country) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  ungroup()%>%
  ggplot(aes(x=reorder(country, n), y = n)) +
  geom_segment(aes(x=reorder(country,n), xend=reorder(country,n),
                   y=0, yend=n), color="red") +
  geom_point(size=3, color = "blue") +
  scale_y_continuous(breaks = seq(0,20,by=2)) +
  labs(x="Country", y="Appearances",
       title = "Countries With Most Fifa World Cup Appearances") +
  theme(plot.title = element_text(size=15, family="Times", hjust=0.5),
        axis.title = element_text(size=15, family = "Times"),
        axis.text = element_text(size = 10, family="Times")) +
  coord_flip()
plot2
```




```
# the next part will be picking out the top 4 countries with the most total
# goals of all years combined.
topGoalsCountries <- final_tibble %>%
  group_by(country) %>%
  summarise(l=sum(goals)) %>%
  arrange(desc(l)) %>%
  head(4)

# The next part will figure out how many goals each of the top 4 countries with
# the most total goals of all years combined, scored over time, in each
# year.
selecting_countries <- final_tibble %>%
  group_by(country, year) %>%
  summarise(n = sum(goals)) %>%
  filter(country %in% topGoalsCountries$country)
```

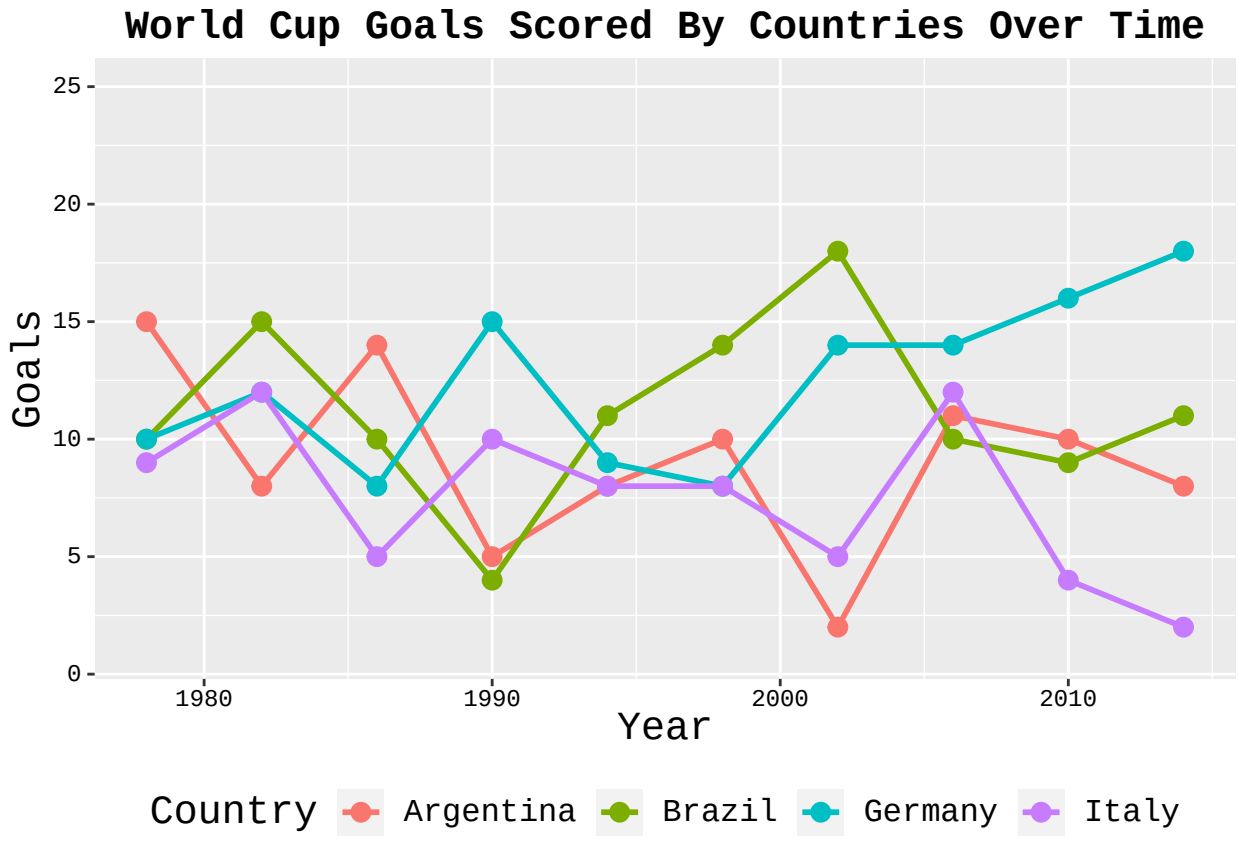
```
## `summarise()` has grouped output by 'country'. You can override using the `.groups` argument.
```

```
plot3 <- selecting_countries %>%
  ggplot(aes(x=year, y=n, color = country)) +
  geom_line(linetype = "solid", size=1) + geom_point(size=3) +
  xlim(1978,2014) +
  labs(x = "Year", y = "Goals",
       title = "World Cup Goals Scored By Countries Over Time",
       color="Country") +
```

```
  theme(plot.title = element_text(size = 15, color="black", family="Courier",
                                  face = "bold",
                                  hjust = 0.5),
        axis.title.x = element_text(size=15, color="black", family="Courier"),
        axis.title.y = element_text(size=15, color="black", family="Courier"),
        axis.text.x = element_text(color="black", family="Courier"),
        axis.text.y = element_text(color="black", family="Courier"),
        legend.text = element_text(size=12,color="black", family = "Courier"),
        legend.title = element_text(size = 15, color="black", family="Courier"),
        legend.position = "bottom")
plot3
```

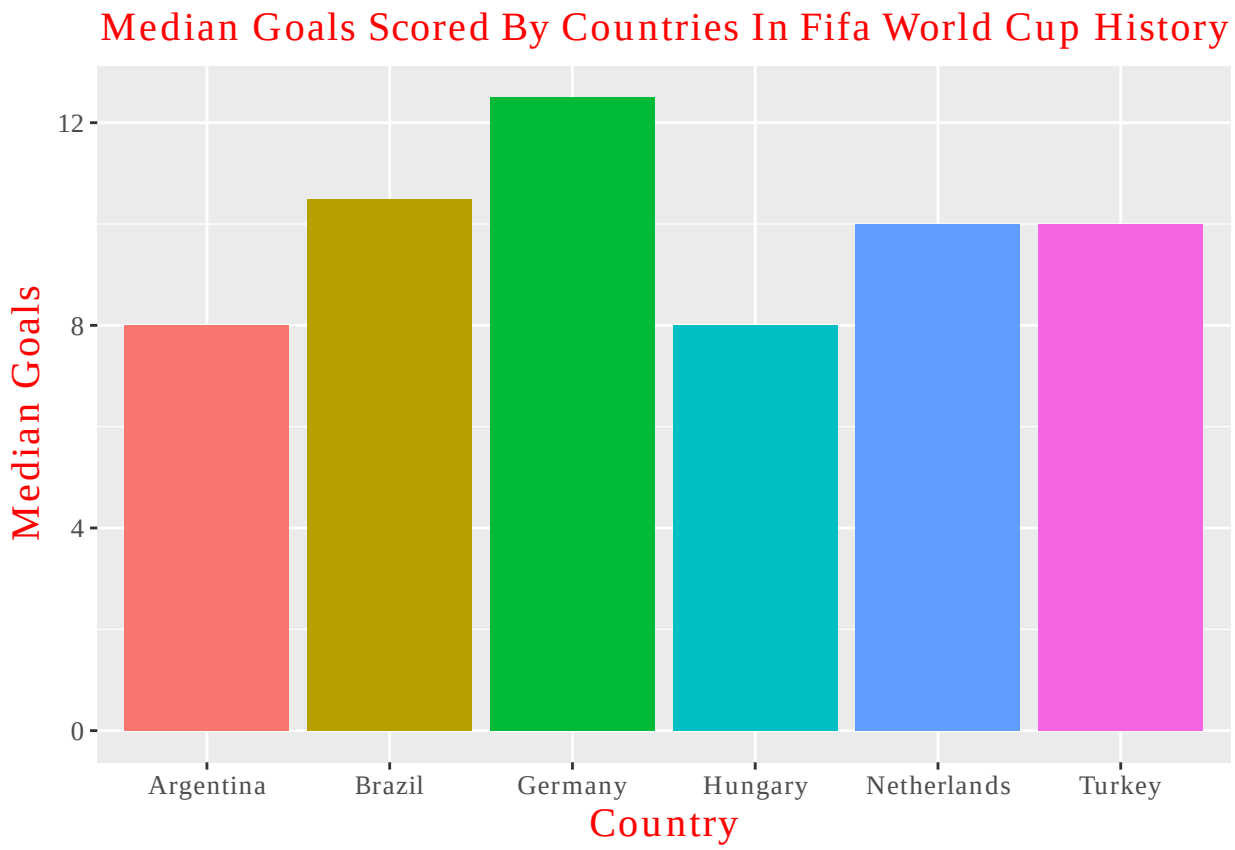
```
## Warning: Removed 32 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 32 rows containing missing values (geom_point).
```




```
plot4 <- final_tibble %>%
  group_by(country) %>%
  summarise(m=median(goals))%>%
  arrange(desc(m)) %>%
  head(6) %>%
  ungroup()%>%
  ggplot(aes(x=country,y=m, fill=country))+geom_bar(stat="identity")+
  guides(fill="none")+
```

```
  labs(x="Country", y="Median Goals",
       title="Median Goals Scored By Countries In Fifa World Cup History") +
  theme(plot.title = element_text(size=15,family="Palatino",face="bold",
                                  hjust=0.5, color = "red"),
        axis.title = element_text(size=15,family="Palatino",face="bold",
                                  color="red"),
        axis.text=element_text(size=10,family="Palatino"))
plot4
```



Conclusion: The main question in this topic is which country is the best at soccer based on its performance in world cup history? This question has been answered by smaller questions which are which countries scored the most goals in world cup history? which ones qualified/participated in the most world cups? how have their performances been over time in terms of goal scoring? Plot 1 looks at the countries that have scored the most goals in world cup history. The top country is Germany which means it has scored the most goals in world cup history, Brazil is second and has scored nearly the same amount as Germany. Argentina has the third most total goals scored, but the plot still shows that it is very far from Germany and Brazil. Plot 2 shows that Brazil has qualified/ participated/appeared in the most world cups, making them the most consistent country in the sporting event. They qualified for 20 world cups, where Italy and Germany are tied at second place, each appearing in 18 world cups. Brazil qualifying for the most does not only show that they are the most consistent, but it also shows that they do the best job in qualifying for the tournament. The debate in the best country at the sport is still between Brazil and Germany, as Germany is very close to Brazil on the plot, meaning they qualified for nearly the same amount of tournaments. Italy is not in the debate as their total goals scored amount is a lot lower that Germany's and Brazil's in plot 1. Plot 3 also shows the greatness of Germany and Brazil as in most of the years, out of the 4 countries that have the most goals in world cup history, the country with the most goals is either Germany or Brazil. There are 2 years where Argentina has had the most goals between the four countries, and there is not a

single year where Italy has had the most goals. There are 4 years where Germany has had the most and 4 where Brazil has had the most. It is also shown that since 1978, the maximum amount of goals scored in a year, or in a world cup, is 18 goals. Germany and Brazil have both reached that number, Brazil reached it in 2002, and Germany reached it in 2014. The plot shows that as each Country is scoring more and more goals over time, that country immediately has a downfall as they score a lot less in the year after. For example Brazil scored more and more goals from 1998 to 2002, and then they scored a lot less in 2006. Germany has been scoring more and more from 2006 to 2014, so based on future predictions from the plot, it is obvious that they may score a lot less in the future years of the plot. Plot 4 shows that Germany has had the highest count for median goals scored in world cup history, with Brazil being second place. It is very surprising to see Turkey on this plot with a high number as they have not been a part of any other plot. Turkey's number of median goals scored in world cup history is 10, meaning that in about half the world cups they participated in, they scored more that 10 goals. Looking back at Plot 3, that is a higher number of goals than how many Italy and Argentina scored in many years. However, this number was probably very easy for Turkey to maintain as Plot 2 shows that they have not appeared in as many tournaments as other countries, which is why they are not part of Plot 2. Since they have not appeared in too many world cups, it would be hard for them to appear in Plot 1 for the most goals scored in world cup history. Overall, Germany and Brazil have been the best countries at soccer in terms of their performance in world cup history. In terms of goal scoring, in plots 1 and 4, Germany has shown to be the best country at the sport. In terms of qualifying, appearing and staying consistent, Brazil has shown to be the best country at the sport in Plot 2. Plot 3 shows that Brazil and Germany are basically even, where Brazil was better at the sport in the past and Germany is better at the sport currently, in terms of goal scoring. Overall, Germany and Brazil are the best countries at soccer as they have had the best performances in world cup history. The plots helped a lot with answering the smaller questions to answer the big question.