**Introduction**

The Dataset which has been chosen for this assignment is entitled *Premier League Matches 1993-2023*. As an avid football fan and lifelong Arsenal supporter, I was intrigued by the potential insights which could be gleaned from the information within. Soccer fans often take pleasure in debating the question, ‘who was the greatest manager in the Premier League era’.

Oftentimes, modern fans, with the aid of certain dose of recency bias, will point to the exploits of Pep Guardiola or Jurgen Klopp as masterminds of the modern game. Those of us of a certain vintage, however, will not forget the buccaneering alure of Arsene Wenger’s wonderful Arsenal side from the early 2000’s, nor the ruthless efficiency and attacking prowess of Sir Alex Ferguson’s Manchester United team.

In this assignment, I intent to carefully explore the statistics relating to each manager to gain a clear insight into who was the greatest manager in the Premier League era.

**Business questions**

Form a business perspective, the analysis which will be undertaken will have real world utility. The act of analysing managerial performance to uncover who is the most suitable candidate for a job is an example of this. If a club was interested in hiring a world class coach, the analysis of statistics and data would play a key role in the decision-making process. By answering the following questions, we will gain a clear insight into which of our selected managers was the best Premier League coach.

**Questions:**

1. ***Which manager had the most experience?***
2. ***Who had the highest win percentage?***
3. ***Which manager had the highest home win percentage?***
4. ***Which manager had the highest away win percentage?***
5. ***Who was the most consistent home and away?***
6. ***Who was the greatest manager in the Premier League era?***

**The Dataset: Exploring the data**

This dataset lists every Premier League Game between 1993 and 2023. The layout provides the reader with information relating to the fulltime score of each game in this period via the *Home Goal*, *Away Goal* and *Full Time Result* *(FTR),* and *Winning Team* columns. Columns relating to the season in which games took place *(SeasonEndYear),* and match dates (*Dates*) are also found within.

**Formatting**

Upon a brief examination of the dataset, it became apparent that the formatting did not look correct. To fix this, CTRL+A was pressed, the format option was chosen, and the row and column width autofitted.

It was also apparent that the date column was in the wrong format. An adjustment was made by clicking format cells to alter them to ‘date’ format.

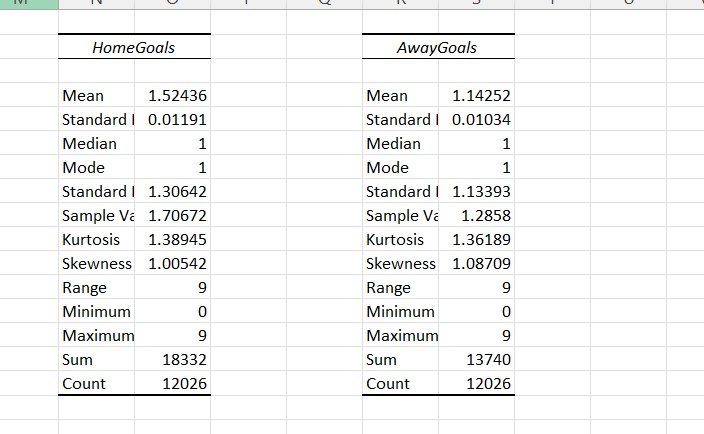
A screenshot of a computer

Description automatically generated

*Example of incorrect general format and date format.*

**Data Exploration using Descriptive Statistics**

To ensure that the data within the dataset was accurate, the *data analysis tool pack* was employed, and *Descriptive Statistics* was selected in the dropdown list. This allowed us to inspect that the total number of goals scored in a single game was ‘9’. This goalscoring record is shared by Liverpool, Leicester City and Manchester United and is elucidated under ‘Range’. This data is accurate and did not contain any noticeable outliers.



*Data pertaining to Home and Away goal statistics in the Premier League*

Data Cleaning

A careful, formulaic approach was then taken to clean the data. The first step in this process was to ensure that the cases of words in the dataset were formatted correctly. To do this, new columns were created and the *Proper Text* formula (e.g, Proper(G2:G12027) was used to convert the first letter of each word to uppercase. The original columns were then deleted leaving only the corrected versions.

**Duplicates**

Duplicates were then removed by clicking firstly on *data* and then on the *remove duplicates* icon.



*Duplicates removed from dataset*

**Blank Cells**

To ensure that no blank cells were present, the entire table was selected followed by *Find and Select*, *Go to Special*, and *Blanks.* On this occasion no offending cells were found. If blanks had been located, empty spaces would have been filled as ‘N/A’.

**Correct Spacing**

Issues pertaining to incorrect spacing were corrected using the trim function. To do this, new columns were formed and a function (e.g. =trim(E2:E12020) was inserted into new cells. This corrected spacing and resulted in a uniform, aesthetically pleasing dataset.

**Header Row Alteration**

The colour of the row header was changed by highlighting it and clicking the paint icon. A blue colour was selected which highlighted the header and provided for a more attractive header row.

A screenshot of a computer

Description automatically generated

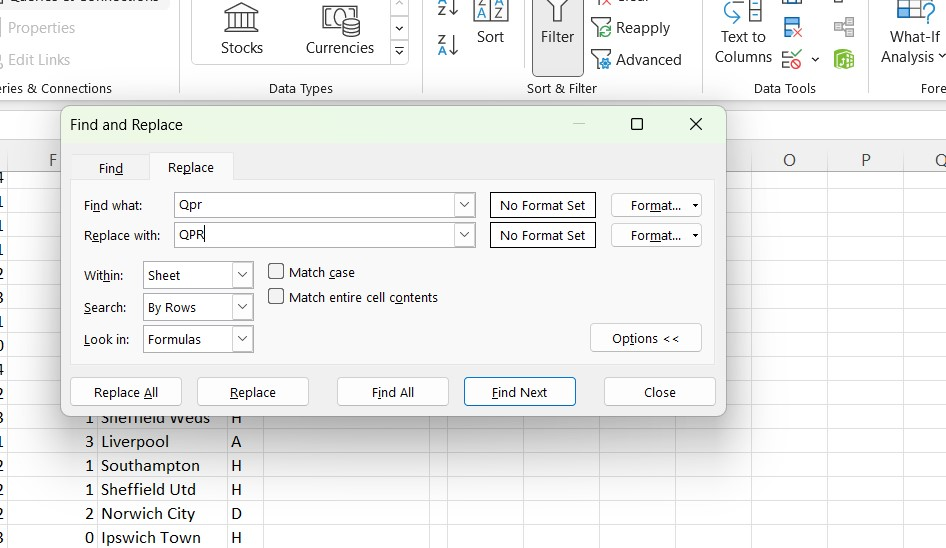
*Altering the colour of the header row*

**Spelling mistakes**

A spellcheck was performed on the data set and found that Nottingham Forest had been abbreviated to Nott’ham Forest.

**Find and Replace**

The find and replace tool was used to correct both this error and an issue with the casing in ‘QPR’. Below is an example of how the issue was found and corrected instantly.

*Using find and replace tool*

**Correcting Headers**

The final act of data cleaning was to change the header of column *B* from Season\_End\_Year to SeasonEndYear. This alteration brought the header into line with others in the dataset.

Data Transformation

To transform the *Premier League Matches 1993-2003* dataset into an analytical base from which the fortunes of various managers could be inspected, certain alterations had to be made to the data.

**Index Column**

Firstly, a match *index* column was added to give each of the 12020 games its own unique identifier. The *Row* function was used for this task. The code ‘row(A1)’ was added to the first cell and each cell was incrementally increased by one thereafter. The inclusion of this column was vital for the implementation of formulas later in the assignment.

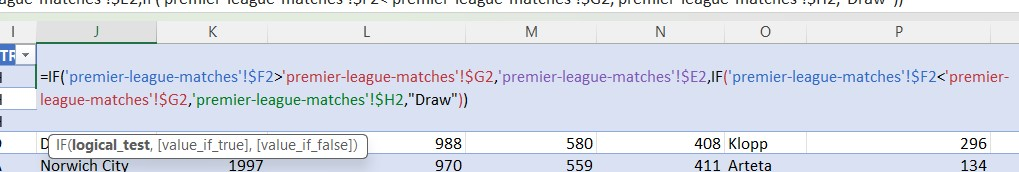
A screenshot of a computer

Description automatically generated

*Index created with row formula*

***Winning team* column**

The next alteration which was made to the dataset was the addition of the *winning team* column. The original dataset had highlighted match results simply as ‘W’ (for *home win*). ‘A’ (for *away win*) and ‘D’ (for *draw*). Adding this new column would be of pivotal importance when calculating win percentages, match numbers, and match victories. The ‘If’ function was deployed for the purpose of displaying the name of the team which scored a higher number of goals (e.g. ‘Arsenal’). If the numbers were equal ‘Draw’ was displayed in the row.



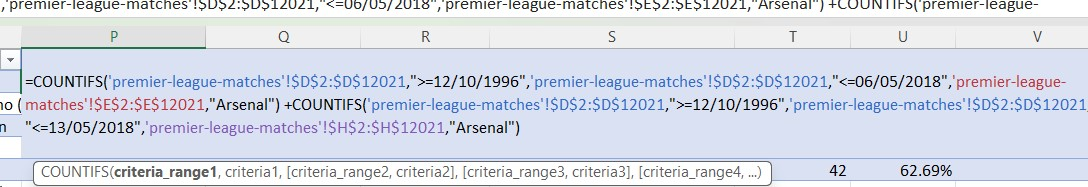
*If function used in the creation of ‘Winning Team’ column*

***Manager* column**

To compare the fortunes of some of the great Premier League managers, a manager column was constructed. Arsene Wenger, Sir Alex Ferguson, Jurgen Klopp, Mikel Arteta, and Pep Guardiola were selected, as were two rows for Jose Mourinho. The first, was included to document his fortunes at Chelsea between 2004-2007- an era in which the Portuguese manager won two Premier league titles. The second, (Mourinho Chelsea 2013-2015) was added to monitor Mourinho’s performance on his eventual return to Chelsea.

***Total Matches* column**

To calculate the total number of matches which each manager oversaw, the first and last game of their respective tenures had to be found. In the case of Arsene Wenger, for instance, his first game took place on 12/10/1996 and his last on 13/05/2018. The *Countifs* formula was used to calculate the number of times that ‘Arsenal’, appeared in the *Home* and *Away* columns between the above dates. This resulted in a count of 828 for Wenger. The same formula was applied for each of the other managers and their total number of matches was calculated and displayed.



*Countifs formula used to calculate total matches per manager*

***Total Matches Won***

To find the total number of matches which each manager won, the *Winning Team* column proved extremely important. Indeed, without the addition of this column these figures would not have been possible to attain. To perform this calculation, the *Countifs* function was used to add the total number of times a given team appeared in the *Winning Team* column between two distinct dates. The formula was again applied to the dates of each manager represented in the dataset.

A screenshot of a computer

Description automatically generated

*Formula used in the creation of the Total Matches Won*

***Win Percentage***

To find the win percentage pertaining to each manager the *Total Matches Won* column was divided by *Total Matches*. The results were displayed in the number format and had to be reformatted as a percentage.



*Calculation of Total Win Percentage*

***Total Home Games Per Manager***

This column was added to count the number of games which took place at a manager’s home ground. The *Countifs* function was used to calculate how many instances of a given team appearing in the *Home* column between selected dates.

***Total Home Games Won***

The number of home games won by each manager was calculated by counting the number of times, for instance, ‘Chelsea’ appeared in the *Home* column in a given period. Below is an example of the formula used to calculate this.



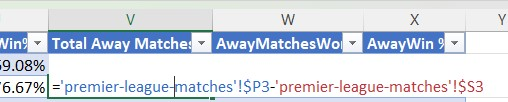
*Calculation of Home Games Won*

***Home Win Percentage***

This was calculated by dividing the number of home games by the total number of games per manager. The results were formatted as a percentage.

***Total Away Matches***

The total number away games per manager was found by subtracting *Total Home Games* from *Total Matches.*



*Use of subtraction in the formation of the Total Away Matches Column*

***Away Matches Won***

To calculate the contents of this column the *Countifs* function is used to count the number of times a team appeared in the *Away* column between selected dates.

***Away Win Percentage***

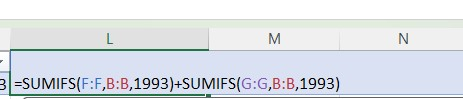
The away win percentage for each manager was calculated by dividing the total number of away matches won by the total number of matches and formatting the answers as percentages.

***Unique year* Column**

The next column which was added was unique years. This column was added to aid calculations relating to goals per year, *Home Goals* *Per Year* and *Away Goals* *Per Year.* The years 1993-2023 are recorded in the column.

***Number of Goals Per Year***

This column records the number of goals scored in a specific season. These goals are calculated using the *Sumifs* function by adding the home and away goals each year.



*Number of Goals Per Year calculation*

***Number of Home Goals Per Year***

The contents of this column were calculated by adding the home goals in each season using the *Sumif* function.

A screenshot of a computer

Description automatically generated

*Number of Home Goals Calculation*

***Goals per year***

To ascertain whether the league has changed in terms of the attacking tendencies of teams over the years, a Goals Per year chart was created. The data from this chart indicates that although a slight tendency toward higher rates of goalscoring existed in the opening three seasons of the premier league, figures have largely remained consistent over time.

*Goals per game chart which indicates similar patterns of goalscoring since the Premier Leagues inception*

Analyse the Data

In this section our initial business questions will be answered to ascertain who the most successful manager on our list.

1. **Which manager had the most experience?**

From the chart below, we see that Arsene Wenger and Sir Alex Ferguson are clear winners in this regard. Both managers amassed more than 800 games each at of their respective clubs and maintained a high level of consistency throughout these years.

Jurgen Klopp, who is third highest on the list, began his Liverpool tenure in 2015 and still sits a remarkable 532 matches behind Wenger. It is also worth noting that Sir Alex Ferguson had been managing Manchester Utd seven seasons prior to the inception of the modern Premier League and that this data is not included in our dataset.

For the purposes of our investigation, however, we can say that Arsene Wenger is the most experienced manager in the Premier League era. **Answer 1: Arsene Wenger**

*Total Matches per manager graph which emphasises the longevity enjoyed by both Wenger and Ferguson*

1. **Who had the highest win percentage?**

The statistics in this column made for interesting reading and point to remarkable success of managers like Pep Guardiola and Jose Mourinho. Guardiola’s win percentage- since his spell at City began in 2016- is 74.06%. The statistic is even more remarkable when we note that Guardiola did not enjoy a successful first season at the club finishing 3rd in the league and amassing only 78 points.

For Jose Mourinho, our stats highlight the dominant first spell which he enjoyed at Chelsea. His initial win percentage of 70.83% is remarkable and contrasts starkly with his win percentage of 59.78% in his second spell at the club.

Sir Alex Ferguson’s win percentage of 65.19% is even more commendable when we factor in the duration of his career and his level of consistency over a prolonged period. It is Guardiola however who comes out on top in the win percentage category. **Answer 2: Pep Guardiola**

A graph of a number of people

Description automatically generated with medium confidence

*Data relating to win percentages which reveal the dominance of Pep Guardiola at Manchester City*

1. **Which manager has the highest Home Win Percentage?**

Our Home Win statistics once again point to the dominance of Guardiola’s Man City outfit. His side have amassed a home win percentage of 78. 95% so far during his tenure.

Again, Jose Mourinho’s average (76.67%) during his first term as Chelsea boss is enough to sit second in our list, while Sir Alex Ferguson produced a fine home win ratio of 75.31% in his time at Manchester Utd. Mikel Arteta’s home win percentage of 62.69% is the least impressive record of any coach in our study. **Answer 3: Pep Guardiola**

*Chart indicating the Home Win Percentages of EPL Managers*

1. **Which manager has the highest Away Win Percentage?**

The top Away win percentage was the 69.17% achieved by Pep Guardiola at Manchester City. Second, was that of Jose Mourinho during his first spell at Chelsea (65.00%) and third Jurgen Klopp of Liverpool with an average of 55.41%.

The pivot table below highlights the dominance of Pep Guardiola; indeed, Guardiola’s away record is a staggering 23.28% higher than that of Arsene Wenger and 4.17% clear of his nearest challenger Jose Mourinho (Chelsea 2004-2007). **Answer 4: Pep Guardiola**

****

*Pivot Table which displays the dominant away form of Pep Guardiola*

1. **Who was the most consistent Home and Away?**

The below graph serves as a good indicator as to which manager was most consistent when we analyse both home and away form. Every manager on our list experienced a significantly inferior away record when compared to their fortunes in home games.

Some managers, however, were more consistent than others. A gap of only 9.77% existed between the home and away form of Pep Guardiola’s Manchester City, this represents the smallest percentage disparity and renders him the most consistent manager in our study. It is also important to note that both Guardiola’s home and away percentages are superior of all other managers in those categories. The greatest disparity in home and away form can be seen Jose Mourinho’s 2013-2015 Chelsea, whose home win percentage was 23.91% greater than their away record. **Answer 5: Pep Guardiola**

**Home/Away Disparity**

*Chart indicating the differing form of manager in home games vs away*

**Some thoughts on the findings**

From analysing the above statistics and charts it is apparent that despite the longevity and experience of Sir Alex Ferguson and Arsene Wenger, Pep Guardiola figures are the most impressive on our list. From a business perspective, Guardiola would be the ideal choice when selecting the best candidate for a new managerial role. Indeed, Guardiola finished top of four out of the five categories in our comparative study and possessed an overall win percentage of 74%.

Despite the fact the Jose Mourinho finished second in several categories due to the form of his first Chelsea side between 2004-2007, Sir Alex Ferguson can be regarded as the second most successful manager in our study due to his consistency over an incredible duration. The Scotsman’s win ratio of 65.19% over 810 matches is a testament to both his longevity and success at Manchester United.

The third most successful manager on our list is Jose Mourinho who amassed over 212 total games as Chelsea manager with an impressive overall win percentage 65.3%.

Jurgen Klopp’s win rates tended to see him placed in the middle of most charts while both Arsene Wenger and Mikel Arteta experienced inferior win percentages compared to their peers.

**Conclusion**

To conclude, the *Premier League Matches 1993-2003* proved to be a useful source of data and provided for some interesting insights. The raw data gathered from Kaggle required cleaning to improve its quality. The data within was vast and encompassed over 12,000 Premier League Matches spanning over 20 years.

To find insightful facts relating to the league, extra columns were added giving the data a new dimension. By obtaining the start and end dates of the EPL managers we were able to add information relating to managerial records and formulate a comparative study. The business questions which were chosen helped to refine the research and focus on the strengths and weakness of each listed manager.

The data was transformed and analysed and led to Pep Guardiola leading four out of five of the comparisons thus making him the most successful manager out of the six selected.

This assignment proved to be a useful study into how statistics can be utilised in the process of managerial selection.