Cleaned up ExploratoryDA

Ethan, Stephen, William

4/19/2021

```
##Connect

con <- dbConnect(drv=RSQLite::SQLite(), dbname="database.sqlite")

##Getting the tables

tables <- dbListTables(con)
tables <- tables[tables != "sqlite_sequence"]

##Reading in SQL DATA

country = dbReadTable(con, "Country")
league = dbReadTable(con, "League")
matches = dbReadTable(con, "Match")
player = dbReadTable(con, "Player")
player_Attributes = dbReadTable(con, "Player_Attributes")
teams = dbReadTable(con, "Team")
team_attributes = dbReadTable(con, "Team_Attributes")

##Disconnect from Database
dbDisconnect(con)
```

Data Source

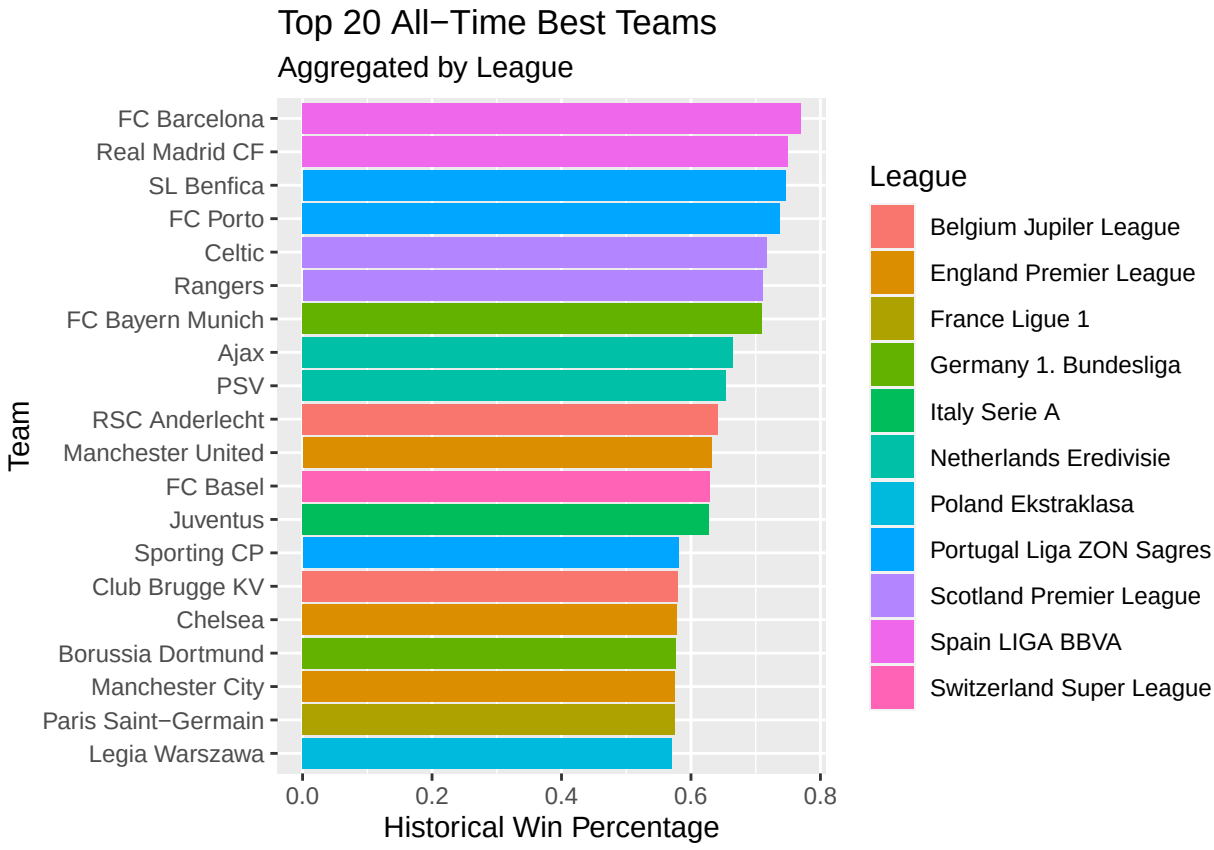
This data is an SQL database containing information for more than 25,000 matches and 10,000 players from European professional football.

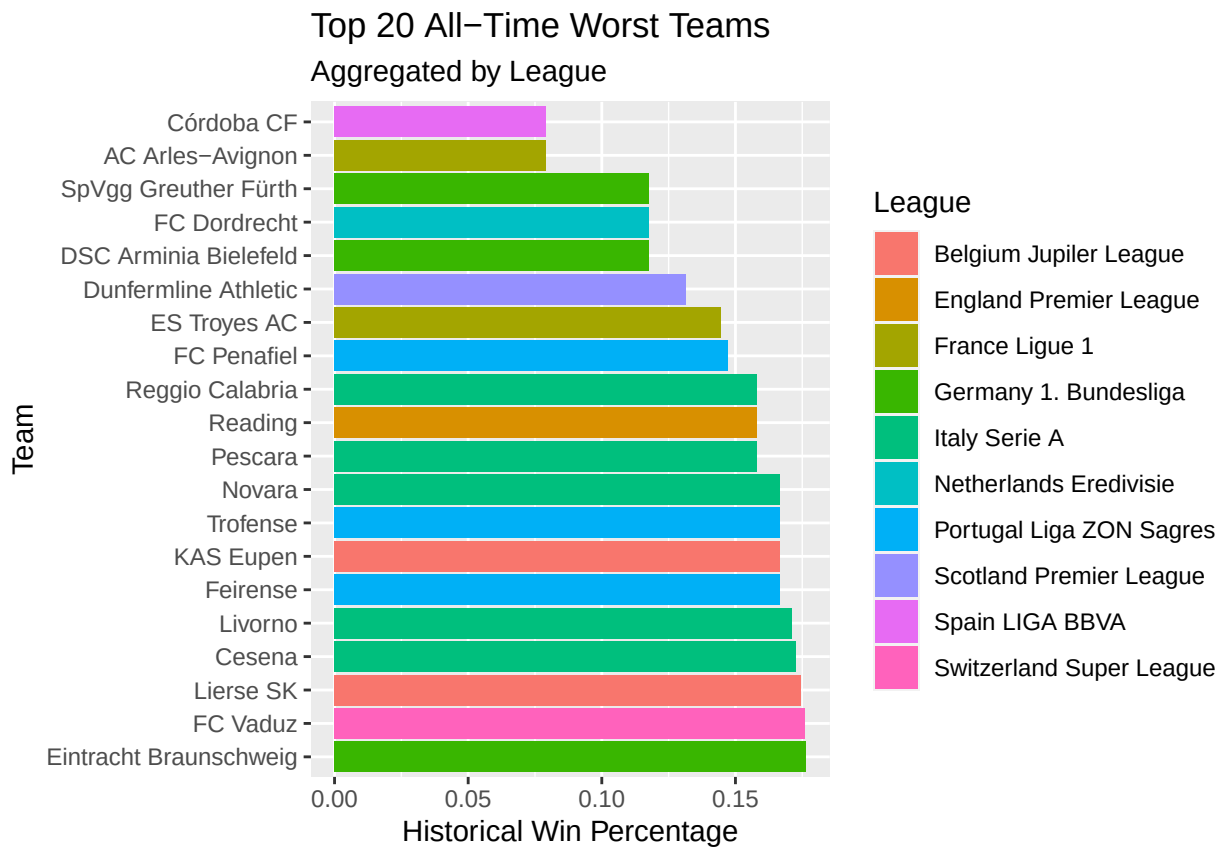
The source of the data originates from:

- http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events
- http://www.football-data.co.uk/ : betting odds.
- http://sofifa.com/ : players and teams attributes from EA Sports FIFA games.

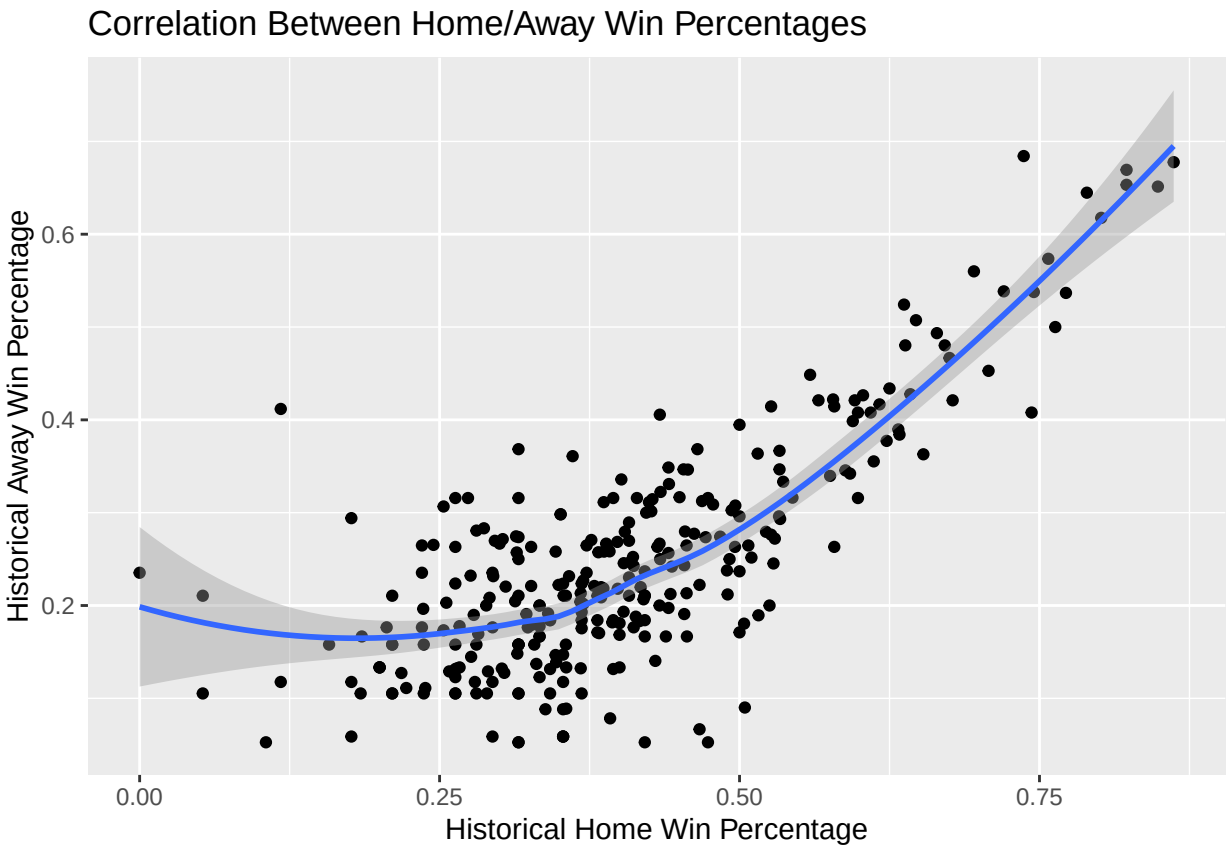
This can be found at https://www.kaggle.com/hugomathien/soccer where these are all compiled into one large dataset.

```
## Joining, by = "country_id"
## `summarise()` has grouped output by 'home_team'. You can override using the `.groups` argument.
## `summarise()` has grouped output by 'away_team'. You can override using the `.groups` argument.
## `summarise()` has grouped output by 'team'. You can override using the `.groups` argument.
```






```
## `geom_smooth()` using formula 'y ~ x'
```



Summary Statstics and Visualizations

Here shows some visualizations of the data, the first showing the top 20 teams in terms of win percentage and the second showing the bottom 20 teams in terms of win percentage. This should give us a good idea of what to expect out of these teams by the end of the final project.

This means that the best team, FC Barcelona, should be predicted to do very well and that the worst team, Cordoba CF, should be predicted to do very poorly. If our model does not reflect this, it will be obvious that we have made an error.

In addition to this, we thought it would be interesting to show the correlation between home winning percent and away winning percent. As expected, we see a largely linear line of best fit. This means that a team that wins a lot of home games is likely to also win a lot of away games. However, there were some irregularities at low winning percentage caused by sparse observations and outliers.

Data Description

```
Data_Description<-read_excel('Data Description.xlsx')
print(Data_Description, n=Inf)
```

```
## # A tibble: 37 x 4
##    `Variable Name`   Description        `Possible Values` `Interpretation (If a~
##    <chr>             <chr>              <chr>             <chr>
##  1 id                ID that represent~ Positive Integers <NA>
##  2 team_fifa_api_id  ID that represent~ Positive Integers <NA>
##  3 away_team_api_id  ID that represent~ Positive integers <NA>
##  4 buildUpPlaySpeed  Represents  the s~ Positive Integer~ Higher number implies~
##  5 buildUpPlaySpeed~ Brackets based of~ Slow, Fast, Bala~ <NA>
##  6 buildUpPlayDribb~ Represents the dr~ Positive Integer~ Higher number implies~
##  7 buildUpPlayDribb~ Brackets based of~ Little, Normal, ~ <NA>
##  8 buildUpPlayPassi~ Representation of~ Positive Integers Higher number implies~
##  9 buildUpPlayPassi~ Brackets based of~ Short, Mixed, Lo~ <NA>
## 10 buildUpPlayPosit~ Represents how th~ Organised, Free ~ Organized implies the~
## 11 chanceCreationPa~ Represents the ab~ Positive Integer~ Higher number implies~
## 12 chanceCreationPa~ Brackets based of~ Safe, Normal, Ri~ <NA>
## 13 chanceCreationCr~ Represents the ab~ Positive Integer~ Higher number implies~
## 14 chanceCreationCr~ Brackets based of~ Little, Normal, ~ <NA>
## 15 chanceCreationSh~ Represents the ab~ Positive Integer~ Higher number implies~
## 16 chanceCreationSh~ Brackets based of~ Little, Normal, ~ <NA>
## 17 chanceCreationPo~ Represents how th~ Organised, Free ~ Organized implies the~
## 18 defencePressure   Represents the pr~ Positive Integer~ Higher number implies~
## 19 defencePressureC~ Brackets based of~ High, Medium, De~ High means full press~
## 20 defenceAggression Represents how ag~ Positive Integer~ Higher number implies~
## 21 defenceAggressio~ Brackets based of~ Press, Double, C~ Press implies more ag~
## 22 defenceTeamWidth  Represents how wi~ Positive Integer~ Higher number implies~
## 23 defenceTeamWidth~ Brackets based of~ Narrow, Normal, ~ <NA>
## 24 defenceDefenderL~ Represents playst~ Cover, Offside T~ Cover is the standard~
## 25 season            Shows the soccer ~ 20XX/20YY         <NA>
## 26 match_id          ID representing t~ Positive Integers <NA>
## 27 league_id         ID representing t~ Positive Integers <NA>
## 28 date              Shows the date of~ Year-Month-Day    <NA>
## 29 home_team_api_id  ID representing t~ Positive Integers <NA>
## 30 home_team_goal    Number of goals t~ Positive Integers <NA>
## 31 away_team_goal    Number of goals t~ Positive Integers <NA>
## 32 match_score       Outcome of game f~ Win, Tie, Loss    <NA>
## 33 home_team_name    Full name of the ~ Characters        <NA>
## 34 away_team_name    Full name of the ~ Characters        <NA>
## 35 home_team_name_S  Shortened name of~ Three letter cha~ <NA>
## 36 away_team_name_S  Shortened name of~ Three letter cha~ <NA>
## 37 country           Country in which ~ Characters        <NA>
```

- This was a xlsx that we created in excel that had a description of all of the variables that was made from scratch.
- For clarity we also attached the xlsx document to our submission

---

In order to help with some of the code for data cleaning, we referenced:

- https://www.kaggle.com/abharg16/predicting-epl-team-season-win-percentages/data