Joshua Thomas
Google Data Analytics Professional Certificate Capstone Project
Date Started: June 30, 2022.
Date Completed: July 4, 2022
Note: Used https://fbref.com for open source football (soccer) data

What Went Wrong?

A Statistical Analysis of Manchester United's 21/22 Season

- Image of Cristiano Ronaldo showing his frustration during a match against Watford

Introduction

- Manchester United is considered by many as the biggest football club in England and in the world. Having a record 20 Premier League Titles, the Red Devils have solidified themselves not only on the pitch but as a brand in the footballing world.

- However, a lot of this dominance was in the past under the leadership of legendary manager Sir Alex Ferguson. After the great Scotsman retired in 2013, United have struggled, last winning a trophy in 2017 and currently experiencing a 10 year Premier League title drought.

- This is despite the United hierarchy throwing billions of pounds on transfers since Ferguson's retirement. However, after finishing 2nd in the 20/21 season, there was

cautious optimism. Under the care of United legend Ole Gunnar Solksjaer for a few years, United built a young, promising side.

- They had the veteran Spanish shot stopper in net, David De Gea.
- The defense was led by England's main center back Harry Maguire, and left back Luke Shaw. The midfield was led by the Portuguese talisman, Bruno Fernandes.
- The attack was young and dynamic, with Marcus Rashford, Anthony Martial, and Mason Greenwood scoring goals and creating chances.

- Thus, many United fans and football enthusiasts only had optimism for Manchester United's 'resurgence'. This optimism increased significantly in the summer right before the 21/22 season.
- Manchester United splashed the cash, signing a World Cup and Champions League winning defender in Raphael Varane, and a young, talented, English winger in Jadon Sancho.
- But, nothing topped the last signing, the return of Cristiano Ronaldo to Manchester United, a player considered by many to be the greatest of all time.

- The United hierarchy made the moves to send a message to the footballing world: they want to win.
- Shockingly, what ensued in the 21/22 season was quite the opposite. Despite assembling what many consider to be the best United side on paper since Sir Alex retired, United finished 6th in the Premier League, with their lowest points accumulation in modern Premier League history.

- United were knocked out of every cup competition quite early. Furthermore, United sacked manager Ole Gunnar Solskjaer, and replaced him with interim boss Ralf Rangnick, whose tenure was marred by a toxic dressing room and equally terrible football.

- Manchester United's 21/22 season can only be categorized as a disaster. Now, under the new, young, Dutch manager Erik Ten Hag, United want to learn from their mistakes from the past to ensure a bright future.

- Image of a few Manchester United players before the match

Problem Statement

In order to ensure that they won't continue to make mistakes and give Erik Ten Hag and his squad every opportunity to succeed in the future, Manchester United under new CEO Richard Arnold and new football director John Murtough have asked me (this is a hypothetical just for the sake of this project) to analyze 21/22 Premier League data to determine important metrics that affect total points, and then determine the factors which influence those metrics. By identifying these metrics, Manchester United hope to learn the key performance indicators that affect winning in the Premier League, and thus focus on these indicators in the subsequent seasons to come.

- New Manchester United Manager Erik Ten Hag (left) and Football Director John Murtough (right)

Initial Hypothesis and Rationale

Prior to conducting analysis, I practice generating a hypothesis in order to have a clear vision while conducting my analysis.

In quantifiable goals/points based sports, scoring goals/points for your team and preventing goals/points against your team are practices that yield success.

As someone who has played and watched football for many years, in order to score goals for your team, you must create high quality chances. While it is possible to score goals with limited chances- or even no chances if the other team makes a significant error - I believe that in general, the more high quality chances created, the greater the probability one of those chances goes in. Additionally and conversely, in order for a football team to prevent goals against them, I believe that they have to prevent the opposition from getting high quality chances. I believe that there are 2 ways to logically achieve this: 1) defend in a way in which the opposition has a tough time creating high quality chances and/or 2) maintain ball possession to the point where the opposition doesn't have the chance to even create any chances.

While both scoring goals through high chance creation and preventing goals are both important metrics which influence win rate, I believe the former matters more, because in general, through creating a plethora of high quality chances which create many shots on target, the probability that the opposing team creates many high quality chances goes down, due to having less of the ball.

Thus, I believe that in order to get a high number of points which in part will increase the probability of winning in the Premier League, Manchester United need to prioritize high quality chance creation, which will in part increase their probability of scoring goals and thus, increase their probability of accumulating more points in future seasons.

Statistical Tools and Methodologies

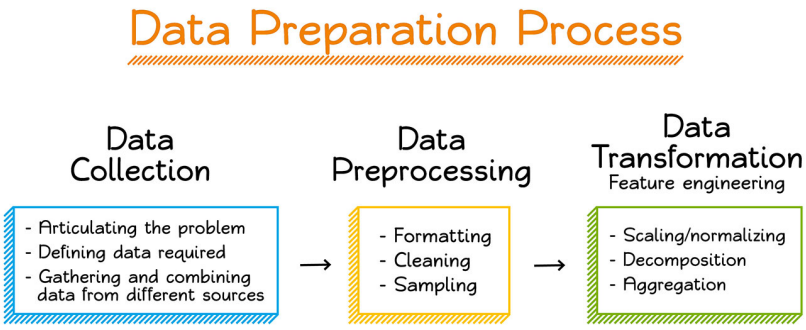
I will be using Spreadsheets and some R Programming for basic data preparation (cleaning and transformation).
I will then use R Programming for Data Exploration and Visualization.

While I would generally use Python for Analytical Programming, since the Google Data Analytics Course taught R, I feel it's appropriate to use that for the Capstone Project (though I will use Python for future projects).

Additionally, regarding the data used in this analysis, I used 21-22 English Premier League data from the site https://fbref.com/en/comps/9/Premier-League-Stats
Their data for the 21-22 English Premier League season, includes the following tables:

League Table
Squad Standard Stats
Squad Goalkeeping
Squad Shooting
Squad Pass Types
Squad Goal and Shot Creation
Squad Defensive Actions
Squad Possession

Data Preparation



My Data Preparation Process

To start the Data Preparation Process of this analysis, I downloaded all the csv files of the data mentioned above from FBRef as Spreadsheets, which totaled to 8 spreadsheets.

Since I plan on taking some data from a majority of the files, and since all the data provided has the same number of rows, I decided to take specific columns necessary and relevant for the hypothesis and analysis from their parent files and paste it into a new spreadsheet for simplification purposes. Then, I will analyze the data for cleaning.

In the subsequent explanations, I will explain which columns I'll be using and the rationale, and my data cleaning process.

Spreadsheet Columns and Rationale

After looking at each of the columns of each table, since my analysis is predicated on chance creation and chance prevention, I used key metrics which I believe are related and associated to chance creation and chance prevention.
For chance creation, I picked the following columns:

- Goals For (GF)
- Expected Goals, which is the probability a shot taken in a certain location results in a goal (xG)
- Expected Assists, which is the probability a pass results in assist (xA)
- Non-penalty Expected Goals (npxG)
- Non-penalty Expected Goals + Expected Assists ('npxG+xA')
- Shots on Target per 90 Minutes (SoTPer90)
- Goals per Shot on Target; i.e Conversion rate (GPerSoT)
- The Average Distance of All Shots Taken By a Team (Dist)
- Total Touches Taken in the Attacking Third (Att3rdTouches)
- Shot Creating Actions per 90 Minutes, which are "the two offensive actions directly leading to a shot, such as passes, dribbles and drawing fouls. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit". (SCA90)
- Goal Creating Actions per 90 Minutes, which are "the two offensive actions directly leading to a goal, such as passes, dribbles and drawing fouls. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit". (GCA90)

I picked these columns for data related to chance creation because all of these metrics are related to number of chances created, the quality of the chances created, where the chances were created, location of ball position, and the effectiveness of those chances.

For chance prevention, I picked the following columns:

- Goals Against (GA)
- Expected Goals Against: the probability a chance from the opposition results in a conceded goal (xGA)
- Clean Sheet Percentage: the number of times a team had no goals conceded in a match (CSPercentage)
- Pressures in the Defensive Third: the amount of times a team applied defensive pressure on the opposition within their own half (Def3rdPressures)

- Pressures in the Attacking Third: the amount of times a team applied defensive pressure on the opposition within the opposition's half (Att3rdPressures)

I picked these columns for chance prevention analysis because these stats indicate the amount of times they completely prevented chances, the probability they are going to concede a chance, and the amount of times they defended in their half throughout the season vs. the opposition half.

I also picked the columns Squad, which is the identifier for teams, Points (Pts) as that is the main focus of the analysis, and Goal Difference (GD), which is the difference between Goals Scored (GF) and Goals Conceded (GA). This metric will be used for correlation in order to determine which of the prior stats correlate strongly with success.

Now, after making this simplified spreadsheet, I analyzed the data in the spreadsheet for potential cleaning. By investigating all the data through using filters and column reorganization, it seems that the FBRef database collected the data quite accurately. There was no missing data as well. Thus, there was no need for cleaning. However, if there was data to be cleaned, I would apply filters and imputation to clean up the data.

The data is now ready for Exploration and Visualization, which I will be doing with the R Programming Language.

Data Exploration and Visualization

Coding and Output

```
# Note: I did my coding work in RStudio.
# Load Packages
> install.packages("tidyverse")
> library(tidyverse)
> library(dplyr)
> library(ggplot2)

# Read data
> data <-
read_csv("/Users/joshuathomas/Downloads/EPL-2122-Capstone.csv")
# Get column names
> colnames(data)
```

```
 [1] "Squad"           "Pts"             "GF"                "GA"
"GD"
 [6] "xG"              "xA"              "npxG"
"npxG+xA"         "xGA"
[11] "SoTPer90"        "GPerSoT"         "Dist"
"Att3rdTouches"   "SCA90"
[16] "GCA90"           "CSPercentage"    "Def3rdPressures"
"Att3rdPressures"
```

Correlation Tests and Visualizations

- Now, before comparing Manchester United to the rest of the league with these key metrics for chance creation and prevention, I first need to determine what columns are the MOST IMPORTANT when it comes to Point Accumulation, as Point Accumulation is how success is quantified in the Premier League.

- This will be done through using correlations in a Pearson Correlation Test.

- I will identify the columns that are strongly correlated with Points, as those columns may be influential when it comes to getting Points.

- After identifying these columns, I will identify other columns which are highly correlated with the most influential columns

- I will provide visualizations of these columns afterwards, comparing Manchester United individually to the other teams.

- For reference, a correlation coefficient (rxy) of $-1 < rxy < -0.8$ or $0.8 < rxy < 1$ indicates a strong correlation, $-0.8 < rxy < -0.5$ or $0.5 < rxy < 0.8$ indicates a moderate correlation, and $-0.5 < rxy < 0.5$ indicates a weak correlation

Most Correlated with Points: Positive and Negative

- After running Pearson Correlation Coefficient Tests between Points and all other columns, the segments of code included in this document indicate the columns of highest importance when it comes to getting Points, as they had the strongest positive and negative linear associations amongst all columns included.

```
# Attach data for easier access to column names
>attach(data)
```

```
# Strong Positive Correlation
# Pts and Gls Pearson Correlation Test
> cor(Pts, GD)
[1] 0.9714088
```

- As expected, there is an extremely strong positive correlation between the number of Points and GD, as a greater positive difference between the number of goals scored and the number of goals conceded is related to an increase in Pts. Also, the columns (highest to lowest) of GF, SoTPer90, GCA90, xG, npxG, 'npxG+xA', SCA90, xA, and Att3rdTouches had strong positive correlations, with a rxy value >= .9

```
# Strong Negative Correlation
# Pts and xGA Pearson Correlation Test
> cor(Pts, GA)
[1] -0.8856865
```

- As expected, there is a strong negative correlation between the number of Goals Allowed and Pts. Additionally, xGA had a strong negative correlation with Points.

*****Interesting Result******

```
# Pearson Correlation Test Between Points and Clean Sheet
Percentage
> cor(Pts, CSPercentage)
[1] 0.09840792
```

- **We know through the previous tests that generally, as Goal Difference (Goals For - Goals Against) increases, Points increases. Additionally, we can see a negative linear association between Points and Goals Against. Therefore, one would think that a team which has a greater number of clean sheets (games where the team doesn't concede) will have more points. But, there is a VERY WEAK, positive correlation between these variables. This is probably because teams with a higher clean sheet percentage played more defensive and thus didn't create as many chances to score goals and thus potentially tie many games as well. Thus, the correlation test indicates that scoring goals is potentially more important than conceding goals to Goal Difference and Point Accumulation.**

Visualization of Key Metrics and Comparison With Manchester United

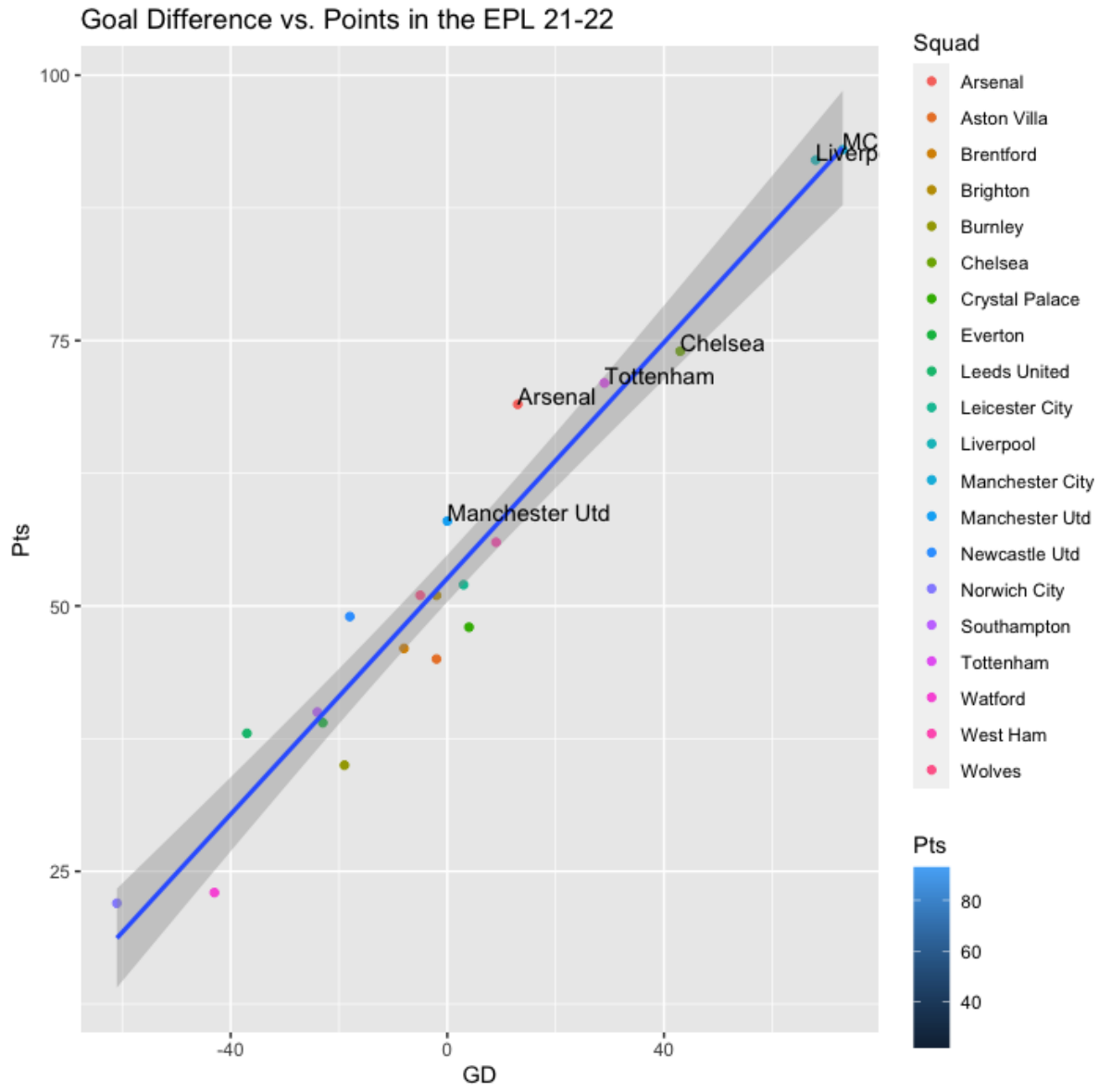
- As shown in our findings, it is clear that the metric of Goal Difference (GD) has the strongest positive relationship with Point Accumulation. Furthermore, the metric of Goals Against has the strongest negative relationship with Points Accumulation. Due to the fact that the correlation between Points and Clean Sheet Percentage was a weak, positive correlation and that Goals For (GF) had a higher absolute value[correlation coefficient]

than Goals Against (.95 vs .88) , we understand that while both Goals For (GF) and Goals Against (GA) matter for Goal Difference (GD), Goals For (GF) matters more when it comes to Point Accumulation than Goals Against (GA). Thus, I'll be visualizing GD, GA, and GF and see how Manchester United's specific metrics compare to the league values.

```
# Now, I will be putting the data table in a dataframe, in order
to get more functionality
> df <- data.frame(data)

# Then I created a dataframe just for Manchester United. This
will be used for comparison in the future.

> df_MUFC <- filter(df, Squad == "Manchester Utd")

# Plot Showing relationship between Pts and GD
> ggplot(data = data) + geom_point(mapping = aes(x = GD, y =
Pts, color = Squad, fill = Pts)) + geom_smooth(aes(GD, Pts),
method = "lm") + labs(title = "Goal Difference vs. Points in the
EPL 21-22") + geom_text(aes(x = GD, y = Pts, label=ifelse(Squad
== "Manchester Utd",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GD, y = Pts, label=ifelse(Squad ==
"Liverpool",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GD, y = Pts, label=ifelse(Squad ==
"Chelsea",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GD, y = Pts, label=ifelse(Squad ==
"Tottenham",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GD, y = Pts, label=ifelse(Squad ==
"Arsenal",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GD, y = Pts, label=ifelse(Squad == "Manchester
City",as.character("MC"),'')),hjust=0,vjust=0)
```

Points vs. GD

- Below is the visualization that follows after executing the code above. I made a scatter plot with a trend line with ggplot2 in R to demonstrate the relationship between Points and GD. I also added labels for Manchester United and the teams that outperformed them in Points in the 21/22 season (Manchester City, Liverpool, Chelsea, Tottenham, and Arsenal)



- As the graph visualizes, there is a strong, positive, linear association between Points and Goal Difference. All the teams that outperformed United (as labeled on the graph) seemed to have a higher GD than United, which makes sense given the relationship.

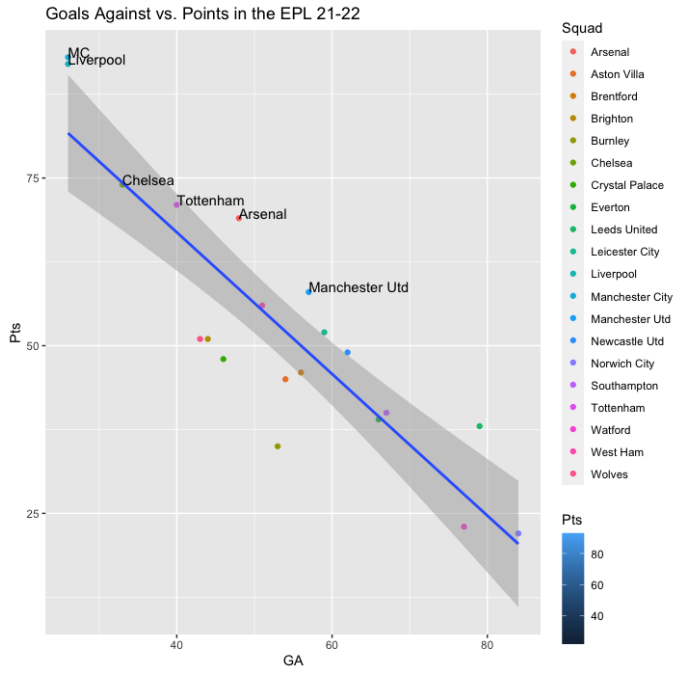
```
# Summary stats for the GD and Utd Comparison
> summary(GD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    -61     -20      -2       0      10      73
> df_MUFC['GD']
  GD
1  0
```

- United's GD was 0 in the 21/22 season, equal to the league average. If Manchester United wants to stop being average and get back to the top, GD needs to improve.

Points vs. GA

```
# Plot Showing relationship between Pts and GA
> ggplot(data = data) + geom_point(mapping = aes(x = GA, y = Pts, color = Squad, fill = Pts)) + geom_smooth(aes(GA, Pts), method = "lm") + labs(title = "Goals Against vs. Points in the EPL 21-22") + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Manchester Utd",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Liverpool",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Chelsea",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Tottenham",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Arsenal",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x = GA, y = Pts, label=ifelse(Squad == "Manchester City",as.character("MC"),'')),hjust=0,vjust=0)
```

- Below is the visualization that follows after executing the code above. I made a scatter plot with a trend line with ggplot2 in R to demonstrate the relationship between Points and GA. I also added labels for Manchester United and the teams that outperformed them in Points in the 21/22 season (Manchester City, Liverpool, Chelsea, Tottenham, and Arsenal)



- As the graph visualizes, there is a strong, negative, linear association between Points and Goals Against. All the teams that outperformed United (as labeled on the graph) seemed to have less GA than United, which makes sense given the relationship.

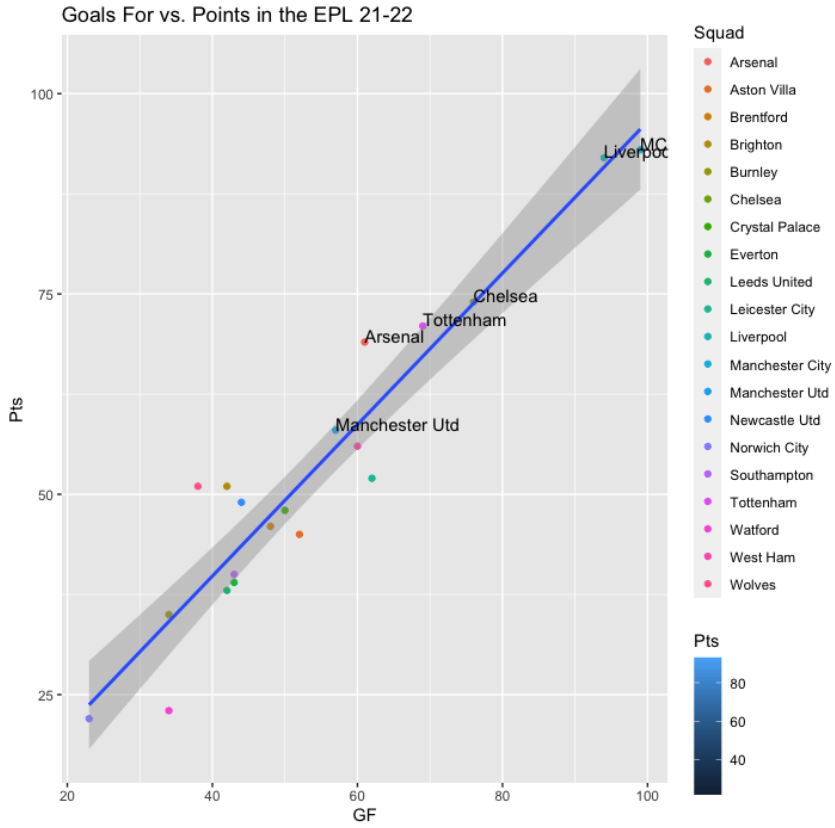
```
# Summary stats for the GA and Utd Comparison
> summary(GA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   43.75   53.50   53.55   63.00   84.00
> df_MUFC['GA']
  GA
1 57
```

- United's GA was 57 in the 21/22 season, greater than the league average and less than the 75% quartile. If Manchester United wants to stop being average and get back to the top, GA needs to lessen, being between the Minimum and 1st Quartile.

Points vs. GF

```
# Plot Showing relationship between Pts and GF
> ggplot(data = data) + geom_point(mapping = aes(x = GF, y =
Pts, color = Squad, fill = Pts)) + geom_smooth(aes(GF, Pts),
method = "lm") + labs(title = "Goals For vs. Points in the EPL
21-22") + geom_text(aes(x = GF, y = Pts, label=ifelse(Squad ==
"Manchester Utd",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = Pts, label=ifelse(Squad ==
"Liverpool",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = Pts, label=ifelse(Squad ==
"Chelsea",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = Pts, label=ifelse(Squad ==
"Tottenham",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = Pts, label=ifelse(Squad ==
"Arsenal",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = Pts, label=ifelse(Squad == "Manchester
City",as.character("MC"),'')),hjust=0,vjust=0)
```

- Below is the visualization that follows after executing the code above. I made a scatter plot with a trend line with ggplot2 in R to demonstrate the relationship between Points and GF. I also added labels for Manchester United and the teams that outperformed them in Points in the 21/22 season (Manchester City, Liverpool, Chelsea, Tottenham, and Arsenal)



- As the graph visualizes, there is a strong, positive, linear association between Points and Goals For. All the teams that outperformed United (as labeled on the graph) seemed to have more GF than United, which makes sense given the relationship.

```
# Summary stats for the GF and Utd Comparison
> summary(GF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   42.00   49.00   53.55   61.25   99.00
> df_MUFC['GF']
  GF
1 57
```

- United's GF was 57 in the 21/22 season, greater than the league average and less than the 75% quartile. If Manchester United wants to stop being average and get back to the top, GF should be somewhere in between the 75% quartile and the Max GF.

```
# Absolute value of Pearson Correlation Coefficient: GF vs GA in terms of Pts
> abs(cor(Pts, GF))
```

```
[1] 0.9505839
> abs(cor(Pts, GA))
[1] 0.8856865

# Absolute value of Pearson Correlation Coefficient: GF vs GA in
terms of GD
> abs(cor(GD, GF))
[1] 0.9570231
> abs(cor(GD, GA))
[1] 0.9375828
```

- **Since GD is the stat that is the most correlated with Pts, out of the stats that comprise GD, GF is more correlated with both Pts and GD than GA. While improving defense is important as GD, it can be recommended for teams to prioritize scoring goals as it is more related and thus to and thus more valuable for GD, which will impact Points**

- Thus, in the next part of this analysis, I will compare the other columns to determine which metric is most important in terms of maximizing Goals For.

Most Correlated with GF

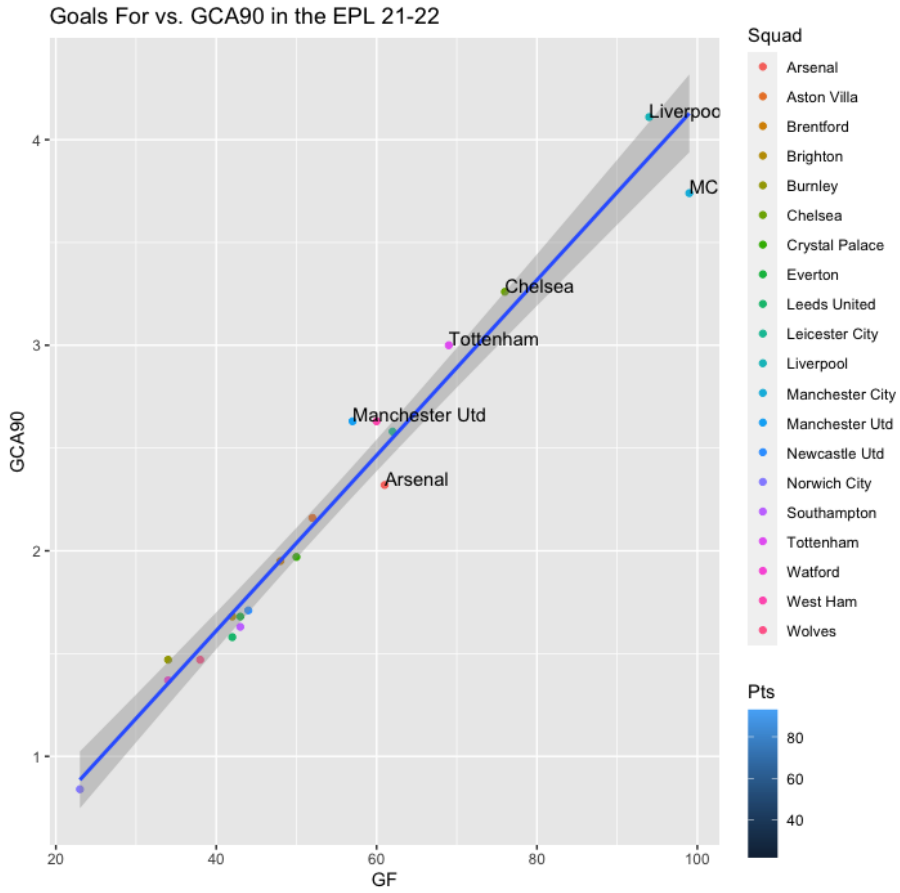
```
# Pts and Gls Pearson Correlation Test
> cor(GF, GCA90)
[1] 0.9838553
```

- **After running Pearson Correlation Tests for GF and other columns, it is clear that the strongest positive linear association is between GF and GCA90, at a staggering rxy value of approximately .98. Goal Creating Actions per 90 Minutes refers to "the two offensive actions directly leading to a goal, such as passes, dribbles and drawing fouls in a 90 minute match. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit". In other words, Goal Creating Actions are the direct passing, tackles, shots, etc. that lead to a goal.**
- I'm now going to visualize GCA90 and provide summary stats to see how Manchester United did in this category compared to the rest of the Premier League- particularly, the teams who got more points than them.

```
# Plot Showing relationship between GF and GCA90
> ggplot(data = data) + geom_point(mapping = aes(x = GF, y =
GCA90, color = Squad, fill = Pts)) + geom_smooth(aes(GF, GCA90),
```

```
method = "lm") + labs(title = "Goals For vs. GCA90 in the EPL
21-22") + geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Manchester Utd",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Liverpool",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Chelsea",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Tottenham",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Arsenal",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GF, y = GCA90, label=ifelse(Squad ==
"Manchester City",as.character("MC"),'')),hjust=0,vjust=0)
```

- Below is the visualization that follows after executing the code above. I made a scatter plot with a trend line with ggplot2 in R to demonstrate the relationship between GF and GCA90. I also added labels for Manchester United and the teams that outperformed them in Points in the 21/22 season (Manchester City, Liverpool, Chelsea, Tottenham, and Arsenal)



```
# Summary stats for the GCA90 and Utd Comparison
> summary(GCA90)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.840   1.617   1.960   2.189   2.630   4.110
> df_MUFC['GCA90']
  GCA90
1  2.63
```

- **Manchester United's GCA90 was higher than Arsenal and in the 75%, but again, that is not good enough compared to the top 4 (Tottenham, Chelsea, Liverpool, and Manchester City), who had higher GCA90. While United had a fairly high GCA90 per GF ratio, United scored less goals than the top 6, which leads me to think that while their GCA90 was good, they didn't create as many chances compared to the top 6.**
- I will investigate this next, through identifying what key factor is related to GCA90

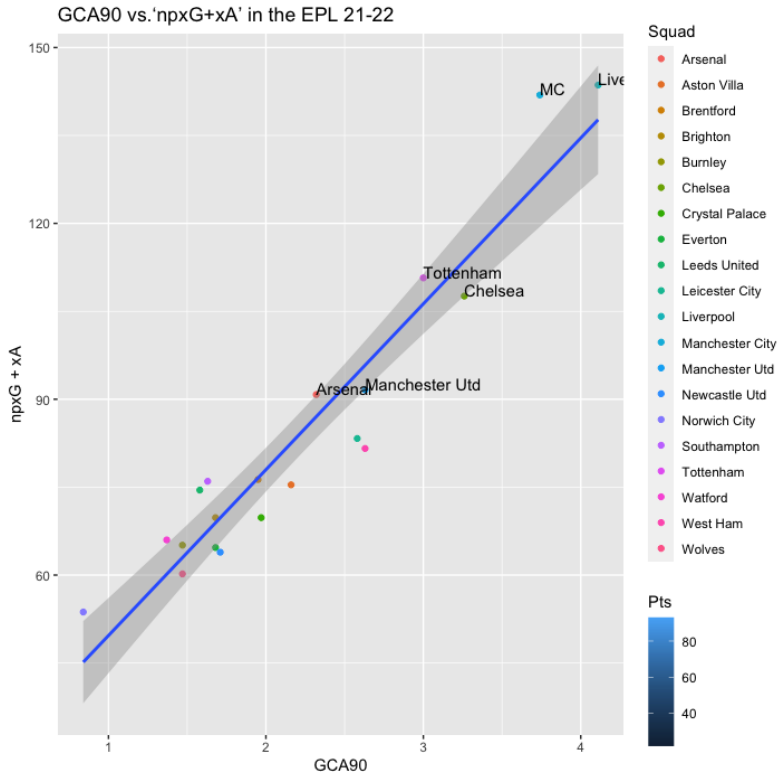
```
# Pearson Correlation Test for Metrics Related to GCA90
# Metrics with Strong Associations
# GCA90 and npxG Pearson Correlation Test
> cor(GCA90, npxG)
[1] 0.9466762

# GCA90 and xA Pearson Correlation Test
> cor(GCA90, xA)
[1] 0.9540922

# GCA90 and 'npxG+xA' Pearson Correlation Test
> cor(GCA90, `npxG+xA`)
[1] 0.9535296
```

- **Based on my Pearson Correlation Tests, it is clear that there is a strong relationship between GCA90 and xA and npxG, with xA mattering a tad bit more. Since xA and xG refers to the probability a pass or a shot leads to a goal or assist rather than the conversion of that chance, it is clear that getting high quality chances which have a high probability for a goal/assist impact GCA90, which is expected.**

- Now, let's use visualizations and summary statistics to see how Manchester United compares to the league with these specific metrics.

```
# Plot Showing relationship between GCA90 and 'npxG+xA'
> ggplot(data = data) + geom_point(mapping = aes(x = GCA90, y =
npxG+xA, color = Squad, fill = Pts)) +
geom_smooth(aes(GCA90,npxG+xA), method = "lm") + labs(title =
"GCA90 vs.'npxG+xA' in the EPL 21-22") + geom_text(aes(x =
GCA90, y = npxG+xA, label=ifelse(Squad == "Manchester
Utd",as.character(Squad),'')),hjust=0,vjust=0) + geom_text(aes(x
= GCA90, y = npxG+xA, label=ifelse(Squad ==
"Liverpool",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GCA90, y = npxG+xA, label=ifelse(Squad ==
"Chelsea",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GCA90, y = npxG+xA, label=ifelse(Squad ==
"Tottenham",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GCA90, y = npxG+xA, label=ifelse(Squad ==
"Arsenal",as.character(Squad),'')),hjust=0,vjust=0) +
geom_text(aes(x = GCA90, y = npxG+xA, label=ifelse(Squad ==
"Manchester City",as.character("MC"),'')),hjust=0,vjust=0)
```

- Below is the visualization that follows after executing the code above. I made a scatter plot with a trend line with ggplot2 in R to demonstrate the relationship between 'npxG+xA' and GCA90. I also added labels for Manchester United and the teams that outperformed them in Points in the 21/22 season (Manchester City, Liverpool, Chelsea, Tottenham, and Arsenal)



```
# Summary stats for the 'npxG+xA' and Utd Comparison
> summary(`npxG+xA`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  53.80   65.70   75.75   83.33   91.03  143.50
> df_MUFC['npxG.xA']
  npxG.xA
1    91.7
```

- **Similar to GCA90, while United had a good 'npxG+xA' (and one better than 5th place Arsenal), United were still well under the likes of Tottenham, Chelsea, Liverpool, and Manchester City. If United want to be back in the top with these teams, they need to improve this statistic. The way to improve 'npxG+xA' is to get more high quality chances in a game. By getting more high quality chances, the probability of a goal/assist will increase- i.e higher xG and xA. Thus, as I stated in my initial hypothesis, high quality chance creation is necessary for goals, which will influence higher point accumulation.**

- Picture of Ten Hag training the team before 22/23 Pre-Season. One of his goals is improving ball creativity, possession, and chance creation ;)

Conclusion and Recommendation

- To conclude, my initial hypothesis seems to be correct. While correlation doesn't equal causation, since these statistics are highly related and make sense given football (soccer) context, there is indication of a direct, positive relationship between Point Accumulation and Goal Difference, Goal Difference and Goals For, Goals For and GCA90, and GCA90 and 'npxG+xA'.
- If United prioritize high quality chance creation, ball creativity, and transferring in players who will improve high quality chance creation, I ensure that United will be extremely competitive in the Premier League and Europe in 1-3 years.
- Since Ten Hag is a coach who prioritizes this type of football in his training and games, I think United's trajectory is certainly upwards.
- This concludes my statistical analysis

- 12/13 Season: Last time United won the Premier League. Hopefully, the good times will return!