Final Team Project

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

2020-05-31

Contents

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. 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. 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

```r
# 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()
```

```r
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)

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>
```

```r
head(EPL_19)
```

```
## # A tibble: 6 x 45
##   Team  category total_points general_league_~ finance_live_ga~ `finance _tv_re~
##   <chr> <chr>           <dbl>            <dbl>            <dbl>            <dbl>
## 1 Manc~ Champio~           98                1               26        150986355
## 2 Live~ Champio~           97                2               29        152425146
## 3 Chel~ Champio~           72                3               25        146030216
## 4 Tott~ Champio~           71                4               26        145230801
## 5 Arse~ Europa ~           70                5               25        142193180
## 6 Manc~ Europa ~           66                6               27        142512868
## # ... with 39 more variables: general_matches_played <dbl>, general_won <dbl>,
## #   general_draw <dbl>, general_lost <dbl>, attack_scored <dbl>,
## #   defence_goals_conceeded <dbl>, general_goal_difference <dbl>,
## #   general_points <dbl>, general_squad_size <dbl>,
## #   general_squad_average_age <dbl>, general_squad_foreigners <dbl>, `finance
## #   _team_market` <dbl>, `finance _market_average` <dbl>, attack_passes <dbl>,
## #   attack_passes_through <dbl>, attack_passes_long <dbl>,
## #   attack_passes_back <dbl>, attack_crosses <dbl>, attack_corners_taken <dbl>,
## #   attack_shots <dbl>, attack_shots_on_target <dbl>,
## #   attack_goals_headed <dbl>, attack_goals_penalty <dbl>,
## #   attack_goals_box <dbl>, attack_goals_outsidebox <dbl>,
## #   general_card_yellow <dbl>, general_card_red <dbl>,
## #   attack_goals_counter <dbl>, attack_goals_freekick <dbl>,
## #   defence_saves <dbl>, defence_blocks <dbl>, defence_interceptions <dbl>,
## #   defence_tackles <dbl>, defence_tackles_last_man <dbl>,
## #   defence_clearances <dbl>, defence_clearances_headed <dbl>,
## #   defence_penalty_conceeded <dbl>, attack_posession <dbl>,
## #   attack_pass_accuracy <dbl>
```

```r
head(EPL_20)
```

```
## # A tibble: 6 x 19
##   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 8 more variables: chances_missed <dbl>, Tackles <dbl>,
## #   Offsides <dbl>, Dispossessed <dbl>, CS <dbl>, Saves <dbl>, GC <dbl>,
## #   GD <dbl>
```

```r
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> ...
```

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

```
## Rows: 10
## Columns: 45
## $ Team                         <chr> ...
## $ category                     <chr> ...
## $ total_points                 <dbl> ...
## $ general_league_position      <dbl> ...
## $ finance_live_games_televised <dbl> ...
## $ `finance _tv_revenue`        <dbl> ...
## $ general_matches_played       <dbl> ...
## $ general_won                  <dbl> ...
## $ general_draw                 <dbl> ...
## $ general_lost                 <dbl> ...
## $ attack_scored                <dbl> ...
## $ defence_goals_conceeded      <dbl> ...
## $ general_goal_difference      <dbl> ...
## $ general_points               <dbl> ...
## $ general_squad_size           <dbl> ...
## $ general_squad_average_age    <dbl> ...
## $ general_squad_foreigners     <dbl> ...
## $ `finance _team_market`       <dbl> ...
## $ `finance _market_average`    <dbl> ...
## $ attack_passes                <dbl> ...
## $ attack_passes_through        <dbl> ...
## $ attack_passes_long           <dbl> ...
## $ attack_passes_back           <dbl> ...
## $ attack_crosses               <dbl> ...
## $ attack_corners_taken         <dbl> ...
## $ attack_shots                 <dbl> ...
## $ attack_shots_on_target       <dbl> ...
## $ attack_goals_headed          <dbl> ...
## $ attack_goals_penalty         <dbl> ...
## $ attack_goals_box             <dbl> ...
## $ attack_goals_outsidebox      <dbl> ...
## $ general_card_yellow          <dbl> ...
## $ general_card_red             <dbl> ...
## $ attack_goals_counter         <dbl> ...
## $ attack_goals_freekick        <dbl> ...
## $ defence_saves                <dbl> ...
## $ defence_blocks                <dbl> ...
## $ defence_interceptions         <dbl> ...
## $ defence_tackles               <dbl> ...
## $ defence_tackles_last_man      <dbl> ...
## $ defence_clearances            <dbl> ...
## $ defence_clearances_headed     <dbl> ...
## $ defence_penalty_conceeded     <dbl> ...
## $ attack_posession              <dbl> ...
## $ attack_pass_accuracy          <dbl> ...
```

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

```
## Rows: 10
## Columns: 19
## $ 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> ...
```

Primary Write-up

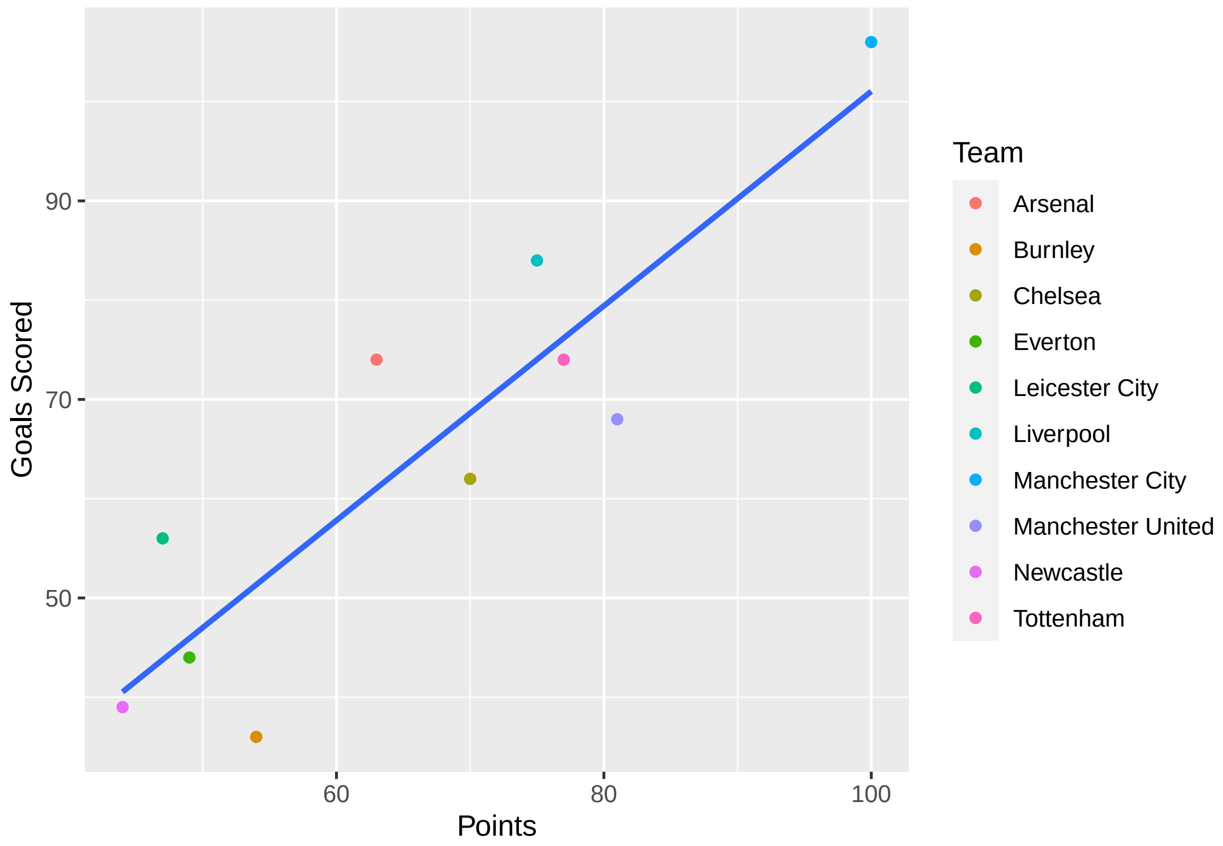
We are mainly using the teams past data in order to predict the team's performance. We used the datasets available from https://www.kaggle.com/thesiff/premierleague1819 and https://www.kaggle.com/idoyo92/epl-stats-20192020, which consists of all the results from 2018-2020, compiled into CSV files. The next step was to discover the dominating factors that interact well with team strength.

Key Outcomes

Summary statistics and visualizations

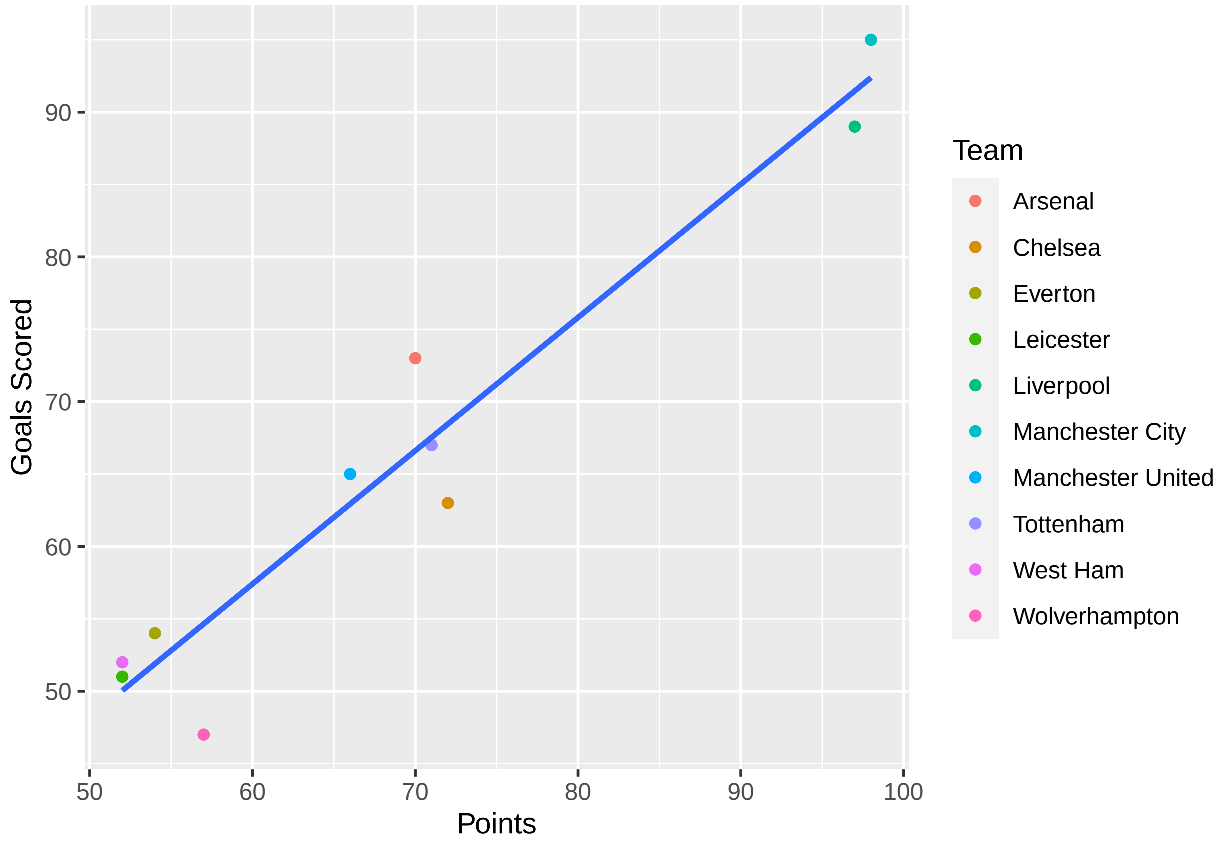
```
# Add these sections as needed, giving them unique names
ggplot(data=EPL_18, aes(x=PTS, y= Goals))+
  geom_point(mapping=aes(color= Team),stat = "identity") +
  xlab("Points") +
  ylab("Goals Scored")+
  geom_smooth(method = "lm", se = FALSE)
```

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



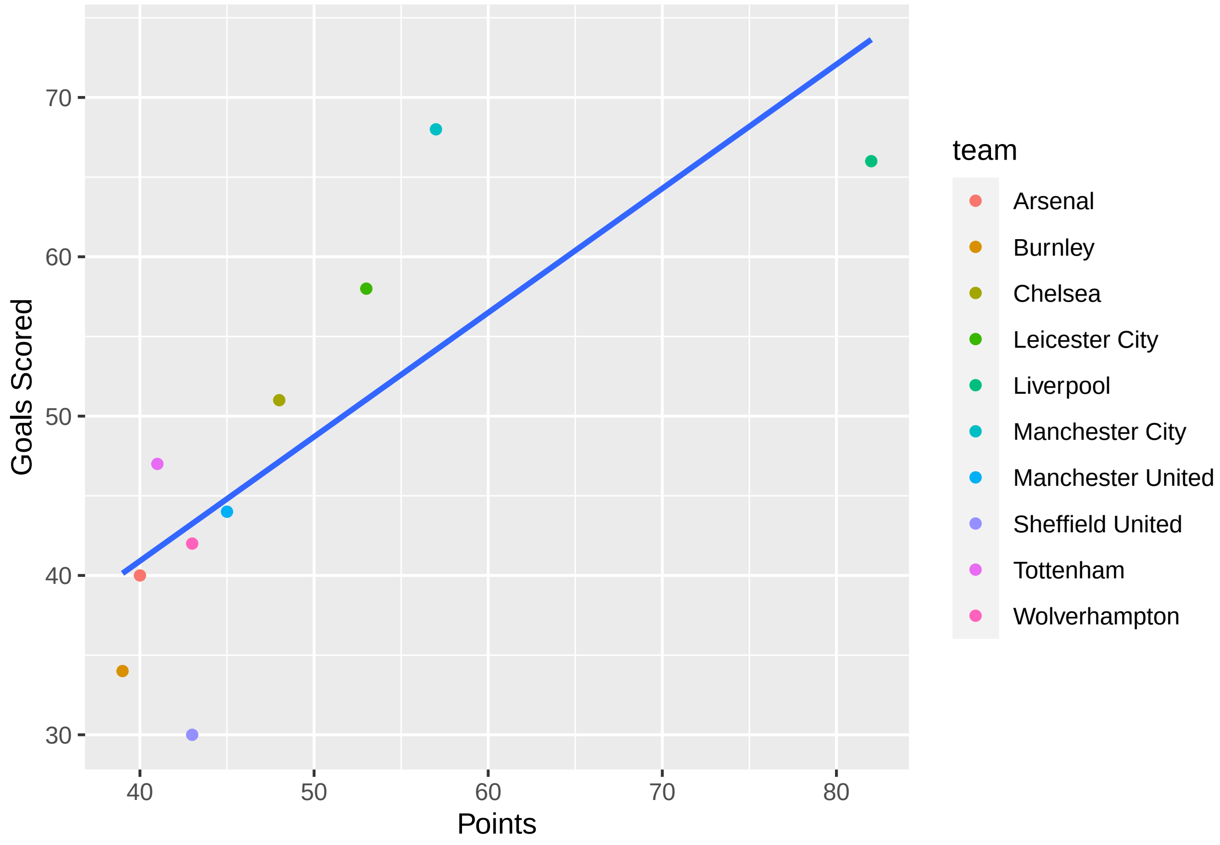
```
# Add these sections as needed, giving them unique names
ggplot(data=EPL_19, aes(x=total_points, y= attack_scored))+
  geom_point(mapping= aes(color= Team), stat = "identity") +
  xlab("Points") +
  ylab("Goals Scored")+
  geom_smooth(method = "lm", se = FALSE)
```

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



```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'
```



## Summarized Results

Critique

Learning

Conclusion and Discussion

Resources Used