Project title: Major League Soccer Dataset Visualization

INFO I-590 Data Visualization

FALL 2021

Team Name: **MLSVisualizers**

Team members: 3

1) Chaitanya Shekhar Deshpande

2) Amit Kumar Patel

3) Abhinav Kumar

Contents

1. ABSTRACT

Major League Soccer is one the most widely watched sport in the United States of America. It has been gathering a large amount of interest not only within the country, but also throughout the world. The sponsors have also been investing a lot in this sport, thus making it rise as a highly promising sporting tournament in the USA apart from NBA and NFL. The dataset that we are downloading for this project, is being taken from Kaggle, with several features like: Goals scored, Goals conceded, Games won or lost, State-wise stadium attendance, etc. We believe that by visualizing these parameters using latest technologies, decision making for investors, football experts, talk shows, and of course team owners and managers could be easier, instead of simply studying the raw data there.

2. INTRODUCTION (MOTIVATION, BACKGROUND, AND OBJECTIVES)

The project is to visualize Major League Soccer, which is loved by many viewers, so it is going to be remarkably interesting. Since plots are based on how teams are performing and based on state-wise attendance, it will be interesting to see much loved the sport is in a particular area, also, we can conclude which team can be supported based on their performance. We further want to increase the popularity of soccer in the United States when compared to European countries. From a financial point of view, investors can also come to know which team can be invested in more heavily, and from a tactical point of view, team management can also find their jobs easier looking at these. Although the United States is home to four major professional sports leagues, the highest level of soccer (the world's most popular sport) is played primarily in Europe. Top international sports leagues, such as the English Premier League, have not only grown their business in their domestic markets, but also across every continent in the world, especially in North America.

Before beginning with the data visualization part, we will first have a look at the history of United States Soccer. From 1954 to 1990, the US was unable to qualify for the FIFA World Cup. However, with hosting the 1994 WC along with good revenue, the MLS was inaugurated. Statistics have shown great improvements since then. The USA qualified for the FIFA World Cup every time except in 2018, with the USA reaching the quarter-finals in 2002 as their highest achievement. MLS is certainly to be thanked for this performance, and with an increase in passion for soccer through MLS, attracting international talent and management skills, soccer is turning out to be one of the sports in extremely high demand here. With the 2026 FIFA World Cup also to be held in the USA, proper investment and skill improvement can lead to miracles for the USA Soccer team.

We have taken the MLS data from 1996 till 2020 from Kaggle, and with some visualizations which we have plotted along with the existing ones, analysts, investors, and fans can be highly motivated to be a part of this sport. It will help the investors to know where they can invest money to get more profit. Along with that, it will help the broadcaster to know which team and which state's people are more interested in the game, which will help them to grow their business more and earn more revenue out of it.

Soccer analysis is one of the most widely used things used by football managers, tacticians, and pundits all around the world. There are several TV shows which show different types of statistical analysis, visualizations, etc. However, our team will be focusing on the visualizations that are already present in Kaggle as the existing work.[1] For instance, this one is available, which plots the goals per 90 minutes for the highest scorer each season along with the goals as a Line plot.

We have used 4 datasets for our visualization which are all_goalkeepers, all_players, all_tables, and matches dataset. The all_goalkeepers and all_players dataset contain data about all the players like goalkeeper Name, Club, Position, Games Played, etc. This also contains the data of wins and losses they have faced during a match played. Furthermore, it also showed the time they played overall. The all_tables dataset contains data related to all the clubs like team name, win, loss, points earned, etc. The matches dataset contains information about the games which was played at a certain date. It has other details like a home club, away club, attendance, venue etc. In this dataset, we were able to see the home and home away scores. Through this dataset we can be able to see which states have larger audiences to watch the soccer game.

The existing visualizations for this dataset are Correlation Heatmaps and Pairplots for all the features between all the datasets, line plots for Goals scored per 90 minutes and total goals scored by the top scorer in each season, and a heatmap.

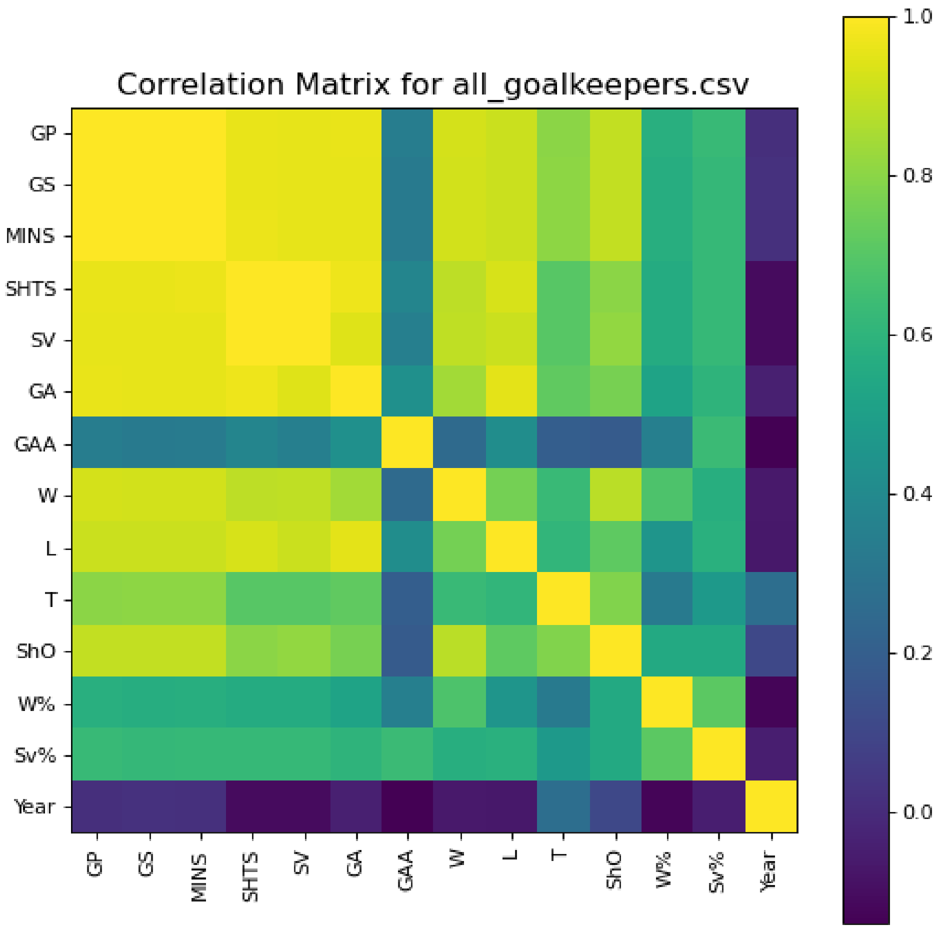


Figure 1: Correlation Matrix of all_goalkeepers

Correlation analysis is used to quantify the degree to which two variables are related. Through the correlation analysis, you evaluate correlation coefficient that tells you how much one variable change when the other one does. Using correlation matrix, we can see the importance of shots are more important while making a goal.

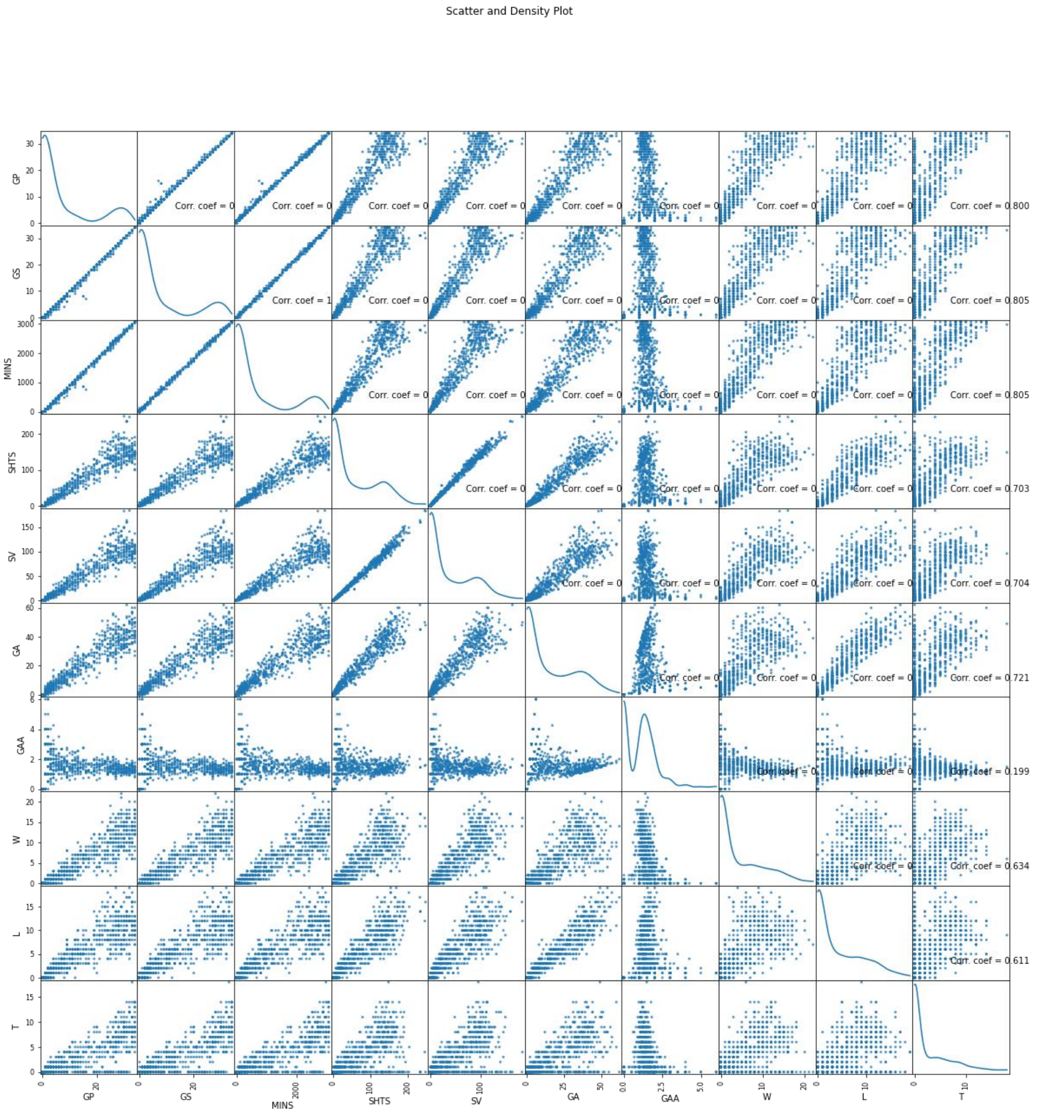
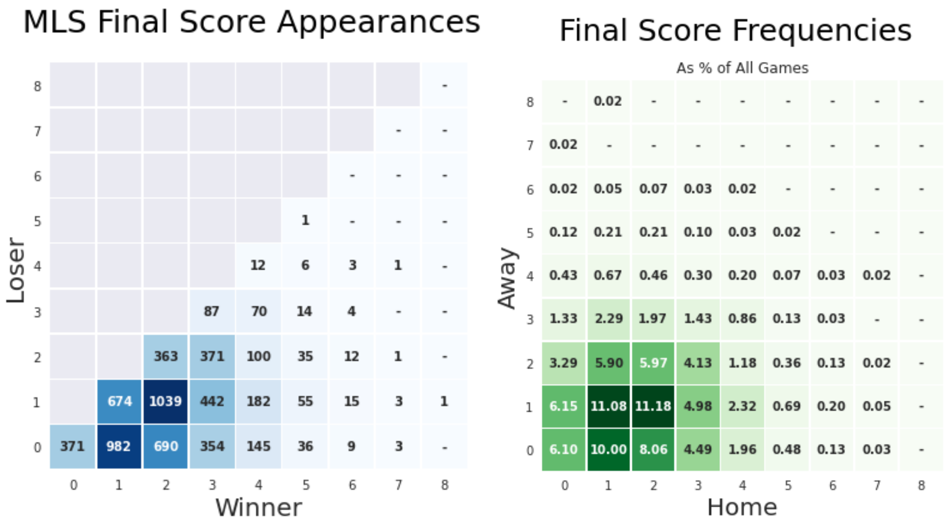


Figure 2: Pair plot

Another visualization example is of sns pair-plot as a scatterplot for all the datasets. Such plots are highly useful for Machine Learning, where we are able to find out the importance and correlation of all the features against each other in the form of scatter plots with the diagonal element of the plot showing the distribution way (normalized, etc.). However, the problem with this plotting method is Scalability, as the number of subplots is squared value of the number of features. With a dataset like matches.csv (defined later), it is impossible to plot the same, as the number of features are 209, so the number subplots here 209*209, making it really clumsy.



Loser	0	1	2	3	4	5	6	7	8
8									-
7								-	-
6							-	-	-
5						1	-	-	-
4					12	6	3	1	-
3				87	70	14	4	-	-
2			363	371	100	35	12	1	-
1		674	1039	442	182	55	15	3	1
0	371	982	690	354	145	36	9	3	-

Away	0	1	2	3	4	5	6	7	8
8	-	0.02	-	-	-	-	-	-	-
7	0.02	-	-	-	-	-	-	-	-
6	0.02	0.05	0.07	0.03	0.02	-	-	-	-
5	0.12	0.21	0.21	0.10	0.03	0.02	-	-	-
4	0.43	0.67	0.46	0.30	0.20	0.07	0.03	0.02	-
3	1.33	2.29	1.97	1.43	0.86	0.13	0.03	-	-
2	3.29	5.90	5.97	4.13	1.18	0.36	0.13	0.02	-
1	6.15	11.08	11.18	4.98	2.32	0.69	0.20	0.05	-
0	6.10	10.00	8.06	4.49	1.96	0.48	0.13	0.03	-

Figure 3: Score frequency heat maps

Another existing visualization is of the heatmaps of the score related attributes. The first one being the frequency of score type, with 2-1 as most frequent score. This approach is really good; however, it doesn't give much information on home-away scores, penalties, etc. So probably it could not provide much insights for us. Similarly, a home-away score frequencies was plotted, but again the frequency values were difficult to interpret, again not providing many insights.

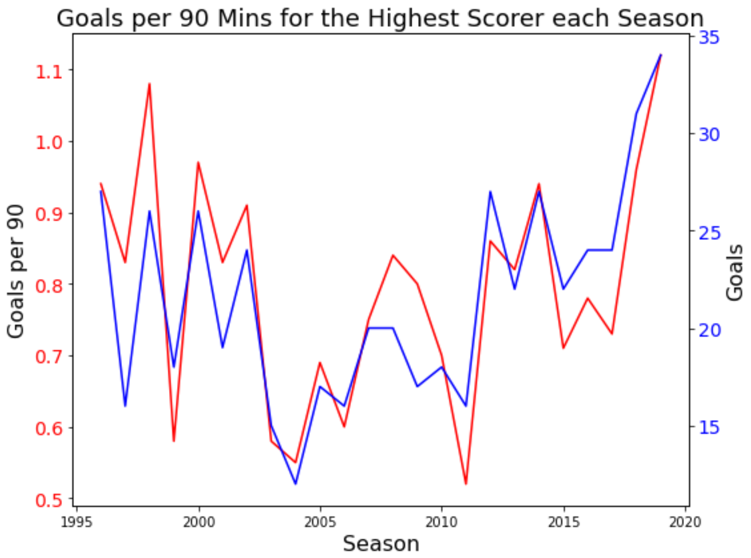


Figure 4: Goals per 90 minutes for Highest Scorer for each season

Another existing plot is of tracking the line plot of Goals scored per 90 minutes and total goals scored by the best player in the season. The advantage of this plot is that it talks about the effectiveness of the highest goal scorer every season, which is a highly useful attribute. In almost all cases, the highest goal scorer is scoring a goal every 90 minutes, except in a couple of seasons- 2004 and 2011. There is no critique against this plot as it is highly insightful, however, statisticians must note that a similar plot can be made for assists made by the midfielders, as they also play a highly important role in the buildup of the game.

With our plots and visualizations, we wish to display the time series analysis of several aspects of the game, choropleth geoplots representing the attendance and total goals scored in the form of a heatmap, frequency related analysis of several aspects such as Fouls Committed, Yellow Cards, etc., pie chart representing categories of the players, regression plots between highly related features, correlation plots of most important and similar features, and most notably, historical analysis of team performances, positions, goals conceded and scored. This can be utilized by soccer analysts, investors, soccer managers, etc. to improvise their business, improve team performance, etc. For example, if a particular managers performance has to be rated, so the time series analysis of the performance for a particular club for a particular time period, where the manager managed the team can be tracked. If consistently, they were ranked high up in the table, it means that he could be a great choice. Similarly, we can track the attacking/defensive mentality of the team based on the analysis of goals scored, conceded, etc. Through the choropleth heatmaps, we can find out which states have higher attendance in soccer, so marketing agencies can use that information to maximize profits in those states. It can also in improving the performance of a particular state with further analysis. Grouping similar values for correlation and pairplots would help Machine Learning Engineers and Data Scientist with feature engineering. Plus the scope of our project is infinite, so more and more insights can also be added to the same (some discussed in the last section).

3. PROCESS

For the new data visualization process, we decided to focus on updating the parameters on the existing ones, and then generate new plots using various attributes based on the 4 datasets that we have.

First, we will mention updating the existing methods. In the previous section, we saw about how a multi-line plot of total goals and goals per 90 minutes was represented for the best goal scorer every season. However, statisticians must also note the importance of assisting a goal, which is equally important especially because it is the most important part of the build-up behind a goal. So, our model will be focusing on the assist providers as well. It is generally the goal scorer who is in the spotlight, but it is time people also started considering and appreciating the importance of midfielders for award ceremonies as well. Hence, we have provided a plot of assists, and assists per 90 mins for the best assist provider across all seasons as well.

Also, regarding the correlation matrix and pair plot for each feature, it is highly essential for Data Scientists to analyze all the features, as they play a huge role in Feature Engineering and improving Machine Learning model accuracy. However, from a statistician's and analyst's point of view, we need to focus on important and relevant features. So, we have decided to split the pair plots into important features, for example like we group the data for stuff related to fouls: Yellow Cards, Red Cards, Fouls committed, Fouls suffered, and several other features as other methods.

Now we will discuss our niche plots and new plots. We decided to split our visualizations into 4 sections, based on our 4 datasets. We will discuss the plots one by one.

1. All Players Dataset: As mentioned above, this dataset contains the largest number of features above consisting of statistics such as Goals Scored, Assists Provided, Player Data, Fouls related data, etc, all related to outfield players. This visualization is one with endless possibilities, however, we decided to plot only the necessary ones for this project. We first began with pie chart with division done based on the type of player- Defender, Midfielder, Forward or a combination of any of those. Then, we have plotted the total number of goals per season by all teams as a line plot to show the trends of the attack performance of the MLS. Then we have plotted the pair plots and heatmap based correlation matrix of the relevant features of this dataset. Then we tried to analyze the goals scored, fouls committed through line plots. After this, we decided to implement a sns box-whisker plot-based visualization on the foul-related features like Fouls Committed, Yellow Cards per club etc. And finally, we plotted the sns lmplot showing the regression relation between Yellow Cards and fouls committed.
2. All Goalkeepers Dataset: In this dataset we focused on plotting line and bar plots for Saves by best goalkeeper in a season, Minimum goals conceded by the best goalkeeper per season, and a yearly average Goals Saved percentage for a combination of all teams, showing the goalkeeping performance of the MLS goalkeepers.
3. Matches Dataset: In this dataset, our focus was on plotting choropleth Geoplots for a state-wise analysis of stadium attendance and goals scored, using a heatmap-based approach. Along with this, the total attendance throughout MLS throughout the USA is also plotted.
4. All Tables Dataset: This dataset focuses largely on the time-series analysis of a particular team based on their Positions per season, Goals scored, and conceded per season. For this visualization, we have written Python functions, where we can simply pass the name of the team about whom the analysis information is required. This is highly useful to identify trends in performance across years and to analyze the key factors behind the success/failure of a particular team for a particular time period.

While implementation of the visualizations and plots for the above values, we faced several issues, failures, etc., out of which we decided to plot only the successful ones. Several updates were also made to clean the data like removal of NaN values, etc.

1. For All players plots, the issue with the dataset was that players had their nationality-related information also provided with it, and that too getting placed in the Club column, due to which needlessly the country-related performances were also getting plotted, in the box-whisker plots and lmplots.

2. For All Goalkeepers, the problem was the clubbing of different goalkeepers of the same club in a particular season, thereby causing quite a trouble to get perfect results. Hence we decided only to plot based on the Goals Conceded and Save% for the best goalkeeper throughout the season.
3. For Matches dataset, large amount of preprocessing was required. To avoid too many complications such as bankrupt teams, incorrect names of the clubs, etc., we only decided to plot the attendance and goals on the choropleth map only for 2019 season. We used groupby and pandas merge functionality to achieve the result. One of the failed attempts was to distinguish internal between the same state, like California had 2 clubs in the same city: LA Galaxy and LAFC both in Los Angeles, however, we couldn't differentiate between the same city on the USA map scale, so decided to take the average in this case, again by applying the group by the tactic.
4. For All tables dataset, the issues we faced were incorrect team names, teams no longer playing in MLS, teams which have been found recently, etc. So, plotting the performance of all teams in a single plot led to many inconsistencies, so we had to discard the same. Hence, we only decided to analyze teams that have been a part of the MLS from 1996 till 2020.

Thus, these kinds of preprocessing methods were used with several inconsistencies. The below section of Results and Insights will explain how our visualizations turned out and how we could infer information out of those.

4. RESULTS AND INSIGHTS

4.1 All Player's dataset

First, we will be beginning with the all_players.csv dataset. Just like how it is done in Exploratory Data Analysis, we thought of plotting the ratios of the number of Defenders, Midfielders, Forwards, Combinations of each type using a pie chart.

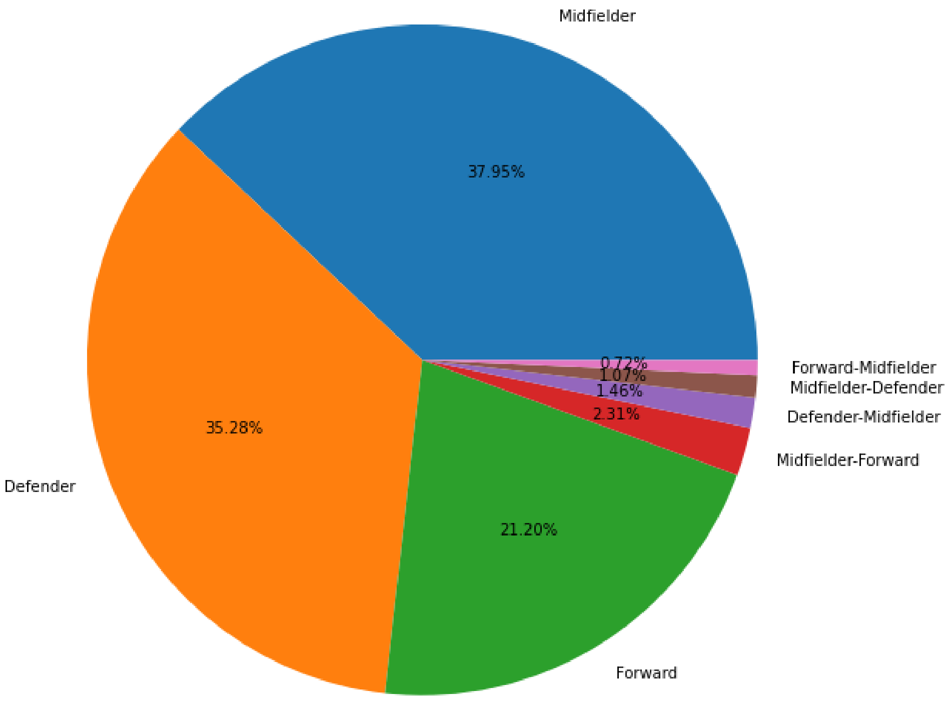


Figure 4: All Players Dataset

It turns out that most MLS players are Defenders and Midfielders, totaling around 73%, whereas the combination of player types is extremely low. If we consider international standards, players must be flexible enough to take up more than 1 role, because that can add balance to the team, which could lead to higher performances. Also, more focus must be placed on Forwards as well, to improve Goal Scoring.

Next, we have a time series line plot of total number of goals scored throughout the MLS season.

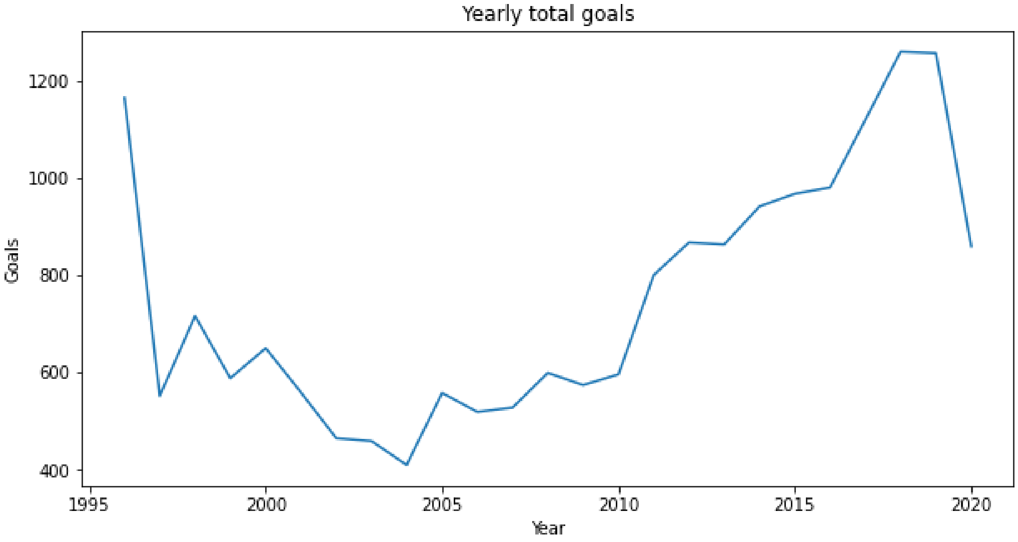


Figure 5: Yearly total goals

It turns out that the last 2-3 seasons have had more than 1200 goals scored, showing tremendous increase in the attacking play of all teams, which means that the MLS is on correct track towards attacking football.

As discussed in the previous section, we decided to only use the key parameters for all players towards goals scoring and plot the correlation matrix and pairplots for those specific features.

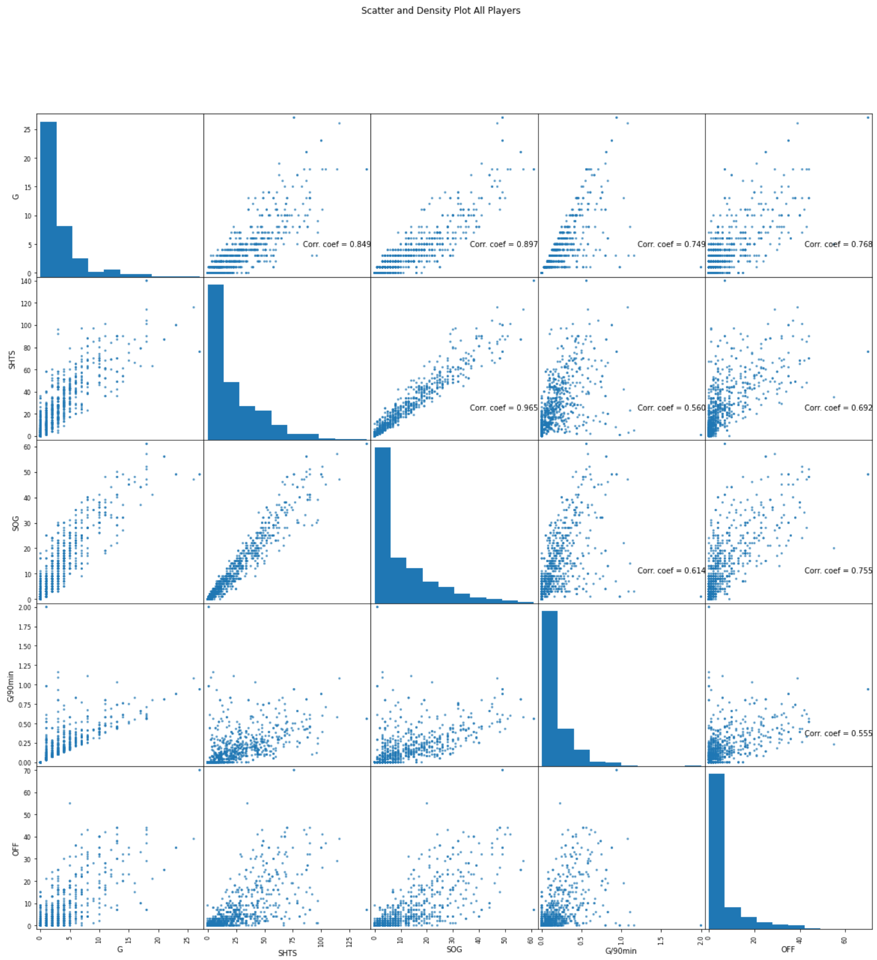


Figure 6: Pair Plot

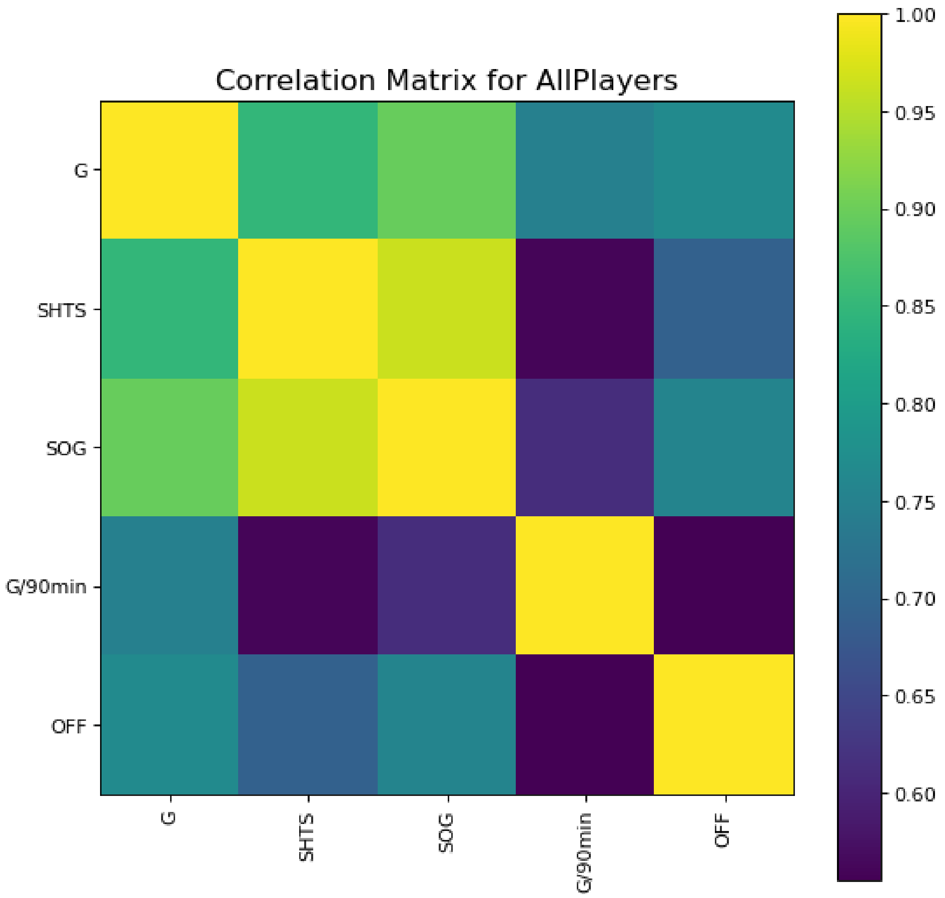


Figure 7: Correlation Matrix for AllPlayers

We chose the key parameters for goal scoring such as Goals, Shots, Shots on Goal, Goal per 90 minutes and Offsides. Offsides are missed chances if there is no opposition player between the goal keeper and the striker before the decisive pass was made, hence it has also been included in the same. This can be used to build great machine learning models, as it can be used in feature engineering.

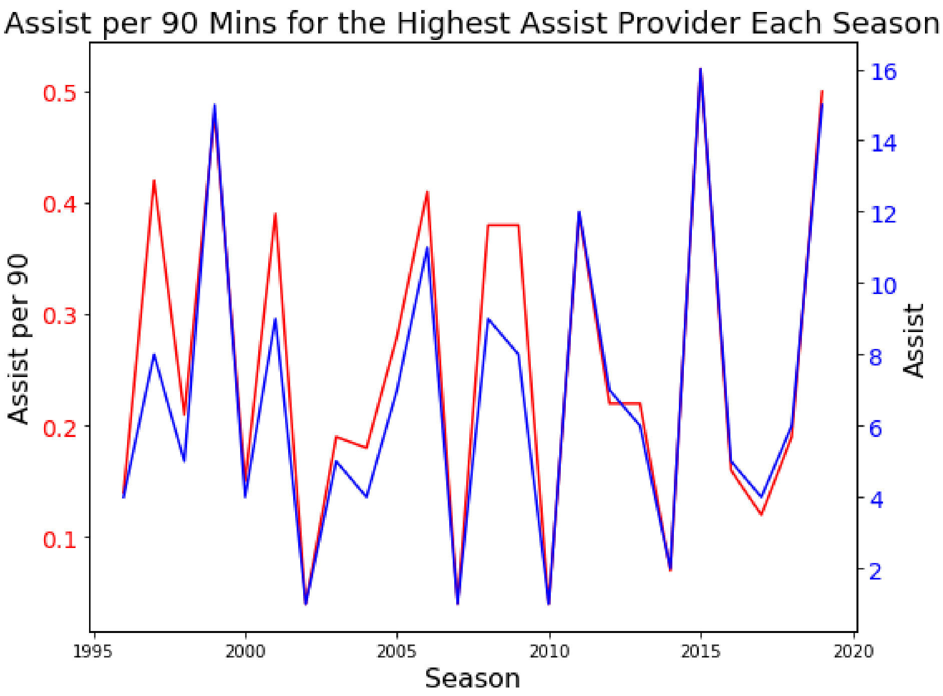


Figure 8: Assist per 90 minutes for the highest assist provider each session

The above plot is an update of the existing visualization, this time we assessed the assists and assist per 90 minutes for each season for the highest assist provider. It turns out for many seasons that, generally in every 2/3 games, the highest assist provider provides the maximum number of assists, because the average value oscillates between 0.25 and 0.5, so if we consider 0.5 assist per match, it basically means 1 assist per 2 matches.

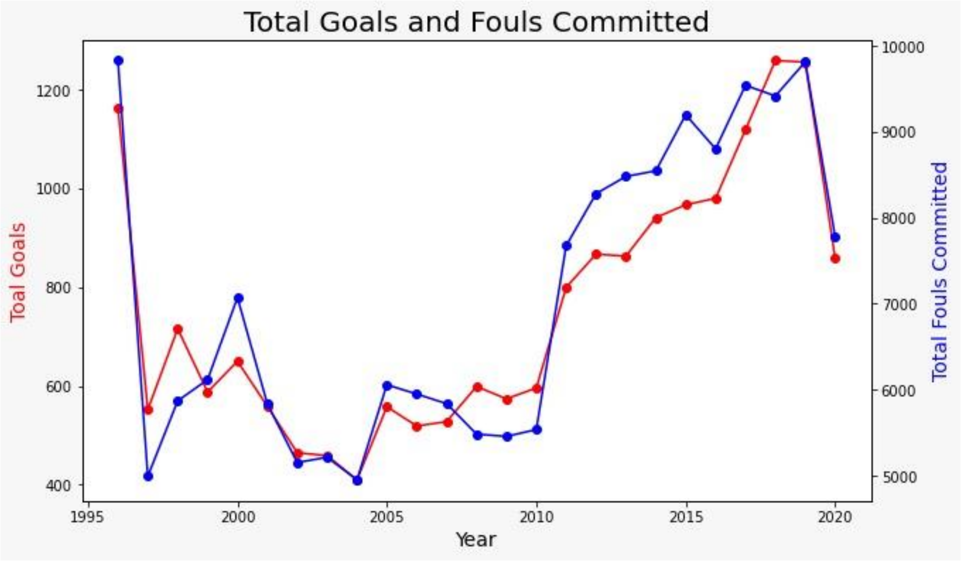


Figure 9: Total Goals and Foul Commited

The above visualization captures two metrics in the same plot. The line plot in blue represents the total fouls committed over years. Similarly, the line plot in red represents the total goals over years. We can observe that the number of fouls committed, and total goals are proportionally increasing after the year 2010. It is expected as it is in line with the increasing number of matches played. As the number of matches played increases so is the number of fouls committed and total goals.

Now, we decided to plot some distribution related values, so we went ahead with box-whisker plots for each club throughout all the seasons that they played. However, the problem with this one was that international teams, like USA, Panama were termed as “Clubs” for a particular player, which is incorrect.

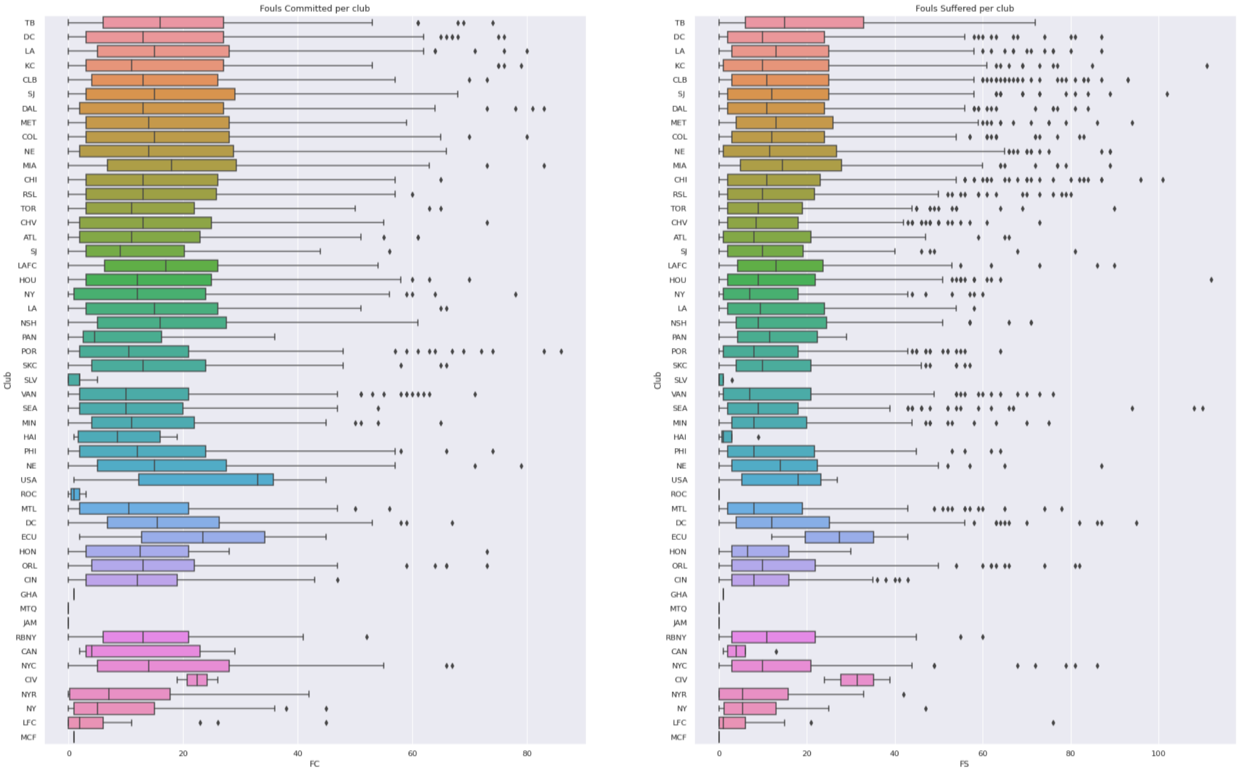


Figure 10: Foul committed per club vs Foul suffered per club

The left side graph shows the data of the fouls committed per club and the right-side graph shows the data the fouls suffered per player, per season. We can see that the San Jose club committed the maximum number of fouls. On the other hand, Tampa Bay suffered the maximum number of fouls. By this plot, we can see which are the most efficient clubs of all. Nation wise USA seem to commit maximum fouls, whereas Ecuador seems to suffer maximum fouls.

Next, we repeated the plot for Yellow Cards per club and Nation.

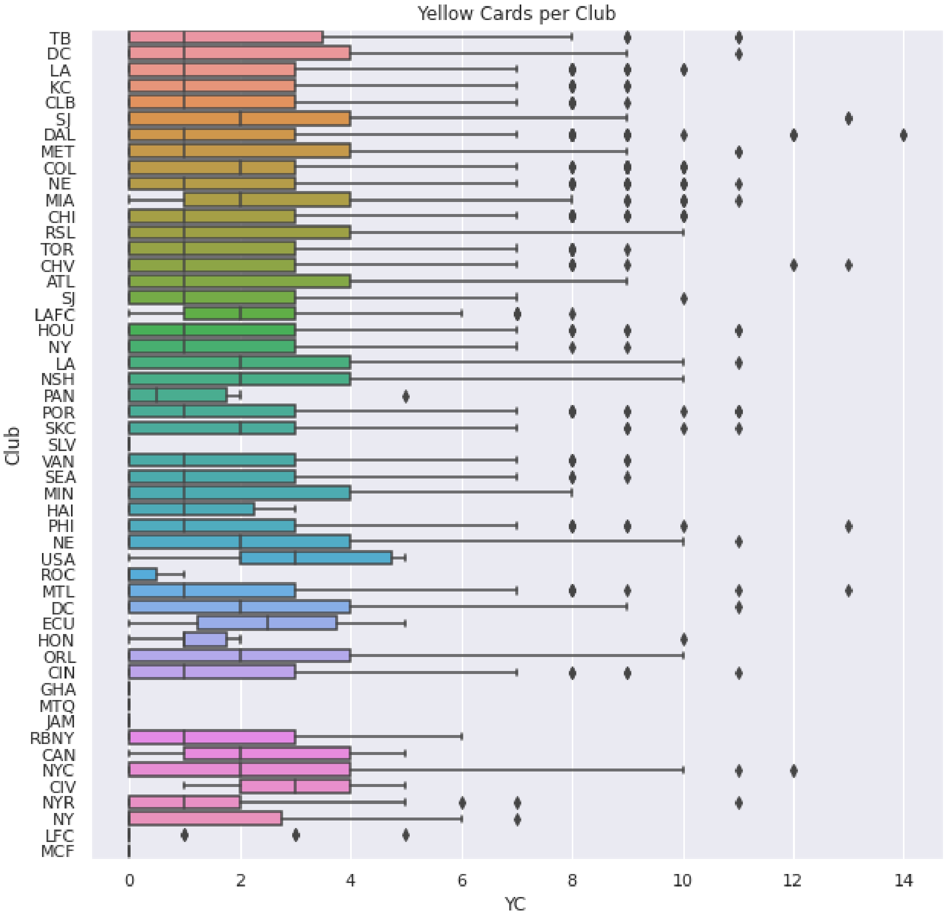


Figure 11: Yellow cards per club

Many clubs such as DC United, San Jose Earthquakes, LA Galaxy have very high number of Yellow Cards per player per season.

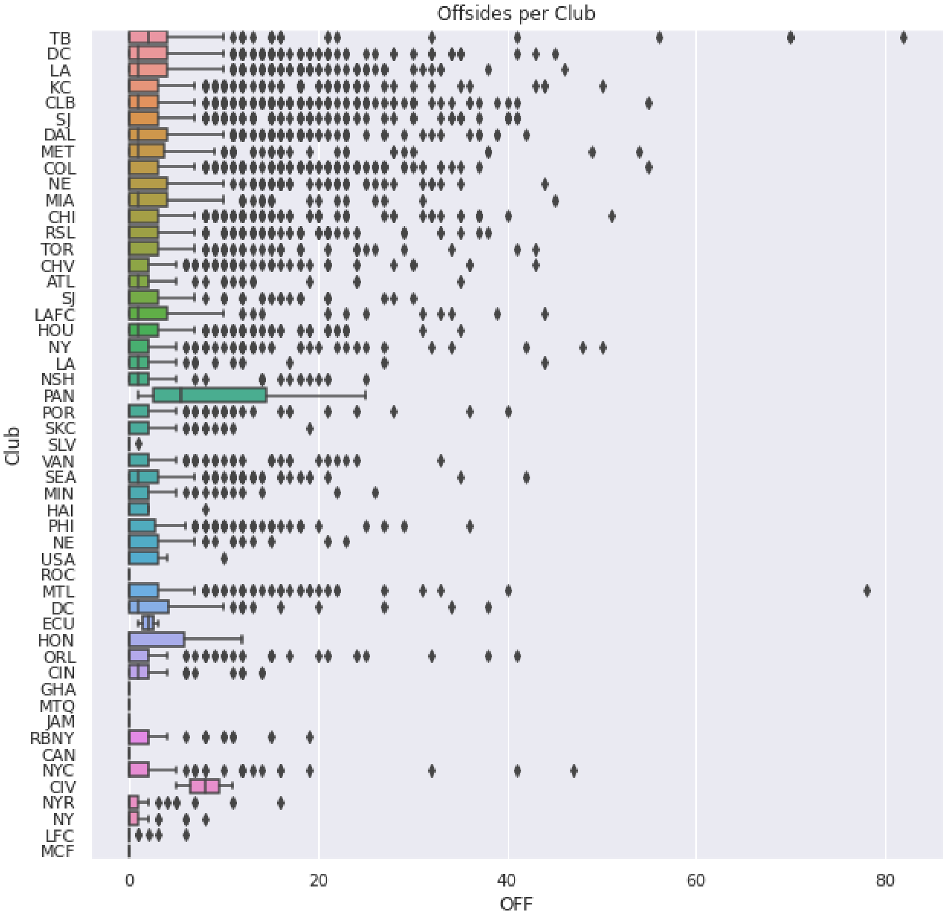


Figure 12: Offsides per club

Next, we evaluated the Offsides per club. Tampa Bay, Dallas, LA Galaxy, Dallas and Miami seem to have many offsides calls against them.

As a final plot, we decided to perform a regression analysis-based plot of sns in Python called the lm plot. We chose 2 features- Yellow Cards and Fouls Committed per club per season, which are almost directly relevant to each other.

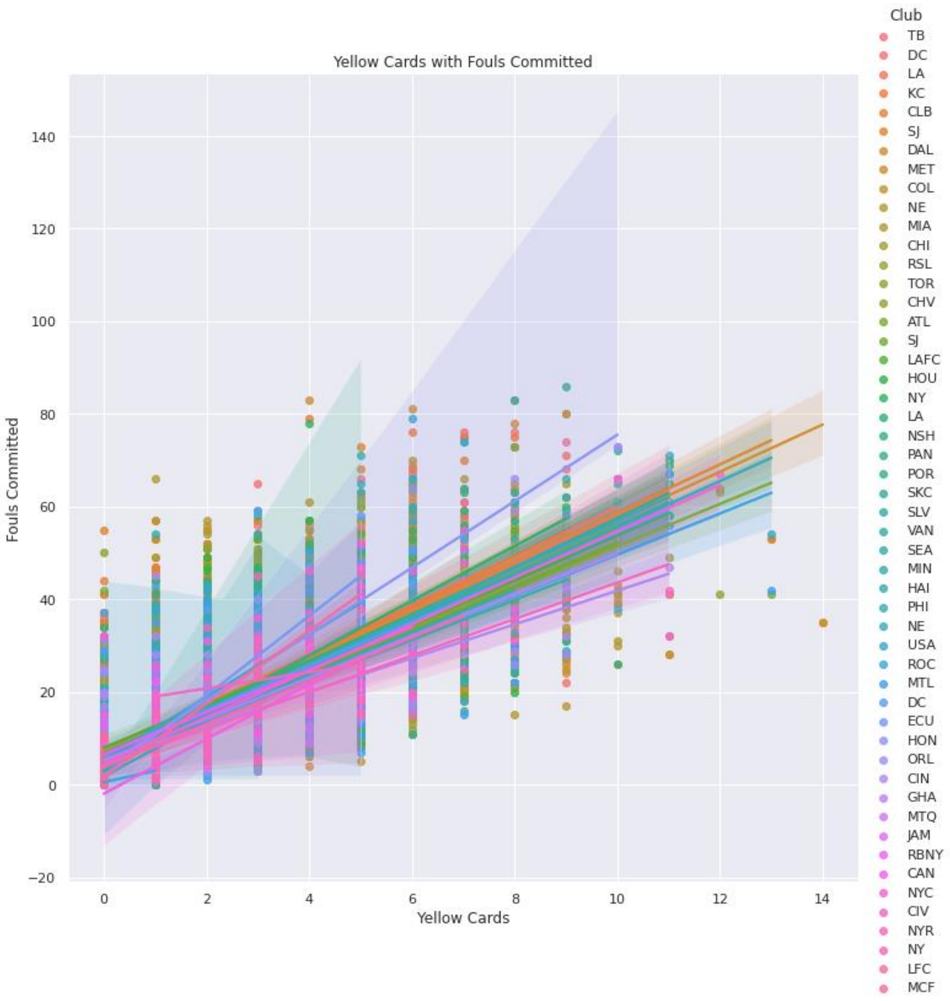


Figure 13: Yellow cards with foul committed

It turns out to be a very nice plot, showing a really nice regression-based line along with the scatter plots for the Yellow Cards against the Fouls Committed. This visualization can also be used to analyze the referees for a particular game.

4.2 All Goalkeepers dataset

Next, we decided to focus on the all_goalkeepers.csv dataset. The first ones were again a part of the exploratory data analysis. For this dataset, we chose Games Won, Lost, Tied, Saves and Goals against as the key features.

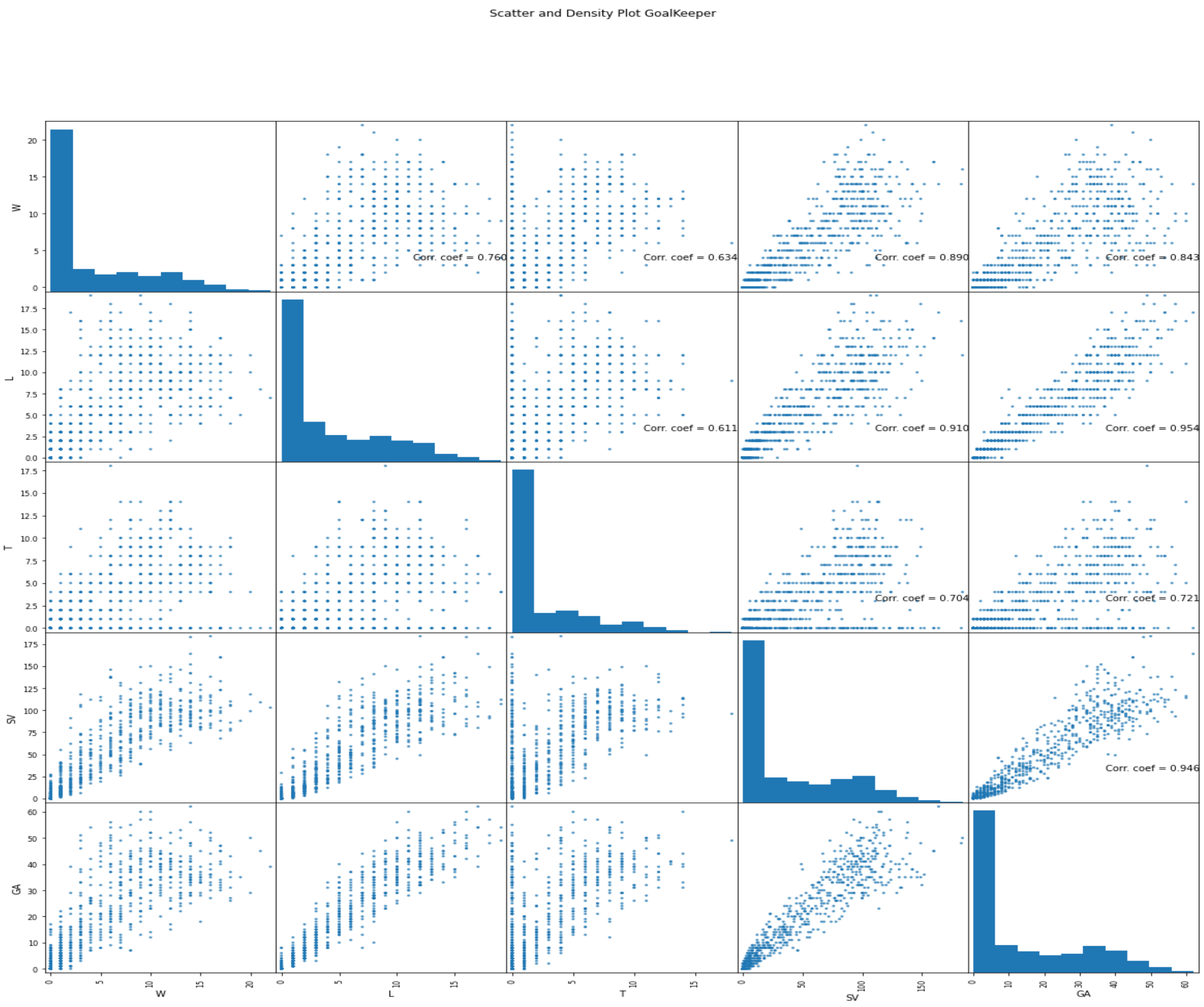


Figure 14: Pair Plot Goalkeeper

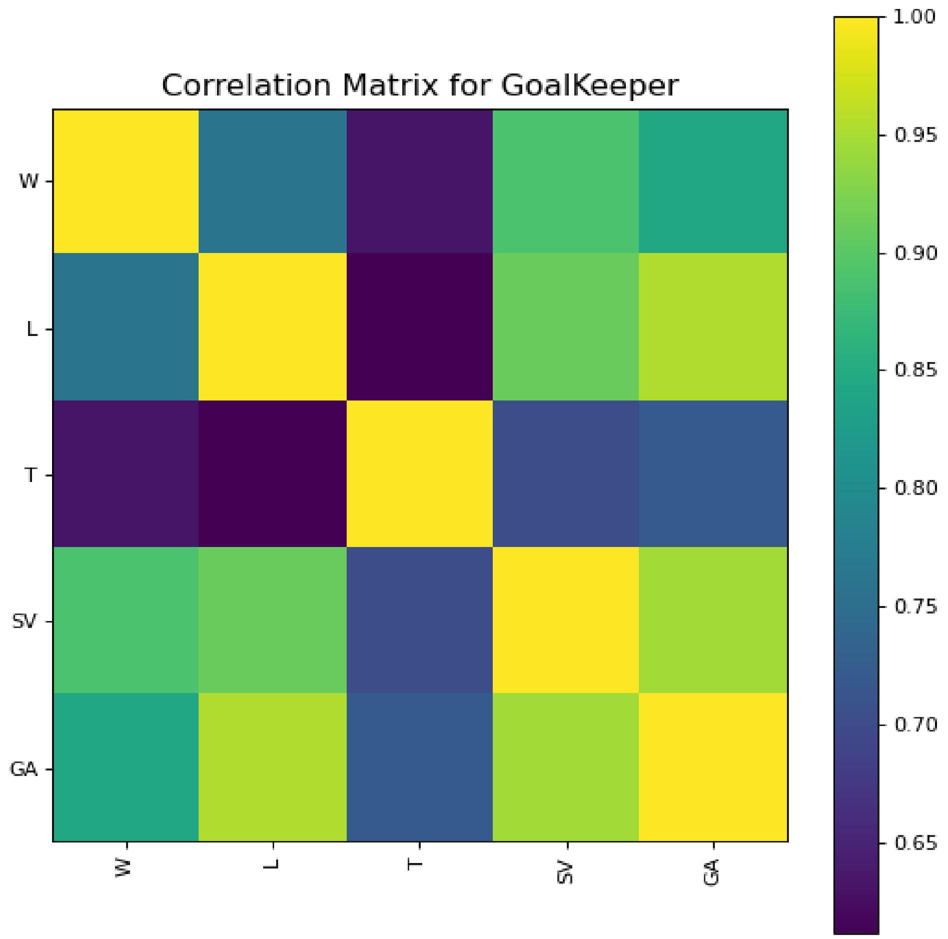


Figure 15: Correlation of Matrix for Goalkeeper

After that, we decided to focus on the best goalkeepers in each season, and the minimum average of goals conceded by a goalkeeper per season.

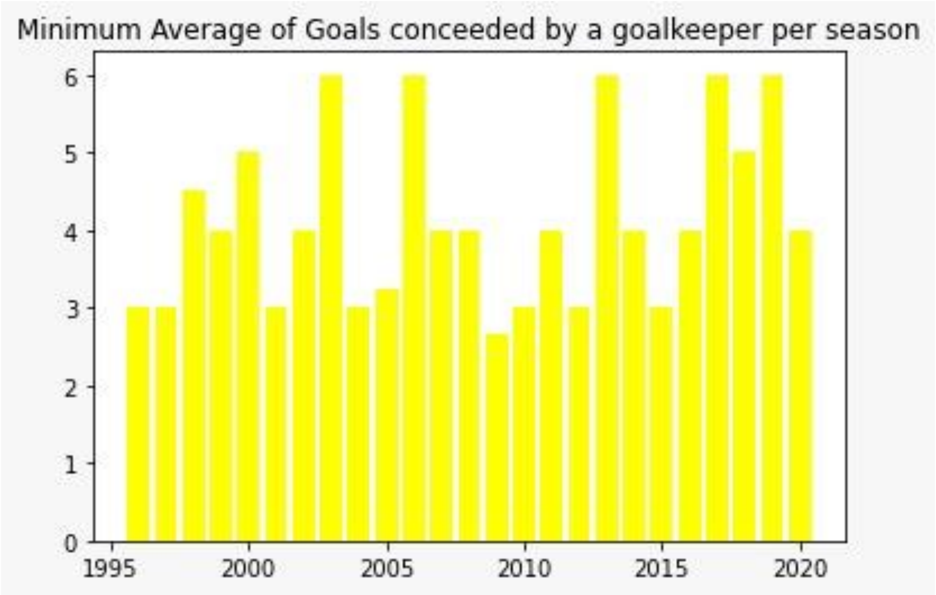


Figure 16: Minimum Average Goals conceded by a goalkeeper per season

The visualization shows the yearly minimum average of goals conceded by a goalkeeper per season. For most of the years the minimum average is around 3 and average 6 goals for some of the years in between.

Similarly, we plotted the bar graph for the total number of saves by the best goalkeeper per season.

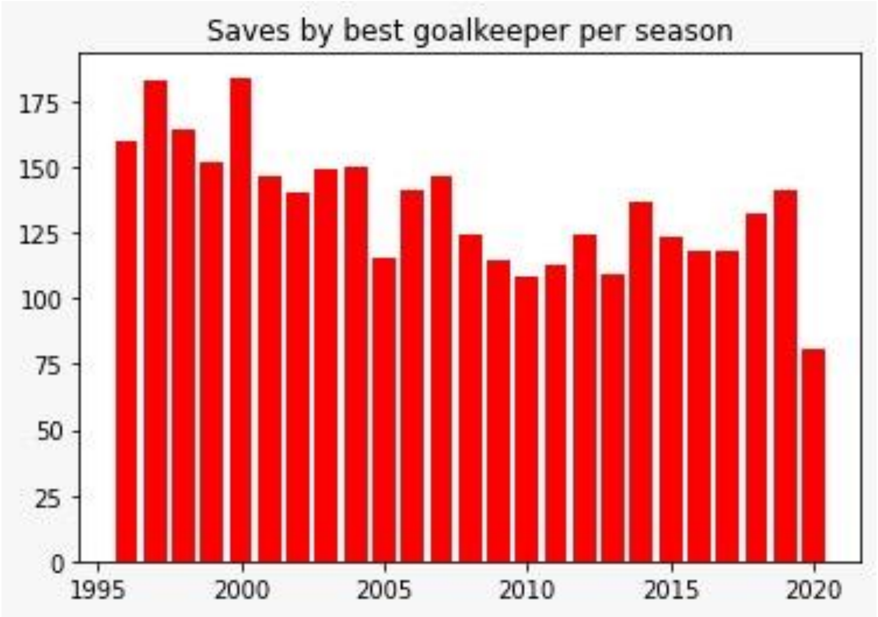


Figure 17: Saves by best goalkeeper per season

The saves by best goalkeeper per season seems to follow a bimodal distribution. The yearly trend is high for the starting years, which then has a dip in the middle and is observed to have again increased for the latest years. The last data point which year 2020 has the lowest of all the years because of the incomplete year data.

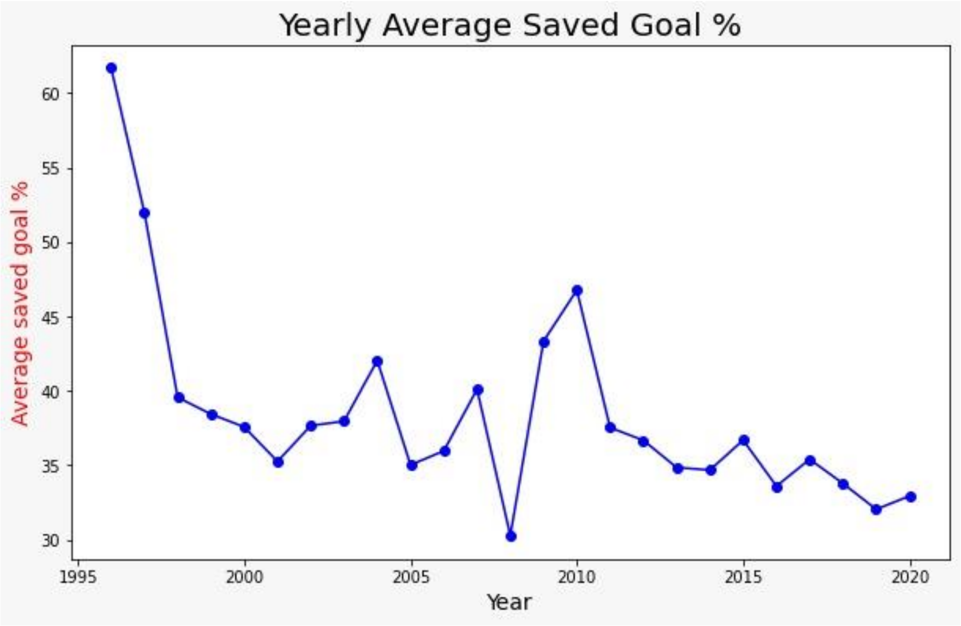


Figure 18: Yearly Average Saved Goal %

The above visualization just captures the yearly trend of average saved goal percentage. As we have seen earlier that the number of matches played has been increasing since the year 2010, however, the yearly average saved goal percentage does not have a high lift.

4.3 Matches dataset

After this, we decided to focus on the matches.csv dataset.

Next, we decided to go with a niche idea, something which we had great expectations for, but unfortunately, it could really work in the best way. We decided to use plotly.express choropleth geoplots for this dataset for a state wise analysis. [5] The 1st geoplot is of the average stadium attendance across all stadiums, with the colors of the map indicating the attendance values. Darker the color of the scale, lower the attendance, whereas lighter the color higher the attendance. We only chose to plot for 2019 for higher accuracies, again because of inconsistent data from last few years, teams getting bankrupt, etc.

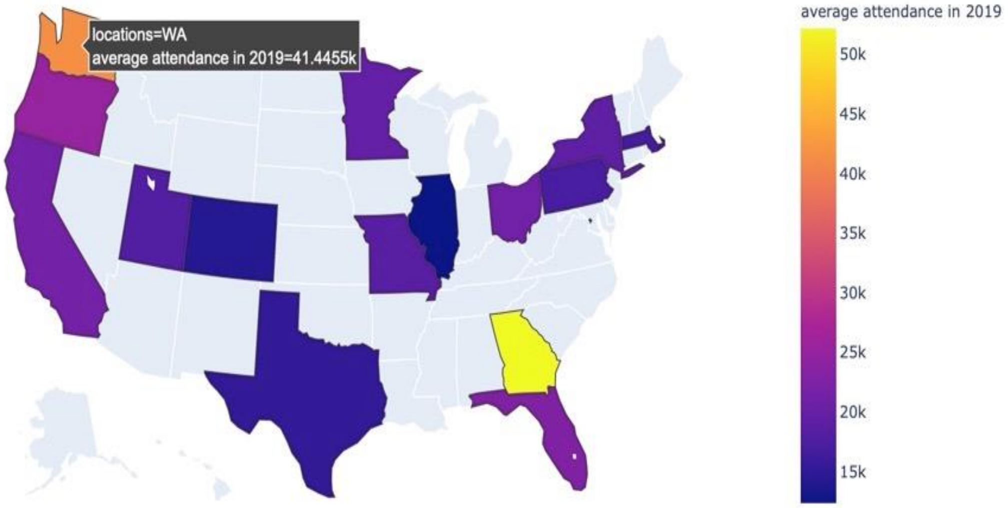

locations=WA
average attendance in 2019=41.4455k
average attendance in 2019
50k
45k
40k
35k
30k
25k
20k
15k

Figure 19: Average attendance in 2019

It turns out that states like Georgia, Washington have really high average attendance, whereas states like Texas have lesser attendance. This analysis can used by investors and marketing experts to try and increase the revenue of such states. They could perform research as to why states like Georgia and Washington have higher average attendance (40,000+).

Next, we plotted the number of goals scored state wise in 2019. For states like California, where 3 different clubs play, we have taken the mean value of all clubs into 1, for a fair comparison between all states.

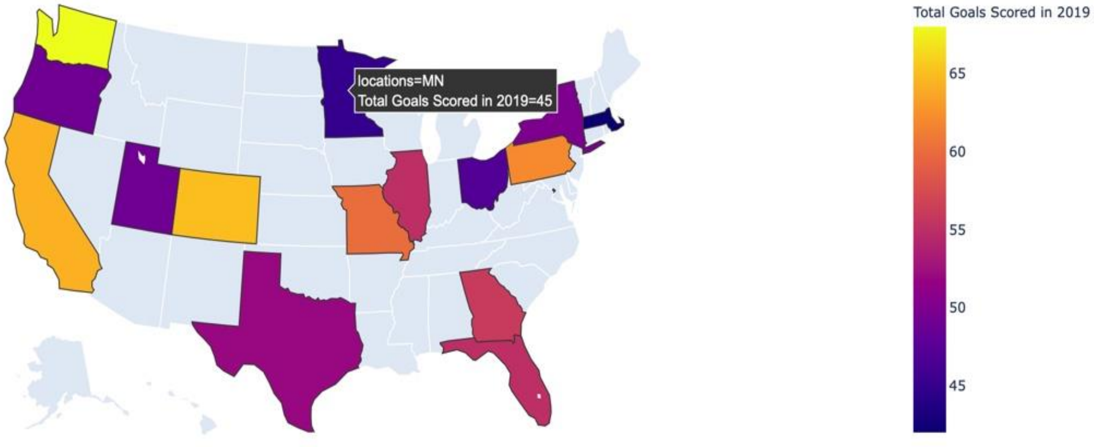


Figure 20: Total Goals Scored in 2019

It turns out that Washington state wins. They had the maximum number of goals (around 70) scored by the club in that state, with the next best being Colorado and California. This concludes that these clubs are excellent in attacking football. States like Minnesota had the least number of goals scored by their team, meaning that more focus must be done in their goal scoring skills.

Since there wasn't any data about defensive statistics in matches.csv, we could only plot the attendance and goals scored for clubs in that state.

Even though the idea was really nice, and plot is beautiful, the big issue with this one is the lack of states. Only 14 out of 50 states got colored there, meaning a lot of soccer needs to be still developed in the USA. Our dream would be to see each map being filled with some color in the future.

Next, we have plotted a time series analysis of the total attendance of MLS for each year.

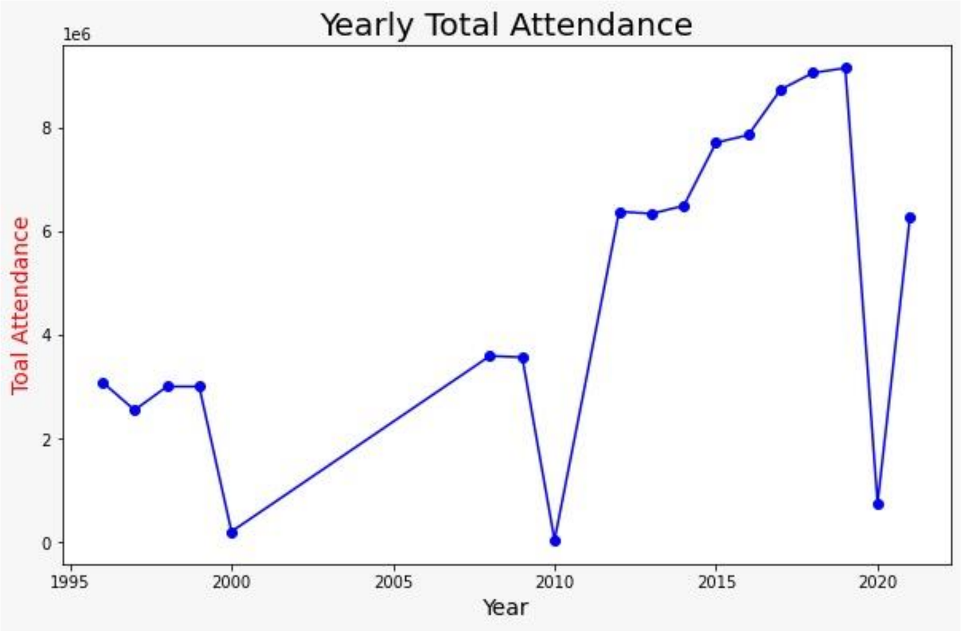


Figure 22: Yearly Total Attendance

The above visualization shows the yearly trend of total attendance. The dip in the years 2000, 2010, and 2020 is because of the data not captured in the data set. It is also not advisable to directly impute the data for the missing years with the mean value or median value as it will not capture the true representation because of the increasing number of matches over the years. The total attendance is substantially increasing every year, making it a great sign for soccer in the USA.

4.4 All Tables dataset

First, we decided to focus on the standard EDA which we have done for all the datasets with only the key features. The features we chose are Games Played, Won, Lost, Goals For, Goals Against and the Year.

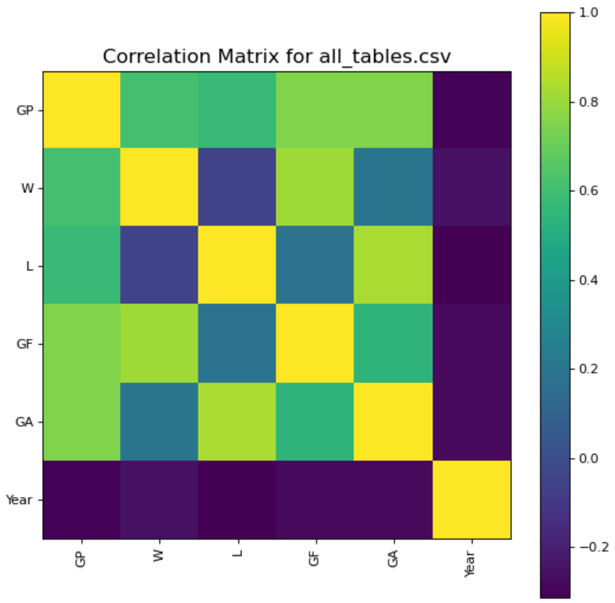


Figure 23: Correlation Matrix for all_tables

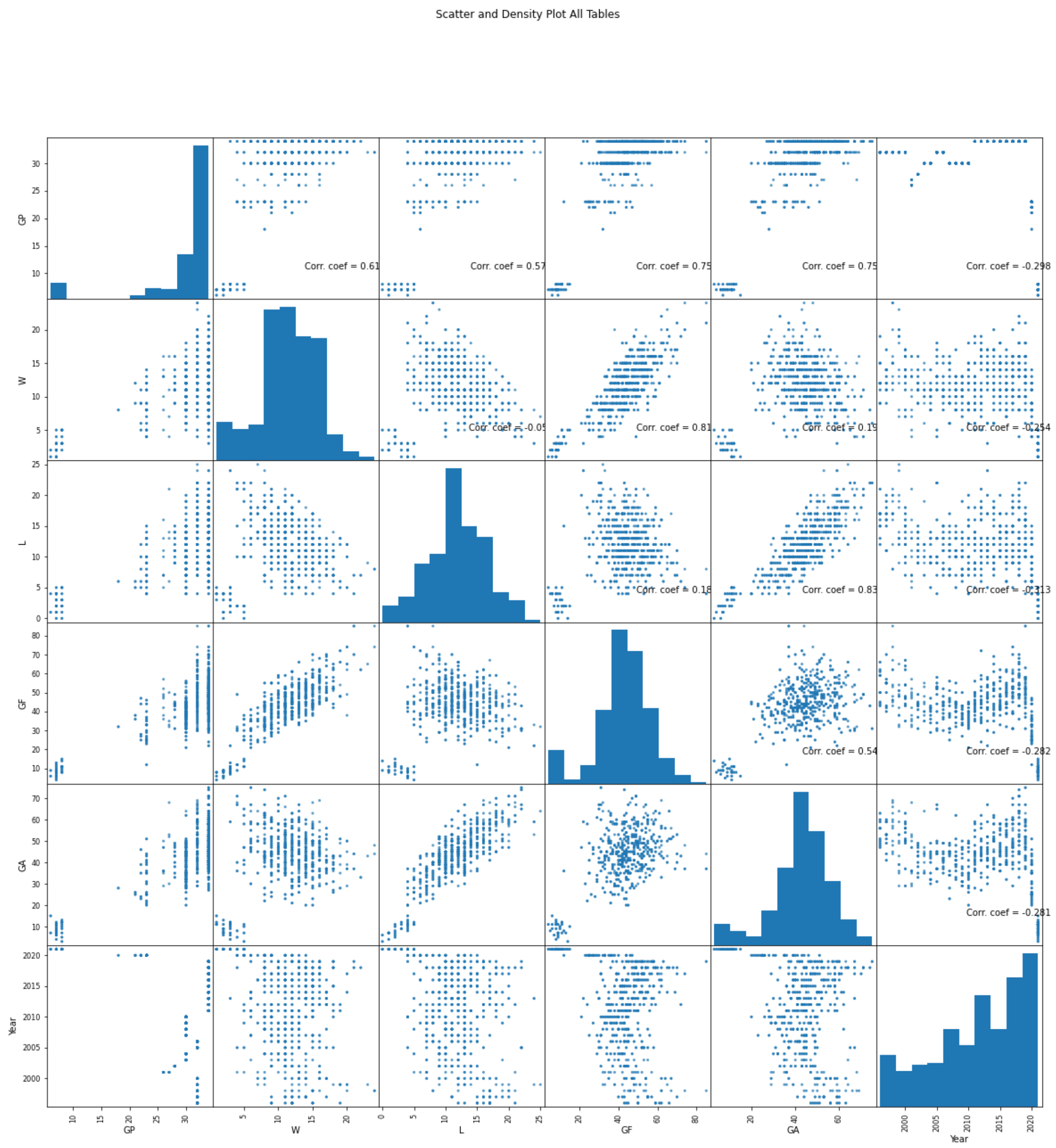


Figure 24: Pair Plot for All Tables

Later, the visualization that we chose to track is the time-series of the position of a club in all seasons. This is a really nice way to visualize the consistency level of a club across several seasons. If for example, under a manager or a star player, the team was consistently finishing at a good position, that particular time period's part of the visualization can be used. Lower the values, better the defensive performance and vice-versa.

Since all teams have not been a part of all MLS seasons, and some of them NaN values, we decided to run the code only for 4 teams: D.C United, Columbus Crew, Colorado Rapids and New England Revolution.

One thing to note here is that MLS is played in Eastern and Western Conference, with each Conference having its position. So, it is possible that 2 teams end up at the same position, because it means that both those teams have ended up in the same position in their respective conferences.

FINAL CONFERENCE POSITIONS PER SEASON

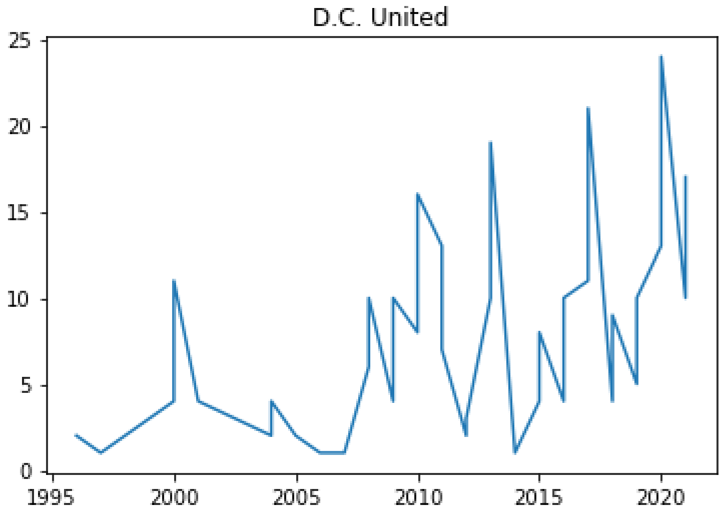


Figure 25: Final Conference Positions per season for D.C United

For D.C United, their performance has been pretty inconsistent. However, there was a phase during 2001 to 2008, where their performances were superb (due to low positions). In recent years, their performance has gone down though.

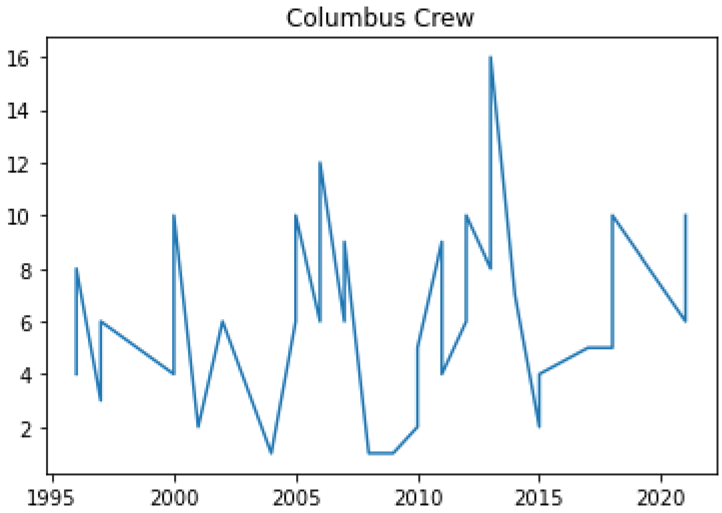


Figure 26: Final Conference Positions per season for Columbus Crew

For Columbus Crew, 2004 and 2015 were great seasons in the conferences. However, 2013 was their worst show.

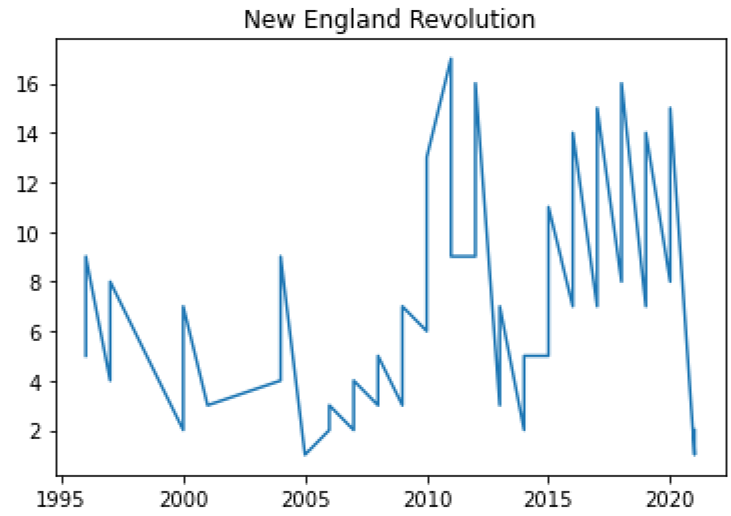


Figure 27: Final Conference Positions per season for New England

New England Revolution in the recent times have struggled really badly. But 2005 and 2014 were great seasons for them. 2011 and 2013 have been bad.

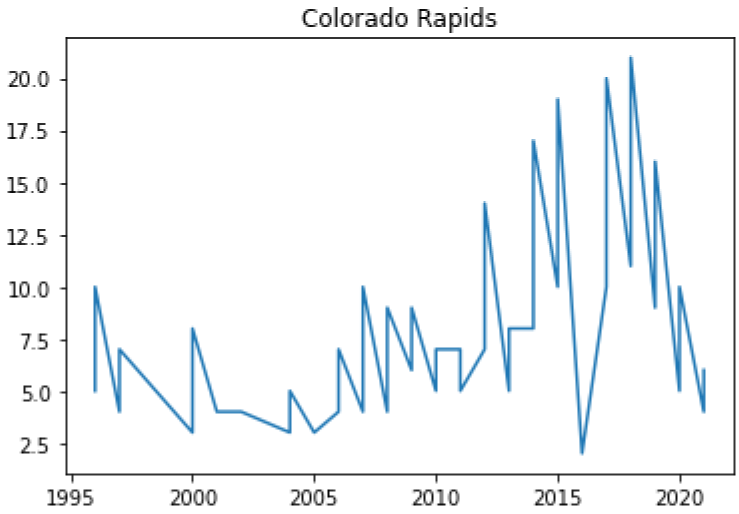


Figure 28: Final Conference Positions per season for Colorado Rapids

Colorado Rapids had a great 2016 season, otherwise their performances have been below par in the last few seasons.

The next visualization that we chose is to track the time-series of the number of goals scored by a club in all seasons. This is a really nice way to visualize the attacking performance of a club across several seasons. Higher the values, better the attacking performance of that team in that season.

GOALS SCORED PER SEASON

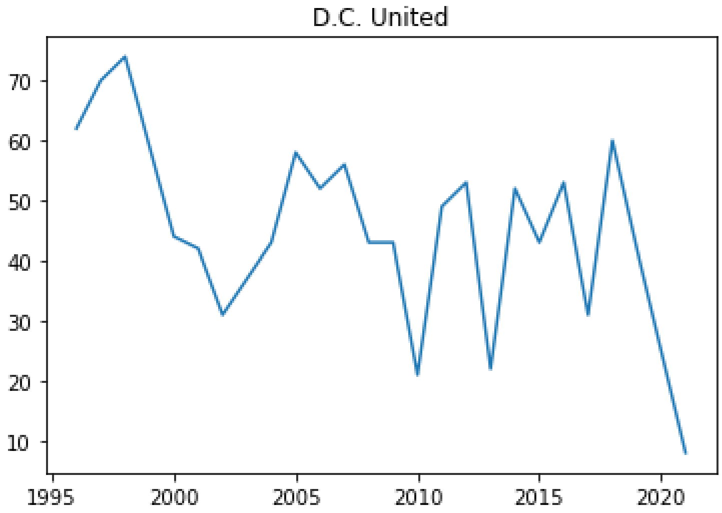


Figure 29: Goal per season D.C United

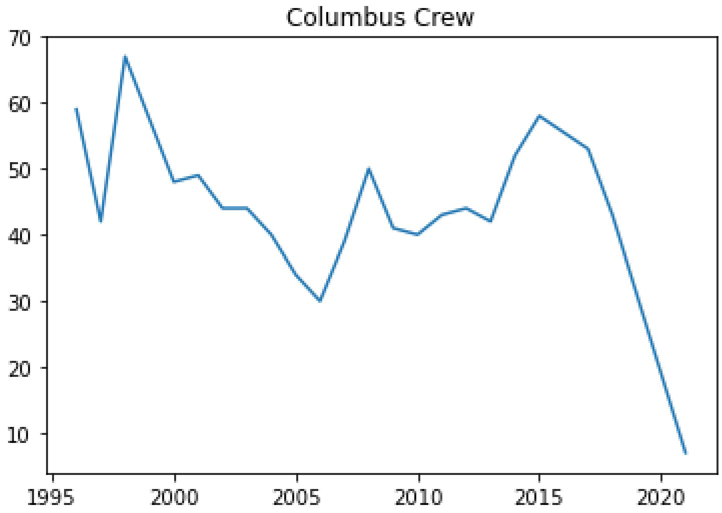


Figure 30: Goal per season Columbus Crew

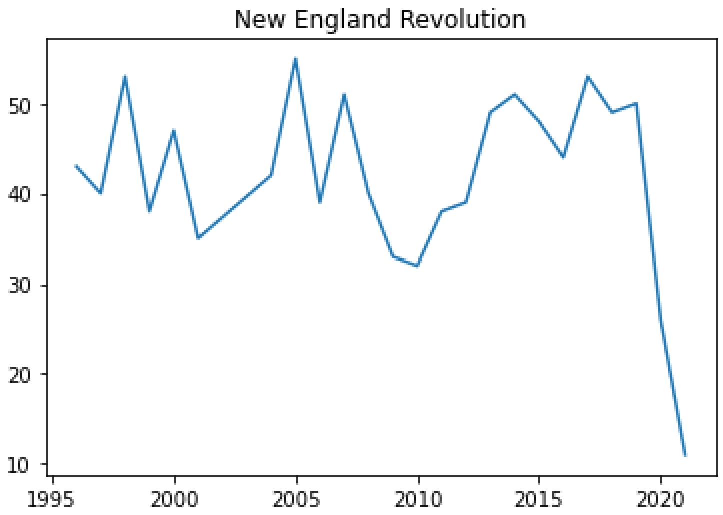


Figure 31: Goal per season New England Revolution

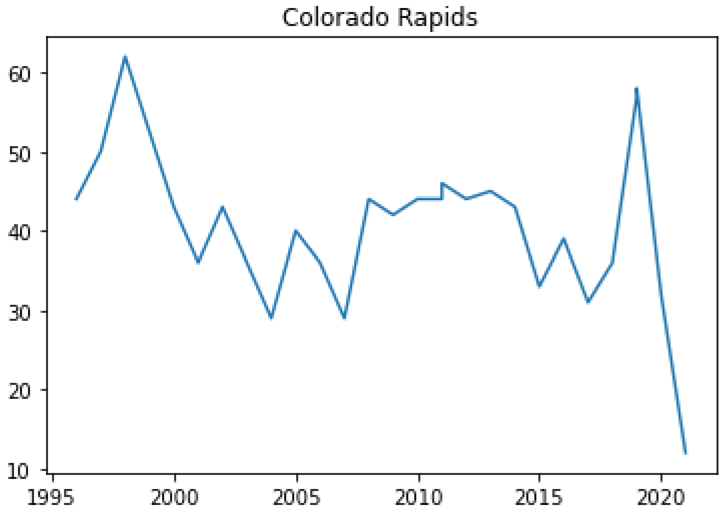


Figure 32: Goal per season Colorado Rapids

The reason why 2020 is low for all teams is because the data is incomplete. So, we will only consider till 2019 for this plot. Columbus Crew are good goal scorers consistently, with a downfall in recent seasons, whereas Colorado Rapids and New England Revolution are scoring a lesser number of goals. More attacking players and midfielders must be signed by such clubs.

The next visualization that we chose to track the time-series of the number of goals conceded by a club in all seasons. This is a really nice way to visualize the defensive performance of a club across several seasons. Lower the values, better the defensive performance and vice-versa.

GOALS CONCEDED PER SEASON

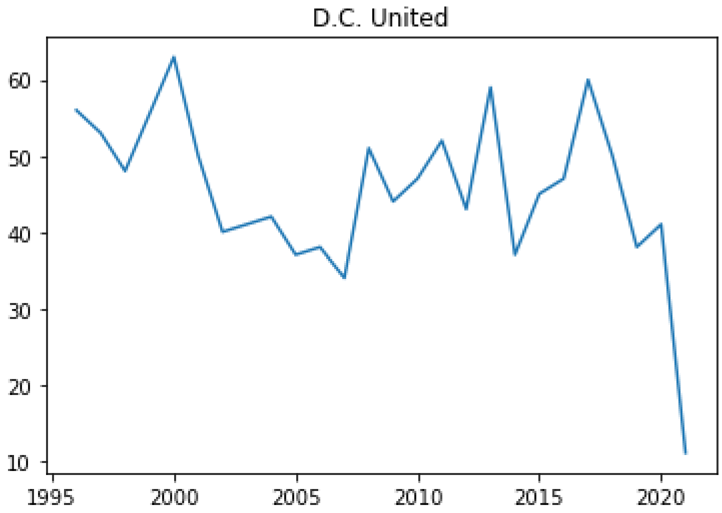


Figure 33: Goals Conceded Per Season D.C United

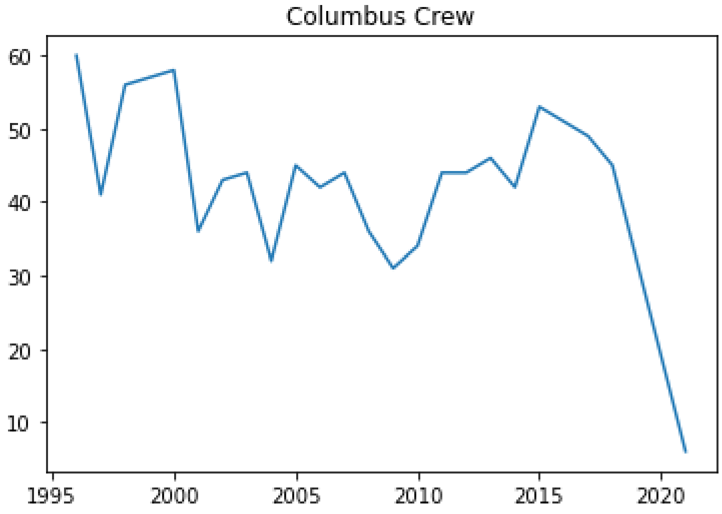


Figure 34: Goals Conceded Per Season Columbus Crew

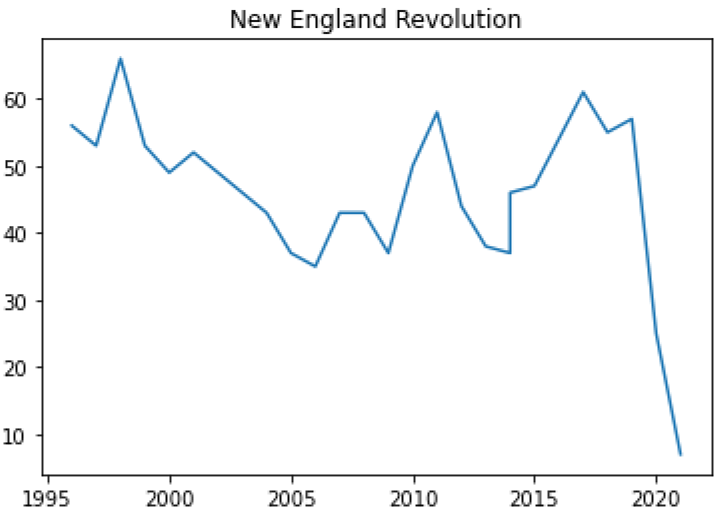


Figure 35: Goals Conceded Per Season New England

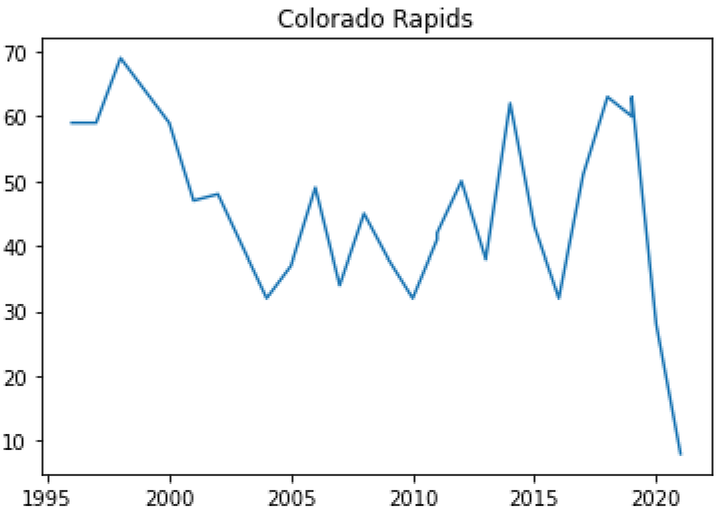


Figure 36: Goals Conceded Per Colorado Rapids

Again, we only consider till 2019 for this particular section. Colorado Rapids had a great improvement from 1997 to 2004 and have been consistent till 2014. Columbus Crew and D.C. United seem to be pretty inconsistent, and New England Revolution are kind of average in defense.

Thus, in conclusion, based on a team’s requirements, we can easily plot any of the above plots. We have written functions in our code, where we simply need to pass the team's name, and an appropriate time series can be generated.

5. DISCUSSION, CONCLUSION, AND FUTURE WORK

After completing the above visualization, along with its improvements, we now have a different kind and set of plots as compared to the existing ones. These combined with the existing ones, along with some machine learning models, and 3-dimensional plots can be used to provide excellent insights for soccer analysts, statisticians, investors, soccer managers, and of course the fans. The analysis has been done in 4 different levels: Team-wise analysis, State-wise Analysis, Player wise analysis, and Overall MLS analysis. The trends and insights which we were able to find were: How much dense is the average attendance per state throughout the United States, how much dense was the number of goals scored per state throughout the US, how each team has been performing throughout all the seasons, based on the number of goals scored (highlighting their attacking performance per season), number of points achieved, number of goals conceded (highlighting their defensive performance per season), trends in yearly attendance, yearly saved goals, yearly scored goals, suffered fouls, committed fouls, red cards, yellow cards, box-whisker plots showing how the distribution of Fouls Suffered, Committed, Yellow Cards, Red Cards with respect to the Clubs, lmplots showing the regression analysis of Yellow Cards and Fouls Committed with respect to each club (useful for Machine Learning Engineers). Based on these visualizations, effective decisions can be made to ensure maximum profit and even to improve the performance level of the teams.

Some limitations and caveats that we faced in the project are:

1. Unclean Data: There were many NaN values and dropping all were leading to some issues. The best way to solve this problem could be to use a resource that will consistently fetch proper data.
2. Countries named as Clubs in Dataset: Countries like the USA, Panama have been mentioned under the clubs' column in the dataset, which caused us several issues, and preprocessing the same was also equally hard. So, we had to consider those values as well, as dropping them might have caused big issues.
3. Clubs that no longer play in the MLS/ Invalid names: There are several clubs that went bankrupt, or have been incorrectly named in the dataset, which caused too many issues there, which is why he had to bring down the number of plots showcasing the trends to a very few of them because not all of them have played for all 25 seasons.
4. The biggest issue in our visualization was the lack of states and a few states from Canada also playing the MLS. Our main goal was to plot the intensity of goals/attendance (like it is done during elections) per state, which could have been an amazing geoplot. However, it turns out that only 14 states play in MLS. Such visualizations can be a great motivation for investors and teams to start introducing soccer in the other states and improving the resources of the already existing ones.

In conclusion, it was really fun to explore the MLS dataset project. The trends, insights, choropleth, lmplots, etc. were an immense joy to the plot. One of the best things about this project is its infinite scope. With more research, we can expand our analysis to even higher levels. The project report and code do not have all the visualizations which can be done but based on a user requirement, different plots can be generated.

As a future goal, an application can be developed, in such a way that based on the user requirements, different plots can be plotted for a particular team, a particular attribute to searched, player profiles (useful for scouts/talent management), team performances, etc., using a query-based approach. A dashboard or a portal with hand sensors can be a great technique to implement the same. These dashboards can also be easily implemented using stronger visualization tools like Tableau. Also, an internal state-wise assessment can be performed, find out the stronger and the weaker zones per state in particular aspects of soccer.

6. REFERENCES

1) https://www.kaggle.com/josephvm/major-league-soccer-dataset

2) https://fisherpub.sjfc.edu/cgi/viewcontent.cgi?article=1141&context=sport_undergrad

3) https://en.wikipedia.org/wiki/United_States_men%27s_national_soccer_team

4) https://en.wikipedia.org/wiki/Major_League_Soccer

5) https://plotly.com/python/choropleth-maps/