FIFA World Cup

Lauren Schmiedeler

Create Functions

```
library(tidyverse)

# create a function that plots a bar graph
plot_bar_graph <- function(data, y_var, fill_var, title, x_lab, y_lab, fill_lab, color_1, color_
2) {
  ggplot(data, aes(x = reorder(country, -get(y_var)), y = get(y_var), fill = get(fill_var))) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  scale_fill_manual(values = c(color_1, color_2)) +
  labs(x = x_lab, y = y_lab, title = title, fill = fill_lab) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
        legend.position = "top")
}

# create a function that plots a world map
plot_world_map <- function(data, fill_var, title, low_color, high_color) {
  data <- data %>% mutate(country = as.character(country)) %>%
    mutate(country = case_when(country == "United States" ~ "USA",
                               country == "England" ~ "UK",
                               country == "China PR" ~ "China",
                               country == "Republic of Ireland" ~ "Ireland",
                               T ~ country)) %>%
    arrange(country) %>%
    mutate(country = factor(country))

  map_data("world") %>%
    merge(data, by.x = "region", by.y = "country", all.x = T) %>%
    arrange(group, order) %>%
    ggplot(aes(x = long, y = lat, group = group, fill = get(fill_var))) +
    geom_polygon(color = "white", linewidth = 0.2) +
    theme_minimal() +
    scale_fill_gradient(low = low_color, high = high_color, na.value = "gray90") +
    labs(title = title) +
    theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blank
(), legend.title = element_blank())
}

# create a function that combines the statistics for two countries
combine_stats <- function(data, country_1, country_2) {
  combined <- data %>% filter(country == country_1) %>% select(-country) +
    data %>% filter(country == country_2) %>% select(-country)
  combined$country <- country_1
  rbind(data %>% filter(country != country_1, country != country_2), combined)
}
```

```
library(paletteer)

# create a color palette
pal <- paletteer_c("viridis::viridis", 10)
```

Load Data

```
# load the data
data <- readRDS("data.RData")

# create a "matches" data frame
# convert the necessary variables to factors
matches <- data[[1]] %>% mutate(country = factor(country), city = factor(city), stage = factor(s
tage), home_team = factor(home_team), away_team = factor(away_team), outcome = factor(outcome),
win_conditions = factor(win_conditions), winning_team = factor(winning_team), losing_team = fact
or(losing_team), month = factor(month), dayofweek = factor(dayofweek))
# fix a spelling error
matches$winning_team[matches$winning_team == "Portagul"] <- "Portugal"
summary(matches)
```

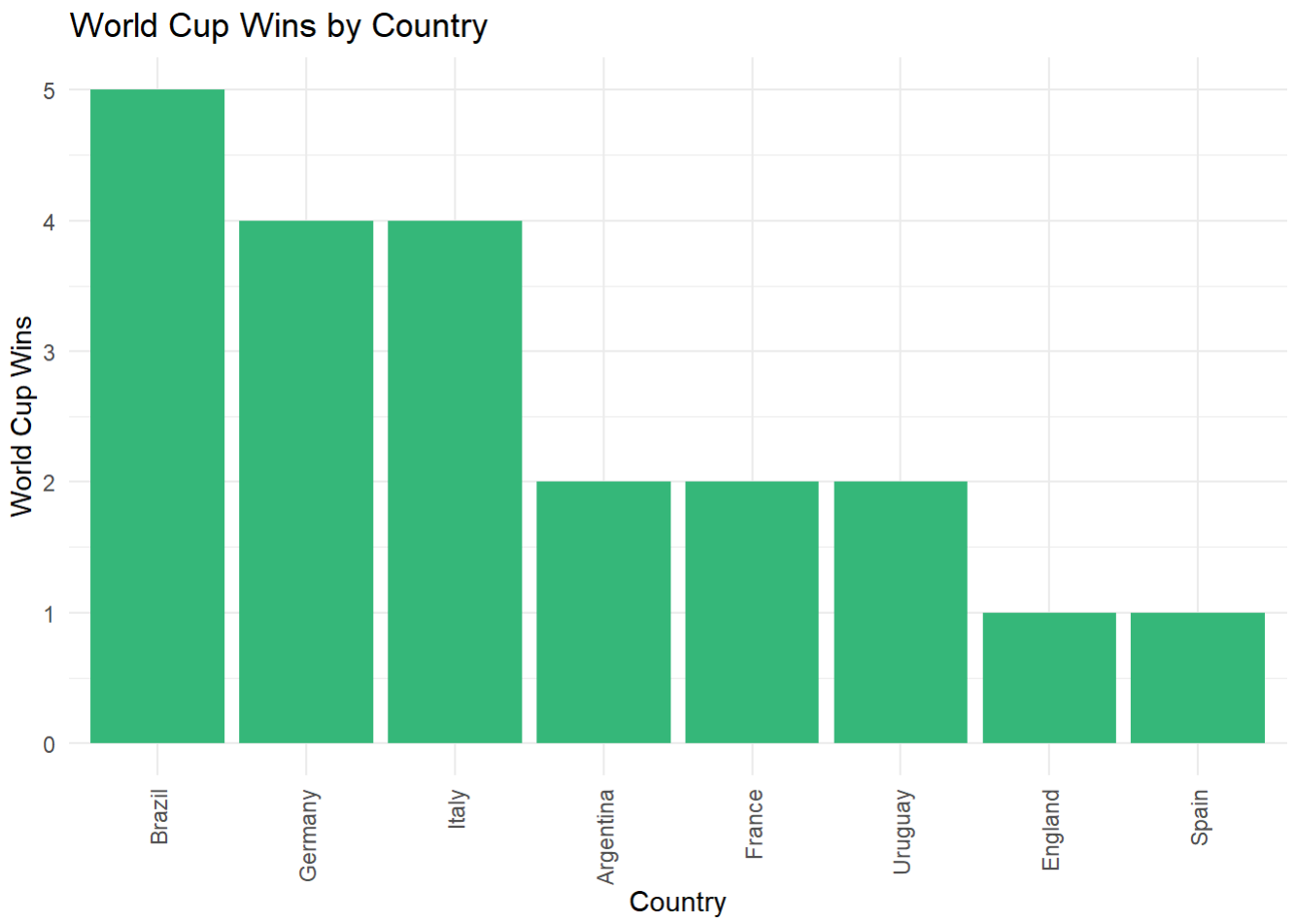
```
##       year         country                city                stage
##  Min.   :1930   Germany:102   Mexico City   : 23   Round of 16  : 89
##  1st Qu.:1970   Brazil : 86   Montevideo    : 18   Group B      : 69
##  Median :1990   Mexico : 84   Guadalajara   : 17   Group A      : 68
##  Mean   :1987   France : 82   Johannesburg  : 15   Quarterfinals: 66
##  3rd Qu.:2006   Italy  : 69   Rio de Janeiro: 15   Group 1      : 59
##  Max.   :2018   Russia : 64   Buenos Aires  : 12   Group C      : 57
##                 (Other):413   (Other)       :800   (Other)      :492
##         home_team        away_team     home_score       away_score    outcome
##  Brazil      : 84   Uruguay   : 37   Min.   : 0.000   Min.   :0.000   A:302
##  Argentina   : 60   Italy     : 34   1st Qu.: 0.000   1st Qu.:0.000   D:169
##  Italy       : 49   England   : 33   Median : 1.000   Median :1.000   H:429
##  West Germany: 43   Spain     : 33   Mean   : 1.569   Mean   :1.262
##  France      : 40   Mexico    : 31   3rd Qu.: 2.000   3rd Qu.:2.000
##  England     : 36   Yugoslavia: 28   Max.   :10.000   Max.   :8.000
##  (Other)     :588   (Other)   :704
##                              win_conditions       winning_team    losing_team
##  Italy won in AET                   :  5    Brazil      : 76   Mexico   : 29
##  Argentina won in AET               :  3    Argentina   : 47   Argentina: 24
##  England won in AET                 :  3    Italy       : 46   England  : 22
##  Argentina won in penalties (4 - 3):  2    West Germany: 40   France   : 21
##  Belgium won in AET                 :  2    France      : 36   Spain    : 21
##  (Other)                            : 47    (Other)     :486   (Other)  :614
##  NA's                               :838    NA's        :169   NA's     :169
##       date             month         dayofweek
##  Min.   :1930-07-13   Jul:150   Friday   : 92
##  1st Qu.:1970-06-14   Jun:727   Monday   : 82
##  Median :1990-06-23   May: 23   Saturday :152
##  Mean   :1987-05-20             Sunday   :196
##  3rd Qu.:2006-06-19             Thursday :111
##  Max.   :2018-07-15             Tuesday  :119
##                                 Wednesday:148
```

```
# create a "cups" data frame
# convert the necessary variables to factors
cups <- data[[2]] %>% mutate(host = factor(host), winner = factor(winner), second = factor(secon
d), third = factor(third), fourth = factor(fourth))
summary(cups)
```

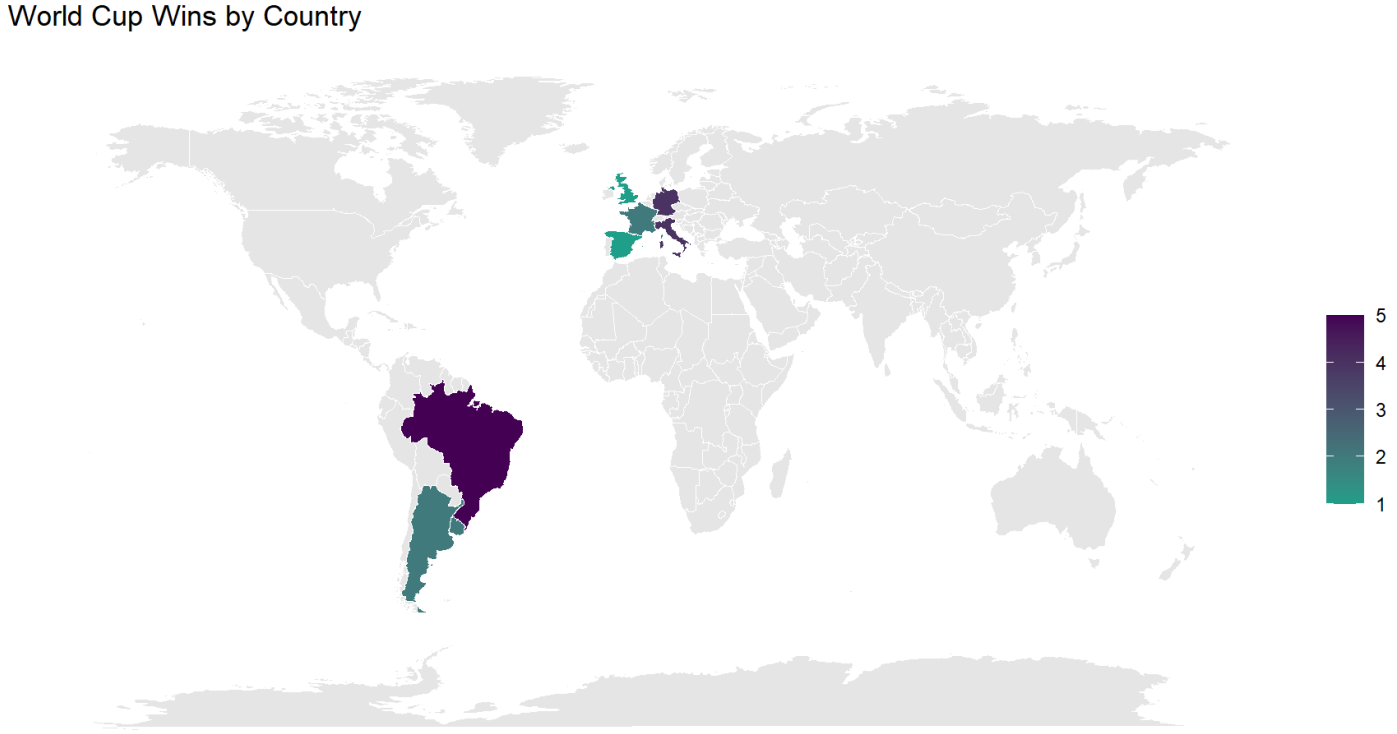
```
##       year             host             winner             second      third
##  Min.   :1930   Brazil   : 2   Brazil      :5   Argentina     :3   Germany:3
##  1st Qu.:1958   France   : 2   Italy       :4   Netherlands   :3   Brazil :2
##  Median :1978   Germany  : 2   West Germany:3   West Germany  :3   France :2
##  Mean   :1977   Italy    : 2   Argentina   :2   Brazil        :2   Poland :2
##  3rd Qu.:1998   Mexico   : 2   France      :2   Czechoslovakia:2   Sweden :2
##  Max.   :2018   Argentina: 1   Uruguay     :2   Hungary       :2   Austria:1
##                 (Other)  :10   (Other)     :3   (Other)       :6   (Other):9
##         fourth    goals_scored       teams           games
##  Uruguay   : 3   Min.   : 70.0   Min.   :13.00   Min.   :17.00
##  Brazil    : 2   1st Qu.: 89.0   1st Qu.:16.00   1st Qu.:32.00
##  England   : 2   Median :126.0   Median :16.00   Median :38.00
##  Yugoslavia: 2   Mean   :121.3   Mean   :21.76   Mean   :42.86
##  Austria   : 1   3rd Qu.:146.0   3rd Qu.:32.00   3rd Qu.:64.00
##  Belgium   : 1   Max.   :171.0   Max.   :32.00   Max.   :64.00
##  (Other)   :10
##    attendance
##  Min.   : 395000
##  1st Qu.: 943000
##  Median :1774022
##  Mean   :1898122
##  3rd Qu.:2724604
##  Max.   :3568567
##
```

Analyze World Cup Wins

```
total_cup_wins <- cups %>% group_by(winner) %>%
  summarize(total_cup_wins = n()) %>%
  rename(country = winner)

# combine the statistics for "Germany" and "West Germany"
total_cup_wins <- combine_stats(total_cup_wins, "Germany", "West Germany") %>%
  arrange(-total_cup_wins)

ggplot(total_cup_wins, aes(x = reorder(country, -total_cup_wins), y = total_cup_wins)) +
  geom_bar(stat = "identity", fill = pal[7]) +
  theme_minimal() +
  labs(x = "Country", y = "World Cup Wins", title = "World Cup Wins by Country") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



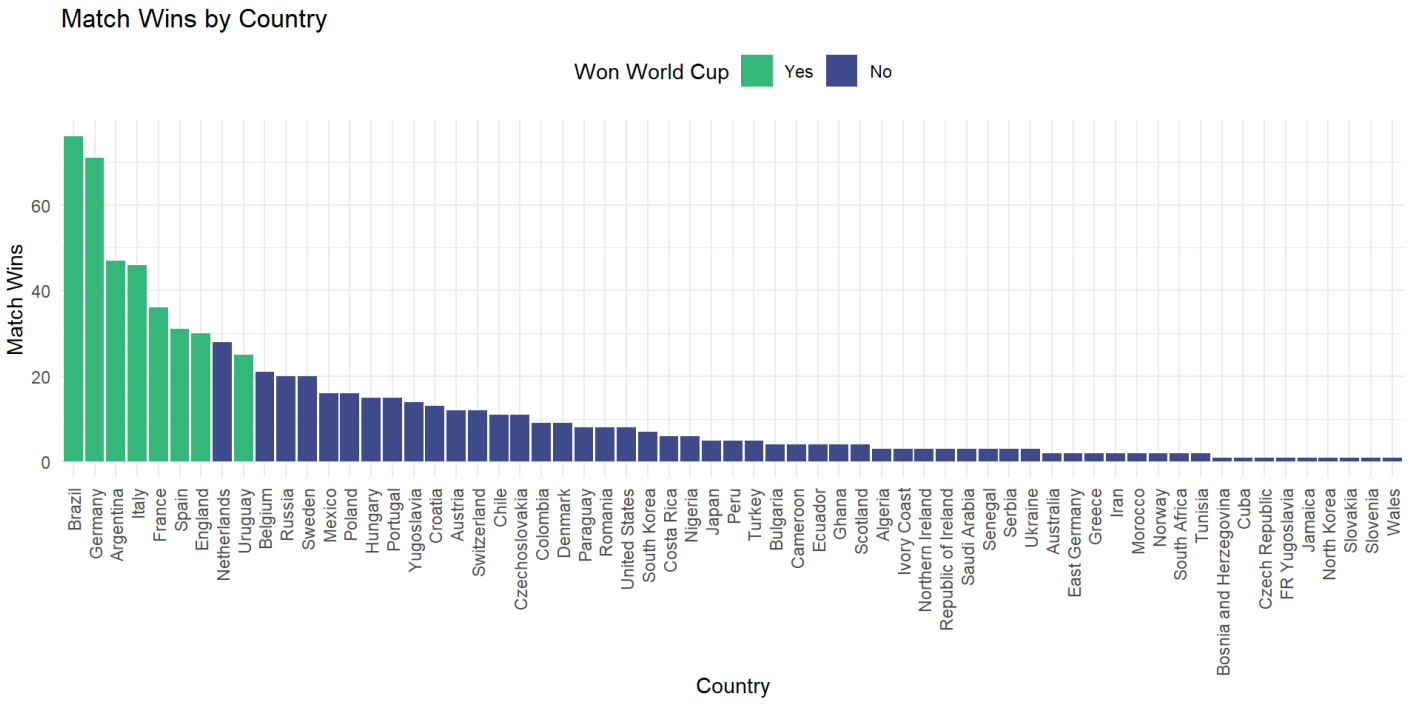

```
plot_world_map(total_cup_wins, "total_cup_wins", "World Cup Wins by Country", pal[6], pal[1])
```



Analyze Match Wins

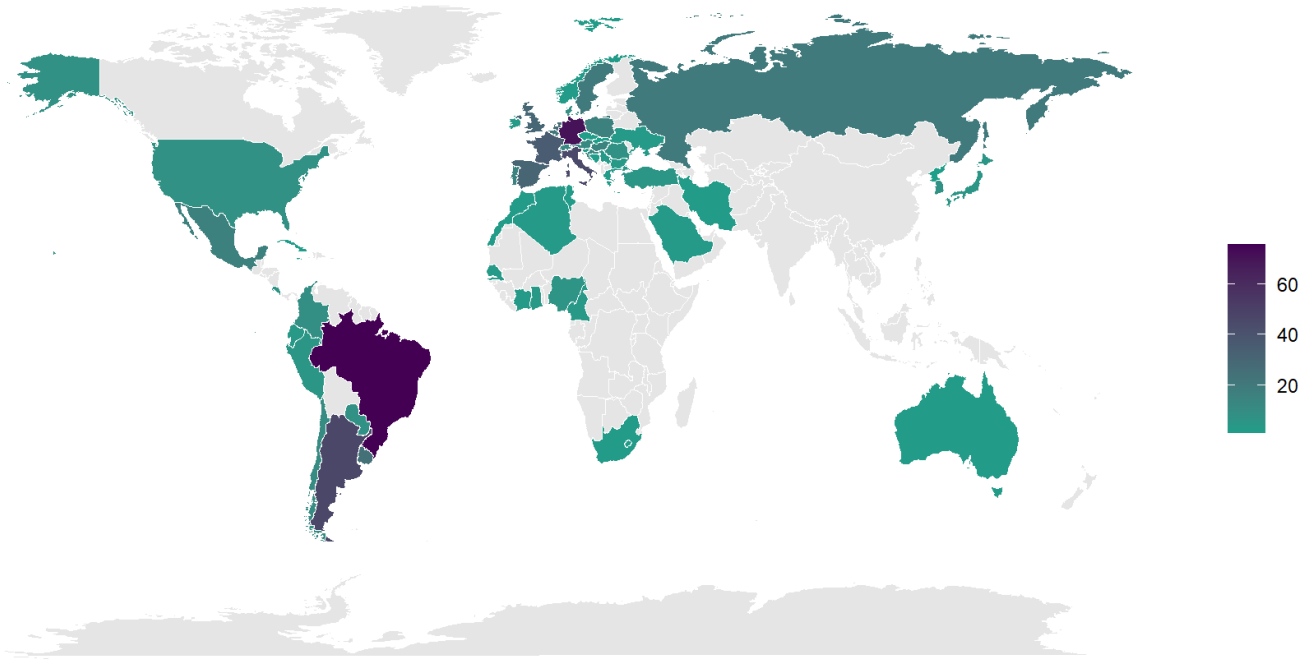
```
total_match_wins <- matches %>% group_by(winning_team) %>%
  summarize(total_match_wins = n()) %>%
  na.omit() %>%
  mutate(won_world_cup = ifelse(winning_team %in% total_cup_wins$country, 1, 0)) %>%
  rename(country = winning_team)

# combine the statistics for "Germany" and "West Germany"
total_match_wins <- combine_stats(total_match_wins, "Germany", "West Germany")

# combine the statistics for "Russia" and "Soviet Union"
total_match_wins <- combine_stats(total_match_wins, "Russia", "Soviet Union") %>%
  arrange(country)

total_match_wins <- total_match_wins %>% mutate(won_world_cup = ifelse(won_world_cup == 1, "Ye
s", "No"))
total_match_wins$won_world_cup <- factor(total_match_wins$won_world_cup, levels = c("Yes", "N
o"))
```

```
plot_bar_graph(total_match_wins, "total_match_wins", "won_world_cup", "Match Wins by Country",
"Country", "Match Wins", "Won World Cup", pal[7], pal[3])
```




```
plot_world_map(total_match_wins, "total_match_wins", "Match Wins by Country", pal[6], pal[1])
```

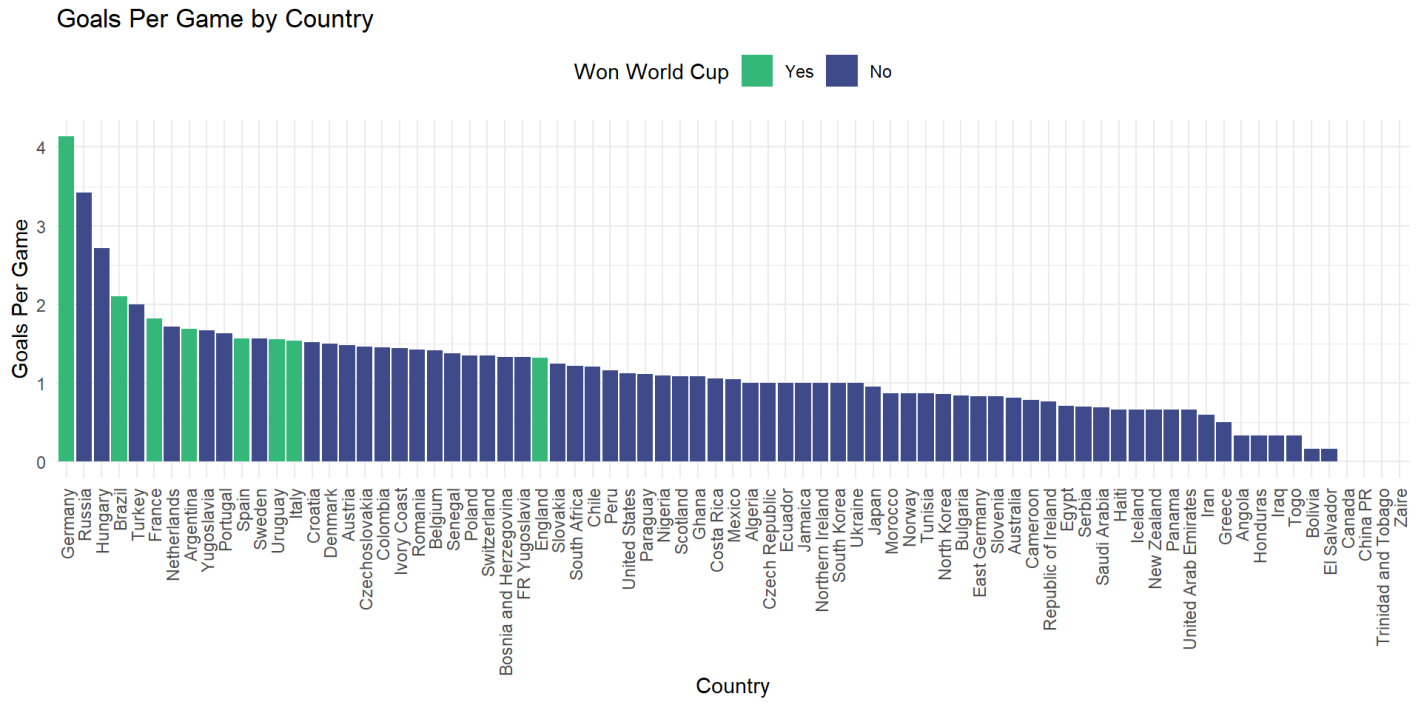
Match Wins by Country



Analyze Goals Per Game

```
home_goals <- matches %>% group_by(home_team) %>%
  summarize(total_goals_home = sum(home_score), total_games_home = n()) %>%
  mutate(team = home_team) %>%
  select(team, total_goals_home, total_games_home)

away_goals <- matches %>% group_by(away_team) %>%
  summarize(total_goals_away = sum(away_score), total_games_away = n()) %>%
  mutate(team = away_team) %>%
  select(team, total_goals_away, total_games_away)

goals <- merge(home_goals, away_goals) %>%
  mutate(total_goals = total_goals_home + total_goals_away,
         total_games = total_games_home + total_games_away) %>%
  select(team, total_goals, total_games) %>%
  mutate(goals_per_game = total_goals / total_games,
         won_world_cup = ifelse(team %in% total_cup_wins$country, 1, 0)) %>%
  rename(country = team)

# combine the statistics for "Germany" and "West Germany"
goals <- combine_stats(goals, "Germany", "West Germany")

# combine the statistics for "Russia" and "Soviet Union"
goals <- combine_stats(goals, "Russia", "Soviet Union") %>%
  arrange(-goals_per_game)

goals <- goals %>% mutate(won_world_cup = ifelse(won_world_cup == 1, "Yes", "No"))
goals$won_world_cup <- factor(goals$won_world_cup, levels = c("Yes", "No"))
```

```
plot_bar_graph(goals, "goals_per_game", "won_world_cup", "Goals Per Game by Country", "Country",
"Goals Per Game", "Won World Cup", pal[7], pal[3])
```




```
plot_world_map(goals, "goals_per_game", "Goals Per Game by Country", pal[6], pal[1])
```