Comparing Passing in Major League Soccer and the English Premier League

Motivation

One of the basic fundamentals of soccer is passing. Passing refers to one player pushing the ball with a part of their body to another player on their team. In this day and age of soccer, passing has become such an asset for top teams in the world and every player on a team needs to be able to pass the ball well, including the goalkeeper.

According to **globalfootballrankings.com**, the English Premier League (EPL) is ranked #1 in the world whereas Major League Soccer (MLS) is ranked #16. A lot of different factors could affect these rankings but taking a deep dive into passing statistics analysis can help paint a small picture of what makes these leagues so different.

This research will compare passing data from the 2023 MLS season with data from the 2022-23 EPL to see what differences can be found.

Some questions that can be answered:

- What makes the EPL such a high quality league?
- What variables are different between MLS and the EPL?
- Can a correlation be found between variables such as attempted passes and league position?

Data Sources

All of the data sets for this research were collected from **fbref.com**. This website contains statistics, scores, and history for 100+ men's and women's club and national team competitions

For MLS, the overall league standing was collected from **wikipedia.com**. This had to be done because the league is split into two conferences, East and West, and fbref.com only has the standings for each conference.

Description of Workflow

Import and Install Libraries: essential libraries needed for data manipulation and visualization were imported and installed, such as pandas, numpy, sklearn, seaborn, and matplotlib

Load Data: The datasets were loaded using pandas.read_html. Every dataset was embedded as an html table which allowed for easy import

Cleaning Data: Datasets were cleaned after import. Many of the datasets had multi-leveled headers that had to be reduced to a single level. Also, many columns were renamed for clarity

Data Type Conversion: Data types of columns were converted, mainly from strings to integers, to allow for easy merging or analysis

Merging Data: Different datasets were merged in order to gather all necessary variables

ANALYSIS AND VISUALIZATIONS: BASIC STATISTICS

Overview

First, basic stats between the MLS and EPL were observed to see if there were any obvious differences between the leagues. The stats compared were broad stats with little obvious explanations for their occurrence. These stats involved *Position, Average Possession, Goals Scored, and Goals Against*.

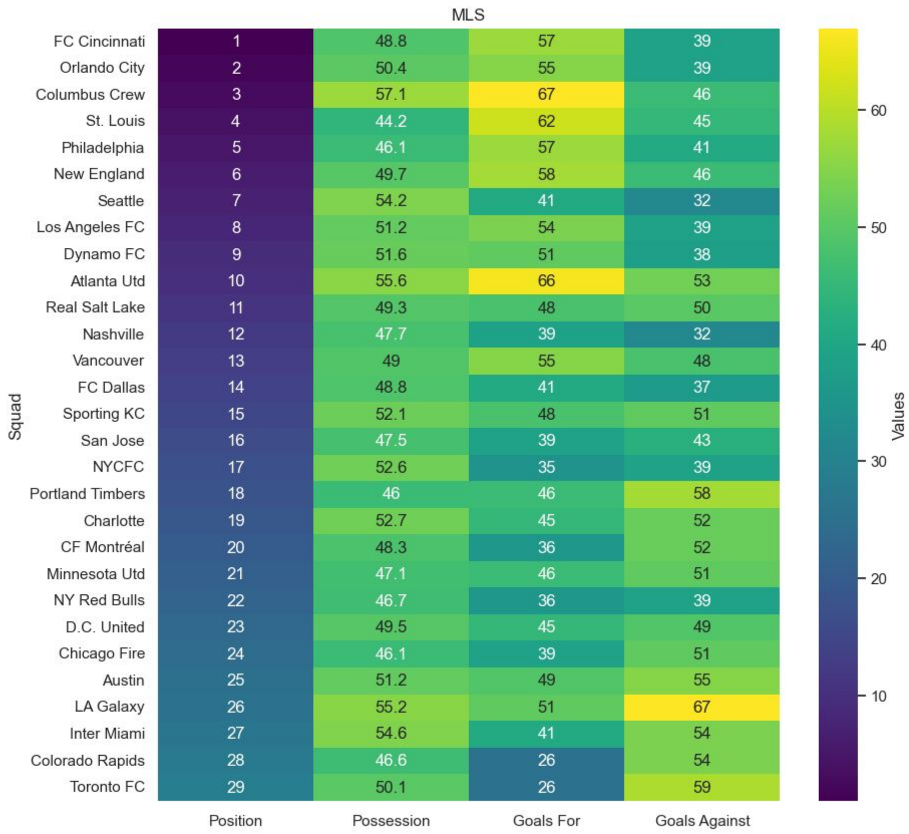
For each league, the process for creating the dataframe used to make the heatmap was very similar. For both leagues, The data was collected from fbref.com. For the EPL data, the ‘Squad’ and ‘Possession’ columns from the ‘Squad Standard Stats’ table were stored in a dataframe. Then, ‘Rk’, ‘Squad’, ‘GF’, and ‘GA’ columns from the ‘Regular season’ table were stored in another dataframe. After this, the two dataframes were merged on ‘Squad’ to get the full table.

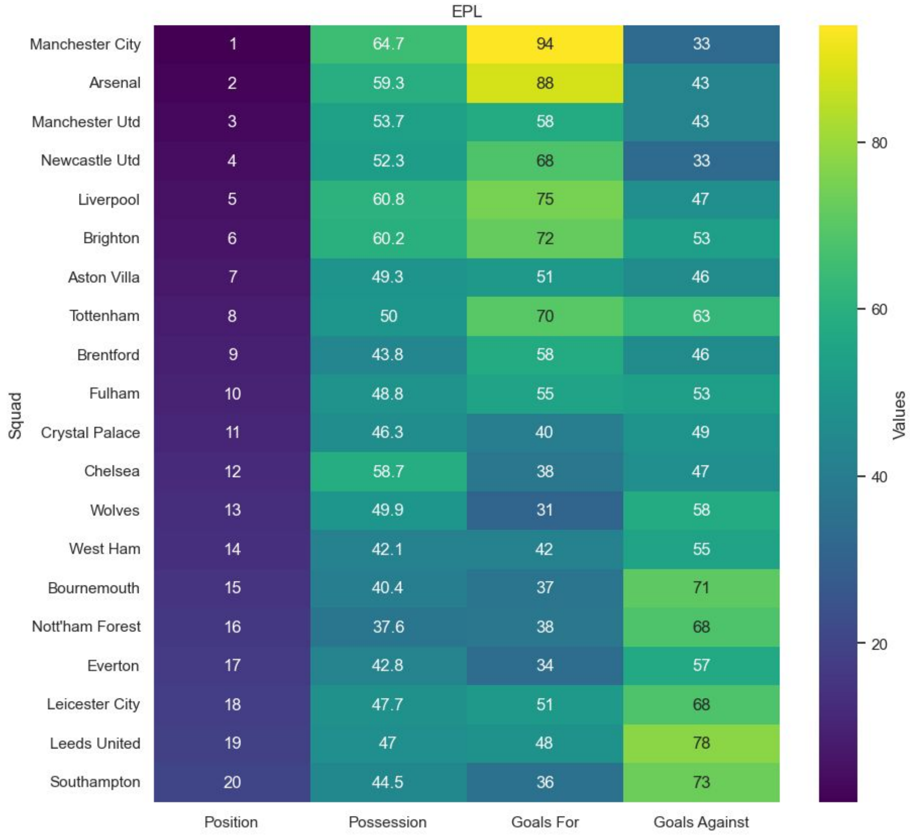
The process for the MLS dataframe was a bit more challenging because the league standings are split into 2 conferences on fbref, and overall standings was needed. The solution to this problem was finding an overall standings table on Wikipedia and manually mapping the rank to each team. Besides this, the process was the same as for the EPL table.

Once the dataframes for each league was created, a heatmap was created using seaborn that put the squad names on the y-axis and the remaining variables on the x-axis with annotations set to True.

Key Insights

1. There is a significant difference in the average possession trends for teams in the EPL and MLS. In the EPL, the top 6 teams all averaged more than 50% possession and the rest of the teams in the league all averaged less than 50% possession with the only exceptions being Tottenham and Chelsea. In the MLS, possession is all over the place. The #1 team, FC Cincinnati, averaged 48.8% whereas LA Galaxy, the #26 team, averaged 55.2%. It seems that in the EPL, better teams will possess the ball and this could influence their chances of winning. Alternatively, teams in the MLS can possess the ball less than their opponent and still win games
2. As expected, top teams in both leagues scored the most goals and gave up the fewest goals. An interesting team from the MLS was the Seattle Sounders, the #7 team, who only scored 41 goals. However, they were tied for the best defence in the league. Since they played 34 games, this means they averaged a little more than one goal scored a game and averaged less than a goal a game, showing how close every game was for them.





ANALYSIS AND VISUALIZATIONS: PASSING STATISTICS

Overview

This section includes the first in-depth comparison for passing statistics between the EPL and MLS. The aim was to see if there were any significant differences between these statistics to get an idea of how teams play in both leagues. As seen in the previous slide, the distribution of possession stats relative to league position was very different between the leagues. This analysis takes a deeper dive into passing which can be argued to have the biggest impact on possession

Four key variables of passing were collected from fbref.com- 'Att', 'Cmp%', 'TotDist', and 'PrgDist'. The variables were renamed to 'Attempted', 'Completed %', 'Total Distance (Yards)', and 'Progressive Distance (Yards)'. Each league has a passing stats section and from this section, the 'Squad Passing' table, which contains the variables we wanted, was extracted. These stats pertained to each team in the league.

After collecting the data, box plots of each variable (next page) were created, with league on the y-axis and value on the x-axis. Box plot was the choice of visualization because it clearly showed how the data was distributed and showed any potential outliers. Creating box plots for each league and showing them on the same graph allowed for an easy visualization of the differences between the leagues.

Note: Instead of plotting Attempted Passes, Attempted Passes per 90 was plotted. This is because teams in the EPL play 38 games a season whereas teams in MLS play 34 games a season. Plotting the average per 90 minutes seemed like a more fair comparison. This transformation was also performed on Progressive Distance.

Key Insights

1. **Attempted Passes Per 90**
   a. This variable describes the averages number of passes attempted during a game by a team. As seen in the first box plot (Figure 1) the range for passes is significantly larger (approx 360-700) in the EPL compare to MLS (approx 400-560). This is mainly due to the fact that many of the top teams in the EPL play a possession-based style. Teams such as Manchester City, Liverpool, and Brighton & Hove Albion were known to wear teams down by passing the ball around them. The teams toward the bottom of the standings most likely attempted fewer passes because they didn't possess the quality of players capable of playing a possession-based style. In MLS, teams were closer together in terms of attempted passes. It's interesting that the team with the most attempted passes per 90 in MLS would barely be above the 75th percentile in the EPL.
2. **Progressive Distance (Yards)**
   a. This variable describes the average distance of completed passes per 90 that traveled towards the opponent's goal, so any forward or diagonal pass. Similar to the box plot for attempted passes per 90, the second box plot (Figure 2) shows a large range for EPL teams compared to MLS teams. It would make sense that the ranges of the boxplots are similar because if a team attempts more passes, they should cover more progressive distance given the passes aren't sideways or backwards. For this variable, there are two outliers. For the EPL, Manchester City is a right outlier and for MLS, the New York Red Bulls are a left outlier.
3. **Average Distance Per Pass (Yards)**
   a. This variable describes the average distance of a completed pass. It was calculated by dividing the total distance of completed passes by the total number of completed passes. As seen in the third box plot (Figure 3), the ranges for average distance per pass do not differ significantly between the EPL and MLS, however the range between the 25th and 75th percentiles are significantly different. The 75th percentile for the EPL is equivalent to the 25th percentile of the MLS, indicating that the majority of teams in the EPL complete shorter passes and the majority of the teams in MLS complete longer passes.
4. **Completed %**
   a. This variable describes the percentage of attempted passes that were completed, meaning they traveled from one player to another with no interference from an opponent. As seen in the fourth box plot (Figure 4), the ranges for completed % is not significantly different between the EPl and MLS. For MLS, the New York Red Bulls are a left outlier. What's interesting from this data is that the top teams in the EPL had a higher completed % compared to MLS. Manchester City won the league and had an 87.3 completed % whereas St. Louis SC topped the Western Conference but had the lowest (besides Red Bulls) completed % of 72.1 and LA Galaxy finished second to last with the best completed % in the league at 85.2.

Passing Statistics

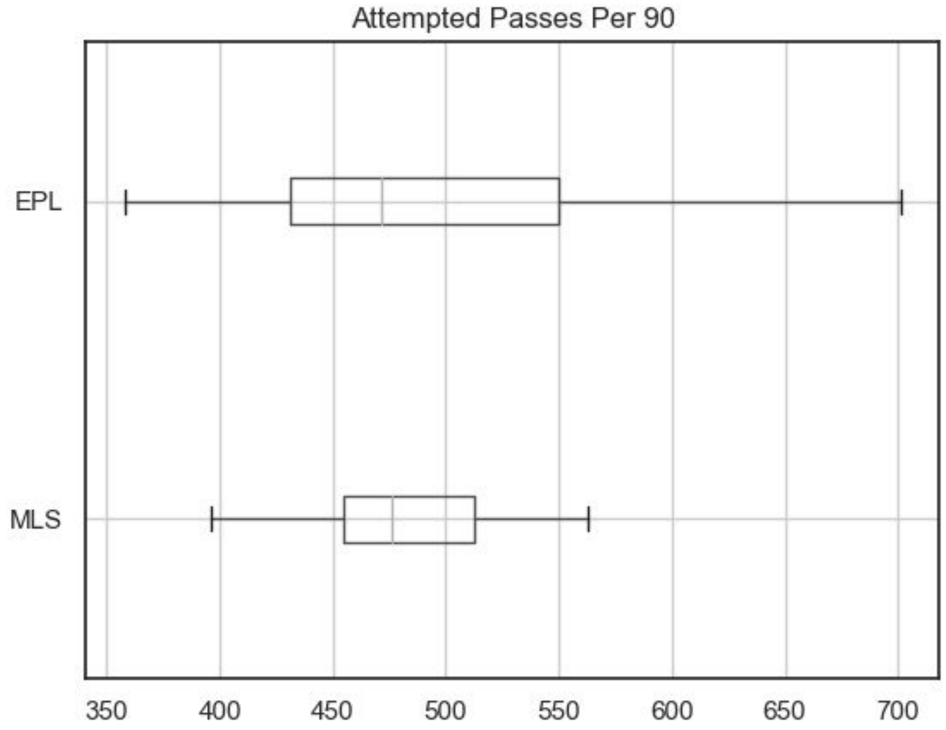


Figure 1: *Attempted Passes Per 90*

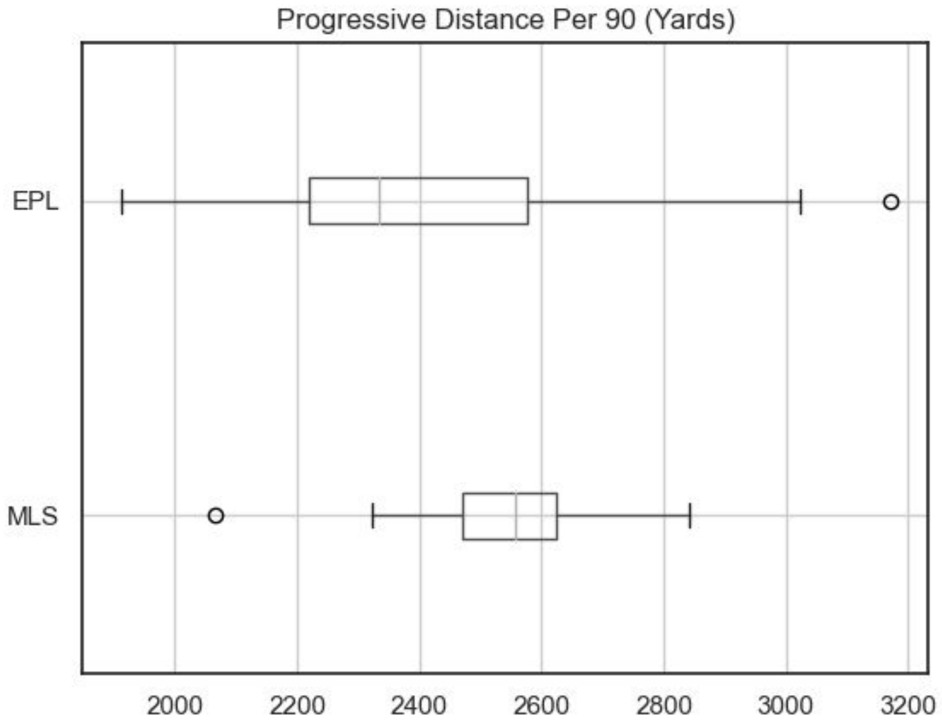


Figure 2: *Progressive Distance Per 90 (Yards)*

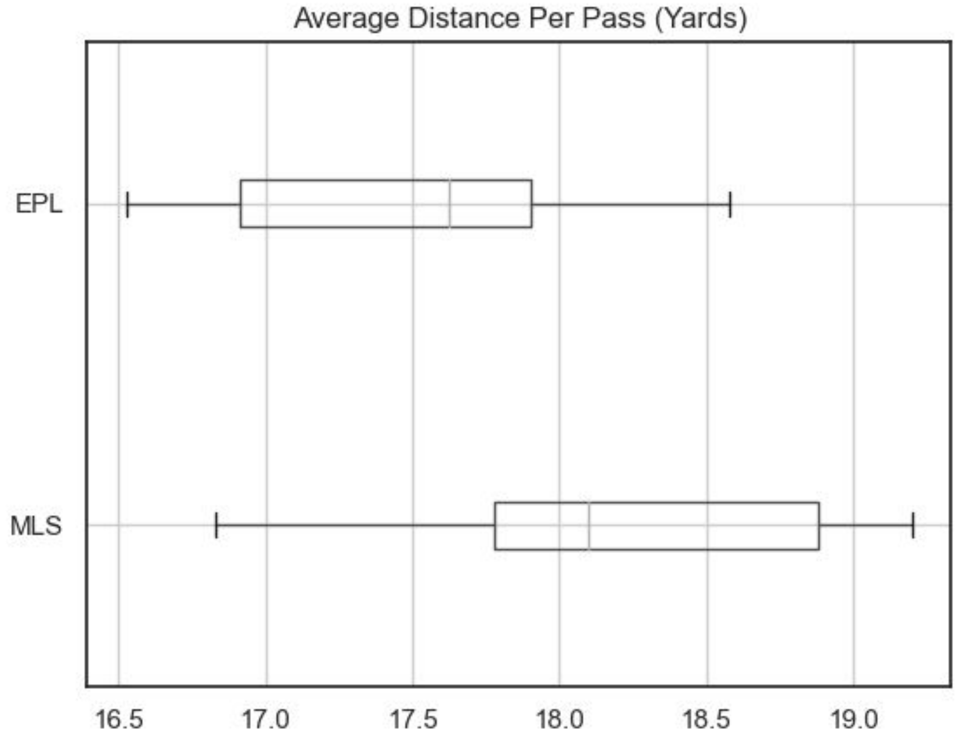


Figure 3: *Average Distance Per Pass (Yards)*

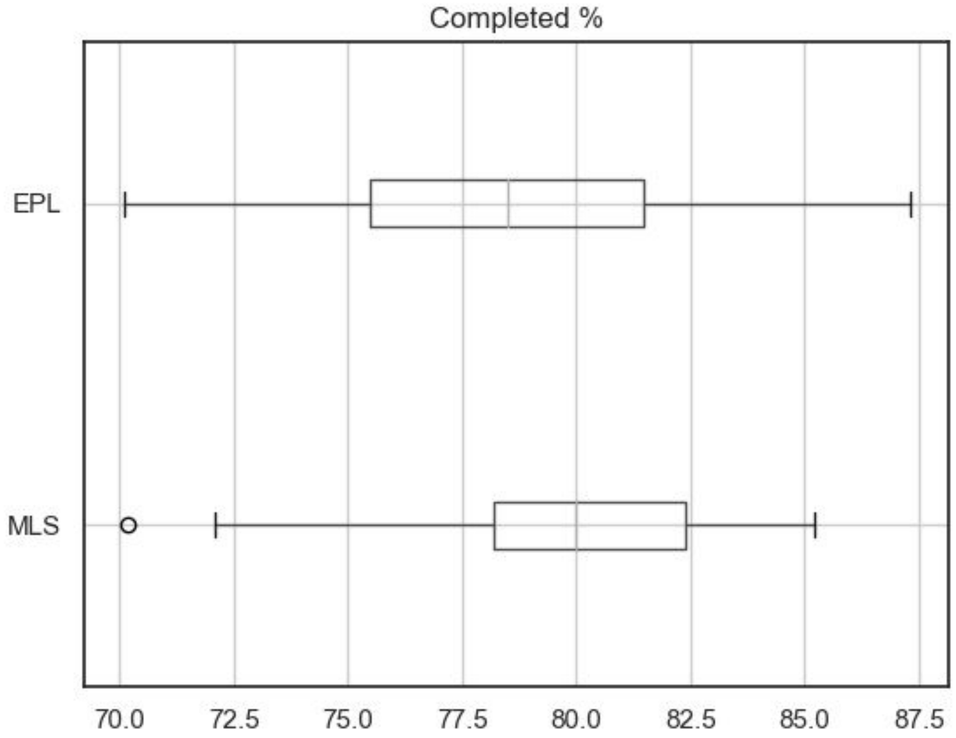


Figure 4: *Completed % (Yards)*

ANALYSIS AND VISUALIZATIONS: CORRELATION ANALYSIS

Overview

The goal for this analysis was to investigate the relationships between different passing statistics. These variables included 'Position', 'Attempted Passes', 'Completed Passes', 'Completed %', Total Distance, and Progressive Distance. A correlation analysis was done to analyze the relationships between these different passing statistics.

A correlation matrix was used to conduct the correlation analysis and was created using .corr(). The heatmap function from the Seaborn library was then called to create a visual representation of the correlation analysis. Values within the squares of these heatmaps represent R-squared values that indicate the strength of the relationship between the boxs' corresponding variables. This process was done for both the EPL and MLS.`

Key Insights

1. Figures 5 and 6 to the right are correlation matrices for passing statistics along with league position in the EPL and MLS. An important thing to note before suggesting any insights is the legend for each heatmap. Both heatmaps utilized the same colormap, however the range for values in the EPL heatmap is significantly larger. Therefore, squares that share a similar color between the heatmaps may not necessarily share a similar value
2. The most significant insight that can be made about these correlation matrices is that passing statistics have a significant impact on league position in the EPL but not in the MLS. The R-squared value for each variable with Position was at least -0.67, indicating a strong negative relationship. This means that as league position increases (the higher the league position value the worst the team placed), all passing statistics tended to decrease. For MLS, none of the squares that involved *Position* has an absolute value greater than 0.07, indicating that they all have close to no impact on where a team finishes in the league. This shows how different the two leagues are in terms of passing.
3. Looking at all the variables beside *Position*, they all have a strong correlation with each other. Every square in each heatmap has an R-squared value of at least 0.8. However, For each square in the EPL heatmap, the corresponding square in the MLS heatmap has a lower R-squared value. This is an interesting finding and can reinforce the idea that passing in the EPL is more efficient than in MLS

Notes

While the purpose of the heatmaps is to show the correlation between passing variables and league position, it's import to note that correlation does not imply causation. This concept mainly applies to the EPL. Just because a team performs well in passing metrics, it does not mean they will automatically finish. higher in the table. There are many other variables that play into how a team performs like shots, goals, and defensive actions just to name a few.

Figure 5: *EPL Correlation Matrix*

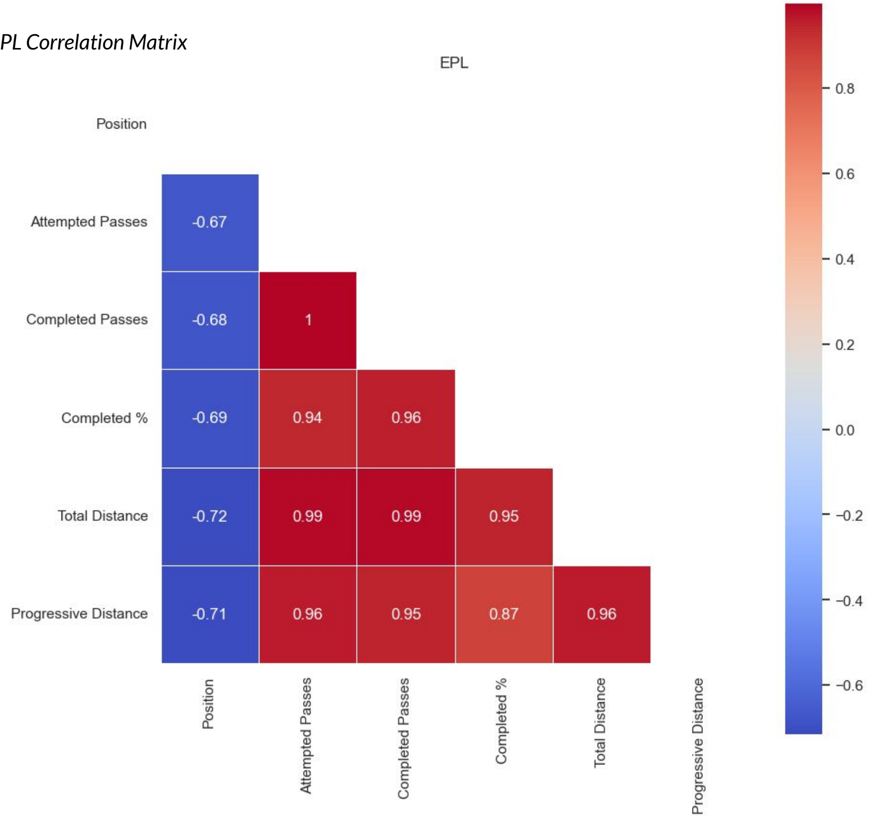
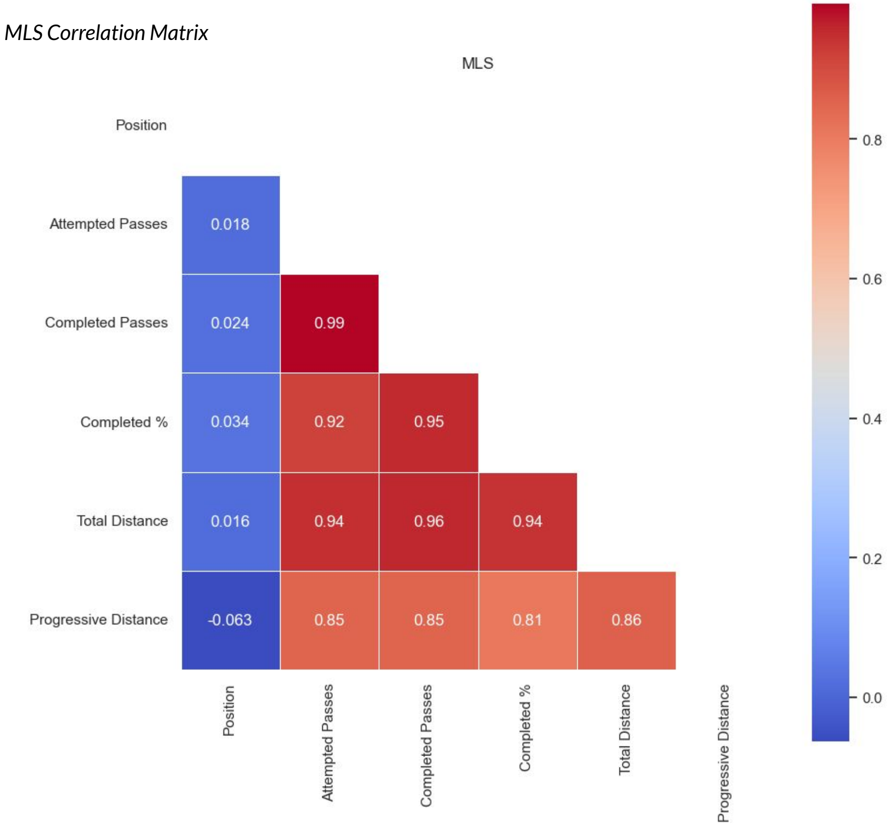


Figure 6: *MLS Correlation Matrix*



ANALYSIS AND VISUALIZATIONS: MULTIVARIATE REGRESSION

Overview

The goal for this analysis was to determine any relationships between passing statistics and see if they are correlated with league position. Finding any significant relationships could help a team figure out what areas of passing they need to improve on to improve their performance in the league.

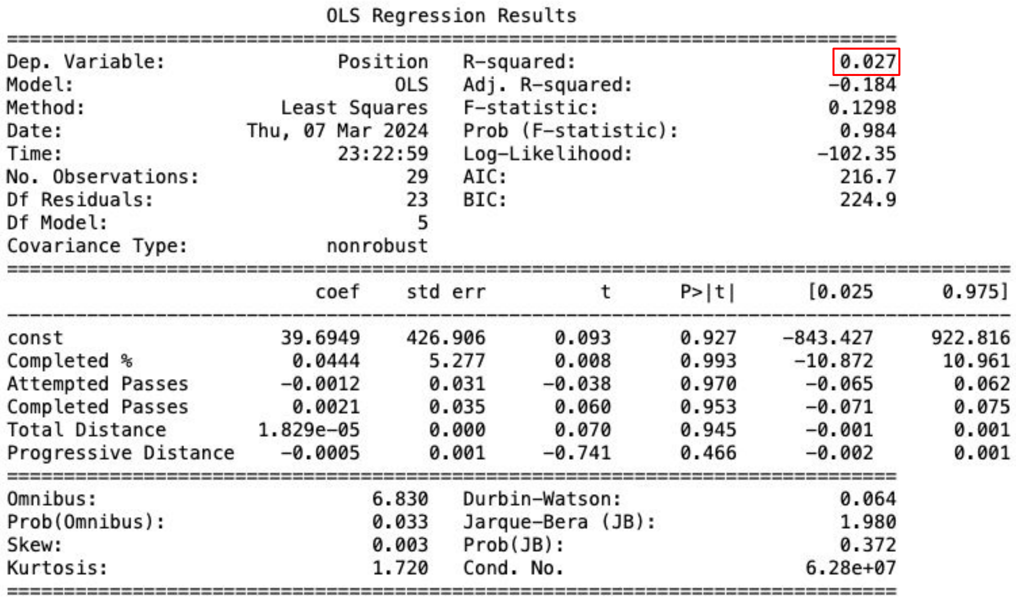
A multivariate regression model was chosen because there were multiple independent variables that may collectively influence a single dependent variable. In this model, the independent variables were *Completed Passes*, *Attempted Passes*, *Completed %*, *Total Distance, and Progressive Distance*. The dependent variable was league position.

This model was created for both the EPL and MLS. These two leagues are very different in terms of play style and comparing passing statistics can provide an idea of how teams try to play.

Key Insights

1. Figures 7 and 8 below are the summary results of the multivariate regression model for MLS and the EPL, respectively. As seen by the R-squared values for each model highlighted by the red boxes in each figure, passing statistics have a stronger impact on league position in the EPL compared to in MLS.
2. An R-squared value of 0.027 for the MLS model indicates that passing statistics have practically no impact on how a team finished in the league table. A team that finishes towards the top could have low values for all passing statistics whereas a team that finishes towards the bottom could have high values for all passing statistics. As discussed in the Key Insights section on Page 4, St. Louis SC and LA Galaxy are perfect examples of this statement. Ultimately, passing statistics do not help a team perform better
3. An R-squared value of 0.638 for the EPL model indicates that passing statistics have a very strong impact on how a team finished in the league table. This tells a lot about the quality of the league. The competition level in the EPL is very high and in order to be at the top a team needs to have a good foundation of passing, which is a very basic fundamental of the game. Manchester City is the best team in the EPL and it could be argued that they win games simply because they are better at passing than every other team in the league

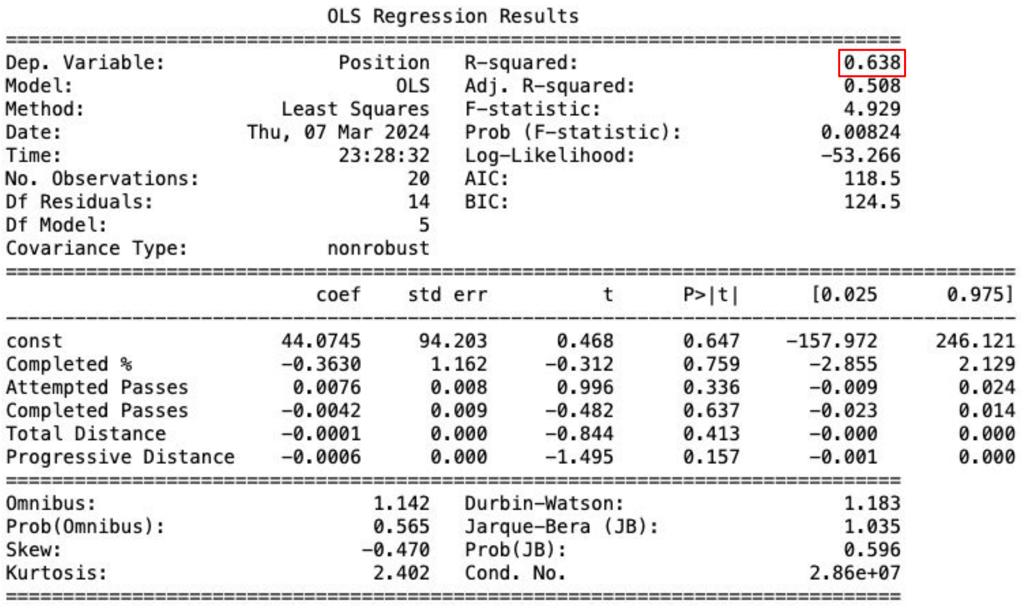
MLS



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Position   R-squared:                       0.027
Model:                            OLS   Adj. R-squared:                 -0.184
Method:                 Least Squares   F-statistic:                    0.1298
Date:                Thu, 07 Mar 2024   Prob (F-statistic):              0.984
Time:                        23:22:59   Log-Likelihood:                -102.35
No. Observations:                  29   AIC:                             216.7
Df Residuals:                      23   BIC:                             224.9
Df Model:                           5
Covariance Type:            nonrobust
==========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   39.6949    426.906      0.093      0.927    -843.427     922.816
Completed %              0.0444      5.277      0.008      0.993     -10.872      10.961
Attempted Passes        -0.0012      0.031     -0.038      0.970      -0.065       0.062
Completed Passes         0.0021      0.035      0.060      0.953      -0.071       0.075
Total Distance        1.829e-05      0.000      0.070      0.945      -0.001       0.001
Progressive Distance    -0.0005      0.001     -0.741      0.466      -0.002       0.001
==============================================================================
Omnibus:                        6.830   Durbin-Watson:                   0.064
Prob(Omnibus):                  0.033   Jarque-Bera (JB):                1.980
Skew:                           0.003   Prob(JB):                        0.372
Kurtosis:                       1.720   Cond. No.                     6.28e+07
==============================================================================
```

Figure 7: *Multivariate Regression Model for MLS*

EPL



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Position   R-squared:                       0.638
Model:                            OLS   Adj. R-squared:                  0.508
Method:                 Least Squares   F-statistic:                     4.929
Date:                Thu, 07 Mar 2024   Prob (F-statistic):            0.00824
Time:                        23:28:32   Log-Likelihood:                -53.266
No. Observations:                  20   AIC:                             118.5
Df Residuals:                      14   BIC:                             124.5
Df Model:                           5
Covariance Type:            nonrobust
==========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   44.0745     94.203      0.468      0.647    -157.972     246.121
Completed %             -0.3630      1.162     -0.312      0.759      -2.855       2.129
Attempted Passes         0.0076      0.008      0.996      0.336      -0.009       0.024
Completed Passes        -0.0042      0.009     -0.482      0.637      -0.023       0.014
Total Distance          -0.0001      0.000     -0.844      0.413      -0.000       0.000
Progressive Distance    -0.0006      0.000     -1.495      0.157      -0.001       0.000
==============================================================================
Omnibus:                        1.142   Durbin-Watson:                   1.183
Prob(Omnibus):                  0.565   Jarque-Bera (JB):                1.035
Skew:                          -0.470   Prob(JB):                        0.596
Kurtosis:                       2.402   Cond. No.                     2.86e+07
==============================================================================
```

Figure 8: *Multivariate Regression Model for the EPL*

ANALYSIS AND VISUALIZATIONS: HYPOTHESIS TESTING

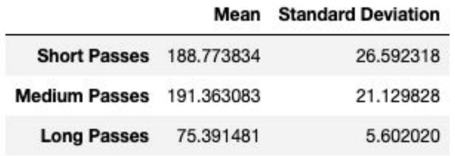
Overview

In this analysis, the goal was to see if there's any statistically significant difference between the means of different passing metrics in the EPL and MLS. If any metrics had means that were deemed different by these tests then it would be fair to claim that the two leagues are different in a meaningful and non-random way.

The metrics being observed for these tests were *Short Passes Attempted*, *Medium Passes Attempted*, *Long Passes Attempted*, *Short Completed %*, *Medium Completed %*, *and Long Completed %*. These metrics refer to the distance of passes made by a team. A short pass is any pass that travels between 5 and 15 yards. Medium is between 15 and 30 and Long is more than 30.

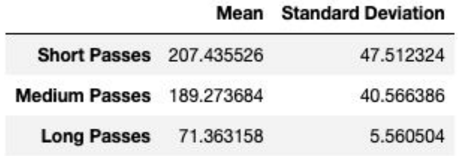
The two tables below show the mean and standard deviation for the mentioned pass lengths. It is initially clear that there are more short passes per game in the EPL but more medium and long passes per game in the MLS. Despite this, a t-test can be used to determine if these differences are statistically significant.

MLS



	Mean	Standard Deviation
Short Passes	188.773834	26.592318
Medium Passes	191.363083	21.129828
Long Passes	75.391481	5.602020

EPL



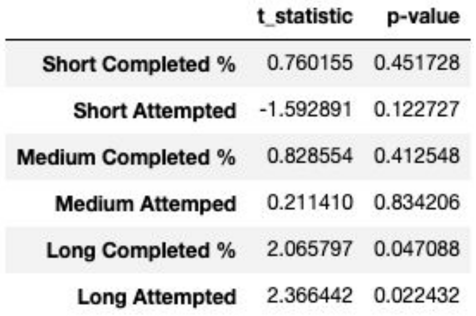
	Mean	Standard Deviation
Short Passes	207.435526	47.512324
Medium Passes	189.273684	40.566386
Long Passes	71.363158	5.560504

Hypothesis testing with a Welch's t-test was used for each passing metric. Welch's t-test was the chosen test because it is less restrictive compared to the original Student's test. This test allows the user to assume that the variance is different between two groups. An alpha of 0.05 will be used.

For each hypothesis test, the statistical hypotheses were as follows:

- Null Hypothesis (H_O): There is no significant difference in mean {passing metric} between the EPL and MLS
- Alternative Hypothesis (H_A): There is a significant difference in mean {passing metric} between the EPL and MLS

The t-statistic and p-value for each test was collected using stats.ttest_ind()



	t_statistic	p-value
Short Completed %	0.760155	0.451728
Short Attempted	-1.592891	0.122727
Medium Completed %	0.828554	0.412548
Medium Attemped	0.211410	0.834206
Long Completed %	2.065797	0.047088
Long Attempted	2.366442	0.022432

Figure 9: *Results of Welch's t-tests*

Key Insights

1. Figure 9 above shows the t_statistics and p-values from each t-test performed on the passing metrics. A positive t-statistic indicates that the mean for MLS is greater than the mean for the EPL and a negative t-statistic indicates that the mean for the EPL is greater than the mean for MLS. The only metric that had a greater mean in the EPL was *Short Attempted*. This would make sense as teams such as Manchester City, Liverpool, Arsenal, and Brighton & Hove Albion played possession-based styles that focused on stringing short passes together to break teams down.
2. The p-values for *Short Completed %*, *Short Attempted*, *Medium Completed %*, and *Medium Attempted* were all above the alpha threshold of 0.05, indicating a failure to reject the null hypothesis for each hypothesis test. Any difference between the means of the EPL and MLS groups may likely be due to random sampling variability rather than a true difference. The p-values for *Long Completed %* and *Long Attempted* were both below 0.05, allowing the null hypothesis to be rejected. The difference in the means of long passing metrics between the EPL and MLS is statistically significant. Although these tests indicates a significant difference, it doesn't explain why these differences exist. It is interesting that teams in the MLS attempt more long passes and are more successful with long passes on average.

SUMMARY

Conclusion

In this research, many different passing metrics were analyzed and compared between the EPL and MLS. After comparing basic passing metrics with league position in both leagues, it's clear that the approach for winning games is very different. Top teams in the EPL are those that perform well across all passing metrics whereas top teams in the MLS focus on other areas of the game to produce winning outcomes.

There are numerous factors that play into the success of a soccer team no matter what league they are in. Teams like Manchester City and Liverpool are among the best in the world and their play styles revolve around dominating possession. It would be interesting to see if any teams in MLS are capable of replicating their play styles. It's understandable that building a team of players with passing abilities similar to Rodri, Trent-Alexander Arnold, and Ruben Dias would be close to impossible for teams in MLS but if clubs tried to put more of an emphasis on passing they might be able to improve performance.

When adding players to the 1st team, MLS clubs could try to use passing metrics to decide who to add. Whether it be from the academy or singing a new player, instilling the importance of passing could be a vital aspect. Taking a look at the Houston Dynamo of MLS could help paint a picture of how improving passing could help a team perform better. In 2022, the Dynamo finished 2nd to last in the Western Conference. 75% of the way through the season, they signed veteran midfielder Hector Herrera and in the offseason, they signed another midfielder in Artur. In 2023, Houston finished 4th in the western conference and attempted almost 1,000 more passes than in 2022. Herrera was by far the best player for Houston and his passing ability was his most vital asset. He provided 11 assists and ranked 2nd in the league for passes attempted. If teams in the league try and sign players with passing abilities similar to Herrera, they could potentially see an improvement in performance. However, there are still other factors that impact performance.

References

1. 2023 MLS Squad Standard Stats. Retrieved from https://fbref.com/en/comps/22/2023/2023-Major-League-Soccer-Stats
2. 2023 MLS Passing Stats. Retrieved from https://fbref.com/en/comps/22/2023/passing/2023-Major-League-Soccer-Stats
3. 2023 MLS Overall Standings. Retrieved from https://en.wikipedia.org/wiki/Template:2023_Major_League_Soccer_season_table
4. 2023 Ranking for Soccer Leagues in the World. Retrieved from https://www.globalfootballrankings.com/
5. 2023 EPL Squad Standard Stats. Retrieved from https://fbref.com/en/comps/9/2022-2023/2022-2023-Premier-League-Stats
6. 2023 EPL Passing Stats. Retrieved from https://fbref.com/en/comps/9/2022-2023/passing/2022-2023-Premier-League-Stats