

Predicting MLS Team League Position

By Tobin Mathew



Introduction

What is MLS?

- Major League Soccer (MLS) is the professional soccer league in the United States and Canada. Founded in 1993, MLS has grown significantly, becoming a vital part of the North American sports landscape. It consists of teams from both countries and has a structured competition format leading to playoffs and the MLS Cup.



Project Objectives

Goal: The primary goal of this project is to determine the league position of MLS teams based on statistical analysis.

Why: By leveraging statistical methods and data analysis, we aim to gain deeper insights into the performance metrics of MLS teams. Understanding these metrics can provide valuable information about team standings and performance trends over multiple seasons.



Data Collection, Preprocessing & Tools


Data collection :

- Source: Official MLS website (<https://www.mlssoccer.com/stats/>) - Seasons (2011-2024)
- Scope: The dataset includes a wide range of metrics such as goals scored, assists, shots on goal, fouls committed, and various team statistics.

Data Preprocessing :

- Cleaning Procedures: Handled missing values, Converted data types as necessary (e.g., numeric, categorical) to facilitate effective data manipulation and analysis
- Standardization: Normalized data to ensure consistency in statistical metrics across different seasons. This step is crucial for fair comparisons and accurate analysis of team performance trends.

Tools:

- Python, Pandas, Matplot, Seaborn, Sklearn and Dask
- 

Exploratory Data Analysis

Metrics Analysis: Goals, Attack, Defense, and Possession

Goals Metrics:

- Features include goals scored, goals conceded, and specific goal types (e.g., penalties, box goals).
- Insights into scoring efficiency and defensive vulnerabilities.

Attack Metrics:

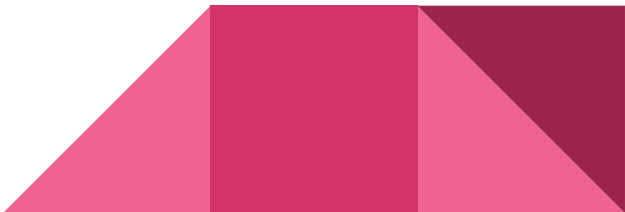
- Features include total scoring attempts, on-target attempts, and set-piece effectiveness.
- Indicates offensive prowess and ability to convert scoring opportunities.

Defense Metrics:

- Features include metrics like saves, clean sheets, and interceptions.
- Highlights defensive capabilities and resilience against opponent attacks.

Possession Metrics:

- Features include possession percentage, pass accuracy, and successful dribbles.
- Reflects team control and ability to maintain and utilize possession effectively.



EDA

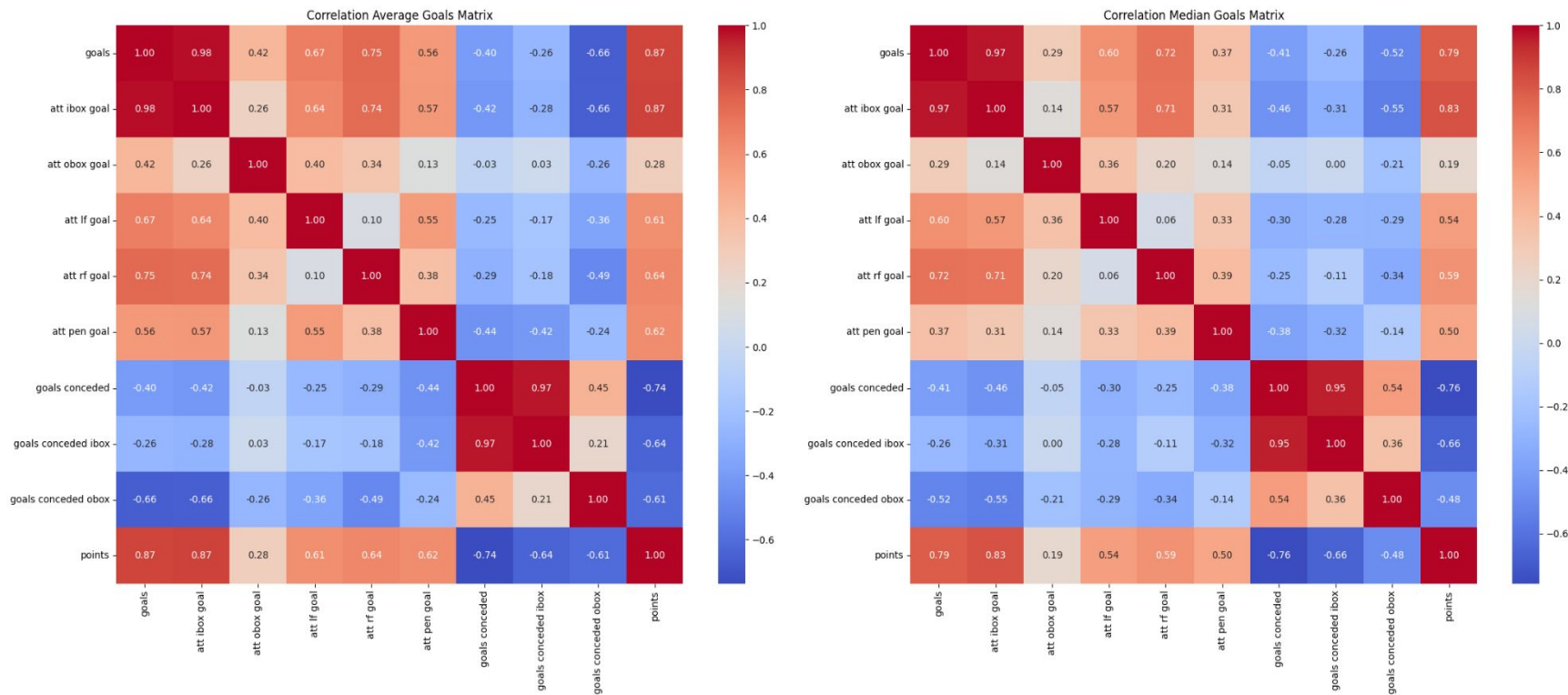
Correlations Analysis:

- Calculated correlation coefficients between each metric and points.
- Identified metrics strongly correlated with team performance (points).



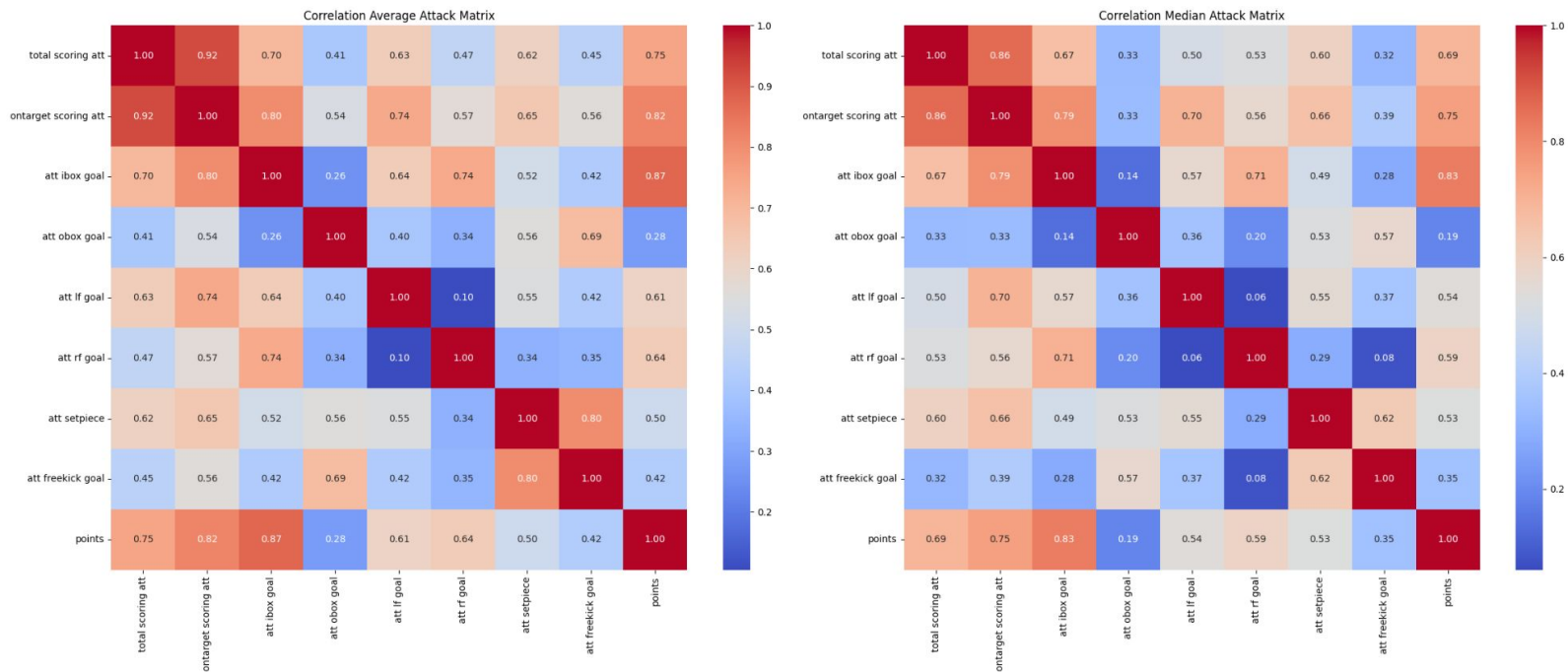
Correlation Goal HeatMaps

Goals METRICS



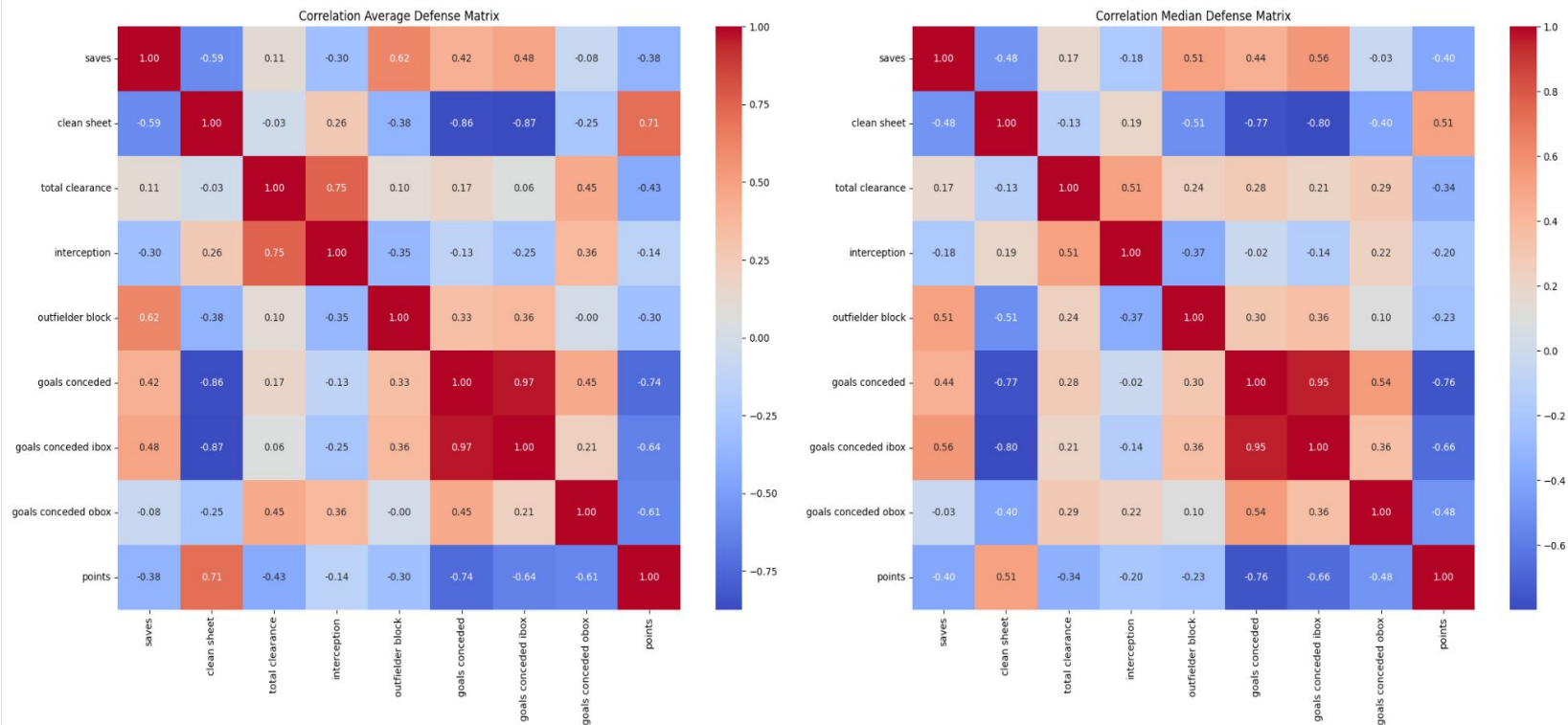
Correlation Attack HeatMaps

ATTACK METRICS



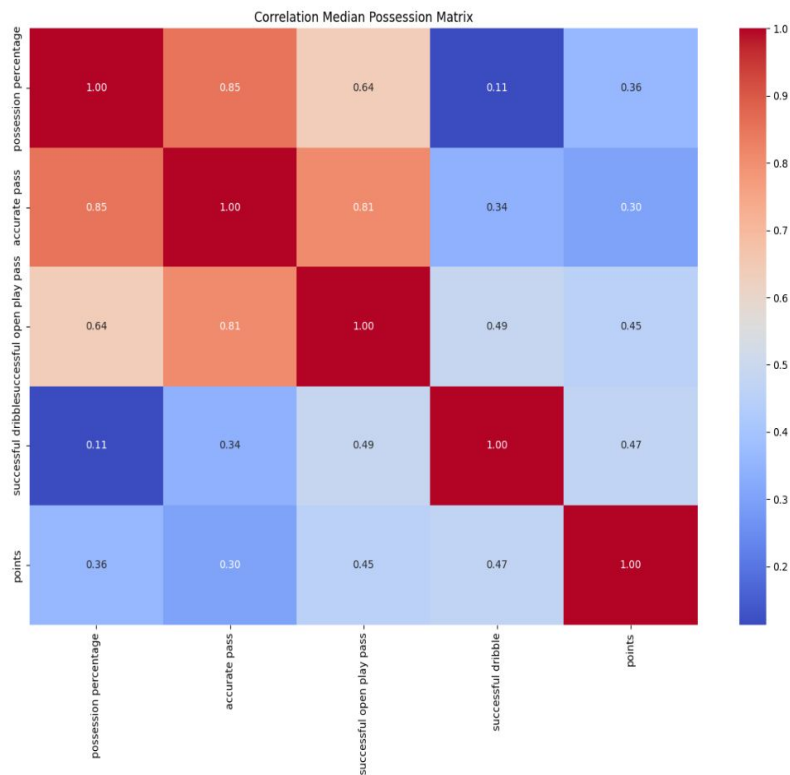
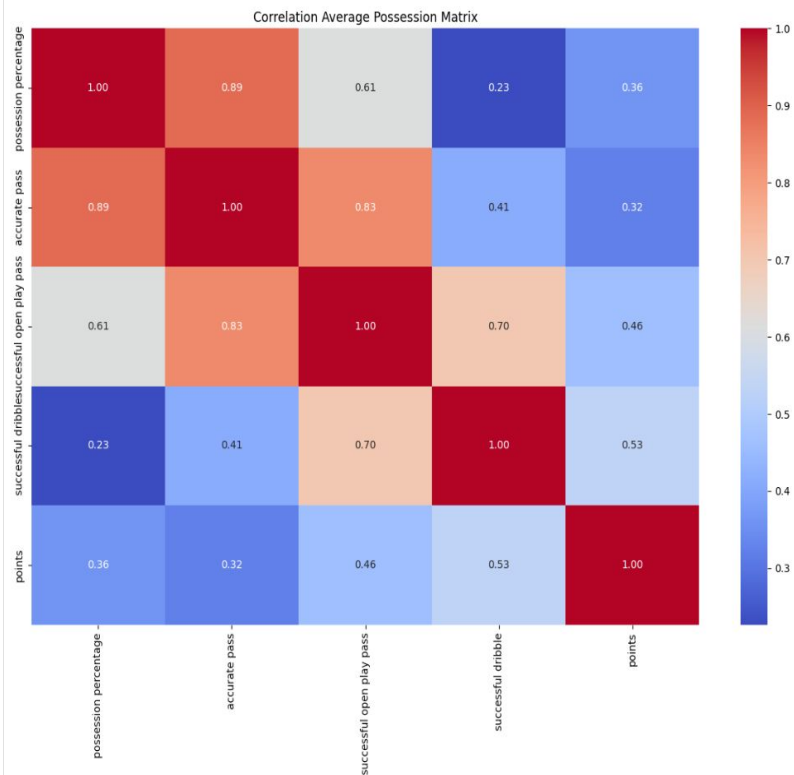
Correlation Defence HeatMaps

Defence METRICS

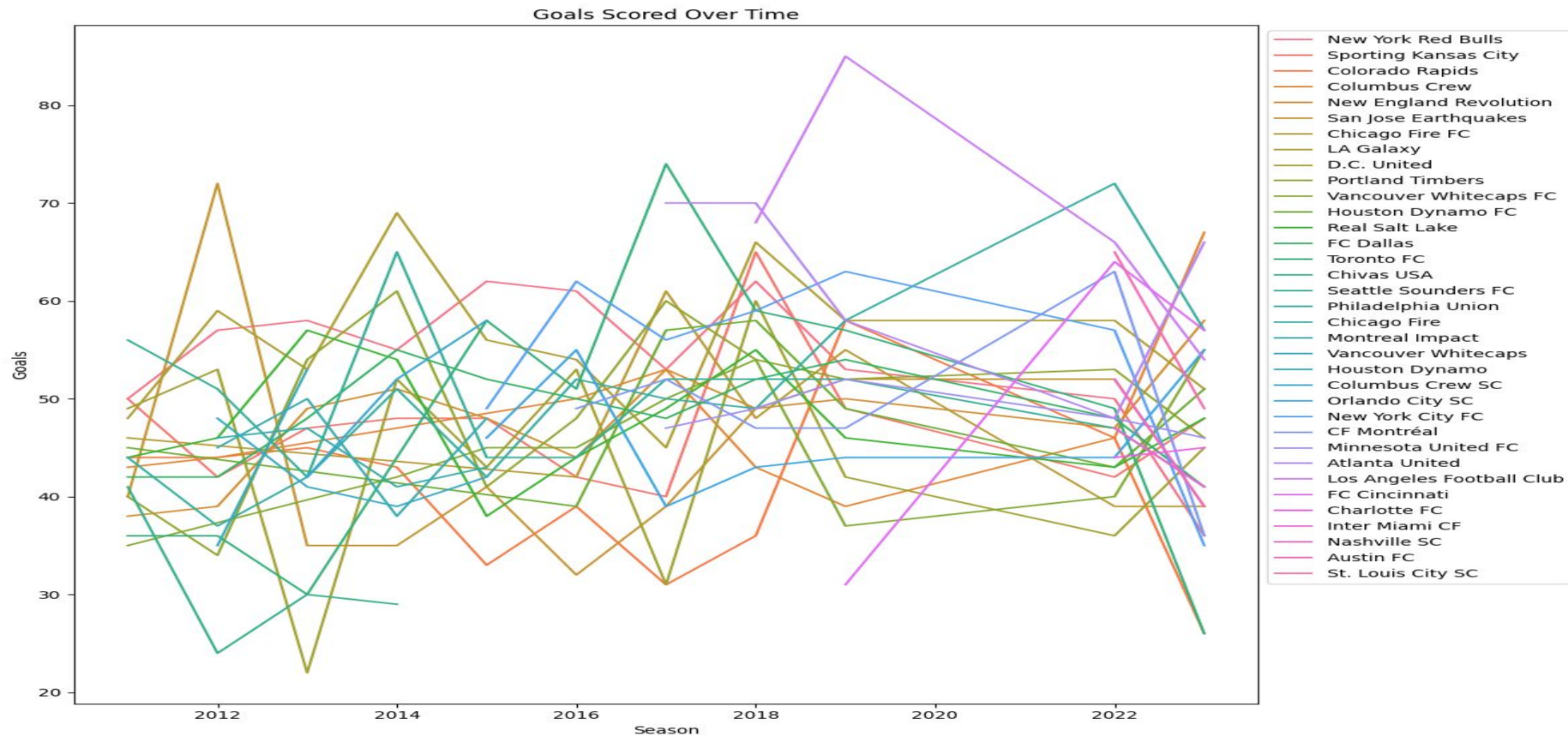


Correlation Possessions HeatMaps

Possessions METRICS

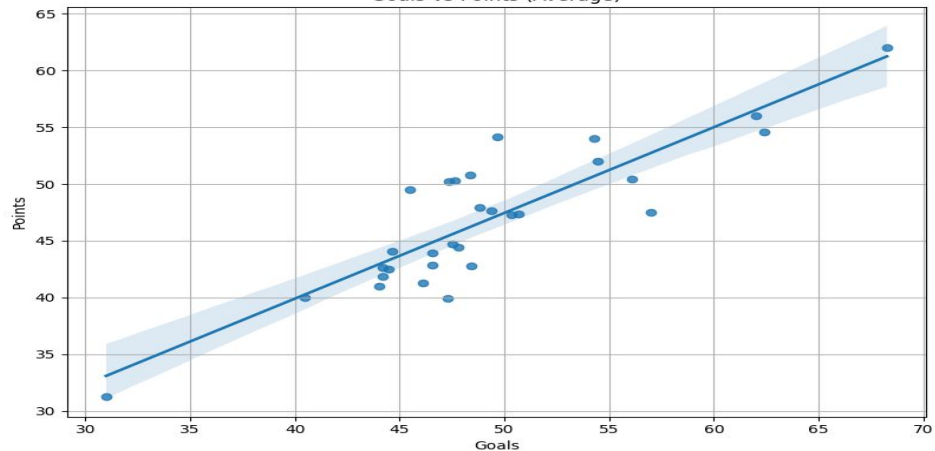


LinePlot

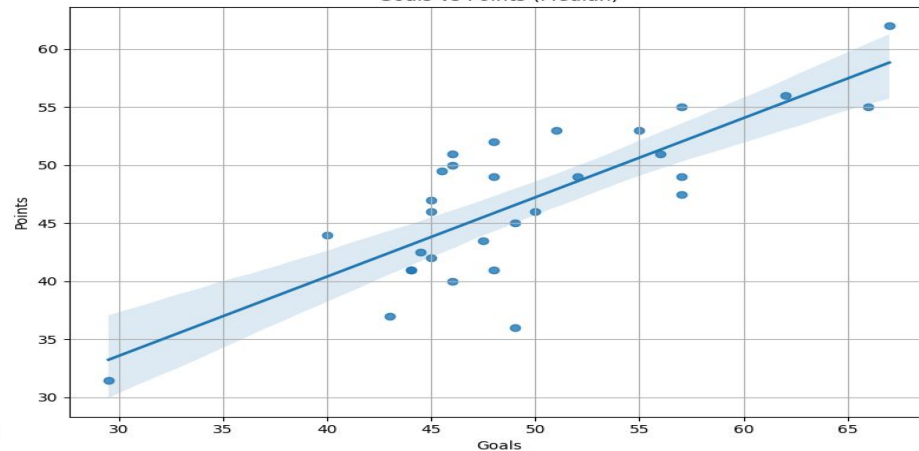


Scatter Plot

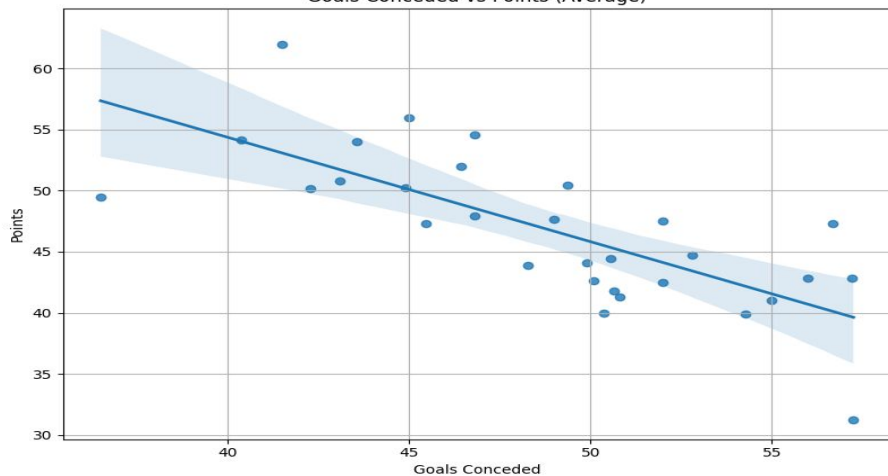
Goals vs Points (Average)



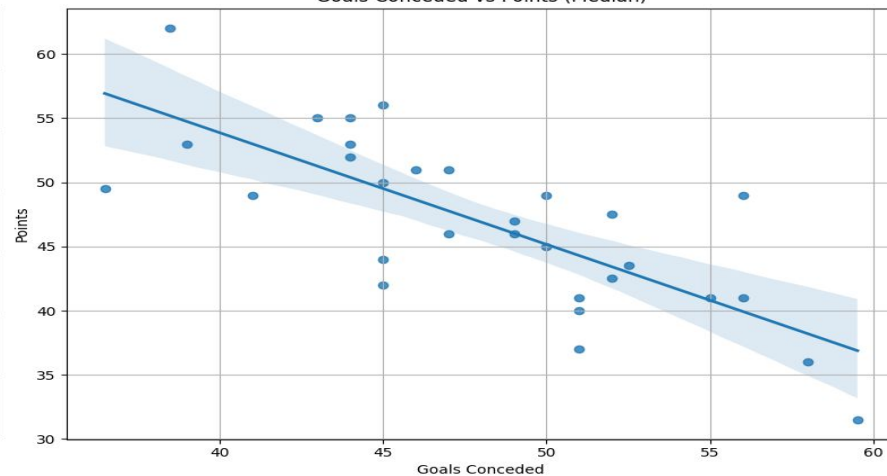
Goals vs Points (Median)



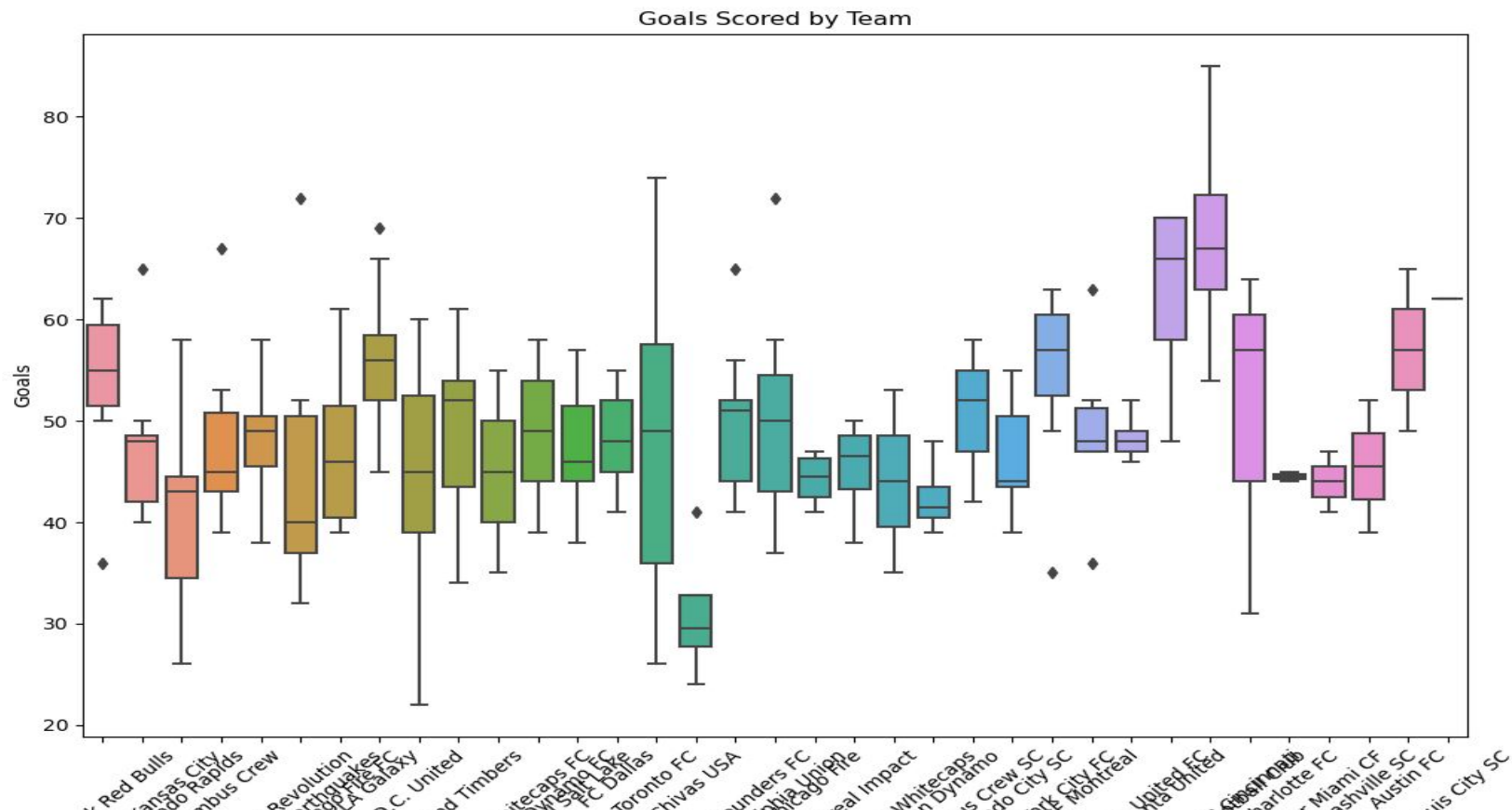
Goals Conceded vs Points (Average)



Goals Conceded vs Points (Median)



Box Plot



Machine Learning

Linear regression:

- Why: linear regression is chosen for its simplicity, interpretability, and ability to provide meaningful insights into the relationship between MLS team performance metrics and their standings

Weight Assignment:

- Weights were assigned to metrics within each category (goals, attack, defense, possession) to emphasize their relative importance in predicting team points.
- This approach allows for a nuanced analysis where certain metrics contribute more significantly to team success than others

```
goals_features = [  
    'goals', 'goals conceded', 'own goals'  
]  
  
attack_features = [  
    'on target scoring att', 'total scoring att', 'open play pass', 'blocked scoring att'  
]  
  
defense_features = [  
    'saves', 'total clearance', 'interception'  
]  
  
possession_features = [  
    'successful open play pass', 'successful short pass', 'accurate pass'  
]  
  
# Define weights for each feature within subsets (adjust as needed)  
weights = {  
    'goals': [0.80, -0.80, -0.10],  
    'attack': [0.15, 0.20, 0.10, 0.10],  
    'defense': [0.15, 0.10, 0.15],  
    'possession': [0.25, 0.25, 0.25]  
}
```

Model Performance Evaluation

Merged Dataset	Team Averages	Team Median																														
Mean Squared Error (Merged of Teams): 23.1479266906127	Mean Squared Error (Average of Teams): 2.536776088156992	Mean Squared Error (Median of Teams): 2.6335112921996484																														
Coefficients:	Coefficients:	Coefficients:																														
<table><tr><th></th><th>Coefficient</th></tr><tr><td>goals</td><td>0.752162</td></tr><tr><td>attack</td><td>0.000787</td></tr><tr><td>defense</td><td>0.008591</td></tr><tr><td>possession</td><td>-0.000050</td></tr></table>		Coefficient	goals	0.752162	attack	0.000787	defense	0.008591	possession	-0.000050	<table><tr><th></th><th>Coefficient</th></tr><tr><td>goals</td><td>0.691315</td></tr><tr><td>attack</td><td>0.012009</td></tr><tr><td>defense</td><td>0.048412</td></tr><tr><td>possession</td><td>-0.001247</td></tr></table>		Coefficient	goals	0.691315	attack	0.012009	defense	0.048412	possession	-0.001247	<table><tr><th></th><th>Coefficient</th></tr><tr><td>goals</td><td>0.721874</td></tr><tr><td>attack</td><td>-0.008910</td></tr><tr><td>defense</td><td>0.031992</td></tr><tr><td>possession</td><td>0.001453</td></tr></table>		Coefficient	goals	0.721874	attack	-0.008910	defense	0.031992	possession	0.001453
	Coefficient																															
goals	0.752162																															
attack	0.000787																															
defense	0.008591																															
possession	-0.000050																															
	Coefficient																															
goals	0.691315																															
attack	0.012009																															
defense	0.048412																															
possession	-0.001247																															
	Coefficient																															
goals	0.721874																															
attack	-0.008910																															
defense	0.031992																															
possession	0.001453																															

Weighted Score

Goals Metrics:

- Features: 'goals', 'goals conceded'
- Weights: [0.80, -0.80].

Attack Metrics:

- Features: 'ontarget scoring att', 'total scoring att', 'open play pass'
- Weights: [0.15, 0.20, 0.10]

Defense Metrics:

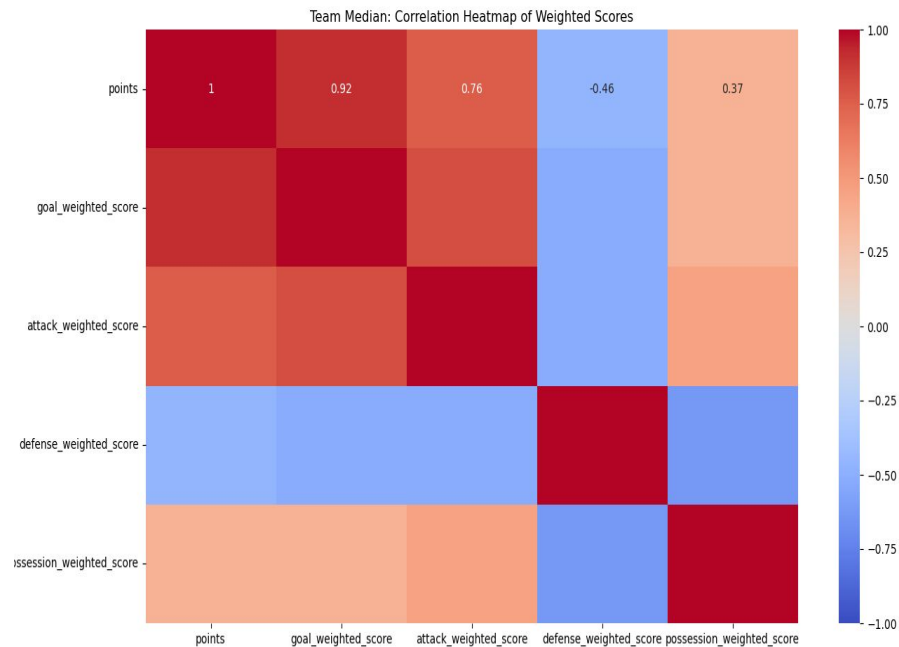
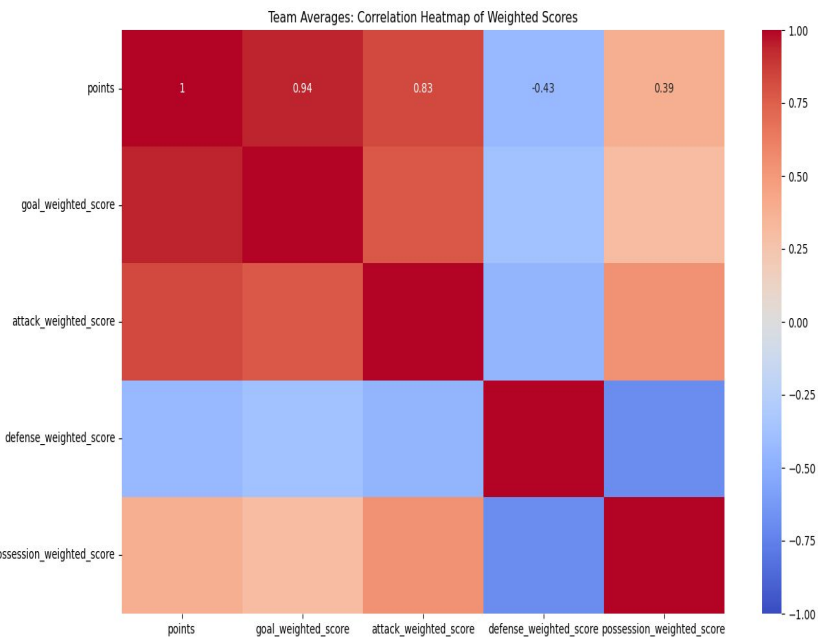
- Features: 'saves', 'total clearance', 'interception'
- Weights: [0.15, 0.10, 0.15]

Possession Metrics:

- Features: 'successful open play pass', 'successful short pass', 'accurate pass'
- Weights: [0.25, 0.25, 0.25]



Correlation of Weighted Score



ML with weighted score

Team Median

Team Median Model Performance:
MSE: 6.902673867950711
R-squared: 0.9131435654504629

Team Average

Team Averages Model Performance:
MSE: 0.8226640476756374
R-squared: 0.9899373017026717

Predicting current season team position in the league based on points

Current season data had to be normalized and weighted score for each metric have to be created.

Data as of July 22

Predicted team position:

club id	club name
8	Columbus Crew
12	LA Galaxy
32	Inter Miami CF
18	Houston Dynamo FC
19	Real Salt Lake
15	CF Montréal
23	Seattle Sounders FC
30	FC Cincinnati
21	Toronto FC
26	Orlando City
28	New York City FC
6	Sporting Kansas City
16	Vancouver Whitecaps FC
29	Atlanta United
9	New England Revolution
14	Portland Timbers
31	Los Angeles Football Club
20	FC Dallas
7	Colorado Rapids
11	Chicago Fire FC
33	Nashville SC
261248	Austin FC
24	Philadelphia Union
5	New York Red Bulls
1406988	Charlotte FC
27	Minnesota United
10	San Jose Earthquakes
13	D.C. United
2326207	St. Louis CITY SC

Limitations, Roadblock Encountered & Future work


Limitations

- **Seasonal Predictions:** Predictive models are more reliable during ongoing seasons when more matches have been played, providing a larger dataset for analysis and reducing the impact of early-season variability.

Roadblocks Encountered

- **Performance of Dask:** The usage of Dask for data processing was suboptimal due to the dataset size being relatively small for its distributed computing capabilities. This resulted in slower processing compared to more conventional tools like pandas, which could handle the data efficiently within memory.

Future Work

- **Incorporation of Player Statistics**
 - **Consideration of Team Form**
 - **Predictive Analytics for Match Outcomes**
- 

Q & A

