match_data_team2

2024-12-06

#Create df with average match data by country

```
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
```

```
df <- read.csv(
  "match_data_clean.csv",
  header=TRUE)

#Create df to store country avg stats
avg_country_df <- data.frame()

#Create df for match data for each individual team (for use later)
match_data_team <- data.frame()


#Create new df with info summarized by country
countries <- unique(df$team1)
for (country in countries) {

  #Create subset df for games when country is team 1
  df_ss1 <- df %>% filter(team1 == country)
  df_ss1 <- df_ss1 %>% rename(fouls.committed = fouls.against.team1)
  df_ss1 <- df_ss1 %>% rename(fouls.drawn = fouls.against.team2)
  #Only keep variables related to country (remove team 2 vars)
  df_ss1 <- df_ss1 %>% select(!contains("2"))
  #Remove 1 from colnames
  colnames(df_ss1) <- gsub("1", "", colnames(df_ss1), "team_num")
  df_ss1$team_num <- 1

  #Vice versa
  df_ss2 <- df %>% filter(team2 == country)
  df_ss2 <- df_ss2 %>% rename(fouls.committed = fouls.against.team2)
  df_ss2 <- df_ss2 %>% rename(fouls.drawn = fouls.against.team1)
  #Only keep variables related to country (remove team 2 vars)
  df_ss2 <- df_ss2 %>% select(!contains("1"))
  #Remove 1 from colnames
```

```
  colnames(df_ss2) <- gsub("2", "", colnames(df_ss2), "team_num")
  df_ss2$team_num <- 2

  #Combine both ss dfs
  df_combined <- rbind(df_ss1, df_ss2)

  #Remove columns from which we can't take the average
  df_combined <- df_combined %>%
    select(-c("date", "hour", "category", "elimination"))

  #Create df for match data for each individual team (for use later)
  match_data_team <- rbind(match_data_team, df_combined)

  #Take means of columns
  df_summary <- df_combined[,-which(names(df_combined) %in% "outcome")] %>% group_by(team) %>%
    summarise_all(mean)

  #Add country row to df
  avg_country_df <- rbind(avg_country_df, df_summary)

}

#Add columns for longitude and latitude of capital cities
#Import this data
capitals <- read.csv("country-capital-lat-long-population.csv", header=TRUE)
#Convert to upper case and change column names to match data formats
capitals$Country <- toupper(capitals$Country)
colnames(capitals)[1] <- "team"

#Need to make some manual additions to capitals df
#Some country names are inconsistent
#Not all teams in the World Cup are actualy countries
capitals <- capitals %>% add_row(team="ENGLAND", Capital.City="London",
                                 Latitude=51.5085, Longitude=-0.1257,
                                 Population=9046485,Capital.Type="Capital")
capitals <- capitals %>% add_row(team="KOREA REPUBLIC", Capital.City="Seoul",
                                 Latitude=37.5683, Longitude=126.9778,
                                 Population=9963497,Capital.Type="Capital")
capitals <- capitals %>% add_row(team="UNITED STATES",
                                 Capital.City="Washington, D.C.",
                                 Latitude=38.8951, Longitude=-77.0364,
                                 Population=5206593,Capital.Type="Capital")
capitals <- capitals %>% add_row(team="WALES",
                                 Capital.City="Cardiff",
                                 Latitude=51.481583, Longitude=-3.179090,
                                 Population=372089,Capital.Type="Capital")
capitals <- capitals %>% add_row(team="IRAN",
                                 Capital.City="Tehran",
                                 Latitude=35.6944, Longitude=51.4215,
                                 Population=8895947,Capital.Type="Capital")

#Merge
coords_df <- merge(capitals, avg_country_df, by="team")
```

```r
write.csv(
  coords_df,
  "avg_team_data.csv",
  row.names = FALSE)

#Add column to match_team_data indicating result of the match
match_data_team$result <- as.factor(ifelse(match_data_team$outcome==0, 0,
                    ifelse(match_data_team$outcome==match_data_team$team_num, 1, 2)))
write.csv(
  match_data_team,
  "match_data_team.csv",
  row.names = FALSE)
```