Final Team Project

Daisy Asiimwe, Eleni L., and Emi Torres-Vera

2020-06-04

Contents

Introduction **1**

Data **2**

Primary Write-up **5**

Key Outcomes . . . 6
Summary statistics and visualizations . . . 6
Here we have the prediction model, for the final standing of the English Premier League season 2019-2020. . . . 8
Here you can see the prediction results we found as a conclusion: . . . 13
Summarized Results . . . 13
Critique . . . 13
Learning . . . 14
Conclusion and Discussion . . . 14
Resources Used . . . 14

Introduction

The *English Premier League* is the top league of England's football, with **20 teams** fighting to be crowned the English champions. Premier League was founded in 1992, and it became widely known around the world due to its strong financial support and global broadcasting (Premier League, 2020). One season in The *English Premier League* begins in August and ends at the end of May. Each team plays with each other

on home and away games. For each game, each team gets awarded 3 points for a win, 1 point for a draw, and no points for a defeat. Manchester United has won the title 13 times, Chelsea has won it 5 times, and Manchester City has won it 4 times so far. The top 4 teams of the season are promoted to compete in the Champions League (Premier League, 2020). The next 3 are competing for the participation in the Europa League competition, and the last 3 teams are moved to relegation.

In our data science project, we will analyze data from the last 2 seasons, one of which has not been completed yet. The 2 seasons are 2018-2019 and 2019-2020 season. Our idea is to work on predicting the standings of the clubs at the end of the season 2019-2020 for the *English Premier League*. Some of the research ideas we are going to include are, which team is most likely to win the current season of the *Premier League*, and which are most likely to be the top 5 teams in the standing. In addition to that, we are going to closely study whether scoring the most goals get titles for the team and if there exists a home advantage for the games played on the home field.

Our team chose to work on a topic about The English Premier League because football is a sport that we have invested personal interest in, as we enjoyed playing and watching football. Additionally, due to COVID-19, like other activities, the *English Premier League* 2019-2020 season has been put on hold which has left football fans suspenseful. As enthusiastic football fans, we would like to use the knowledge that we have acquired in the data science class to predict outcomes that would have happened in the 2019-2020 season of *The English Premier League* if there had not been a global pandemic of the COVID-19.

Data

This is where we imported the data. We imported 5 files. 4 of these files were datasets which helped us decide the outcome of the 2020 season. We had data from the 2018, 2019, 2020 premier league season. We also had a dataset for games left for the top 10 teams in 2020. For these datasets we had to clean the data. We had datasets from these seasons that had missing information and some very important information for predicting the winner of the 2020 season. We had to delete the uneccesary columns and keep the important data. We had missing information that was very key for finding our final answers such as goal differentials, total points, and chances missed and much more. We found the information needed and made a new spreadsheet and filled in the data to meet our needs and this was how we cleaned the data to get what we needed.

```
# Add needed libraries, import all needed data, and show glimspses of all data here
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```
library(readxl)
EPL_18<-read_excel("2018 Top 10.xlsx")
EPL_18$GD <- c(EPL_18$Goals-EPL_18$GC)
EPL_19<-read_excel("2019 Top 10.xlsx")

EPL_20<-read_excel("2020 Top 10.xlsx")
EPL_20$GD <- c(EPL_20$Goals-EPL_20$GC)
EPL_20$Results <- c(EPL_20$PTS)

Games_left <- read_excel("Games Left Top 10.xlsx")
```

EPL_18 dataset is containing the table from the 2018 season. This season Manchester City won with 100 points. No other team seems to be close to the champions of this season. This table shows the stats from the top 10 finishing teams from the 2018 season.

```
head(EPL_18)
```

```
## # A tibble: 6 x 20
##   Team  Matches   PTS  Wins  Draw Losses Touches    YC    RC Goals Passes Shots
##   <chr>   <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 Manc~      38   100    32     4      2   35130    59     2   106  28241   665
## 2 Manc~      38    81    25     6      7   27525    64     1    68  20064   512
## 3 Tott~      38    77    23     8      7   29412    50     2    74  21660   623
## 4 Live~      38    75    21    12      5   30324    44     1    84  22962   638
## 5 Chel~      38    70    21     7     10   28728    42     4    62  21264   606
## 6 Arse~      38    63    19     6     13   30635    57     2    74  23524   594
## # ... with 8 more variables: chances_missed <dbl>, Tackles <dbl>,
## #   Offsides <dbl>, Dispossessed <dbl>, CS <dbl>, Saves <dbl>, GC <dbl>,
## #   GD <dbl>
```

EPL_19 dataset is showing once again that Manchester City finihsed with 98 points in the season which made them champions. Liverpool was only 1 point behind the champions meaning the title was between them two all season.

```
head(EPL_19)
```

```
## # A tibble: 6 x 22
##   Team    PTS   POS Matches   Won  Draw  Lost Goals    GC    GD Points    YC
##   <chr> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 Manc~    98     1      38    32     2     4    95    23    72     98    44
## 2 Live~    97     2      38    30     7     1    89    22    67     97    38
## 3 Chel~    72     3      38    21     9     8    63    39    24     72    49
## 4 Tott~    71     4      38    23     2    13    67    39    28     71    56
## 5 Arse~    70     5      38    21     7    10    73    51    22     70    72
## 6 Manc~    66     6      38    19     9    10    65    54    11     66    73
## # ... with 10 more variables: RC <dbl>, Touches <dbl>, Shots <dbl>,
## #   Offsides <dbl>, chances_missed <dbl>, Tackles <dbl>, Clearances <dbl>,
## #   Dispossessed <dbl>, CS <dbl>, Saves <dbl>
```

EPL_20 dataset is showing the current standings of the 2020 season. This season was supposed to be over a while ago but due to Covid-19 the season was delayed and will begin June 17th. This table shows the amount of games each team has played and their stats. Liverpool is on top with 82 points with a big gap between second place team Manchester City with 57 points. We are using this tables stats to compare them to the past seasons to predict the outcome for this table.

```
head(EPL_20)
```

```
## # A tibble: 6 x 20
##   team     GP   PTS  Wins Losses Touches    YC    RC Goals Passes Shots
##   <chr> <dbl> <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 Live~    29    82    27      1   23560    26     1    66  18043   452
## 2 Manc~    28    57    18      7   24172    51     3    68  18926   541
## 3 Leic~    29    53    16      8   20974    31     1    58  14973   405
## 4 Chel~    29    48    14      9   23523    53     0    51  17458   481
## 5 Manc~    29    45    12      8   20360    56     0    44  14596   429
## 6 Wolv~    29    43    10      6   18207    45     2    42  12496   372
## # ... with 9 more variables: chances_missed <dbl>, Tackles <dbl>,
```

```
## #   Offsides <dbl>, Dispossessed <dbl>, CS <dbl>, Saves <dbl>, GC <dbl>,
## #   GD <dbl>, Results <dbl>
```

Explanation of Abbriviations

- PTS means points
- YC means Yellow cards (Discipline actions for a misconduct)
- RC means Red cards (Discipline actions for a misconduct)
- CS means clean sheets (How many games the team has won without conceding any goal)
- GC means goals conceded in total
- GD means the goal difference (It is the difference between the goals scored and the goals conceded)

Here are the types of data we are using for the season 2017-2018:

```
glimpse(EPL_18, width = 3)
```

```
## Rows: 10
## Columns: 20
## $ Team           <chr> ...
## $ Matches        <dbl> ...
## $ PTS            <dbl> ...
## $ Wins           <dbl> ...
## $ Draw           <dbl> ...
## $ Losses         <dbl> ...
## $ Touches        <dbl> ...
## $ YC             <dbl> ...
## $ RC             <dbl> ...
## $ Goals          <dbl> ...
## $ Passes         <dbl> ...
## $ Shots          <dbl> ...
## $ chances_missed <dbl> ...
## $ Tackles        <dbl> ...
## $ Offsides       <dbl> ...
## $ Dispossessed   <dbl> ...
## $ CS             <dbl> ...
## $ Saves          <dbl> ...
## $ GC             <dbl> ...
## $ GD             <dbl> ...
```

Here are the types of data we are using for the season 2018-2019:

```
glimpse(EPL_19, width = 3)
```

```
## Rows: 10
## Columns: 22
## $ Team           <chr> ...
## $ PTS            <dbl> ...
## $ POS            <dbl> ...
## $ Matches        <dbl> ...
## $ Won            <dbl> ...
## $ Draw           <dbl> ...
## $ Lost           <dbl> ...
```

```
## $ Goals          <dbl> ...
## $ GC             <dbl> ...
## $ GD             <dbl> ...
## $ Points         <dbl> ...
## $ YC             <dbl> ...
## $ RC             <dbl> ...
## $ Touches        <dbl> ...
## $ Shots          <dbl> ...
## $ Offsides       <dbl> ...
## $ chances_missed <dbl> ...
## $ Tackles        <dbl> ...
## $ Clearances     <dbl> ...
## $ Dispossessed   <dbl> ...
## $ CS             <dbl> ...
## $ Saves          <dbl> ...
```

Here are the types of data we are using for the season 2019-2020:

```
glimpse(EPL_20, width = 3)
```

```
## Rows: 10
## Columns: 20
## $ team           <chr> ...
## $ GP             <dbl> ...
## $ PTS            <dbl> ...
## $ Wins           <dbl> ...
## $ Losses         <dbl> ...
## $ Touches        <dbl> ...
## $ YC             <dbl> ...
## $ RC             <dbl> ...
## $ Goals          <dbl> ...
## $ Passes         <dbl> ...
## $ Shots          <dbl> ...
## $ chances_missed <dbl> ...
## $ Tackles        <dbl> ...
## $ Offsides       <dbl> ...
## $ Dispossessed   <dbl> ...
## $ CS             <dbl> ...
## $ Saves          <dbl> ...
## $ GC             <dbl> ...
## $ GD             <dbl> ...
## $ Results        <dbl> ...
```

Primary Write-up

We are mainly using the teams past data in order to predict the team's performance. We used the datasets available for seasons 2017-2018, 2018-2019, and 2019-2020 compiled into CSV files.

- For season 2017-2018 we got the data from Kaggle.com (Nalla, 2018).
- For season 2018-2019 we got the data from Premier League website (JohnDoe, 2019).
- For season 2019-2020 we got the data from Kaggle.com (Ido, 2020).

The next step is for us to discover what factors show that a team is going to win the title.

Key Outcomes

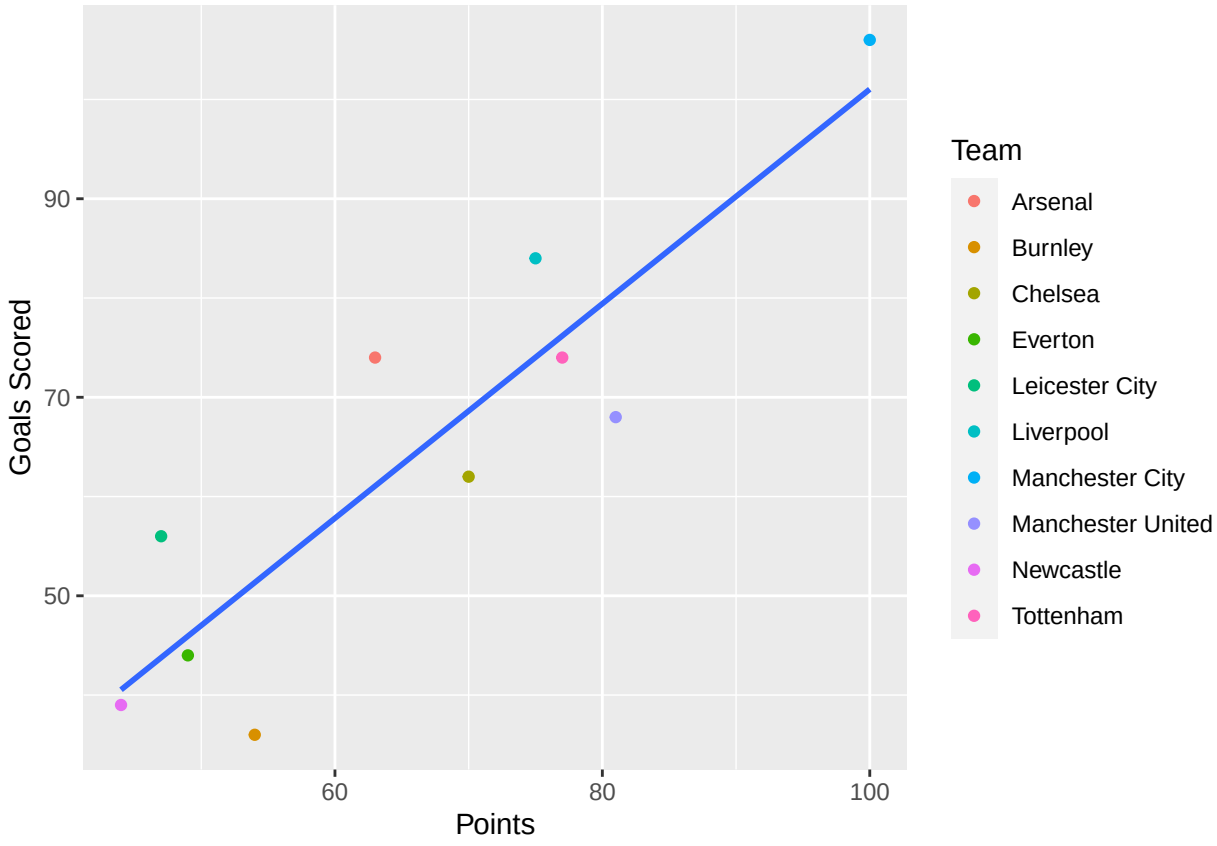
Summary statistics and visualizations

For Season 2017-2018

The graph shows a positive correlation between the goals scored and the points earned by the top ten football teams in the European Premier League in the 2017- 2018 season. As illustrated by the graph, the higher the number of goals scored by the team, the more points earned by that team. For example, the graph shows that Manchester City had the highest number of goals scored and points earned compared to Burnley which had the lowest number of goals scored and points earned in 2017-2018 Premier League Season.

```
# Add these sections as needed, giving them unique names
ggplot(data=EPL_18, aes(x=PTS, y= Goals))+
  geom_point(mapping=aes(color= Team),stat = "identity") +
  xlab("Points") +
  ylab("Goals Scored")+
  geom_smooth(method = "lm", se = FALSE)
```

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

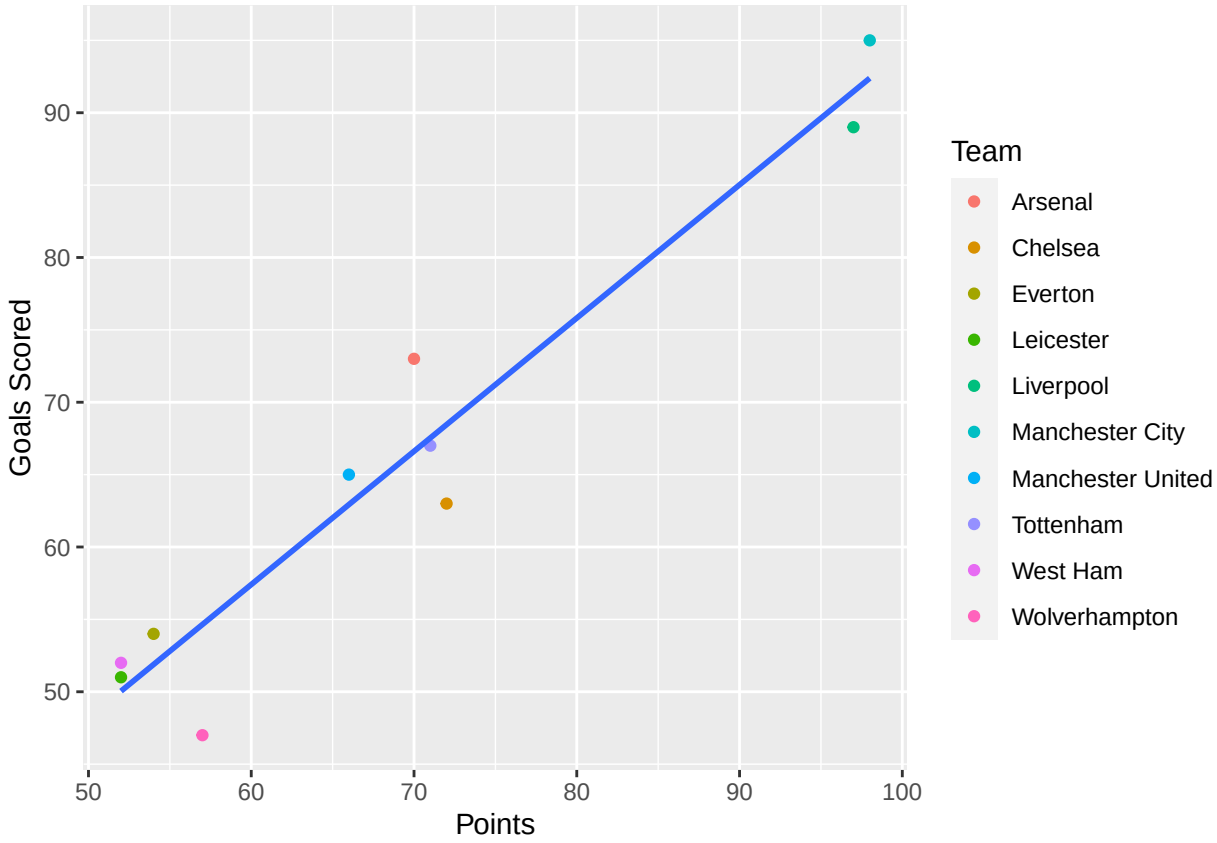


For Season 2018-2019

The graph shows a positive correlation between the goals scored and the points earned by the top ten football teams in the European Premier League in the 2018- 2019 season. As illustrated by the graph, the higher the number of goals scored by the team, the more points earned by that team. For example, the graph shows that Manchester City had the highest number of goals scored and points earned compared to Wolverhampton which had the lowest number of goals scored and points earned in 2018-2019 Premier League Season.

```
# Add these sections as needed, giving them unique names
ggplot(data=EPL_19, aes(x=PTS, y= Goals))+
  geom_point(mapping= aes(color= Team), stat = "identity") +
  xlab("Points") +
  ylab("Goals Scored")+
  geom_smooth(method = "lm", se = FALSE)
```

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

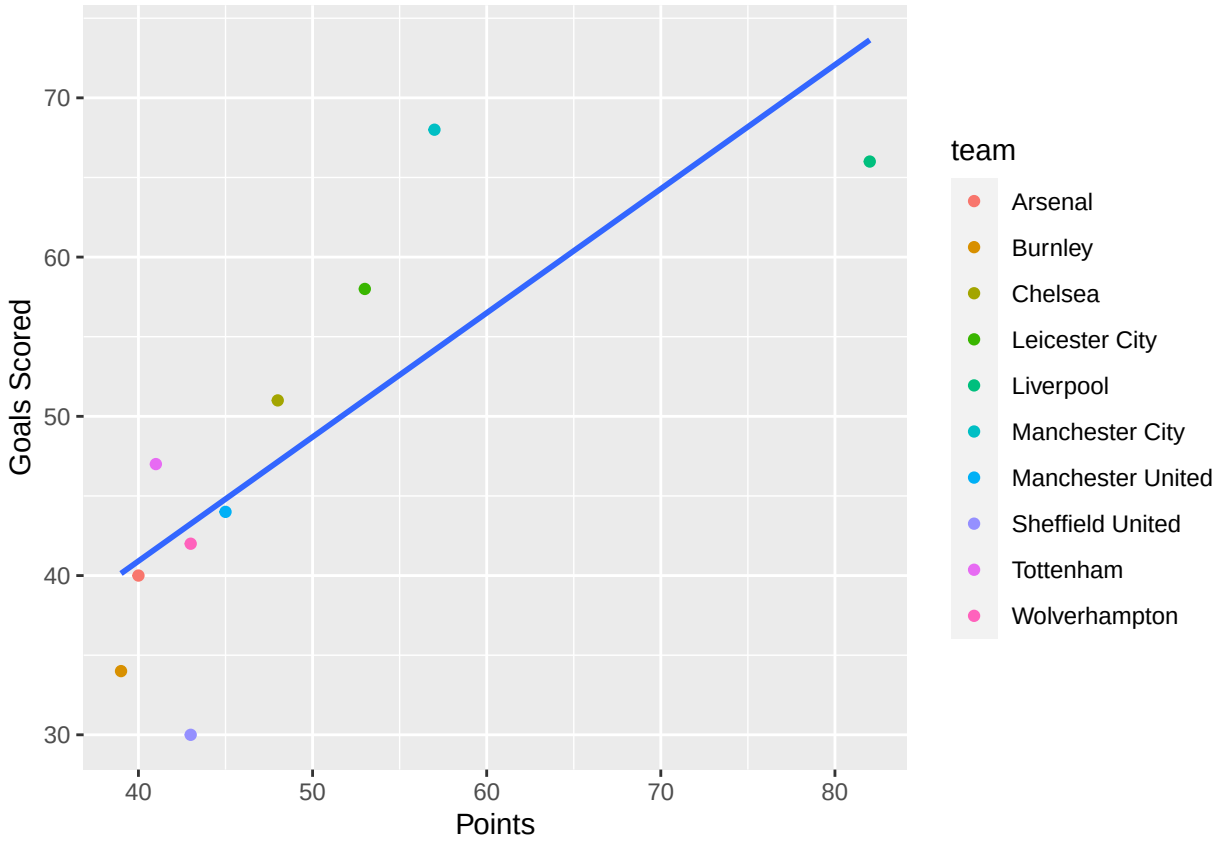


For Season 2019-2020

The graph shows there will be a positive correlation between the goals scored and the points earned by the top ten football teams in the European Premier League in the 2019- 2020 season. As illustrated by the graph, the higher the number of goals scored by the team, the more points earned by that team. For example, the graph shows that Manchester City will have the highest number of goals scored and points earned compared to Sheffield United which will have the lowest number of goals scored and points earned in the 2019-2020 Premier League Season.The model demonstrated by the graph shows that 2019-2020 season ranking will be Manchester city, Liverpool, Leicester city, Chelsea, Tottenham, Manchester united, Wolverhampton, Arsenal, Burnley and Sheffield united respectively.

```r
# Add these sections as needed, giving them unique names
ggplot(data=EPL_20, aes(x=PTS, y= Goals))+
  geom_point(mapping= aes(color= team), stat = "identity") +
  xlab("Points") +
  ylab("Goals Scored")+
  geom_smooth(method = "lm", se = FALSE)
```

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



Here we have the prediction model, for the final standing of the English Premier League season 2019-2020.

First, we assigned the values of the average goals scored for each of the team at their Home and on an Away field. Then, we worked on how many goals were conceded by each team either Home or Away.

The current standing of the teams goes as following, and we are going to number the teams as their current place (Apaar, 2019). The data collected are from two websites, premierleague.com and footystats.org (FootyStats, 2020) & (Premier League, 2020).

1. Liverpool
2. Manchester City
3. Leicester City
4. Chelsea
5. Manchester United
6. Wolverhampton
7. Sheffield United
8. Tottenham

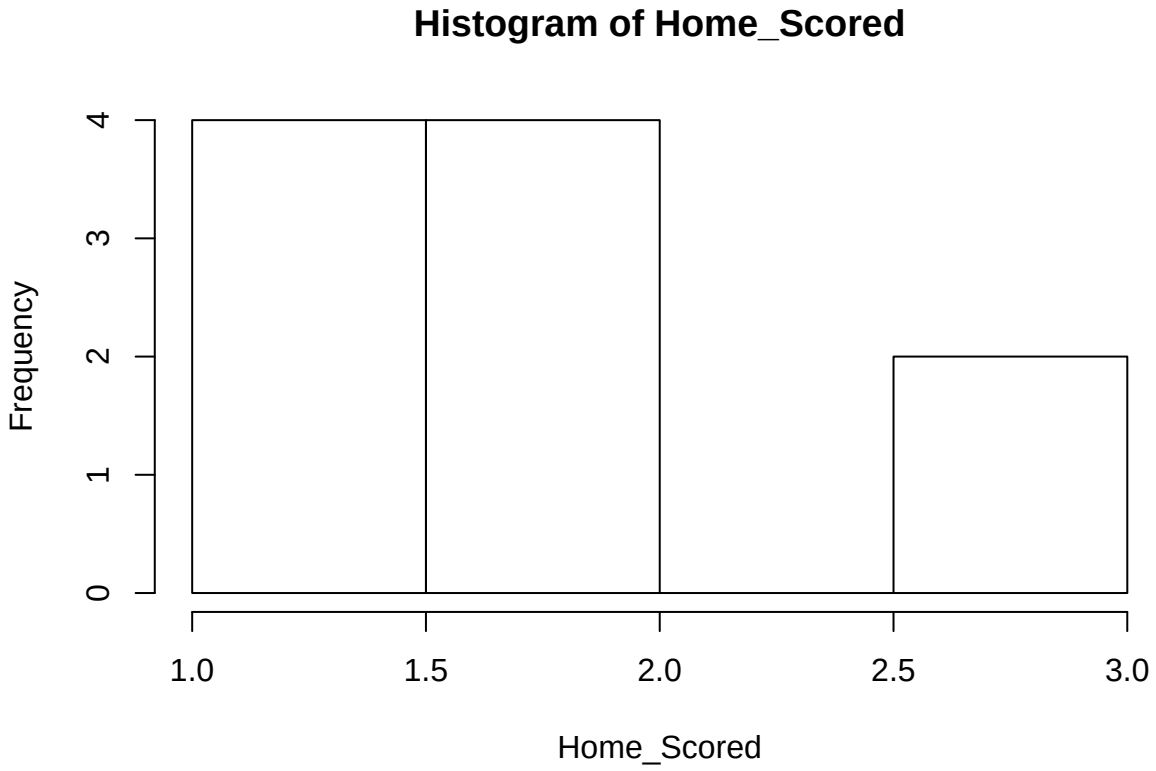
9. Arsenal
10. Burnley

```
# Add these sections as needed, giving them unique names
#The order goes as follow Liverpool (1), Man City (2), Leicester City (3), Chelsea (4), Man U(5), Wolve

#Home Score Average number of goals scored in home matches
Home_Scored <- c(2.67, 2.54, 2.00, 1.47,1.93, 1.40, 1.13,1.93,1.73, 1.33)
#Home Conceded Average number of goals conceded in home matches
Home_Conceded <- c(0.80, 0.92, 1.00, 1.00, 0.80,1.13, 0.87,1.07, 1.33, 1.27)
#Away Score Average number of goals scored in away matches
Away_Scored <- c(1.86,2.33, 2.00,2.07,1.07, 1.43,1.00, 1.33, 1.08, 1.00)
#Away Conceded Average number of goals conceded in away matches
Away_Conceded <- c(0.64, 1.27, 0.93,1.71,1.29, 1.21, 0.92,1.67,1.23, 1.50)
```

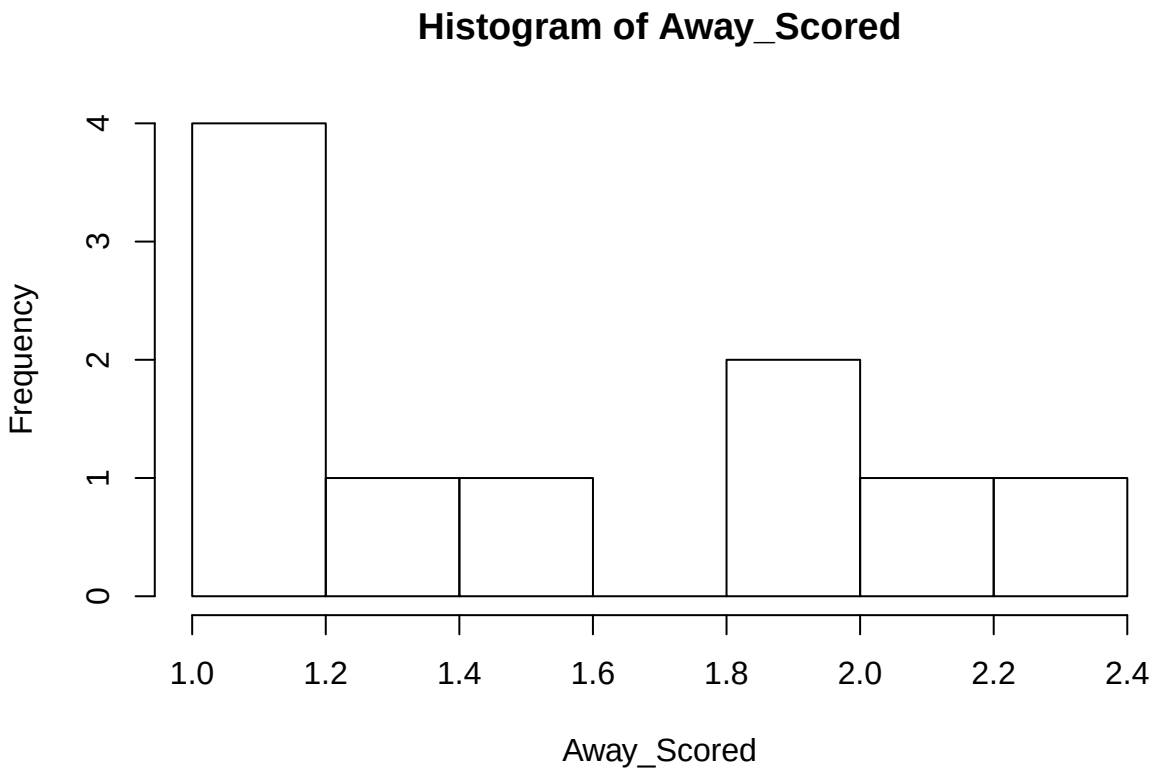
The histogram shows a distribution of how often goals are scored by the home teams in the English Premier League. **Scenario to explain histogram:** If Berea College soccer team played against Barnard College soccer at the Berea College stadium, the histogram demonstrates that Berea College soccer team is more likely to score more goals at home (College stadium).

```
hist(Home_Scored)
```



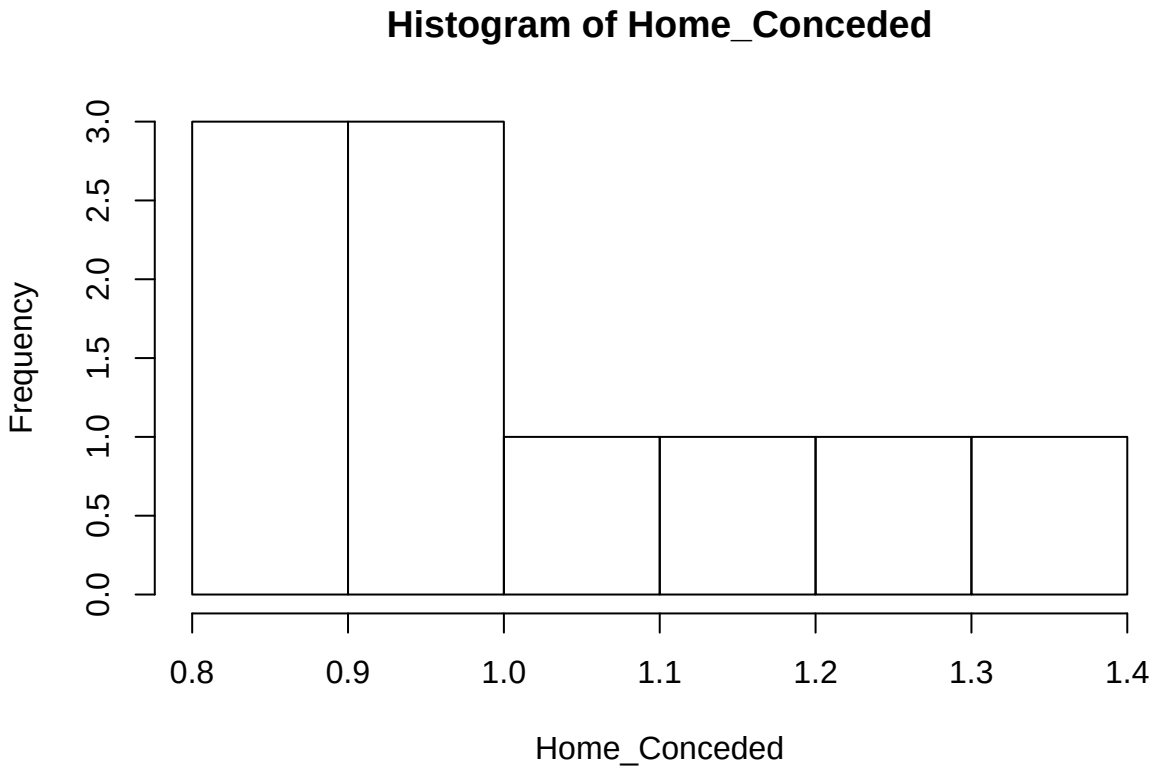
The histogram shows that the distribution of the goals scored away is roughly symmetrical and bimodal. The goals scored away are relatively spread out. The EPL teams playing away from how are less likely to score goals. **Scenario:** If Berea College soccer team plays at the Barnard College soccer stadium, they are less likely to score.

```
hist(Away_Scored)
```



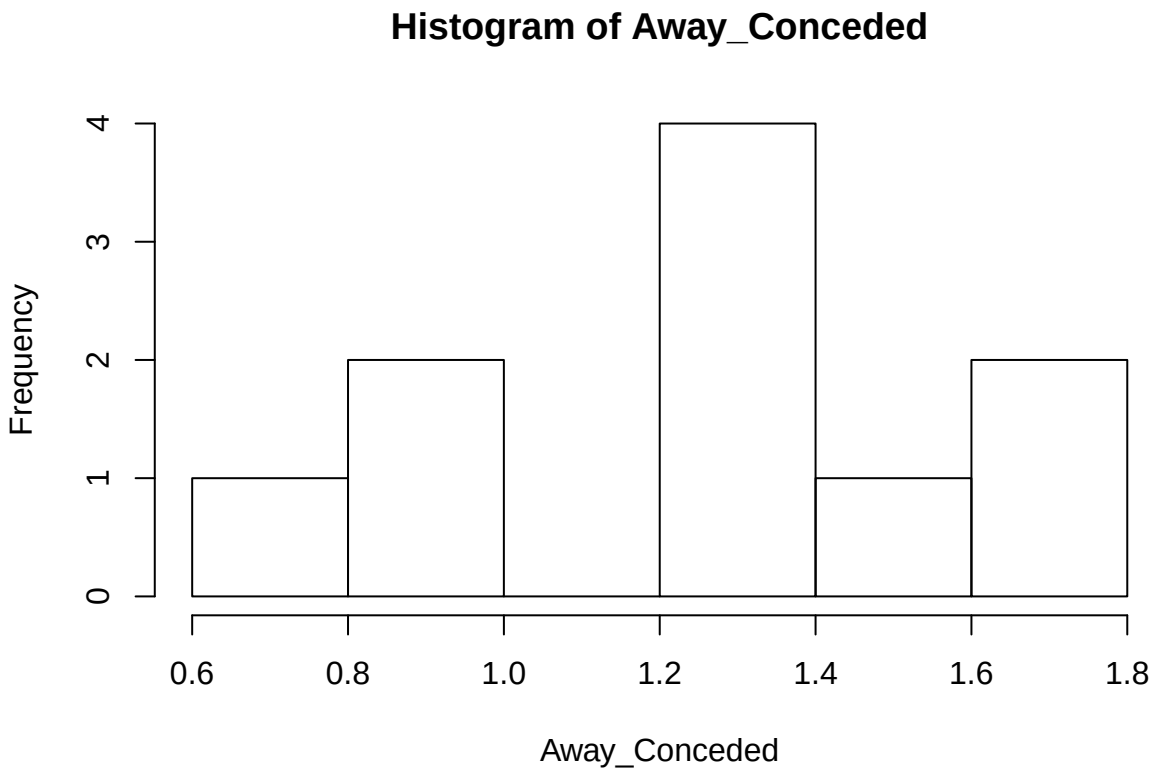
The histogram shows the goals that the away team scored at the home team stadium in the EPL. The distribution shows that the away team is more likely not to score. **Scenario:** If Barnard College soccer is playing at the Berea College stadium, they are less likely to score at the Berea College stadium.

```
hist(Home_Conceded)
```



The histogram shows a distribution that is symmetrical. The distribution shows the likelihood of away teams to score at the home team stadium. **Scenario:** If Berea College was playing at the Barnard College stadium, they are relatively less likely to score at the Barnard College stadium.

```
hist(Away_Conceded)
```



Based on the graphs above we can see a corelation between the goals of a team and its chances of winning the title. Therefore, we have to calculate the mean value of all the scored goals at home, and away. In addition, we calculate the mean of all the conceded goals.

The attack strength shows how strong the team is by making more offensive plays. So, the home attack strength is the ratio of the averaged scored goals of a team in an amount of games to the average scored goals of the complete league in an amount of games. Similarly, defense strength is how strong is the team defensively and we calculate that by dividing the goals conceded by the average of goals conceded of all teh teams in the league (Apaar, 2019).

Then, we calculate the attack strength of each team by dividing the goals scored at home by the general average of the league. We calculate the defense strength of each team by dividing the goals conceded at home by the general average of the league (Apaar, 2019). Lastly, we introduced the variables Overall_Goals_Home which is calculating the general factor of how many goals are scored and conceded on average by each game at the home stadium. Similarly we created the Overall_Goals_Away which is calculating the general factor of how many goals are scored and conceded on average by each game for the away team.

```
League_Home_Scored <- mean(Home_Scored)
League_Away_Scored <- mean(Away_Scored)

League_Home_Conceded <- mean(Home_Conceded)
League_Away_Conceded <- mean(Away_Conceded)

Home_Attack_Strength <- c(Home_Scored/League_Home_Scored)
Home_Defense_Strength <- c(Home_Conceded/League_Home_Conceded)
```

```
Away_Attack_Strength <- c(Away_Scored/League_Away_Scored)
Away_Defense_Strength <- c(Away_Conceded/League_Away_Conceded)

Overall_Goals_Home <- (League_Home_Scored+League_Away_Conceded)/2
Overall_Goals_Away <- (League_Away_Scored+League_Home_Conceded)/2
```

Here we calculate the general probability of a team going to score and we compare this to the opposite team. We used this method based on a project made by Gupta Apaar. The parameter Home and Away are the team number based on the current order. We can now formulate the lambda parameter as follows: (Apaar, 2019)

```
Calculate_lambda <- function(Home, Away){
  LambdaH <- Home_Attack_Strength[Home] * Away_Defense_Strength[Away] * Overall_Goals_Home
  LambdaA <- Home_Attack_Strength[Away] * Away_Defense_Strength[Home]* Overall_Goals_Home
Lambdas <- c(round(LambdaH), round(LambdaA))
return(Lambdas)
}

Calculate_lambda
```

```
## function(Home, Away){
##   LambdaH <- Home_Attack_Strength[Home] * Away_Defense_Strength[Away] * Overall_Goals_Home
##   LambdaA <- Home_Attack_Strength[Away] * Away_Defense_Strength[Home]* Overall_Goals_Home
## Lambdas <- c(round(LambdaH), round(LambdaA))
## return(Lambdas)
## }
```

This function shows the prediction between the two teams. On this block we are finding the score based on the probability of who is going to win. We are getting the results for the two teams based on the previous function. If we could change this project and we had more time we would focus more on the additional statistical analysis. This function takes 3 parameters, Home is the team number which plays as a home game, Away is the team number of the team which plays away, and the score vector is there in order to keep track of the score.

In order to calculate the results we are assigning points for each game, each team gets awarded 3 points for a win, 1 point for a draw, and no points for a defeat. This idea is based on how points are being calculated in the English Premier League.

The home team with the highest probability to win gets 3 points for a win. If the probability is the same both teams are getting 1 point. Lastly, if the away team has the highest probability to win get 3 points. Here we were supposed to calculate the poisson distribution and find the scores of the specific games.

```
Predict_Game <- function(Home, Away, score_vector){
  Lambdas <- Calculate_lambda(Home, Away)
  #Lambda1 would be the score for the home team
  #Lambda2 would be the score for the away team
  Lambda1 <- Lambdas[1]
  Lambda2 <- Lambdas[2]
  if (Lambda1 == Lambda2){ #If there is a tie between the teams
    score_vector[Home] <- score_vector[Home] + 1
    score_vector[Away] <- score_vector[Away] + 1
    }
  else if (Lambda1 > Lambda2){ #If Home wins
    score_vector[Home] <- score_vector[Home] + 3 }
  else { #If Away wins
    score_vector[Away] <- score_vector[Away] + 3 }
  return(score_vector)
```

```
}
```

Here is the r loop that analyzes the remaining games of the current season, and passes the parameters of the team number to the previous two functions. Here i represents the game number and it can go up to 20, since there are only 20 remaining games between the top 10 teams.

```
for(i in (1:20)){
  EPL_20$Results <- Predict_Game(Games_left$Home[i], Games_left$Away[i], EPL_20$Results) }

final_standing <- data.frame(EPL_20$team, EPL_20$Results)
```

Here you can see the prediction results we found as a conclusion:

```
final_standing[rev(order(final_standing$EPL_20.Results)),]
```

```
##          EPL_20.team EPL_20.Results
## 1          Liverpool             94
## 3     Leicester City             63
## 2    Manchester City             63
## 7   Sheffield United             49
## 6      Wolverhampton             49
## 4            Chelsea             49
## 5  Manchester United             47
## 8          Tottenham             44
## 9            Arsenal             42
## 10           Burnley             41
```

Summarized Results

The summarized results we have found are that there is a linear correlation between the goals scored by a team and whether or not a team gets more points. In addition to that, we also have found out that there is actually a home advantage for the teams playing on their home stadium. This is a good thing to see and it makes sense because most teams would prefer to be cheered on by their own fans on their home ground and this will motivate them a little more as the away team gets the opposite feeling. We saw the relation of goals scored by a team and the points a team had. According to the data, the more points a team had would show the team being at the top of table sitting with the most goals scored. This is a good thing to notice becasue it shows the team with the most goals often dominated their opponents through the season thus leading them to the title. The histograms is showing how often a team would score home or away. A team would score more at home games more often than scoring in away games. This makes sense because we have decided that a team playing at home has more of an advantage. The other histogram shows how often a team would get scored on. It shows more teams would get scored on more and more often when being the away team and would get scored on less when playing at home. All these stats and data end up giving us information and help us build our predictions. According to the predictions made; Liverpool ends up winning the 2020 premier league title among these top 10 teams.

Critique

The methods we used in our project are logical because we used variables that are important for understanding how teams get points compared to their goals scored. They might not be very accurate as of the fact that we only used the top 10 teams and these are not the only factors that are influencing the results. However, for the short amount of time that we had this was a project that used our knowledge of vectors, functions,

parameters, and distributions. If we could do this project differently and have more time to work on this. We could compare games played the last 10 years and calculate the possible score, we can also resolve score ties by the goal difference between the two teams.

Learning

Emi- This project went very well between our group. The group would show up when needed and did what they were supposed to. According to the data science process I obtained the the data from a bunch of sources to build our own dataset from. I then cleaned the data by getting rid of the uneccesary stuff and keeping the goof information. I would then figure out what we needed and would find the information and fill the data inside the spreadsheet for each of the datasets we used. I learned a lot of things needed for data cleaning from this. My team also helped me learn a lot about coding and we learned how to communicate with each other and use each others skills.

Eleni - In this project, I learned a lot about data science. More specifically I found a topic that I am really interested in and then I learned how to explore the data set in order to get meaningful feedback. During this process I learned how to find significant patterns and trends that are happening in soccer games already played, allowing me to think about how can we predict the upcoming games. In addition to that, I learned how to construct models to predict results for the future.

Daisy - I learned from the process of the project, from coming up with a topic to producing results using the R programming language. I learned the relevance of teamwork and task redistribution in order to come up with a final project. I also learned that you need to be able to not only write your code but to also be able to communicate your code to help other people understand.

Conclusion and Discussion

In conclusion, we managed to predict the winner of the English Premier League based out of the top 10 teams and each teams remaining games. We defintely struggled to find a starting point at the beginning of this project but, once we took our times and decided what everyone was bringing to the table and set out each other ideas we managed to complete the project.

The findings we got from this project are very significant because from just the scores and the average goals scored and conceded for each team we are able to reach a conclusion. We already knew which teams are most likely to be on the top 3 and so based on our prediction we were proven to be right

Resources Used

Yoely, Ido. (2020). English Premier League stats 2019-2020 [Data set]. Kaggle. https://www.kaggle.com/idoyo92/epl-stats-20192020?select=players_1920.csv

Premier League. (2020, May 26). Retrieved from https://en.wikipedia.org/wiki/Premier_League

Gupta, Apaar. (2019, Jan 30). Predicting Premier league standings - putting that math to some use. Medium. https://towardsdatascience.com/predicting-premier-league-standings-putting-that-math-to-some-use-e8de64938d7

Premier League. (2020, June 01). Football News, Fixtures, Scores & Results. Premier League. https://www.premierleague.com/

Premier League Club Stats - Goals conceded. (2020). Retrieved 2020, from https://www.premierleague.com/stats/top/clubs/goals_conceded

Zaeem, Nalla. (2018, July 12). Results of each match and stats of each team from season 2006/2007 to

2017/2018. Kaggle. https://www.kaggle.com/zaeemnalla/premier-league

JohnDoe. (2019, June 1). 2018-2019 Premier League Data. Kaggle. https://www.kaggle.com/thesiff/premierleague1819

2018-2019 Premier League matches - dataset by chas. (2019, September 22). Retrieved from https://data.world/chas/2018-2019-premier-league-matches/workspace/file?filename=2018-2019.csv

Premier League First Team Club Statistics, Team & Player Stats. (2020). Retrieved from https://www.premierleague.com/stats

FootyStats. (2020). Premier League. FootyStats. https://footystats.org/clubs/england/liverpool-fc#:~:text=On%20average%20Liverpool%20score%202.28,of%2021%20goals%20this%20season.