Analysis on the English Premier League

Brian Witarsa

Main Questions

1. Which metrics are the best indicators of success?
2. Which team will win the English Premier League according to the best metrics?

Process

1. Extract
2. Filter
3. Merge
4. Identify top metrics
5. Visualize Results
6. Final Prediction

Extract Data

```
[9]  import pandas as pd
     from google.colab import files
```

```
[15] df = pd.read_html('https://fbref.com/en/comps/9/2020-2021/2020-2021-Premier-League-Stats')
```

```
[5]  for idx,table in enumerate(df):
      print('**************************')
      print(idx)
      print(table)
```

```
df[2]
df[2].to_csv('Premier-League-data-21.csv')
files.download('Premier-League-data-21.csv')
```

Filter

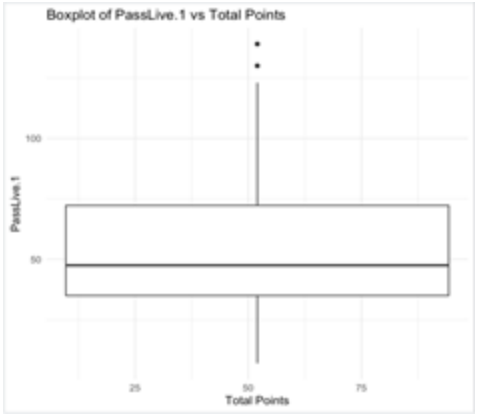
```r
# Reading season data and removing specific columns
season_1819 <- read.csv("~/Downloads/Premier-league-data-18.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
season_1920 <- read.csv("~/Downloads/Premier-league-data-19.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
season_2021 <- read.csv("~/Downloads/Premier-league-data-20.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
season_2122 <- read.csv("~/Downloads/Premier-league-data-21.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
season_2223 <- read.csv("~/Downloads/Premier-league-data-22.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
season_2324 <- read.csv("~/Downloads/Premier-league-data-23.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)

# Read in the Wins, Draws, and Losses data
wl_1819 <- read.csv("~/Downloads/Premier-League-WLdata-18.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]
wl_1920 <- read.csv("~/Downloads/Premier-League-WLdata-19.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]
wl_2021 <- read.csv("~/Downloads/Premier-League-WLdata-20.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]
wl_2122 <- read.csv("~/Downloads/Premier-League-WLdata-21.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]
wl_2223 <- read.csv("~/Downloads/Premier-League-WLdata-22.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]
wl_2324 <- read.csv("~/Downloads/Premier-League-WLdata-23.csv", stringsAsFactors = FALSE)[, c("Squad", "W", "D", "L")]

#defensive actions data
da_1819 <- read.csv("~/Downloads/Premier-league-DAdata-18.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
da_1920 <- read.csv("~/Downloads/Premier-league-DAdata-19.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
da_2021 <- read.csv("~/Downloads/Premier-league-DAdata-20.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
da_2122 <- read.csv("~/Downloads/Premier-league-DAdata-21.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
da_2223 <- read.csv("~/Downloads/Premier-league-DAdata-22.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
da_2324 <- read.csv("~/Downloads/Premier-league-DAdata-23.csv", skip=1, header=TRUE, stringsAsFactors = FALSE)
```

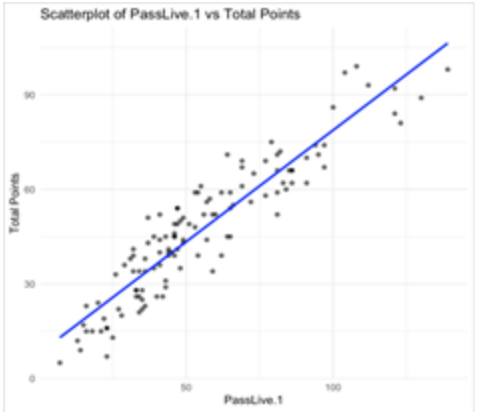
Merge

```r
# Combine all the WL data into one dataframe
wl_data <- rbind(wl_1819, wl_1920, wl_2021, wl_2122, wl_2223, wl_2324)
all_seasons_data <- rbind(season_1819, season_1920, season_2021, season_2122, season_222
all_da_data <- rbind(da_1819, da_1920, da_2021, da_2122, da_2223, da_2324)
all_sca_data <- rbind(sca_1819, sca_1920, sca_2021, sca_2122, sca_2223, sca_2324)


# Merge the WL data with the regular season data by Squad and Season
all_data <- merge(all_seasons_data, wl_data, by=c("Squad", "Season"))
all_data <- merge(all_data, all_da_data, by=c("Squad", "Season"))
all_data <- merge(all_data, all_sca_data, by=c("Squad", "Season"))

all_data$Pts <- all_data$W * 3 + all_data$D
```

Visualizing results

- Correlation analysis
  - Bar plot
  - Boxplots
  - Scatterplots

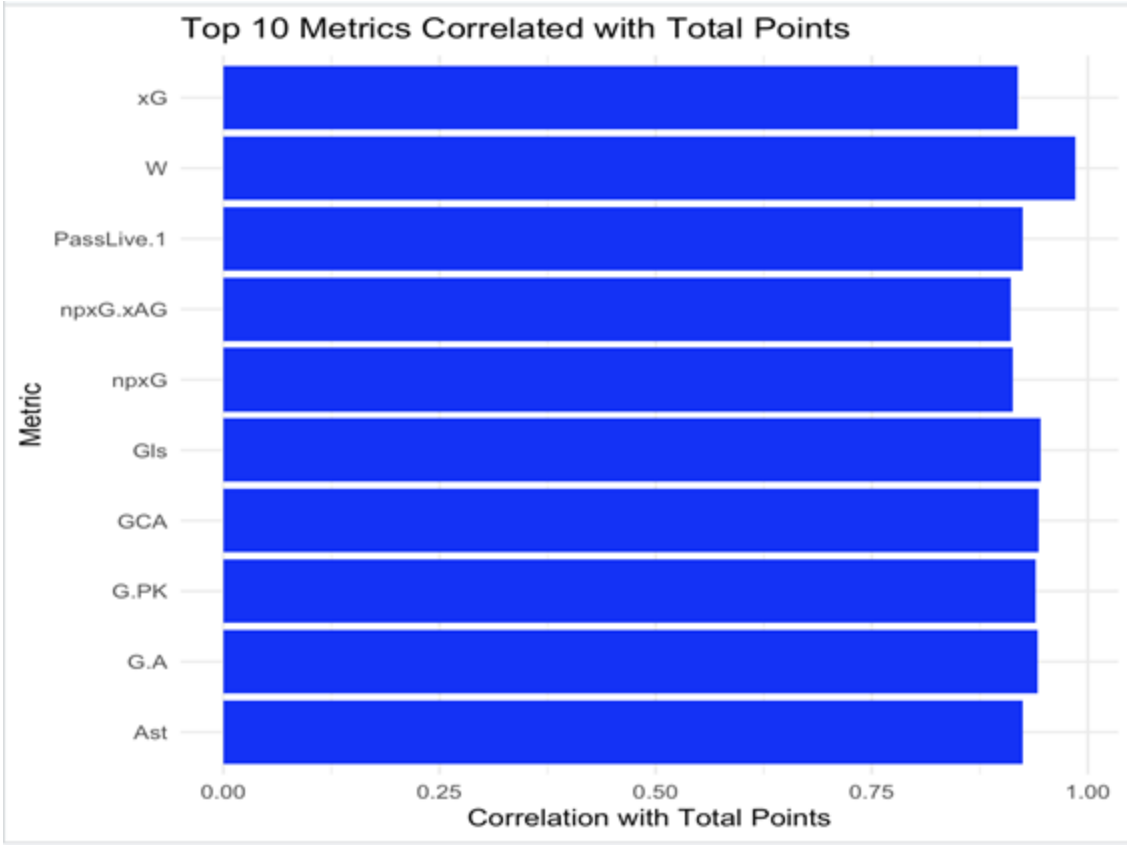




What metrics are the best indicators of success?

1. Wins
2. Goals per 90
3. Goal Creating Actions: Goal Creating Actions (GCA) and Shot Creating Actions (SCA), meaning the two offensive leading to a shot or goal. This includes live-ball passes, dead-ball passes, successful dribbles, shots which lead to another shot, and being fouled.
4. Goals and Assists per 90
5. Non penalty Goals per 90
6. Live Passes: completed live ball passes that led to a shot attempt
7. Assists
8. Expected Goals
9. Non-penalty expected goals
10. Non-penalty expected goals and assists

Best Metrics



Prediction

- Multiple regression model
- Estimate: represents the estimated change in the total points (dependent variable) for a one-unit change in the respective independent variable, with all other variables constant.
- W and npxG. xAG p values are < 0.05

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.9835 -1.9136 -0.3058  1.9096  6.9080

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.09488    0.88730   2.361   0.0200 *
W            2.58212    0.12176  21.206   <2e-16 ***
Gls         -0.33290    0.37703  -0.883   0.3792
GCA          0.17879    0.11371   1.572   0.1187
G.A          0.01007    0.12950   0.078   0.9382
G.PK         0.10770    0.30951   0.348   0.7285
PassLive.1  -0.11293    0.06449  -1.751   0.0827 .
Ast               NA         NA      NA       NA
xG           0.42457    0.33571   1.265   0.2086
npxG         0.58349    0.44382   1.315   0.1913
npxG.xAG    -0.41884    0.19350  -2.165   0.0326 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.954 on 110 degrees of freedom
```

According to the metrics, which team is predicted to win this season?

Past 6 winners of the Premier League	Predicted Winners
18/19: Manchester City	18/19: Manchester City
19/20: *Liverpool	19/20: *Manchester City
20/21: Manchester City	20/21: Manchester City
21/22: *Manchester City	21/22: *Liverpool
22/23: Manchester City	22/23: Manchester City

23/24 Premier League Winners: Manchester City