**CSC 331 Data report**

**Denzel Ntim**

The data set I choose to use for this report is a soccer dataset of the Premier League 2022-2023 season that I found from the Internet and on a site called data.world. I like it because it is quite recent, and it contains information on the last complete soccer season from the league in the world. I'm a you soccer fan and i would like to derive some information from this set, i would love to see the trends that are present throughout the season and understand the game a bit more from a statistical standpoint because even though I watch a lot of games I do not particularly know what the statistics say about the Premier league or even soccer in general so that is one reason why I chose this set. Another reason I choose this set is because I personally like working with more recent data sets that have are at most five-year-old because I feel like when you are working with old data it is a lot of stuff that I cannot pass a judgement on based on what is currently happening with that information. For example, in the past few years a couple important things have happened in the world of soccer like VAR being introduced to the game as technology which helps referees make decisions it helps them to see things that they could not and make their final decision. Some people think it spoilt the game some people think it's a good addition but at the end of the day it's very vital to take it into consideration because the number of cards that are given will definitely be affected by VAR even the goals scored because some awarded penalties need VAR to confirm whether it was a penalty or not. and all of that so I cannot use a set that was collated before VAR to give a conclusion about soccer currently. Therefore, I prefer to use a modern data set that can help me to which is why I chose the last complete data set available which is this of the 2022/2023 season.

**Detailed Description of the Dataset**

This dataset gives us information from possibly all aspects of every game that was played that season. It gives us the time each game was played, the day it was played which team was the home team and the away team. The half time goals scored for the home team (HTHG) and the half time goals scored for their way team (HTAG) then the half time result (HTR). The full-time goals scored for the home team (FTHG) and that of the away team as well(FTAG) then they we have the full-time result (FTR) which is as a result of the full time scoreline. For the full-time and half-time results, we have A, D and H as values to represent, An Away team win (A), a draw (D) and a home team win (H). There is also information and statistics for the referees which officiated each game of the season. There are listed with the cards awarded to the home and away both red and yellow and the variables for those are termed as HY, AY, HR and AR. The number of corners awarded to both teams are listed as HC(home team corners) and AC (Away team corners). Also, the number of fouls the referee called for each game Is listed under HF and AF (home team fouls and away team fouls). The home team’s shots (HS) and the away team’s shots (AS) are given to us in the dataset as well as the home team’s shots on target (HST) and the away team’s shots on target (AST). It is quite a relatively simple set to deal with if you ask me, I feel using my knowledge of SQL coding it would not be that difficult use queries to derive the information that I want to and make the comparisons that I intend to with this dataset.

What I intend to do with this dataset is to find out whether certain stereotypes about the game are true and whether I can prove is disprove those stereotypes with facts and statistics from an entire season. A typical example of this is in the soccer world everybody believes that when a team has a red card in the game it’s very likely to affect them and since they will be playing with 10 men and the other team may still have their eleven men which is an advantage over them it would affect how the game would turn out and they might possibly lose. Another stereotype is that the team playing at home has an advantage over the other team because of the atmosphere in the stadium, knowing the pitch better and them having a lot of support from their home fans since they are the majority of the fans in the stadium. Another stereotype is that usually, the team that scores the first goal or is leading the game as of half time has an upper hand in the game and it's more likely to win the game eventually so regarding this dataset that stereotype equates to the team the half-time result (HTR) being the same as the (FTR).

Analyzing a soccer dataset would help with testing to see if these stereotypes are actually present in the game we love so much and I choose to use a Premier League set because 80% of soccer fans around the world would agree that the Premier League is the most entertaining ‘week-in-week-out’ league in world football. It is believed to be the most entertaining league because it's actually very competitive, other leagues in Europe have a couple top teams that win it back-to-back and it's like there's no there's not much competition from the teams even in the lower part of the top half of those leagues or from the bottom half. But in the Premier League, all teams are competitive with each other even including those in the bottom half of the league. You can even witness this when they play the big teams. The stereotypes that I do agree with are the home team one and the halftime results having a big effect on the full-time results because I feel like when the team is playing at home, they have so much support and they know their home stadium and pitch very well, so it helps them. Also, leading the game at halftime helps in the sense where they could either just go on and finish the game in the second-half or they could sit back and defend to maintain their lead and win so those are the two stereotypes that I believe this dataset would help me prove.

For the first stereotype that I agree with, which is the half time scoreline having an effect on the full time scoreline, I ran the following queries.

**SQL work**



Here, I basically searched for all the games where the half-time result was ‘H’ meaning the home team was winning as of half-time and the full-time result was also ‘H’ meaning the home team won the game. 116 rows being returned means I got 116 results where this was the case out of 380 games played during the season.

I ran another query to see how games played out when the half-time result was a draw.



Here we only had 51 rows returned meaning that, in the other 329 games the half time result did not affect the full-time score because if the majority of the games had ended in a draw, I could conclude that a tie at half time definitely had an effect on the final result.

One more query to assess this stereotype of the ‘half-time scoreline effect’ was ran for the away team leading at half time and this was the result of it.



Here since we have 62 rows returned it shows that the away team leading at half time only had an effect and ended up the same way at full time for 62 games out of 380… the other 318 games ended up either in a draw or a home team win.

These three queries sort of debunk the stereotype of halftime results affecting the full time results because in all the three possible scenarios which is a home team win, an away team win or a draw out of the 380 games played in a season the halftime result and fulltime result only stayed the same in less than 50% for all the different scenarios so the game. Most games had different results for HTR and FTR than they stayed the same, which disproves this stereotype.

Below is a table that shows the number of matches that ended in away team wins, draws and home team wins, respectively.

A screenshot of a computer

Description automatically generated

This table here I believe defends my agreement with the stereotype about the home team having an advantage during their games. Clearly, we can see how they have won the most matches than drawn or the away team winning. The home team wins are 48.4% of all the 380 games played during the 2022-23 season as compared to that of the 87 draws which are 22.9% and the away team wins which make up 28.7%.

I exported this table to excel to represent it in a pie chart form so its easier to visualize the comparison and percentages between the home wins, away wins and draws.

A pie chart with numbers and a few different colored sections

Description automatically generated with medium confidence

I can conclude on this and agree with the stereotype that based on the information given in my dataset, home teams are more likely to win games they play, and it could be for several reasons that give them an upper hand.

To test the last stereotype which is red cards having an effect on the outcome of a game, I ran a query to check on how games played out for the teams that were awarded red cards…Both home and away. I first counted how many home and away red cards were awarded throughout the season.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

So here, we are able to see that the home team received red cards in 17 games and away teams received red cards in only 10 games so then I went on to see how that affected the games where red cards were given.

A screenshot of a computer

Description automatically generated

This table here shows us all the times the match ended In a home team win and the home red cards column was not of value 0 which means there were red cards for the home team in that game I chose to use ‘HR != 0’ in the code because as we know it there was one game where two red cards were awarded to the home team. The table derived from this query shows that out of the 17 games where the home team was awarded a red card(s), they only went on to win 4 of those games, which is less than 25%.

I ran the same query for the away games this time around to see how the cards affected the outcome of the game.

A screenshot of a computer

Description automatically generated

Here we can see that out of the ten games where the away teams had lost a player on the pitch due to a red card, they only went on to win 1. These queries do prove the stereotype right and it is quite evident. During the entire season, anytime both home/away teams got a red card, they mostly went on to lose the game. This doesn’t surprise me though as losing a man on the pitch affects a lot since the opponent team has an advantage over you by having one more player, sometimes it even demotivates the team that lost a player as they may feel its unfair to have lost them due to a harsh decision by the referee.

Out of curiosity I wanted to find out which referee awarded the most red cards in a game and incorporating a grouping into my code, I found out the most red cards given in a single game were by A.Medley and that was when he gave the home team 2 red cards.

A screenshot of a computer

Description automatically generated

One thing I wanted to discover is whether the team which had the most shots on target won the games in which they did or not. This is because to win a game of soccer you need to score goals and to score those goals you need to take shots at goal which are to be on target in order for them to high chance to hit the net, so I wanted to see know often the teams with most shots on target (HST/AST) won those games.

So, I first checked how many games there were in which the home team had more shots on target than the away team and vice versa. These are the results, respectively.





This query and table here below shows that out of the 214 times games in which the home team had a higher shot on target count than the away team, they won 144 of those games which is 67.2% of those games.

A screenshot of a computer program

Description automatically generated

The table below shows that out of 127 games where the away team had more shots on target than the home team, they won 65 of those games which is 51.1%. Even though it just about crosses the 50% mark, it still serves as the majority of those games.

A screenshot of a computer program

Description automatically generatedThese queries prove that the probability of the team which has higher number of shots on target is most likely to go on and win that game because every time a team had a higher shot on target rate, they won more of those games in which they did based on the information provided in this dataset.

**Implications**

Working with this dataset was an interesting experience. I previously worked with a dataset on the Premier League from the year before this season so it was interesting to see how things could change in the span of a single season and also patterns from certain statistics that I saw in that previous data set which are still present in this one. Two changes that I think would help improve the datasets and would have helped me more from this data set is that some of the variable names didn't quite work well with SQL so I had to change the names of some variables because some of them were known as functions and I had to change the them so that I could actually create the columns for the table, I didn't necessarily change them but I had to tweak them a little bit. One more thing as well is that the dataset could have also had a column/columns for possession. Possession is a very important part of the game and it basically shows how much a team is in possession of the ball for a period of time so it gives a percentage of how much that team is holding the ball at a certain point in the game so the set could have had it in the form where it would have a column for half time possession and full time position as it may show us that either the away team or home team was holding the ball for a longer period during whichever time, it could be the first half or the second-half. I suggest this because I do believe that possession affects how the game is played. It is more likely and more expected that the team who has the most possession of the ball will have a better game because it means they are in control of it and therefore they are more likely to win. So, I felt like that would have been a particularly good addition to the data set and it may have been complete then if you ask me.