Middlesex University London

CST3990: Undergraduate Individual Project

Data Analysis on International Football Results over the Last 150 Years and Predicting Future Footballing Results Through Machine Learning

Coursework 3: Final Report
Submission: Sunday 28th April 2024

Student ID:M00810926

Name: Abdullahi Mohamed

CAMPUS: HENDON

Contents

Acknowledgements

I would like to thank my supervisor Carl James-Reynolds for guiding me in this project, the help was necessary for me to progress accordingly and maintain a high quality of work throughout the project.

Abstract

The aim of this project is conducting data analysis on international football results to find meaningful insights and use the data explored to build a hypothesis that is unique. When this stage is complete, two teams with the highest home winning form of all time will be put to the test in machine learning; it will present five previous results to understand the variety of factors leading to victory in football. Then predict five matches of the same two teams and see the differences but also the similarities and the accuracy of the results. The project sheds light to football analytics, machine learning, further research on team performances and valuable insights to stakeholders.

Chapter 1 Introduction

1.1 Introduction

The research aims to explore the link between data analytics, machine learning and international football by focusing on historical trends and predictive modelling to enhance insights for all viable stakeholders. International football has a rich history and vast data to which is going to be presented in analysis. The current integration of data analytics and machine learning into sports has started a range of new patterns, methods of predicting outcomes and gaining a deeper understanding of football and how it can benefit management, supporters, and any of those that have interest within football. The literature review consists of a diverse understanding of analytics, which includes machine learning, data visualization and data science. From player contribution to tournament hosting, home advantage, the literature details the nature of football analysis. With this project, historical data will be used to analyze the best home winning team per decade, the highest and least scoring teams both home and away, the number of games played and an overall summary per decade. Currently football analytics has machine learning and data sciences projects to offer valuable insights to stakeholders, however a gap exists between understanding how historical factors leads to different results as majority of team success is dependent on various reasoning. This research ensures to fill the void in understanding machine learning models, past results, best teams per decade, highest goal scoring team and give an overall contribution to the success of international football. Figuring out the complex links of past results to team success in international is one of the main study challenges presented. The effects of home advantage, hosting tournaments effectiveness and other specific problems that will be answered. The research aims to provide an understanding of football dynamics over time through the analysis of historical data. Objectives include patterns within home victories of a particular decade and evaluating the impact of factors and proposing predictive models. The project will follow a set structured format, progressing from the literature review to the methodology, functional requirements, implementation, evaluation, and testing.

Chapter 2: Literature Review

2.1 introduction

The subject matter is to dissect 150 years of international football results, find historical and predicted results using machine learning and data analysis techniques. The aim of this is to find patterns within data and predict results, focusing on the historical trend, and produce predictive modelling. The concept of the literature is to find data and new developments in football analytics. This study illustrates the information and approach as it tries to conclude the justification of the data sources. Whilst understanding the restrictions of research, this analysis shows the scope and limitations of building an analytical and critical evaluation of football data analysis and predictive modelling. The objective is to identify trends within the football dataset and use predictive modelling to predict results. This research paper seeks to steer the obstacles of historical data analysis whilst clarifying the dynamic nature of football analytics through past trends and predictive analysis. It seeks to conduct a thorough evaluation of the data sources whilst identifying the investigation's restrictions. This introduction acts as a recommendation, directing an exploration of past data, analytical patterns and forecasting results and finding techniques to advance the football industry.

2.2 Relevant Literature

Machine Learning

Baboota & Kaur (2019) sets an outline to predict football matches using strategies that are crucial factors in football. The project is assessed by goals per game, minutes played and a detailed study of the club budget, to illustrates the writers of project went to deliver a quality thesis and how all those elements effect match results. The study portrays the importance of team cohesion, and the roles different play for football teams, financial complication, and how the combination of those can lead to a thesis focusing on the value of team strategy, dynamics, and the financial stability of teams. The evaluation of the literature review includes components of football match predictions. Baboota & Kaur’s (2019) technique is praiseworthy, to which they have provided a way to predict football matches. However, the project lacks the importance of individual player’s performances and how they can decide games alone. The study undermines the impact one player can have on results, ignoring the complexity presented in player contributions, therefore a study on player-specific contributions is necessary. Examples such as the injury list, ranking teams, player statistics, formation are all suggestion that can be taken to enhance the methodology of this project. Including these components not only offers a more complex picture of team dynamics but also recognises the influence that players can have on the outcome of a game. Overall, the project is clear and concise by including these new perspectives can lead to a wider approach and understanding of football predictions.

Differently to Baboota & Kaur's (2019) style, Rodrigues & Pinto (2022) uses a different point of view as it centralises on the players historical match data. The method is significantly different because it focuses how a single player can impact the outcome of matches through their abilities. Moreover, the review explores the player contribution, acknowledging the influence of any athlete can have on a football match. The study persists the importance of players and how they can change the fate of a team. This introduces a more in-depth analysis of the sport move from team to player analysis, which captures the player form, consistency, and skill sets of players. Although it provides a useful player focus perspective, the project has restraint that could hinder it, the lack of team dynamics is the reason for it. Football is a team sport, including team strategies, formations, tactics is vital for match predictions, further studies to investigate football can achieve a balance literature review.

Che Mohamad Firdaus Che Mohd Rosli et al. (2018) conducted a relative study that assesses the reliability of data mining approaches, and the research paper explores the variety of predicting approaches such as decision trees. The study examines the methods that can be used in football, to which illustrates the disadvantages and advantages of both Bayesian networks and decision tree model evaluation. The research has perception to predictive analytics by utilizing these data mining techniques to predict football matches. Despite this admirable project, the study doesn't recall including football dynamics such as team tactics, adaptable such as starting line-up and how it effects the results. The model evaluation is great predictive models, but it is difficult to represent the dynamic of the sport, problems such as overwriting code can delay results or lead to a less precise forecasts. Differently, Pantzalis & Tjortjis (2020) uses past data and advance statistics to predict other for team performance included refer to whether a team is going to have a better chance or not next season. This approach emphasises the use of data and how it can lead to strategic decisions for stakeholders of football and how it can be used for prediction purposes and aligns with the trends of sports analytics. Furthermore, the paper focuses on statistical categories that separates players into ability, the experimental results are based on data available of the beginning season which highlights the effectiveness of results. The study adds to classifying players based on their skills and contributions, giving knowledge of how individuals affect team outcomes. This player viewpoint improves the outcomes and adds to a thorough understanding of assessing team performances.

Alam & Almulla (2020) uses machine learning frameworks to investigate a relationship between player performance and results. The theory highlights the role that individual players determine the outcome of the results, identified patterns is utilized by machine learning techniques. Sports analytics present a how team overall performance affect whether they win a match and how they can improve for team tactics and forecast on player dynamics. Alternatively, the selected methodology has a disregards team variable, particularly the effect of injuries to results and player performances. The dynamics of a team can be easily halted by injuries, illnesses for example, and can change how the fame plays out. To improve the overall effectiveness of their research, a more thorough examination that incorporates individual players as well as team-related aspects that can lead to a more accurate model.

Stubinger, Mangold and Koll (2019) investigated football prediction and found significant returns, the research paper emphasises the potential financial rewards from targeting specific betting strategies. The studies above are all machine learning related to which they all collectively focus on the predictive analysis and offering valuable insights to stakeholders and narrowing certain aspects to achieve a prediction. Differently, to the earlier research, Stubinger examines the financial environment of football and how it can benefit stakeholders, therefore proposing precise predictions can give a betting company a financial advantage. The study creates a body of work that focuses on a variety of areas to improve the precision and relevance of forecasts in football. However, a lot of complexity involves in predicting match outcomes, displaying the necessity of a balanced approach that incorporates both team and player contributions to the results. Sports are a team collaboration, therefore its critical to consider the effectiveness of players, finding a balance between accuracy and a review of an events. Developing predictions model requires a wide range of methods to consider meaning that both team and player should be taking into consideration.

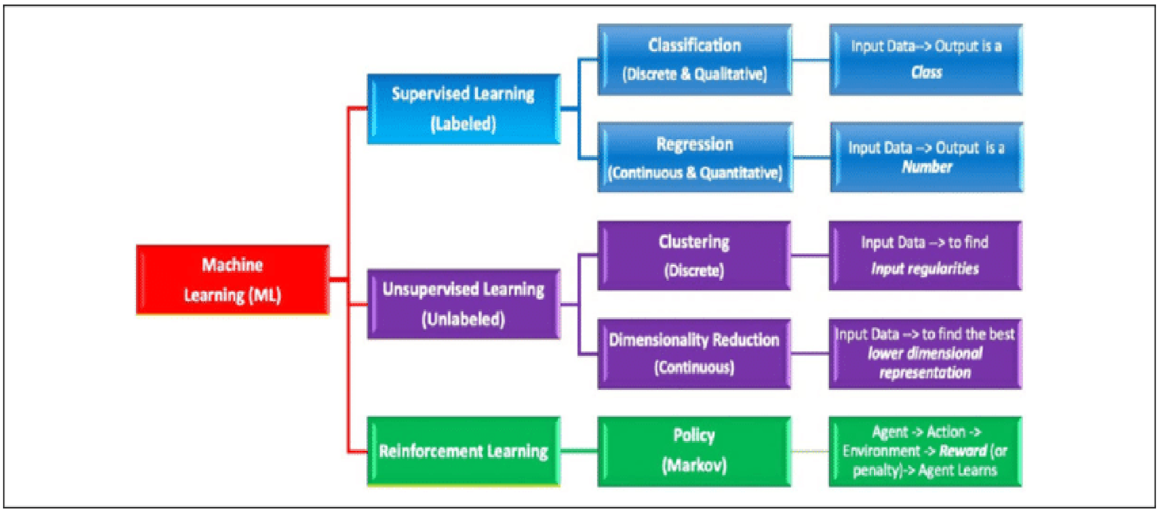

Machine Learning (ML)
Supervised Learning (Labeled)
Unsupervised Learning (Unlabeled)
Reinforcement Learning
Classification (Discrete & Qualitative)
Regression (Continuous & Quantitative)
Clustering (Discrete)
Dimensionality Reduction (Continuous)
Policy (Markov)
Input Data--> Output is a Class
Input Data --> Output is a Number
Input Data --> to find Input regularities
Input Data --> to find the best lower dimensional representation
Agent -> Action -> Environment -> Reward (or penalty)-> Agent Learns

Figure 1

Data visualization

Data visualization is a crucial part of footballing performance analysis as it provides a detailed understanding of games. Gásquez & Royuela (2016) decided to analyse team performances by concentrating on goals scored and displaying it. Exploring differences between team dynamics and visualizing to insight managers, betting's companies, and fan engagement. Their work purpose is to a serve a variety of stakeholders that includes managers that can use insights to enhance their skillset and find method to win more football matches. Betting organizations can try improving odds and fans interests can increase their engagement for the sports as they can compare predicted results to actual results. When the data is visualised, it becomes a great tool to make a well-informed decision and adds values to the sporting industry. Mohamad (2023) focuses on how much statistics affect team performance using data visualization tool. The purpose of this study was to determine whether match statistics influences team victories, the concept is to link performances to real world match results. The goal aims to align the performance measurement and real-world matches, using data visualization can create a connection between statistics and actual outcomes. This approach allows for data easier to understand, but it also makes it easier to comprehend how performance can affect a team's ability to succeed. Football analytics gains a practical understanding of this study as it emphasises on the application of predicting matches. However, its necessary to address wider ranges of team successes for examples data visualized the key formation to winning match helps provide a further understanding of performance enhancement.

Sainan (2023) focuses on sports data visualization, the purpose of the research paper is to visualize and present the playing pattern throughout the game and a focused approach found in football games. The goal of the research paper is to highlight the complex playing patterns that occur during a game, focusing on football matches. The study improves understanding of dynamic character of football game play because it provides analysis of interactions, formations, and strategic movements. On the other hand, neglection of direction can lead to a lacklustre analysis and leaves gaps in knowledge on the implications of the team success. This can lead to an ordinary analysis only to which it forgets the implications of team performances, ignoring this factor could reduce the depth of understanding from visualisations and makes it harder to identify the patterns and for team performance is neglected thoroughly.

Elkins (2017) emphasises to real-world uses, using data analytics techniques to improve football performance. The goal is to give coaches probabilities that are based on data analytics to enhance the skillset of coaches to understand what the common patterns within matches. Although the study has immediate application and provides valuable insights for coaches, its limitation when defining the scope of the project may prevent a comprehensive assessment of all factors that influence performance in international football such as the availability of players and home advantage. These studies all significantly focus on data visualization within football performance, overlooking broader factors that lead to team victories. However, it provides a realistic approach that bridges statistical insights with a broader understanding of how data visualization can enhance football performance analysis and contribute to team victories. The reason being for this is that coaches are crucial for the progression of a football team as they make decisions such as selecting the starting team, formation change throughout and tackling any problems in game.

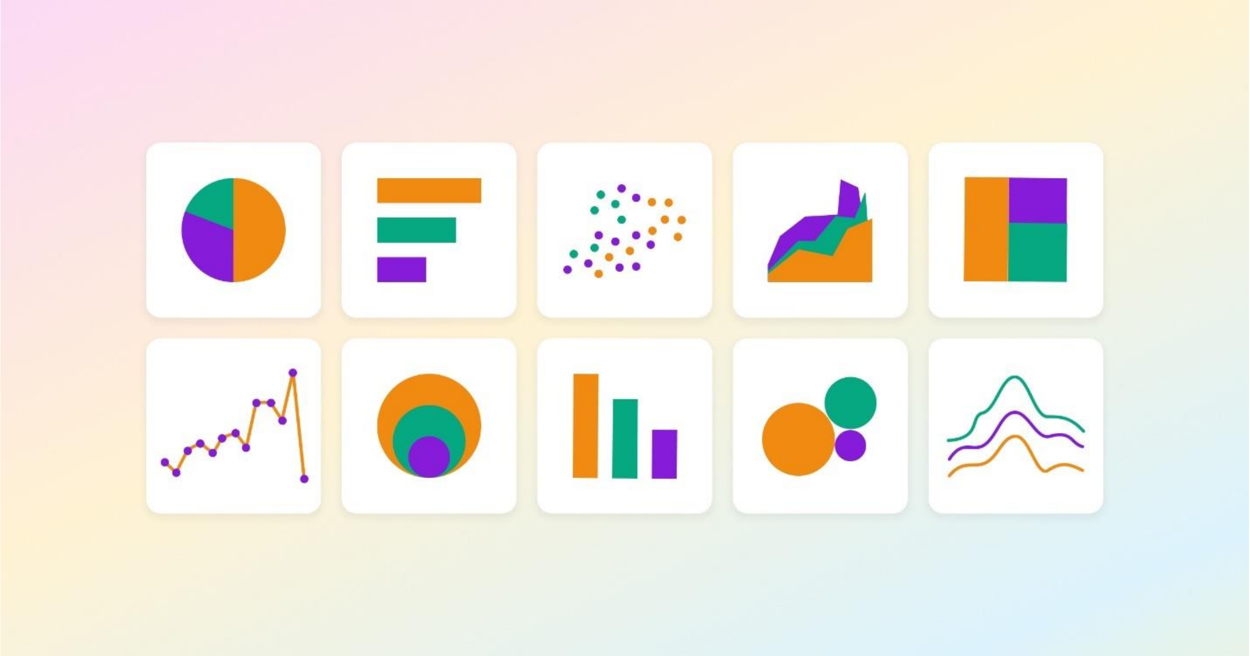


Figure 2

Data Science

Mansoor Alam (2020), the paper presents a data science approach to minimize the time taken in team selections. Using Power BI and Python Pandas to statistical analysis of player performance based on abilities and skills for players can demonstrate and improve managers to select players and help provide knowledgeable data for stakeholders. The benefits of this are to stakeholders such as management can learn what players are beneficial and that can lead to great victories. Using a programming language for predictive analytics helps to forecast results using model evaluation and what field out the best player for team selection in code. Power BI is a great tool for storytelling data insights to which it can be displayed in a dashboard that helps understand the player performance through charts, figures, and line graphs. However, this literature review focuses on the business side of football but the main interest in sports is internal, the abstract should include fan engagement and management strategies as they enhance the sport's popularity. The reason being is that internal factors of football are a major responsibility to the success of the sports so that it is important to acknowledge as stakeholders in this project.

Poojan Thakkar & Manan Shah (2021) literature review aims and delves into the impact of data and the influences of data science of football and seeks to advance and transform the sports. Also, the research paper highlights changes in sports and examines the effects and displays how different strategies, tactics, and overall approach to playing a match can affect. In addition, the competitive edge to gain an advantage over opponents highlights how useful insights can lead to a quality understanding of football. The review also extends into the political and economic state and how it affects various aspects within and beyond sports. The reason these are great points in football because the tactical approach of football can lead to better results, had an inexperienced manager who wants to stick to playstyle can have a read of this research and find the best route to success in football management. Using data science to advance sport is a great move as predictive analysis can lead to actual outcomes in football. On the other hand, the research paper is vague because it generalizes a lot of important topics that need detailing, for example the project can be formulated better if there was a direct aim of the project rather than overview because of the complexity of each topic and it leaves the reader with little understanding of football analytics as it fails to narrow down on one topic.

Hucaljuk and Rakovic's (2011) research uses an exploration of the complex challenge of predicting football match results. It is a crucial resource because it focuses on using machine learning specifically to anticipate results and presents a methodology to maximize predictive powers. It also gathers the different methods that can be used to predict football matches, including logistic regression, highlighting the project using model evaluation to predict results. The aim of the project is to find the differences and difficulties of predicting results using Python. This literature review focuses on stakeholder demand and gives them a guide to use if needing to predict results. The research does, however, also show several limits that should be considered. Furthermore, because it was published in 2011, it might not include the most recent developments or machine learning techniques that are relevant to sports analytics. This study's main goal is to use machine learning approaches to handle the complexity involved in forecasting the results of football matches. Its goal is to provide a software-based solution that improves the predictive capacity for football results by feature selection and classifier models. It acts as a starting point, providing a structure or preliminary method for creating football prediction models. Additionally, it creates a standard by proving better predictive performance than conventional techniques, giving the project's models a benchmark to measure them against.

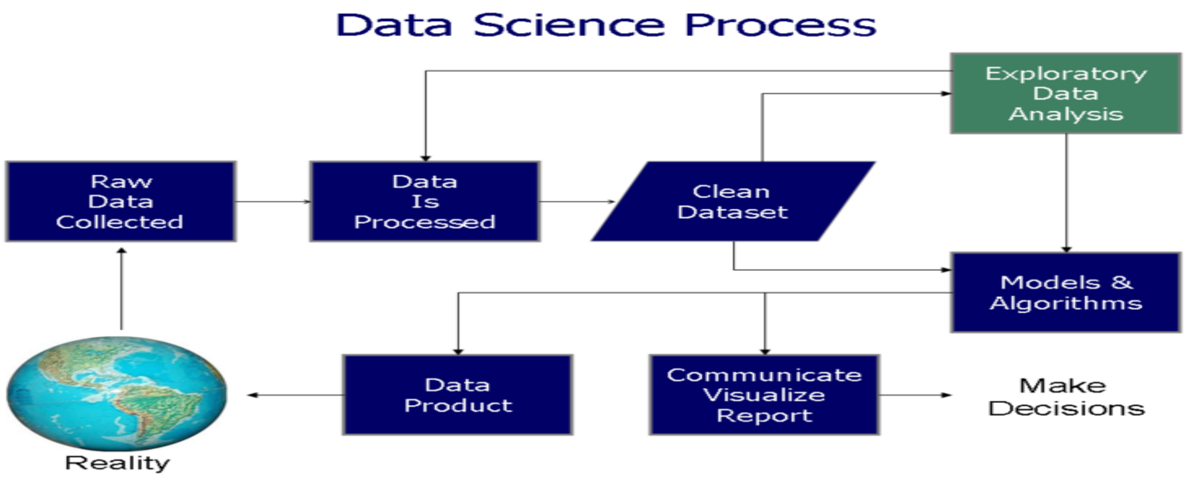


Figure 3

Relevancy to project

The project is in line with the literature review, which supports a variety of machine learning techniques while highlighting player characteristics and team dynamics. Understanding data visualisation techniques offers a way to successfully interpret game dynamics from the past. Furthermore, player performance analysis throughout football periods is enhanced by data science applications for squad selection. Recognising the effects of game scenarios helps predictive models by taking outside factors into account. Finally, the project's emphasis on examining tournament hosting and the impact of friendly matches throughout several footballing eras aligns with the review's need for uniqueness. By adding team within historical settings, these insights enhance the project's predictive models and provide a more thorough knowledge of football outcomes.

Resilience of the team/player's study

Miguel-Angel Gomez (2020) focuses on soccer teams' performances in close game scenarios over various match-status intervals from seventeen Spanish professional league games. There is a notable number of dynamics in football, and each can affect the winning team's play, particularly in the final third of a game. For example, the clubs usually with less ball possession of all parts of game are more likely to lose the match. The goals of this research paper are to understand the resiliency of players and managers and how the results can interpret what scenario are likely to lead players to perform better whether scoring first or scoring later, it covers all aspects game development. Differently the review of this topic still lacks direction and certainty because no matter what order of sequence of a game, there are more factors such as the weather and players fitness that decide the game. The review of this subject nevertheless struggles with uncertainty and lacks a clear direction, despite the thorough investigation. This uncertainty arises external variables such as players' fitness levels and the weather, have a considerable influence on game outcomes, directly dependant of the events that happen during a match. These extra variables add an unpredictable factor that makes it more difficult to determine whether certain patterns can be seen or whether there is a clear relationship between team performances and match scenarios.

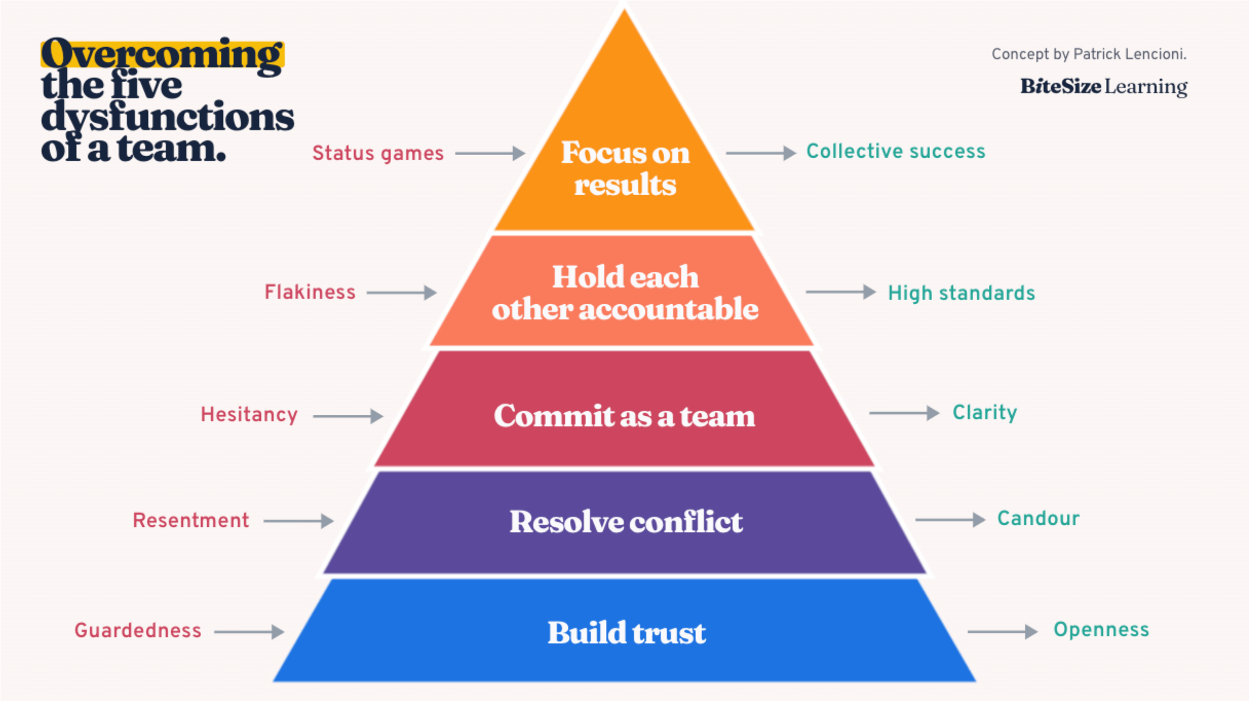


Figure 4

Recommendations to Project

The organised approach to the current research is presented in this overview of the literature review to which a detailed summary of the topic has been presented, however for this project there will be an element of originality as it will review how hosting big tournaments affects a nation's success rate, and how long friendly matches affect team performances—all while incorporating machine learning and data visualisation techniques. A more detailed understanding of football through data analysis will be presented and an analysis on different decades of results as players retire or regress in quality, therefore there will be a significant analysing between timelines of players and how different teams prevail depending on the national team. As such, the study seeks to characterise the complex relationship between time and team success, identifying trends that shed light on how different national teams succeed in changing environments. This project seeks to provide comprehensive insights into the diverse field of football analysis by analysing the complex relationships within football dynamics and shedding light on how the sport has changed over time.

2.3 Conclusion

To conclude this literature review consists of machine learning data visualization and scientific projects that provide a detailed understanding of football analysis to a certain extent. The analysis of the body of research in football analytics demonstrates the predictive power of machine learning and how it could benefit necessary stakeholders such as manager and betting agencies to predict the correct winner from the code. In addition, displaying the data in a dashboard helps create a visual picture of football and how it can benefit fan engagement because they can see the complex patterns of data simplified into bar charts and line graphs allows for non-technical stakeholders understand it. To improve the conversation, further study should go into the unexplored areas mentioned to offer a thorough comprehension of the complex dynamics of football that this project will include and is going to be beyond prediction models and visualisation strategies. This thorough literature study provides a substantial comprehension of football analysis by encompassing a variety of machine learning, data visualisation, and scientific endeavours. It draws attention to the predictive power of machine learning techniques and highlights how important stakeholders, including managers and betting agencies looking to make accurate match predictions. But this investigation opens new avenues for research into football dynamics, indicating that more research is necessary to fully examine these undiscovered areas. This project aims to provide a better understanding of football dynamics by going beyond simple prediction models and visualisation techniques that instead seeks to provide a thorough grasp of present football dynamics.

Chapter 2.4: Initial Steps

Overview

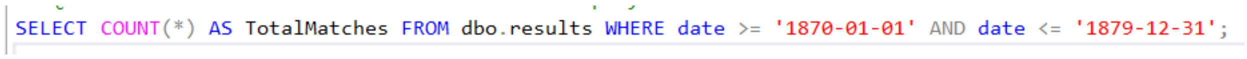
In this chapter, the project will focus on the first stages of date cleaning and preparation on Excel, to which one sizeable dataset will be analyzed and data cleaned for data explore. This will include a dataset on the results of over 45,000 rows of data to be analyzed, all the international results and those matches decided by penalties. When working with datasets, data cleansing is crucial for several reasons, particularly prior to any analysis or exploration as it contributes to accuracy by correcting mistakes, discrepancies, or missing values in the data. Regarding trustworthy judgements and decisions to be made based on the information given, accuracy is essential. Additionally, data cleaning makes the dataset easier to read and understand. Formats are standardized, duplication is eliminated, and outliers are addressed to make the information more arranged and simpler for the average reader to understand. By cleaning the data, you can make sure that any insights you get from it are well-founded and avoid biased or false information that could distort your findings.

2.3 SQL

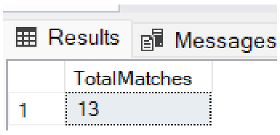
In this section of the project, the data has now been cleaned and prepared, now data is going to be explored using the query language and to find patterns of results, trends of how international teams performed dependant on the scenario of the match, and it will provide an overview of data exploration to football results. Below will be a set questions for the first decade of international football with results.

Results Dataset 13 rows - 1870 - 1879

1) **How many games were played in the 1870s.**



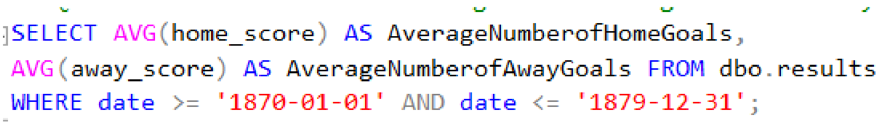
```sql
SELECT COUNT(*) AS TotalMatches FROM dbo.results WHERE date >= '1870-01-01' AND date <= '1879-12-31';
```



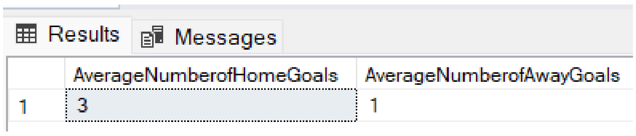
	TotalMatches
1	13

This is the first decade of football and only thirteen matches were played.

2) What is the average number of goals scored by home teams and away teams



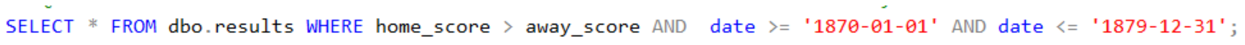
```sql
SELECT AVG(home_score) AS AverageNumberofHomeGoals,
AVG(away_score) AS AverageNumberofAwayGoals FROM dbo.results
WHERE date >= '1870-01-01' AND date <= '1879-12-31';
```



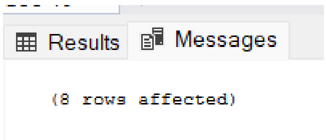
	AverageNumberofHomeGoals	AverageNumberofAwayGoals
1	3	1

The home team is three times as likely to score than the away team.

3) Games were home team won in 1870s.



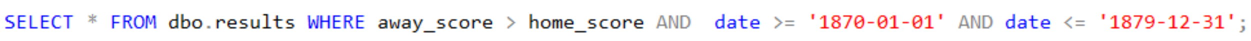
```
SELECT * FROM dbo.results WHERE home_score > away_score AND  date >= '1870-01-01' AND date <= '1879-12-31';
```



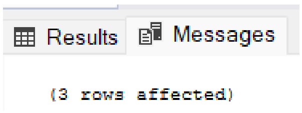
```
(8 rows affected)
```

Eight out of the thirteen matches which shows 60% of the time the home team won.

3b) Games were away team won in 1870s.



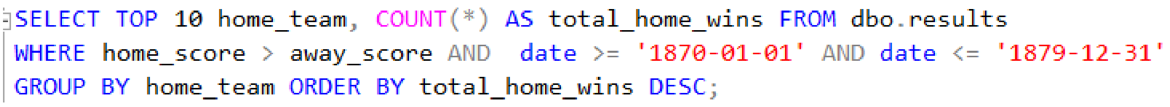
```
SELECT * FROM dbo.results WHERE away_score > home_score AND  date >= '1870-01-01' AND date <= '1879-12-31';
```



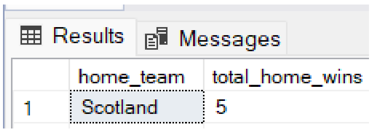
```
(3 rows affected)
```

The away team only won three matches in the 1870s.

4) National team with most home wins in 1870s



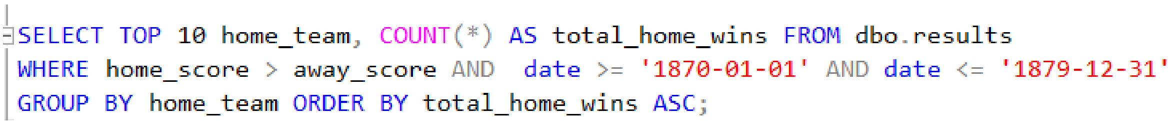
```
SELECT TOP 10 home_team, COUNT(*) AS total_home_wins FROM dbo.results
WHERE home_score > away_score AND  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY total_home_wins DESC;
```



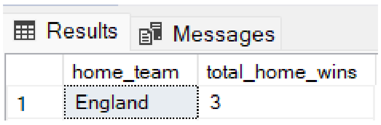
	home_team	total_home_wins
1	Scotland	5

Scotland was the best home team in terms of winning.

4b) National team with least home wins in 1870s.



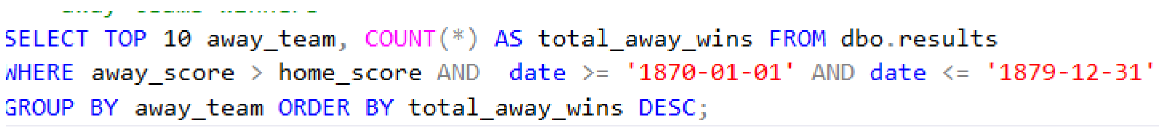
```
SELECT TOP 10 home_team, COUNT(*) AS total_home_wins FROM dbo.results
WHERE home_score > away_score AND  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY total_home_wins ASC;
```



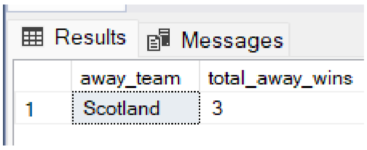
	home_team	total_home_wins
1	England	3

England won the least home matches however only they and Scotland played at home in 1870s.

4c) National team with the most away wins in the 1870s.



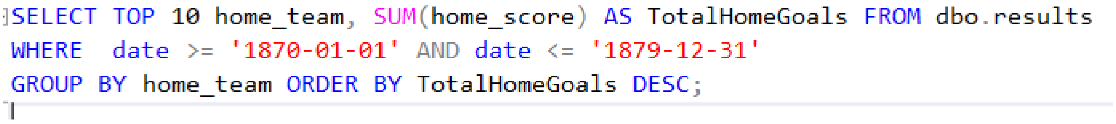
```
SELECT TOP 10 away_team, COUNT(*) AS total_away_wins FROM dbo.results
WHERE away_score > home_score AND  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY away_team ORDER BY total_away_wins DESC;
```



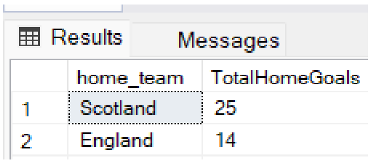
	away_team	total_away_wins
1	Scotland	3

Scotland was the most successful away team in the 1870s.

5) Total Home goals scored in 1870s



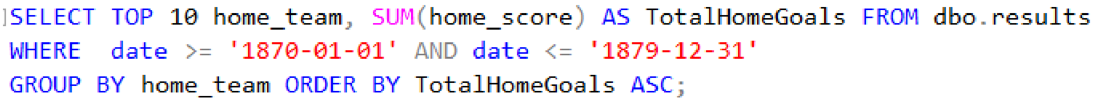
```
SELECT TOP 10 home_team, SUM(home_score) AS TotalHomeGoals FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY TotalHomeGoals DESC;
```



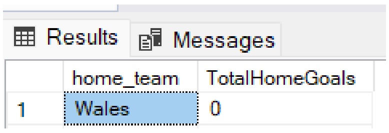
	home_team	TotalHomeGoals
1	Scotland	25
2	England	14

Scotland scored the most goals at home, eleven more than England.

5b) Least Home goals scored in 1870s.



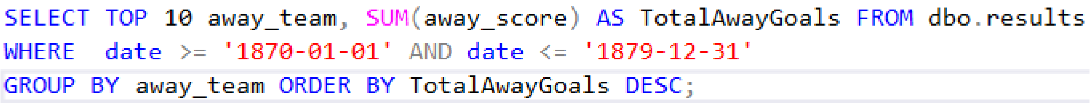
```
SELECT TOP 10 home_team, SUM(home_score) AS TotalHomeGoals FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY TotalHomeGoals ASC;
```



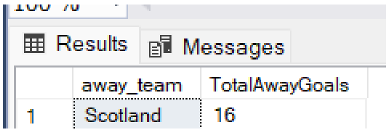
	home_team	TotalHomeGoals
1	Wales	0

Wales failed to score at home in 1870s.

6) Highest away goals scored in 1870s.



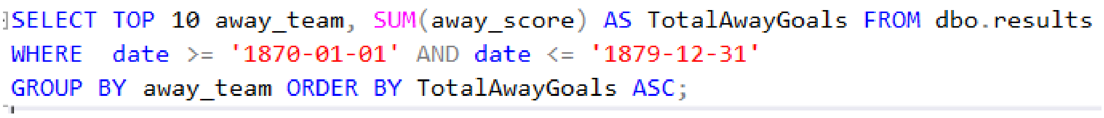
```
SELECT TOP 10 away_team, SUM(away_score) AS TotalAwayGoals FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY away_team ORDER BY TotalAwayGoals DESC;
```



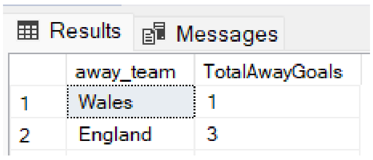
	away_team	TotalAwayGoals
1	Scotland	16

Scotland scored the most away goals in the 1870s.

6b) least away goals scored in 1870s.



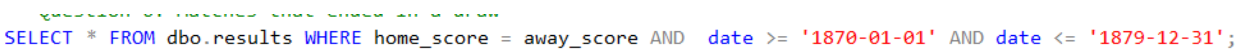
```
SELECT TOP 10 away_team, SUM(away_score) AS TotalAwayGoals FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY away_team ORDER BY TotalAwayGoals ASC;
```



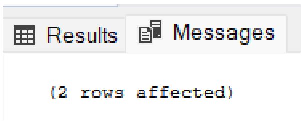
	away_team	TotalAwayGoals
1	Wales	1
2	England	3

Wales scored the least away goals, England being a close second.

7) Matches that ended in a draw in 1870s.



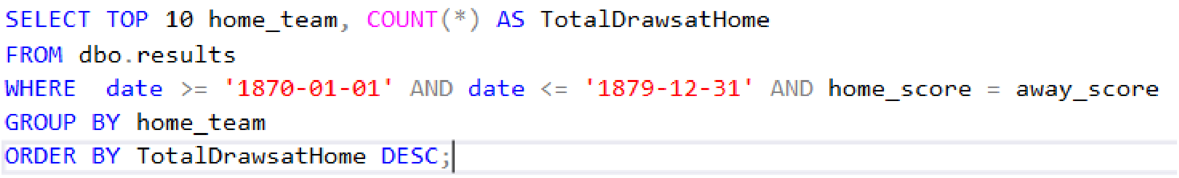
```
  Question 6: Matches that ended in a draw
SELECT * FROM dbo.results WHERE home_score = away_score AND  date >= '1870-01-01' AND date <= '1879-12-31';
```



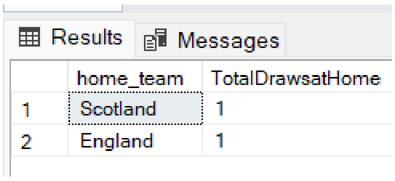
```
(2 rows affected)
```

Only two matches ended in a draw in 1870s.

7b) most home team draws in 1870s.



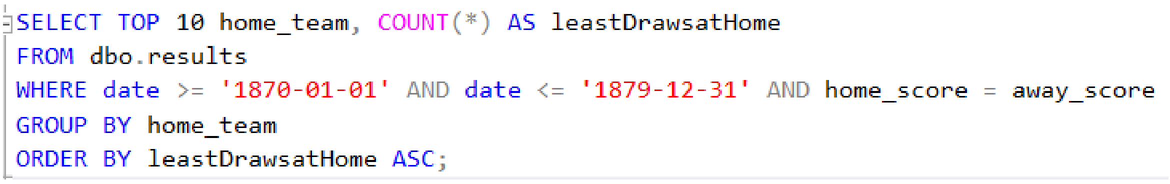
```
SELECT TOP 10 home_team, COUNT(*) AS TotalDrawsatHome
FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31' AND home_score = away_score
GROUP BY home_team
ORDER BY TotalDrawsatHome DESC;
```



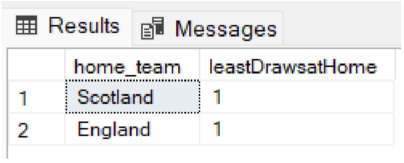
	home_team	TotalDrawsatHome
1	Scotland	1
2	England	1

The joint most draws at home in this era of football is both Scotland and England.

7c) Least draws at home in 1870s.



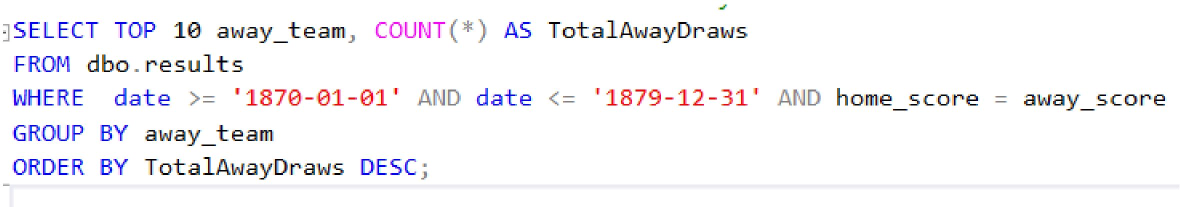
```
SELECT TOP 10 home_team, COUNT(*) AS leastDrawsatHome
FROM dbo.results
WHERE date >= '1870-01-01' AND date <= '1879-12-31' AND home_score = away_score
GROUP BY home_team
ORDER BY leastDrawsatHome ASC;
```



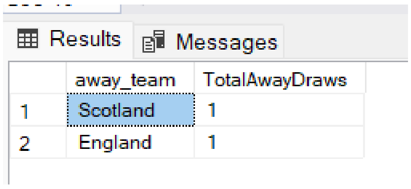
	home_team	leastDrawsatHome
1	Scotland	1
2	England	1

They both also had the least draws due to being the only draw of this decade.

8) Most away draws in 1870s.



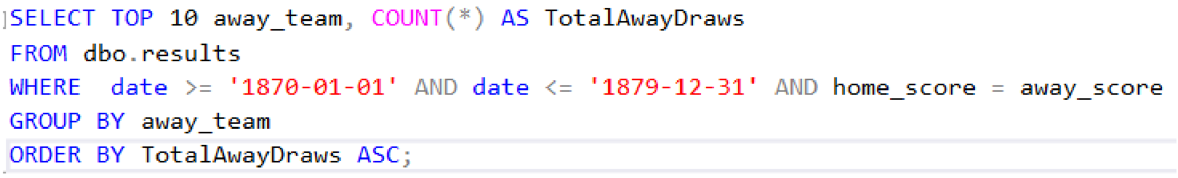
```
SELECT TOP 10 away_team, COUNT(*) AS TotalAwayDraws
FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31' AND home_score = away_score
GROUP BY away_team
ORDER BY TotalAwayDraws DESC;
```



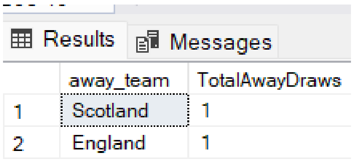
	away_team	TotalAwayDraws
1	Scotland	1
2	England	1

Both England and Scotland and the most away draws with only one.

8b) least away draws in 1870s.



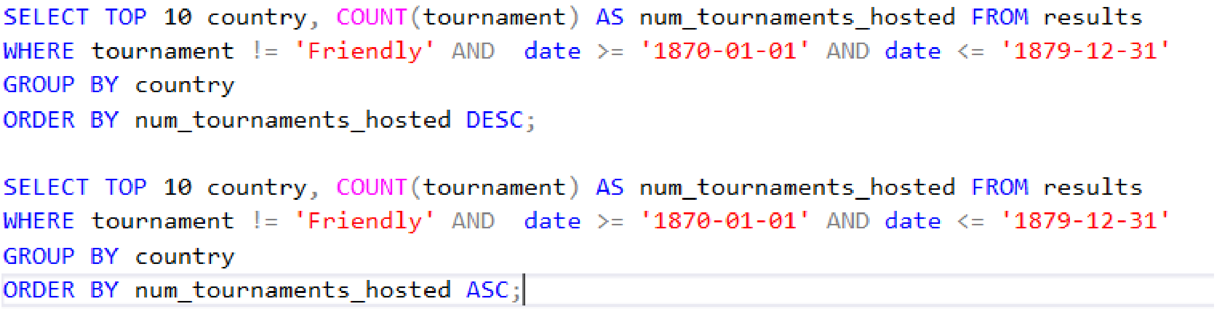
```
SELECT TOP 10 away_team, COUNT(*) AS TotalAwayDraws
FROM dbo.results
WHERE  date >= '1870-01-01' AND date <= '1879-12-31' AND home_score = away_score
GROUP BY away_team
ORDER BY TotalAwayDraws ASC;
```



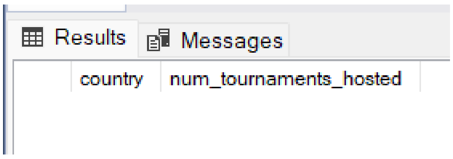
	away_team	TotalAwayDraws
1	Scotland	1
2	England	1

Both England and Scotland and the least away draws with only one.

9) Tournaments hosts in the 1870s.



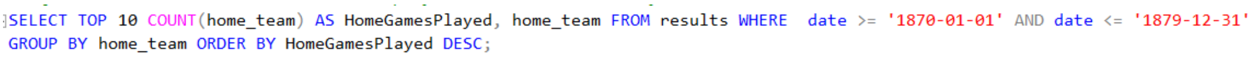
```
SELECT TOP 10 country, COUNT(tournament) AS num_tournaments_hosted FROM results
WHERE tournament != 'Friendly' AND  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY country
ORDER BY num_tournaments_hosted DESC;

SELECT TOP 10 country, COUNT(tournament) AS num_tournaments_hosted FROM results
WHERE tournament != 'Friendly' AND  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY country
ORDER BY num_tournaments_hosted ASC;
```



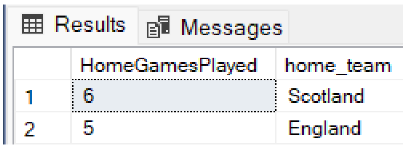
	country	num_tournaments_hosted

No countries hosted a singled tournament in this decade due to it being the first era of football.

10) Home games played in 1870s.



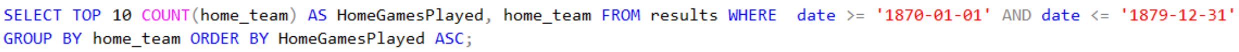
```
SELECT TOP 10 COUNT(home_team) AS HomeGamesPlayed, home_team FROM results WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY HomeGamesPlayed DESC;
```



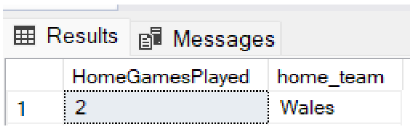
	HomeGamesPlayed	home_team
1	6	Scotland
2	5	England

Scotland played the home games in the 1870s, one more than England.

10b) least home games played in 1870s.



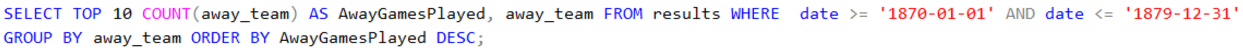
```
SELECT TOP 10 COUNT(home_team) AS HomeGamesPlayed, home_team FROM results WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY home_team ORDER BY HomeGamesPlayed ASC;
```



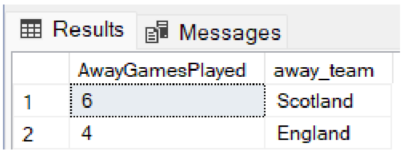
	HomeGamesPlayed	home_team
1	2	Wales

Wales played the least home games, exactly three times less than Scotland.

11) Highest away games played in 1870s.



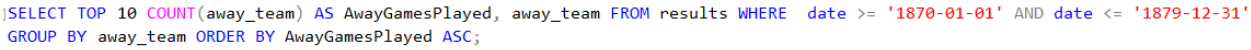
```
SELECT TOP 10 COUNT(away_team) AS AwayGamesPlayed, away_team FROM results WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY away_team ORDER BY AwayGamesPlayed DESC;
```



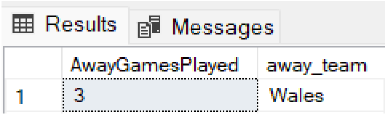
	AwayGamesPlayed	away_team
1	6	Scotland
2	4	England

Scotland played the most away games, two more than second place England.

11b) Least away games played in 1870s.



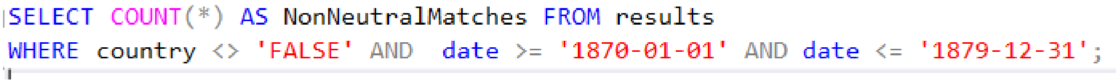
```
SELECT TOP 10 COUNT(away_team) AS AwayGamesPlayed, away_team FROM results WHERE  date >= '1870-01-01' AND date <= '1879-12-31'
GROUP BY away_team ORDER BY AwayGamesPlayed ASC;
```



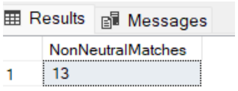
	AwayGamesPlayed	away_team
1	3	Wales

Wales played the least away games played in 1870s.

12) Neutral venue matches in 1870s.



```
SELECT COUNT(*) AS NonNeutralMatches FROM results
WHERE country <> 'FALSE' AND  date >= '1870-01-01' AND date <= '1879-12-31';
```



	NonNeutralMatches
1	13

All matches played a team had an advantage in the 1870s.

<u>Summary of 1870s data explorations</u>

This is start of international football that is document to which only thirteen matches were played. The home team held a clear upper hand as 60% of games ended in them winning. Scotland was both the dominant force at home and away, they scored the most goals overall. On the other hand, Wales failed to find the net whether they played home or away and played the least matches whilst Scotland played the most. Only two matches ended in a draw, to which Scotland and England share the highest draws. This era signifies the home advantage, Scotland dominance, England and Wales finding their feet in the early stage of footballs and limited matches.

Chapter 3 Requirements Specification

3.1 Functional Requirements

Functional requirements are used to determine the behaviors and features a system must have to meet user needs and expectations. Excel uses data cleaning activities such as deleting duplicates, changing date formats, and assuring data accuracy. SQL data exploration as query development for historical football match analysis and statistical calculations. These requirements serve as a guideline for development, ensuring that the system executes its functions properly. Functional requirements are critical for matching development efforts to user needs and assuring system functionality.

Excel Data cleaning Requirements

Remove Duplicates

- Identify and remove duplicated rows from the dataset to ensure it is as accurate as possible.

Change Source Type

- Convert source type of data from text and make sure the date is formatted correctly.

New Date Format

- Transform the date format of the dataset to enable proper date-time.

Capitalize First Word

- Capitalize the first word in text entries to improve readability of the data.

Data Quality Check

- Perform a thorough check to identify and address any errors or inconsistencies in the data, including empty rows or cells.

Find and replace.

- Utilize the find and replace function to correct common formatting issues or errors, such as replacing incorrect date separators with the appropriate format.

Change Column Names

- Rename columns in the dataset to provide clear and descriptive labels that reflect the content of each column, improving understanding and usability.

SQL Data Exploration Requirements

Data Querying

- Develop SQL queries to retrieve and analyse historical football match data from the dataset from 1872 – 2023.
- Ensure the queries are structured to efficiently manage a large dataset to retrieve information based on the criteria.

Statistics

- Calculate the total numbers of games played within a decade and provide an overview of match frequency over the years.

- Present data such as top home team goal scores, number of wins, loses and draws to understand the overall match dynamics within a decade.

Team performance

- Identify the number of games won by home and away teams separately across the data and the significance of home advantage.
- Determine the pattern, trends, and highlight which period of dominance teams had, answers vary as there is over 15 decades in this dataset.

Goals Analysis

- Analyse the distribution of goals scored by home and away teams over the years to identify scoring patterns and trends.
- Calculate metrics such as highest scoring matches, goal away from home and teams with the highest goals.

Match Outcomes and Trends

- Determine the frequency of different match outcomes over a decade and understand the competitiveness.
- Identify trends in match's outcomes over time and changes in percentages of draws, wins and losses.

National Team Performance

- Evaluate the performance of national teams in terms of home and away win, goals scored and tournament participation over the years.
- Compare the performance of different national teams to identify historical successes teams and periods of dominance and produce a summary per decade.

Tournaments Analysis

- Analyse tournament data to identify hosts, participants, match outcomes, and trends over the period.
- Determine the impact of tournaments on team performance and assess the significance of home advantages.

Machine Learning Requirements

Hypothesis

When analysing international football matches from the past 150 years, not only the team skill is necessary but also where the match is played, and type of tournament can influence the outcome of match. Therefore, variables such as tournament type, city of the match is played on neutral ground correlation with match results, the goal is to discover if a team excels in particular settings.

Import necessary variables and display rows.

- Load the dataset and display a few rows.
- Understand structure and content.
- Import correct variables.

Column Types

- Verify data types of each column.
- Ensures are correctly interpreted for analysis.

Display Top cities and countries.

- Top 10 countries to understand which places were frequent played in and use for analysis.
- Matches played in the most cities allows for interpretation and insights.
- Focus on teams and aim to cover every continent.

Check neutral Grounds.

- Ensure matches out of the result to see impact of neutrality.
- Compare the quality of results depending on the venue.

Predict future matches.

- Machine learning models, considering newly engineered features like tournaments significance.
- Evaluate the hypothesis, see the quality and truth to it, and find the patterns seen in it.

Review Hypothesis

- Include factors such as venue, tournament significance and country played in.
- Home advantage influence match outcomes
- Specifically analyse two teams per continent

3.2 Non-Functional Requirements

Non-functional requirements are the quality attributes and limitations that the system must follow to ensure overall performance, dependability, security, and usability. Non-functional requirements have aspects to this project, including the system's performance in handling large datasets for data cleaning and SQL querying, processing, and analyzing data, it measures to protect sensitive information, and its usability to ensure ease of use for analysts and stakeholders. These requirements will drive the system's design, development, testing, and deployment, ensuring that it satisfies the required quality standards and effectively serves the project's objectives.

Performance

- Clean Excel data
- Optimize SQL queries for managing historical football match datasets from 1872 to 2023.
- Machine learning models should be developed to process and analyze datasets efficiently, considering their quantity and complexity.

Reliability

- The data cleaning processes should accurately manage different data formats and types without causing data loss or corruption.
- SQL queries should retrieve accurate information from the dataset without errors.
- Machine learning models should provide consistent and reliable predictions based on the input variables and dataset.

Security

- Access to data is public from a website called Kaggle to which all users can download themselves a copy and view dataset on Excel.

Usability

- The Excel data-cleaning tool should have a user interface with clear instructions and feedback to enhance usability.
- SQL queries should be well documented and structured in a way that is easy for analysts to understand and modify.
- Machine learning predictions should be presented in a format that is easy for stakeholders to interpret and act upon.

Maintainability

- The system components, including the Excel data cleaning tool, SQL queries, and machine learning models, should be well documented to facilitate maintenance and updates.
- Changes to data cleaning rules, SQL queries, or machine learning models should be easy to implement without requiring extensive amount of work.

Compatibility

- The Excel data cleaning tool, SQL queries, and machine learning models should be compatible with the platforms and be deployed easily.

Portability

- The system should be designed to be portable across different environments or platforms, allowing for easy deployment and adaptation.

Chapter 4: Methodology

4.1 Excel Data Cleaning Methodology

1) Remove duplicates

Step: The first step involves identifying and removing any duplicate rows from the dataset.

Tool/Technique: Microsoft Excel's built-in functionality for removing duplicate rows is utilized for this task.

Description: By removing duplicate rows, the dataset's originality was protected, leading to more accurate findings during analysis.

2) Change source type

Step: The source type of the data, initially in text format, was changed to the date format.

Tool/Technique: Microsoft Excel's formatting tools were employed to convert the source type from text to the desired date format.

Description: This conversion is necessary to ensure that the dates in the dataset were properly formatted for analysis, allowing for precise date-time calculations and dividing data.

3) New date format

Step: The date format of the dataset was transformed to ensure consistency and facilitate date-time calculations.

Tool/Technique: Excel's formatting options were used to adjust the date format according to the project requirements.

Description: Adopting a standardized date format across the dataset enabled accurate date-time calculations and sorting, crucial for analyzing historical football match data.

4) Capitalize first word

Step: A process was implemented to capitalize only the first word in text entries.

Tool/Technique: Excel's text functions for formulas were applied to achieve this transformation.

Description: Capitalizing only the first word improved the consistency and readability of text entries in the dataset, enhancing its usability for analysis purposes.

5) Data quality

Step: A thorough check is performed to identify and address any errors or inconsistencies in the data, including empty rows or cells.

Tool/Technique: Excel's data validation features are utilized to identify and rectify data quality issues.

Description: Ensuring data accuracy through quality checks minimized errors and discrepancies, enhancing the reliability of analysis results.

6) Find and replace

Step: Common formatting issues or errors, such as replacing incorrect date separators, were corrected using the find and replace function.

Tool/Technique: Excel's find and replace feature was employed to search for specific patterns and replace them with the desired format.

Description: By standardizing formatting conventions, such as date separators, data consistency was maintained, facilitating easier analysis and interpretation.

7) Change column name

The name of the column has been changed to provide context of what is being analysed.

Step: Column names in the dataset were renamed to provide clear and descriptive labels reflecting the content of each column.

Tool/Technique: Excel's functionality for renaming columns or headers was utilized for this task.

Description: Clear and descriptive column names improved the understanding of the dataset's structure, making it easier to interpret and analyze.

These steps collectively formed the methodology used for cleaning the Excel dataset, ensuring that it was prepared and optimized for further analysis in the subsequent chapters.

4.2 SQL Data Exploration Methodology

To explore the historical international football match data and perform statistical calculations, the following methodology was employed:

Identifying Analysis Objectives: The first step involved identifying the key analysis objectives, such as determining total matches played, average goals scored, home team performance, away team performance, top-performing teams, and tournament hosting statistics.

Query Development: Queries were developed to extract relevant information from the dataset. Each query was designed to address a specific analysis objective.

- Query 1: Find the total number of matches played in international football to give an overview of the dataset.
- Query 2: Calculated the average number of home and away goals scored in international football.
- Query 3: Identified matches where the home team emerged victorious and calculated the percentage of such matches for each decade to draw a conclusion.
- Query 4: Listed the top ten most successful home teams of all time in international football.
- Queries 5-12: Similar queries were formulated to address various analysis objectives, such as identifying successful away teams, teams with the most home/away goals, teams with the least away goals all help formulate the hypothesis.

Efficient Query Structuring: Queries were structured to manage large data and retrieve relevant information. Techniques such as indexing, proper query optimization, and selective querying were employed to enhance efficiency and minimize processing time. Additionally, the queries were optimized to leverage the databases for efficient data retrieval and aggregation.

Refinement: The methodology involved a refinement process, where queries were based on performance feedback and analysis requirements. This iterative approach allowed for the optimization of query performance and analysis outcomes.

Query Structuring Efficiency:

The queries were structured to efficiently oversee large datasets and retrieve relevant information by implementing the following strategies:

Indexing: Proper indexing was applied to columns frequently used in the WHERE clause or involved in join operations. This enhanced query performance by reducing the time required for data retrieval.

Aggregation Techniques: Aggregation functions such as COUNT (), AVG (), and SUM () utilized to perform calculations efficiently.

Query Optimization: techniques, such as using appropriate WHERE clause predicates and avoiding unnecessary calculations, were implemented to streamline query execution and improve efficiency.

By using these strategies, the queries were structured to efficiently manage large datasets and retrieve relevant information, facilitating comprehensive analysis of historical football match data.

4.3 Machine Learning Methodology

- Present the methodology for utilizing machine learning techniques to predict future football match outcomes.
- Discuss the creation of hypotheses, selection of variables, data preprocessing steps, and model evaluation techniques.

Hypothesis Formulation

In the analysis of international football matches spanning the past 150 years, factors beyond team skill, such as match location (city and venue type), and tournament type, significantly influence match outcomes.

Dataset Preparation

- Loading the Dataset only a few rows to preview it and help user visualize the rest of it:
- Utilize appropriate libraries (e.g., pandas) to load the dataset containing historical football match data.
- Display rows to understand the structure and content of the dataset.

Data Understanding and Verification

- Examine the structure and content of the dataset to ensure it aligns with the analysis objectives.
- Verify the data types of each column and ensure correct interpretation for subsequent analysis.

Feature Engineering

- Identify relevant variables such as tournament type, match venue (city), and neutral ground indicator.
- Import these variables from the dataset for further analysis.

Column Types Verification

- Verify the data types of each column to ensure consistency and accuracy.
- Convert data types if necessary to facilitate analysis.

Display Top Cities and Countries

- Identify the top ten countries and cities based on the frequency of matches played.
- Analyze matches played in multiple cities to gain insights into geographical patterns.

Check Neutral Grounds

- Examine matches played on neutral grounds and their impact on match outcomes.
- Compare match results based on venue neutrality to evaluate its significance.

Machine Learning Model Development

- Predicting future matches
- Utilize machine-learning models to predict match outcomes.
- Incorporate engineered features such as tournament significance, match venue, and team performance indicators into the model.

Model Evaluation

- Evaluate the hypothesis by assessing the performance of the machine-learning model.
- Analyze the quality of predictions in relation to the formulated hypothesis and the included factors.

Review Hypothesis and Analysis

- Assessing Hypothesis Factors
- Review the hypothesis considering factors such as venue, tournament significance, and home advantage.
- Determine the influence of each factor on match outcomes based on the analysis results.

Specific Team Analysis

- Conduct a detailed analysis of two teams who from the SQL data explore won the home matches, analyze dynamics and performance trends, and use the Hypothesis.

Conclusion

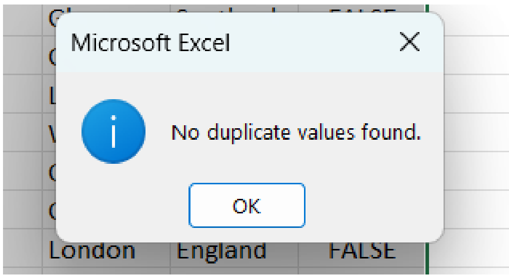
- Summarize the findings from the machine learning analysis.
- Discuss the implications of the results on the formulated hypothesis and broader insights into international football match outcomes.

Chapter 5: Implementation & Testing

5.1 Implementation Details

Excel Data Cleaning

1) Remove duplicates



Microsoft Excel

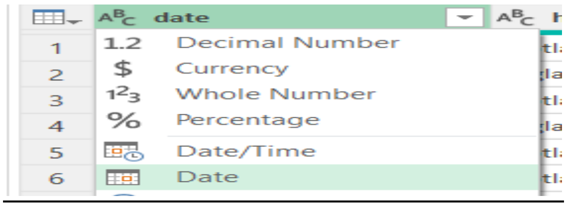
No duplicate values found.

The first step is to proceed with the duplication of any rows to which none is, this provides originality and accurate findings further down.

Issues:

- Challenge: To identify duplicates accurately, especially in large datasets, can be challenging.
- Solution: Use Excel's built-in functionality for accurate identification and removal of duplicates

2) Change source type

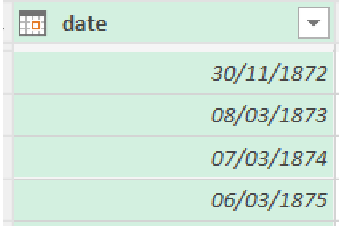


Implementing a change to the source type because it is text based, therefore, it has been changed to the date format.

Issues:

- Challenge: Convey text data to date format may result in errors, especially if the data format is inconsistent that leads to poor analysis.
- Solution: Validate the data format before conversion and oversee any inconsistencies appropriately. Implement error-handling mechanisms to address conversion issues to lead to detailed analysis.

3) New date format



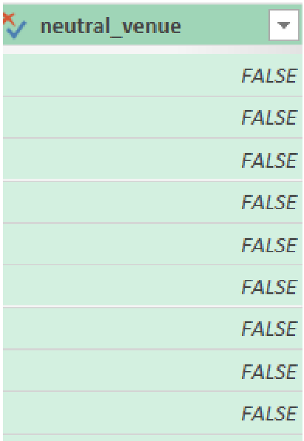
date
30/11/1872
08/03/1873
07/03/1874
06/03/1875

These are a few rows of the date with the correct format to allow the date of matches to be in a familiar format.

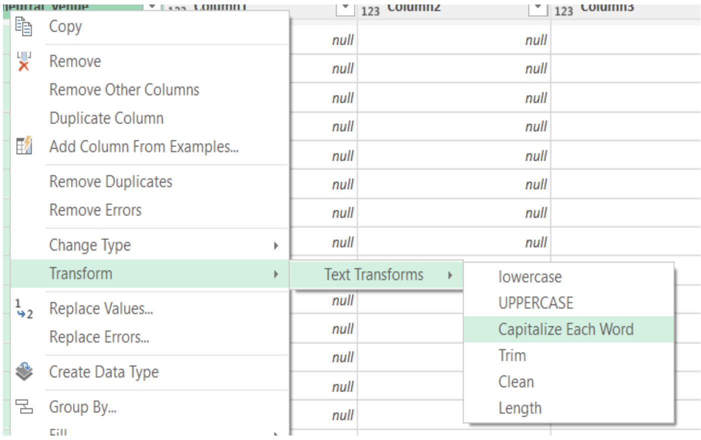
Issues:

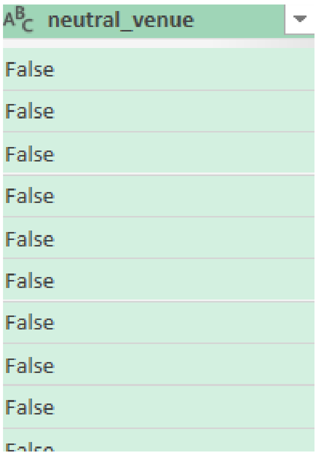
- Challenge: Ensuring new date format across the dataset can be difficult, especially if the original dates are in various formats due to results being played over 150 words ago.
- Solution: Use Standardize to date format using Excel's formatting options and implement validation checks to identify and correct any inconsistencies before going on to the next step.

4) Capitalize first word



neutral_venue
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE





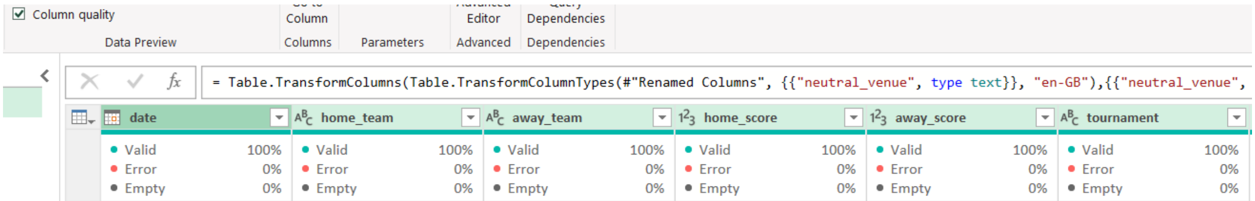
neutral_venue
False
False
False
False
False
False
False
False
False

This is the three-step process for transforming capital words only to just the first word, this allows for it be more user friendly.

Issues:

- Challenge: Capitalizing only the first word in text entries accurately might be challenging, especially if the entries contain exceptional cases or exceptions.
- Solution: Develop text manipulation formulas and assess the capitalization process thoroughly to ensure accuracy.

5) Data quality



```
= Table.TransformColumns(Table.TransformColumnTypes(#"Renamed Columns", {{"neutral_venue", type text}}, "en-GB"),{{"neutral_venue",
```

date		home_team		away_team		home_score		away_score		tournament	
● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%	● Valid	100%
● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%	● Error	0%
● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%	● Empty	0%

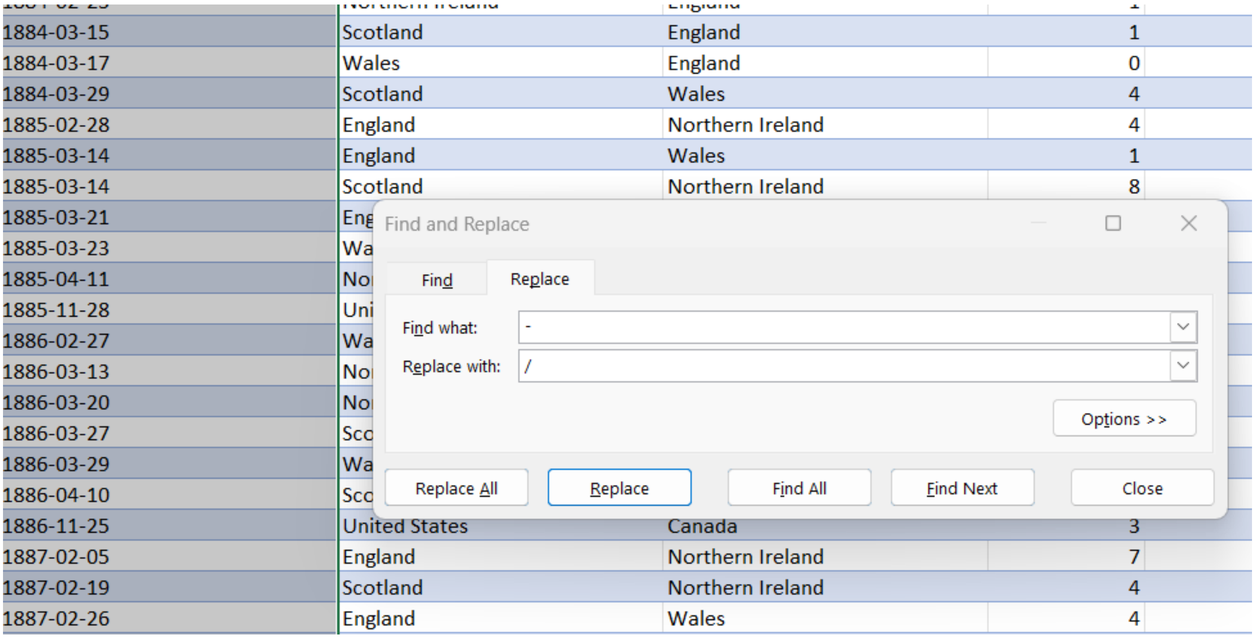
Checking whether the data has any errors or empty rows.

Issues:

- Challenge: Identifying and addressing data errors or empty rows manually can be time-consuming and error prone.
- Solution: Utilize Excel's data validation features and automated scripts to perform comprehensive data quality checks.

6) Find and replace

Replacement of the dash to a slash, as it is the common calendar date separator and is a common method of displaying the date.



1884-02-23	Northern Ireland	England	1
1884-03-15	Scotland	England	1
1884-03-17	Wales	England	0
1884-03-29	Scotland	Wales	4
1885-02-28	England	Northern Ireland	4
1885-03-14	England	Wales	1
1885-03-14	Scotland	Northern Ireland	8
1885-03-21	Eng		
1885-03-23	Wa		
1885-04-11	No		
1885-11-28	Uni		
1886-02-27	Wa		
1886-03-13	No		
1886-03-20	No		
1886-03-27	Sco		
1886-03-29	Wa		
1886-04-10	Sco		
1886-11-25	United States	Canada	3
1887-02-05	England	Northern Ireland	7
1887-02-19	Scotland	Northern Ireland	4
1887-02-26	England	Wales	4

Find and Replace

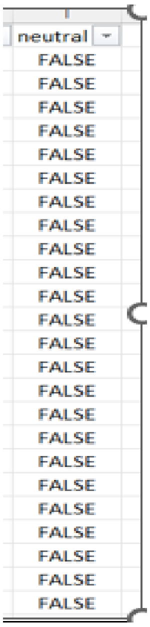
Find what: -

Replace with: /

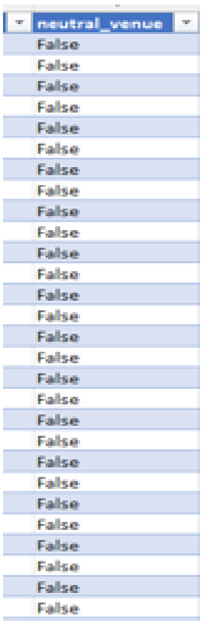
Issues:

- Challenge: Find and replace specific patterns accurately throughout the dataset can be challenging.
- Solution: Use Excel's find and replace functionality cautiously, considering potential variations in the data patterns.

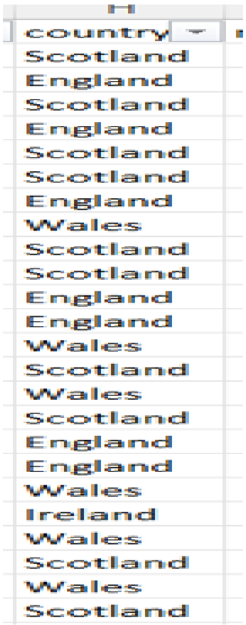
7) Change column name



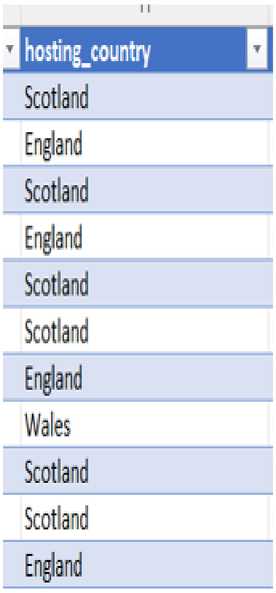
neutral
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE
FALSE



neutral_venue
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False
False



country
Scotland
England
Scotland
England
Scotland
Scotland
England
Wales
Scotland
Scotland
England
England
Wales
Scotland
Wales
Scotland
England
England
Wales
Ireland
Wales
Scotland
Wales
Scotland



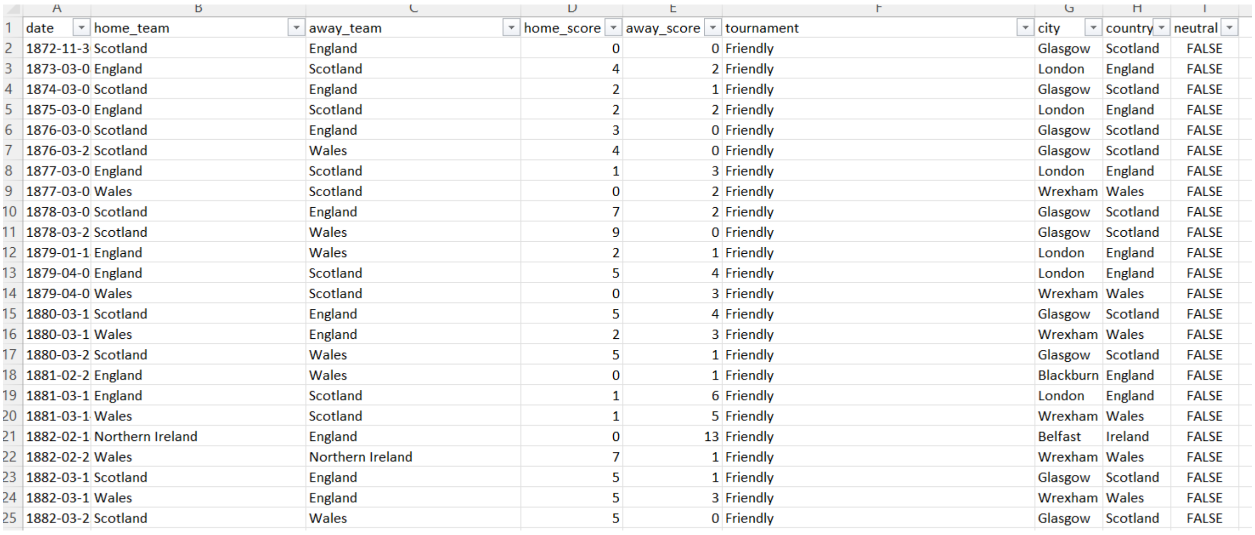
hosting_country
Scotland
England
Scotland
England
Scotland
Scotland
England
Wales
Scotland
Scotland
England

The name of the column has been changed for it to provide better context for the dataset.

Issues:

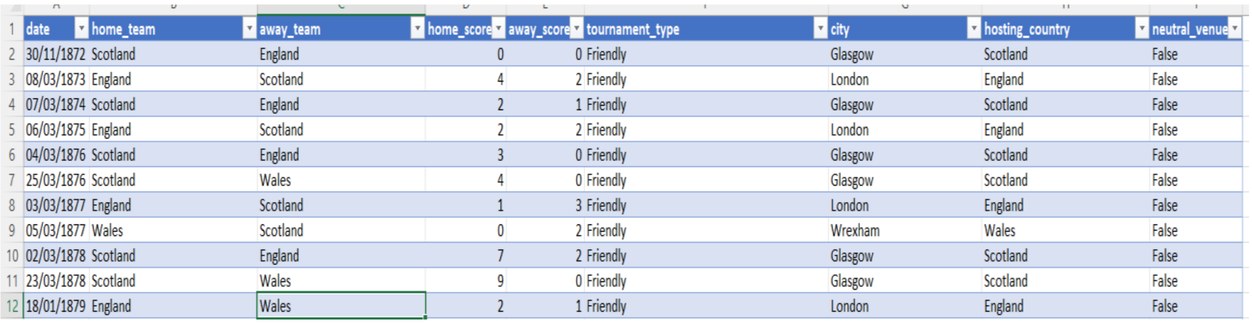
- Challenge: Renaming columns accurately to provide clear context may be challenging, especially if the dataset contains many columns and can lead to confusion.
- Solution: Develop a standardized naming convention and apply it consistently across all columns. Consider automating the column renaming process using Excel macros or scripts to ensure accuracy and efficiency.

Old dataset



1	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
2	1872-11-3	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
3	1873-03-0	England	Scotland	4	2	Friendly	London	England	FALSE
4	1874-03-0	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
5	1875-03-0	England	Scotland	2	2	Friendly	London	England	FALSE
6	1876-03-0	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
7	1876-03-2	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
8	1877-03-0	England	Scotland	1	3	Friendly	London	England	FALSE
9	1877-03-0	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE
10	1878-03-0	Scotland	England	7	2	Friendly	Glasgow	Scotland	FALSE
11	1878-03-2	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	FALSE
12	1879-01-1	England	Wales	2	1	Friendly	London	England	FALSE
13	1879-04-0	England	Scotland	5	4	Friendly	London	England	FALSE
14	1879-04-0	Wales	Scotland	0	3	Friendly	Wrexham	Wales	FALSE
15	1880-03-1	Scotland	England	5	4	Friendly	Glasgow	Scotland	FALSE
16	1880-03-1	Wales	England	2	3	Friendly	Wrexham	Wales	FALSE
17	1880-03-2	Scotland	Wales	5	1	Friendly	Glasgow	Scotland	FALSE
18	1881-02-2	England	Wales	0	1	Friendly	Blackburn	England	FALSE
19	1881-03-1	England	Scotland	1	6	Friendly	London	England	FALSE
20	1881-03-1	Wales	Scotland	1	5	Friendly	Wrexham	Wales	FALSE
21	1882-02-1	Northern Ireland	England	0	13	Friendly	Belfast	Ireland	FALSE
22	1882-02-2	Wales	Northern Ireland	7	1	Friendly	Wrexham	Wales	FALSE
23	1882-03-1	Scotland	England	5	1	Friendly	Glasgow	Scotland	FALSE
24	1882-03-1	Wales	England	5	3	Friendly	Wrexham	Wales	FALSE
25	1882-03-2	Scotland	Wales	5	0	Friendly	Glasgow	Scotland	FALSE

Updated dataset



1	date	home_team	away_team	home_score	away_score	tournament_type	city	hosting_country	neutral_venue
2	30/11/1872	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
3	08/03/1873	England	Scotland	4	2	Friendly	London	England	False
4	07/03/1874	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
5	06/03/1875	England	Scotland	2	2	Friendly	London	England	False
6	04/03/1876	Scotland	England	3	0	Friendly	Glasgow	Scotland	False
7	25/03/1876	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	False
8	03/03/1877	England	Scotland	1	3	Friendly	London	England	False
9	05/03/1877	Wales	Scotland	0	2	Friendly	Wrexham	Wales	False
10	02/03/1878	Scotland	England	7	2	Friendly	Glasgow	Scotland	False
11	23/03/1878	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	False
12	18/01/1879	England	Wales	2	1	Friendly	London	England	False

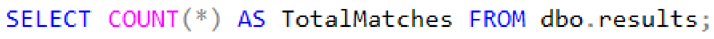
The changes have been made to the dataset, is cleaned, and prepared and is going to be explored in the next chapter.

Issues using Excel:

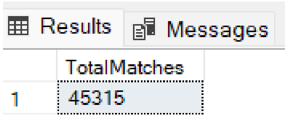
- Challenge: Lack of description can make it difficult for readers to understand the implementation process fully and may reread the screenshots carefully to understand.
- Solution: Review the description to ensure clarity and provide sufficient detail for each step.

SQL Data Exploration - All time results from 1872-2023 findings

1) Total matches played in international football.



```sql
SELECT COUNT(*) AS TotalMatches FROM dbo.results;
```



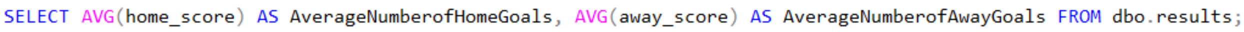
	TotalMatches
1	45315

Description: SQL query to retrieve the total number of matches played in international football to which is a little over 45000.

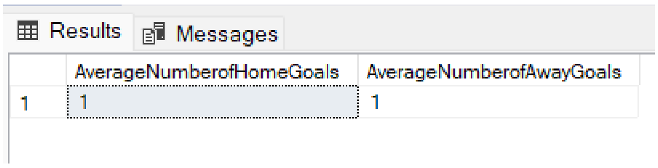
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: AS applied for better wording and count function used to get the number needed for the insight.

2) What is the average number of home and away goal scored in international football?



```sql
SELECT AVG(home_score) AS AverageNumberofHomeGoals, AVG(away_score) AS AverageNumberofAwayGoals FROM dbo.results;
```



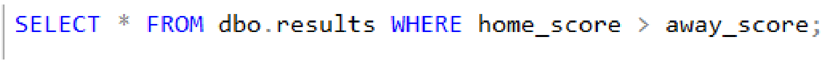
	AverageNumberofHomeGoals	AverageNumberofAwayGoals
1	1	1

Description: SQL query to calculate the average number of home and away goals scored in international football to which the ratio is 1:1 meaning that that both sides are as likely to score a goal.

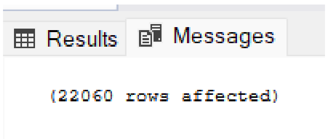
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: Optimized AS conditions for efficient data retrieval and easier for user to understand.

3) Where home team won in international football matches?



```sql
SELECT * FROM dbo.results WHERE home_score > away_score;
```



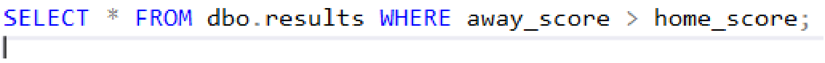
```
(22060 rows affected)
```

Description: SQL query to identify matches where the home team won in international football. This finding proves 49% of matches the home team won of all time, the home advantage.

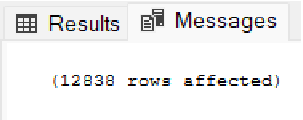
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering home team victories.

3b) where away team won in international football matches?



```sql
SELECT * FROM dbo.results WHERE away_score > home_score;
```



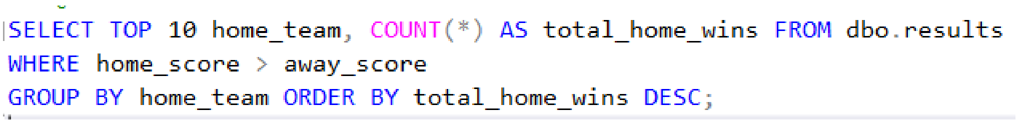
```
(12838 rows affected)
```

Description: SQL query to identify matches where the away team won in international football. 28% of matches the away team have of all time**.**

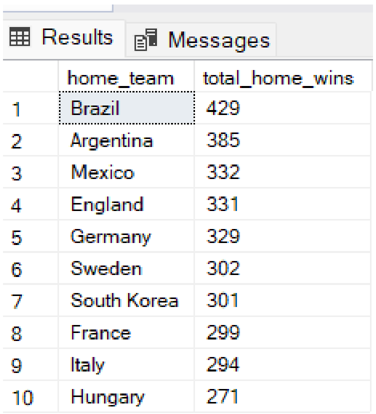
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering home team victories and the greater symbol in favor of the away team to see how many games they won of all time.

4) Top 10 most successful home teams of all time in international football?



```sql
SELECT TOP 10 home_team, COUNT(*) AS total_home_wins FROM dbo.results
WHERE home_score > away_score
GROUP BY home_team ORDER BY total_home_wins DESC;
```



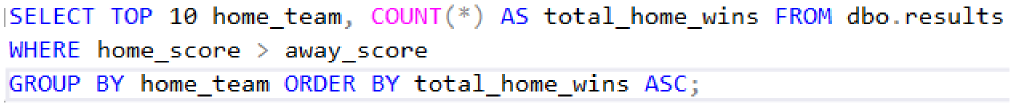
	home_team	total_home_wins
1	Brazil	429
2	Argentina	385
3	Mexico	332
4	England	331
5	Germany	329
6	Sweden	302
7	South Korea	301
8	France	299
9	Italy	294
10	Hungary	271

Description: SQL query to list the top ten most successful home teams of all time in international football. Brazil is the most dominant home winning team of all time to which Argentina is an honourable mention finishing second place. As a result, these two teams will be used for the hypothesis of the machine learning later to prove whether the home advantage, city, tournament have an impact on the result.

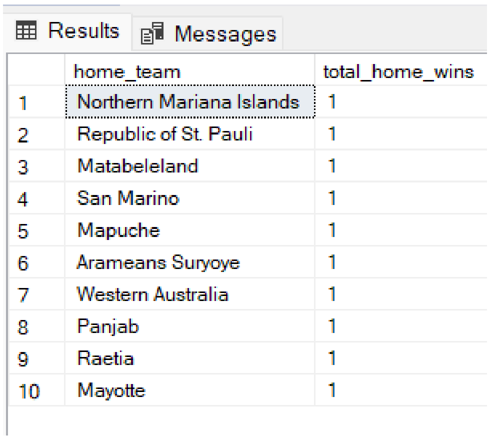
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on home wins, COUNT function used to identify exactly amount of home wins, DESC used to ensure it is in the correct order and the greater symbol to filter out home wins and TOP purpose is to only display results necessary. GROUP BY is used.

4b) Least 10 successful home teams of all time in international football?



```
SELECT TOP 10 home_team, COUNT(*) AS total_home_wins FROM dbo.results
WHERE home_score > away_score
GROUP BY home_team ORDER BY total_home_wins ASC;
```



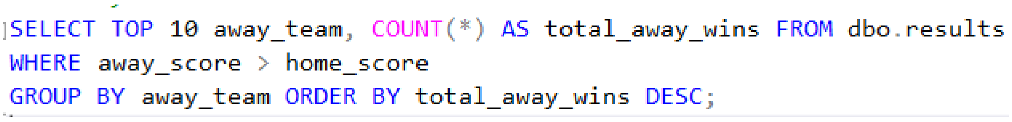
	home_team	total_home_wins
1	Northern Mariana Islands	1
2	Republic of St. Pauli	1
3	Matabeleland	1
4	San Marino	1
5	Mapuche	1
6	Arameans Suryoye	1
7	Western Australia	1
8	Panjab	1
9	Raetia	1
10	Mayotte	1

Description: SQL query to list the top ten least successful home teams of all time in international football. There is a joint least successful home winning teams to which is 10 national teams.

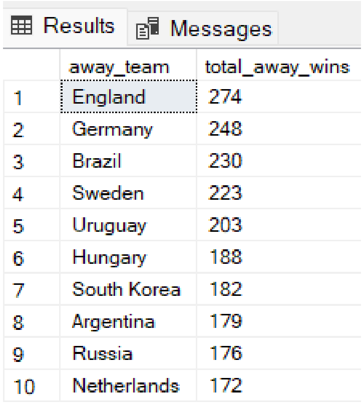
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on least home wins, same steps taken place in question 4a, but the biggest difference is the use of ASC to which it goes from least to the top.

5) Top 10 most successful away teams of all time in international football?



```
SELECT TOP 10 away_team, COUNT(*) AS total_away_wins FROM dbo.results
WHERE away_score > home_score
GROUP BY away_team ORDER BY total_away_wins DESC;
```



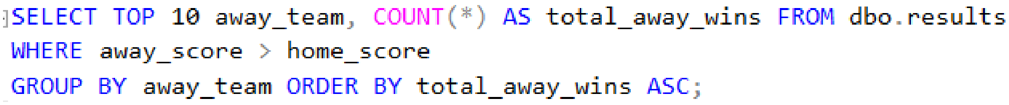
	away_team	total_away_wins
1	England	274
2	Germany	248
3	Brazil	230
4	Sweden	223
5	Uruguay	203
6	Hungary	188
7	South Korea	182
8	Argentina	179
9	Russia	176
10	Netherlands	172

Description: SQL query to list the top ten most successful away teams of all time in international football. England of all time is the most successful away winning team to which they have almost thirty more victories than second place Germany.

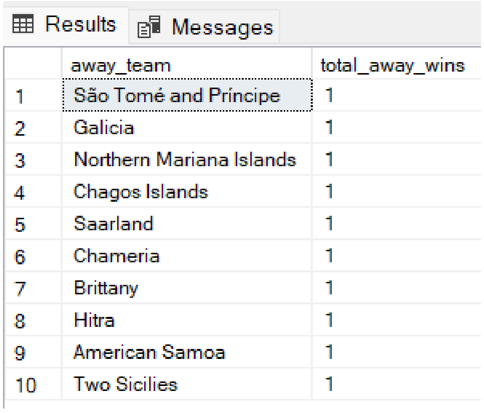
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on away wins, GROUP BY to identify teams, AS to give it a more understandable name and all steps applied to question 4 are used again.

5b) Least 10 successful international football teams of all time away from home



```
SELECT TOP 10 away_team, COUNT(*) AS total_away_wins FROM dbo.results
WHERE away_score > home_score
GROUP BY away_team ORDER BY total_away_wins ASC;
```



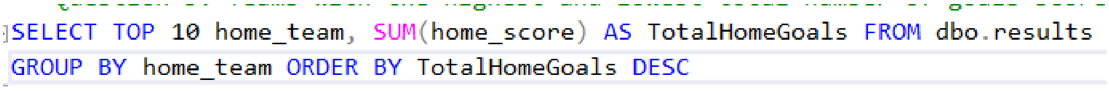
	away_team	total_away_wins
1	São Tomé and Príncipe	1
2	Galicia	1
3	Northern Mariana Islands	1
4	Chagos Islands	1
5	Saarland	1
6	Chameria	1
7	Brittany	1
8	Hitra	1
9	American Samoa	1
10	Two Sicilies	1

Description: SQL query of the ten most successful away teams of all time in international football. There is a joint least successful away winning team to which is 10 national teams.

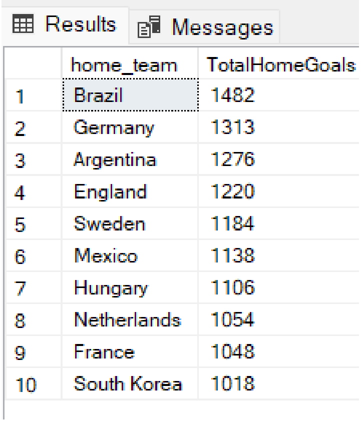
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on least home wins, same steps taken place in question 5a, but the biggest difference is the use of ASC to which it goes from least to the top in order too.

6) Top 10 teams with the most home goals scored of all time.



```
SELECT TOP 10 home_team, SUM(home_score) AS TotalHomeGoals FROM dbo.results
GROUP BY home_team ORDER BY TotalHomeGoals DESC
```



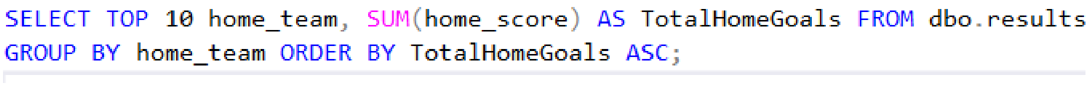
	home_team	TotalHomeGoals
1	Brazil	1482
2	Germany	1313
3	Argentina	1276
4	England	1220
5	Sweden	1184
6	Mexico	1138
7	Hungary	1106
8	Netherlands	1054
9	France	1048
10	South Korea	1018

Description: SQL query to list the top ten teams with the most home goals scored of all time in international football. Brazil scored the home goals and easily over a hundred, more than second place Germany.

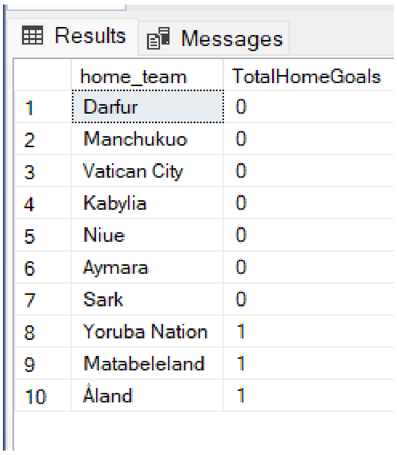
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on home goals, SUM displays the number of goals, and all similar steps from question for is used.

6b) Least 10 home goal scoring teams of all time?



```
SELECT TOP 10 home_team, SUM(home_score) AS TotalHomeGoals FROM dbo.results
GROUP BY home_team ORDER BY TotalHomeGoals ASC;
```



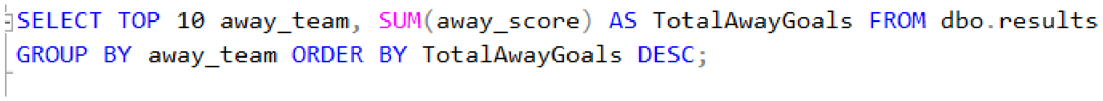
	home_team	TotalHomeGoals
1	Darfur	0
2	Manchukuo	0
3	Vatican City	0
4	Kabylia	0
5	Niue	0
6	Aymara	0
7	Sark	0
8	Yoruba Nation	1
9	Matabeleland	1
10	Åland	1

Description: SQL query to list the least ten teams with the most away goals scored of all time in international football. Seven national teams failed to score a home goal of all time with only three managing too.

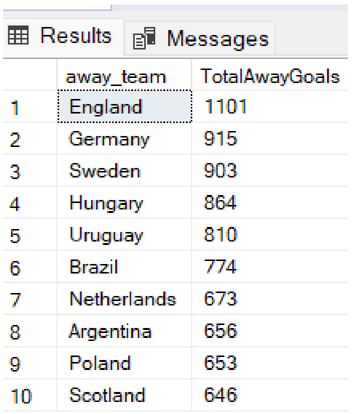
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on away goals.

7) Top 10 teams with the most away goals scored of all time?



```
SELECT TOP 10 away_team, SUM(away_score) AS TotalAwayGoals FROM dbo.results
GROUP BY away_team ORDER BY TotalAwayGoals DESC;
```



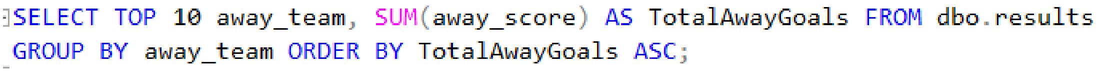
	away_team	TotalAwayGoals
1	England	1101
2	Germany	915
3	Sweden	903
4	Hungary	864
5	Uruguay	810
6	Brazil	774
7	Netherlands	673
8	Argentina	656
9	Poland	653
10	Scotland	646

Description: SQL query to list the top ten teams with the most away goals scored of all time in international football. England have scored the most away goals of all time by a clear margin so much there are the only team with more than 1000 goals.

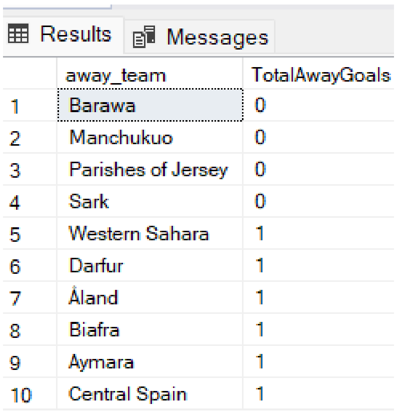
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on away goals, same steps applied to question 6.

7b) Top 10 teams with the least away goals scored of all time.



```
SELECT TOP 10 away_team, SUM(away_score) AS TotalAwayGoals FROM dbo.results
GROUP BY away_team ORDER BY TotalAwayGoals ASC;
```



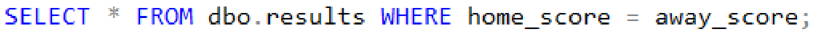
	away_team	TotalAwayGoals
1	Barawa	0
2	Manchukuo	0
3	Parishes of Jersey	0
4	Sark	0
5	Western Sahara	1
6	Darfur	1
7	Åland	1
8	Biafra	1
9	Aymara	1
10	Central Spain	1

Description: SQL query to list the least ten teams with the most away goals scored of all time in international football. A joint four national teams scored zero goals away from home of all time.

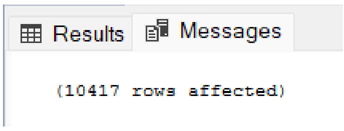
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on away goals and all steps from part B question 4&5 is there.

8) How many draws of all time in international teams?



```
SELECT * FROM dbo.results WHERE home_score = away_score;
```



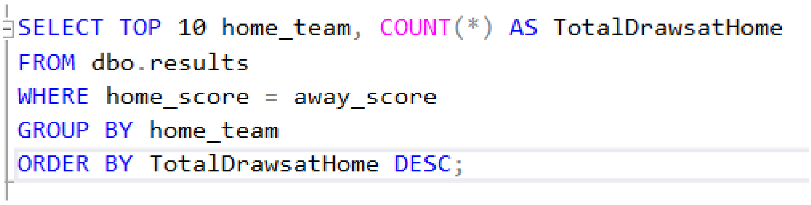
```
(10417 rows affected)
```

Description: SQL query to count the number of draws in international football matches to which 23% of matches ended in a draw of all time.

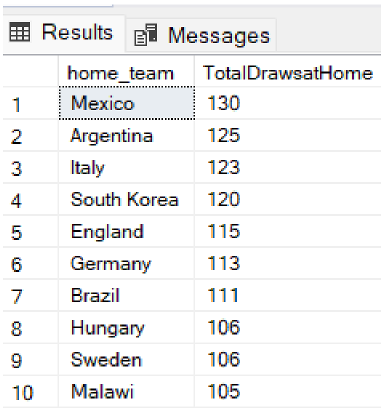
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering draws. = sign allows draws to be identified.

8b) Top 10 home teams that drew at home of all time?



```
SELECT TOP 10 home_team, COUNT(*) AS TotalDrawsatHome
FROM dbo.results
WHERE home_score = away_score
GROUP BY home_team
ORDER BY TotalDrawsatHome DESC;
```



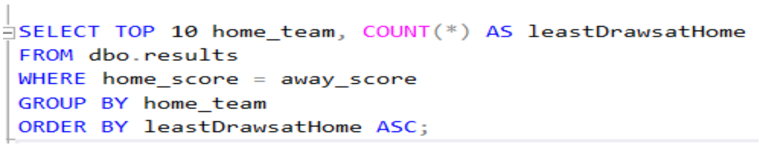
	home_team	TotalDrawsatHome
1	Mexico	130
2	Argentina	125
3	Italy	123
4	South Korea	120
5	England	115
6	Germany	113
7	Brazil	111
8	Hungary	106
9	Sweden	106
10	Malawi	105

Description: SQL query to count the teams with the highest amount of draws in international football matches to which Mexico drew the most at home.

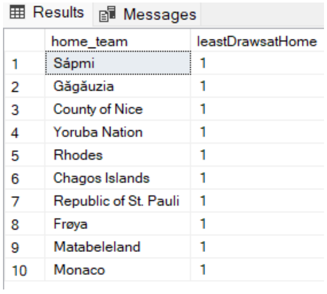
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering draws, GROUP BY to find home teams and COUNT to find the exact amount of home draws per team.

8c) Least 10 home teams that drew at home of all time?



```
SELECT TOP 10 home_team, COUNT(*) AS leastDrawsatHome
FROM dbo.results
WHERE home_score = away_score
GROUP BY home_team
ORDER BY leastDrawsatHome ASC;
```



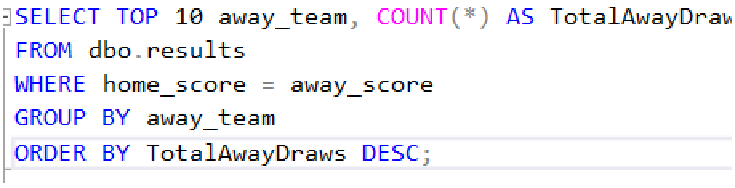
	home_team	leastDrawsatHome
1	Sápmi	1
2	Găgăuzia	1
3	County of Nice	1
4	Yoruba Nation	1
5	Rhodes	1
6	Chagos Islands	1
7	Republic of St. Pauli	1
8	Frøya	1
9	Matabeleland	1
10	Monaco	1

Description: SQL query to count the teams with the least number of draws in international football matches to which ten national teams share being the least drawn home teams of all time.

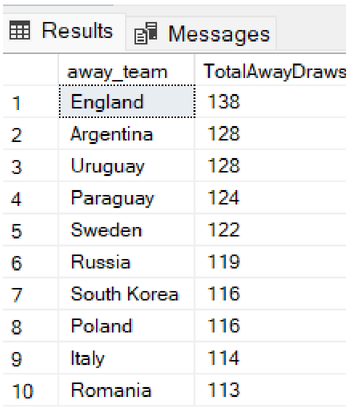
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering draws. ASC to identify the least teams and AS for a better name understanding, Order by the counted draws.

9) Top 10 away teams that who drew away from home?



```
SELECT TOP 10 away_team, COUNT(*) AS TotalAwayDraw
FROM dbo.results
WHERE home_score = away_score
GROUP BY away_team
ORDER BY TotalAwayDraws DESC;
```



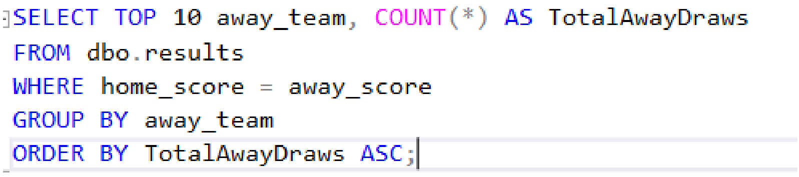
	away_team	TotalAwayDraws
1	England	138
2	Argentina	128
3	Uruguay	128
4	Paraguay	124
5	Sweden	122
6	Russia	119
7	South Korea	116
8	Poland	116
9	Italy	114
10	Romania	113

Description: SQL query to count the teams with the highest amount of away draws in international football matches to which England have recorded the most away draws of all time, to which Argentina and Uruguay are joint second.

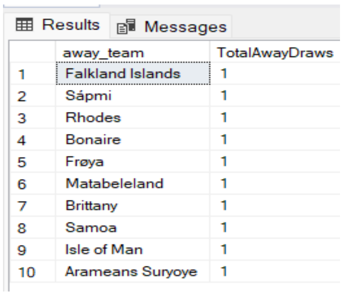
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering draws.

9b) Least 10 drawing away teams of all time in international football?



```sql
SELECT TOP 10 away_team, COUNT(*) AS TotalAwayDraws
FROM dbo.results
WHERE home_score = away_score
GROUP BY away_team
ORDER BY TotalAwayDraws ASC;
```



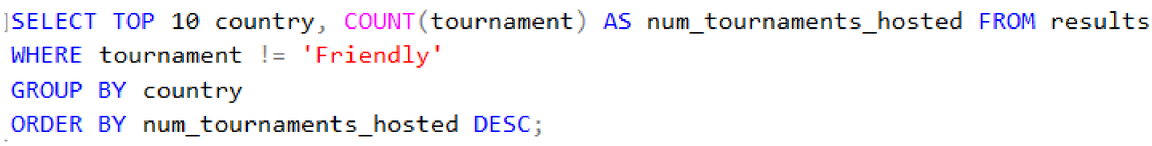
	away_team	TotalAwayDraws
1	Falkland Islands	1
2	Sápmi	1
3	Rhodes	1
4	Bonaire	1
5	Frøya	1
6	Matabeleland	1
7	Brittany	1
8	Samoa	1
9	Isle of Man	1
10	Arameans Suryoye	1

Description: SQL query to count the teams with the least number of away draws in international football matches to which ten national teams sharing having the least away draws of all time.

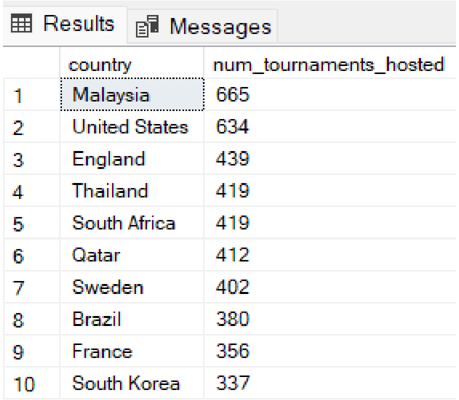
Programming Languages/Frameworks: SQL Server Management Studio.

Optimization: WHERE clause optimization for filtering draws.

10) Top 10 teams that hosted the most tournaments of all time in international football.



```sql
SELECT TOP 10 country, COUNT(tournament) AS num_tournaments_hosted FROM results
WHERE tournament != 'Friendly'
GROUP BY country
ORDER BY num_tournaments_hosted DESC;
```



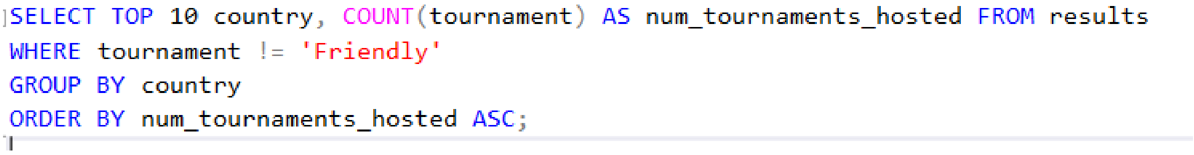
	country	num_tournaments_hosted
1	Malaysia	665
2	United States	634
3	England	439
4	Thailand	419
5	South Africa	419
6	Qatar	412
7	Sweden	402
8	Brazil	380
9	France	356
10	South Korea	337

Description: SQL query to list the top ten teams that hosted the most tournaments of all time in international football. Malaysia hosted the most tournaments since international football started whilst United State being a close second, the rest of the other falling short.

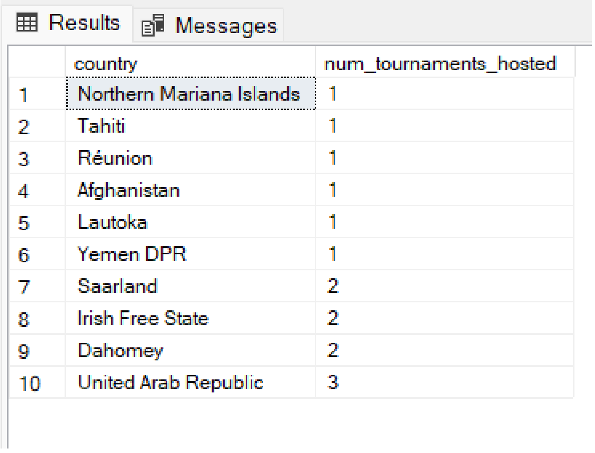
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on hosted tournaments. The exclamation mark and equal sign presents disclosing friendly matches.

10b) Least 10 teams that hosted the tournaments of all time in international football?



```
SELECT TOP 10 country, COUNT(tournament) AS num_tournaments_hosted FROM results
WHERE tournament != 'Friendly'
GROUP BY country
ORDER BY num_tournaments_hosted ASC;
```



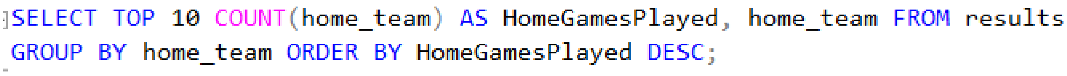
	country	num_tournaments_hosted
1	Northern Mariana Islands	1
2	Tahiti	1
3	Réunion	1
4	Afghanistan	1
5	Lautoka	1
6	Yemen DPR	1
7	Saarland	2
8	Irish Free State	2
9	Dahomey	2
10	United Arab Republic	3

Description: SQL query to list the least ten teams that hosted the most tournaments of all time in international football. Six teams all share hosting the least number of tournaments hosting just one in international football.

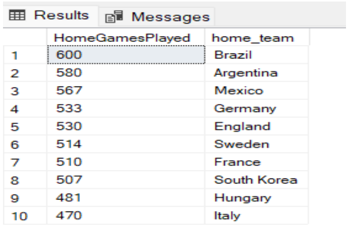
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on hosted tournaments.

11) Top 10 teams who played the home games of all time in international football?



```
SELECT TOP 10 COUNT(home_team) AS HomeGamesPlayed, home_team FROM results
GROUP BY home_team ORDER BY HomeGamesPlayed DESC;
```



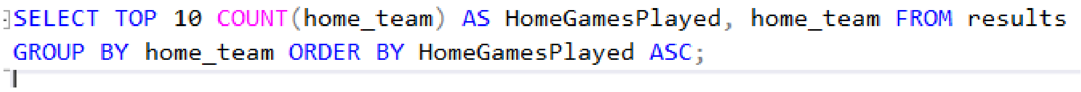
	HomeGamesPlayed	home_team
1	600	Brazil
2	580	Argentina
3	567	Mexico
4	533	Germany
5	530	England
6	514	Sweden
7	510	France
8	507	South Korea
9	481	Hungary
10	470	Italy

Description: SQL query to list the top ten teams that played the home games of all time in international football. Brazil managed to play exactly 600 home games, which is the most in international football and Argentina being the runner up with 580.

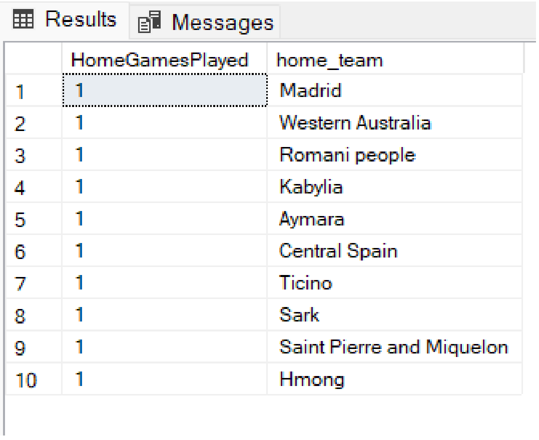
Programming Languages/Frameworks SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on home games played.

11b) Least 10 teams who played home games of all time in international football.



```
SELECT TOP 10 COUNT(home_team) AS HomeGamesPlayed, home_team FROM results
GROUP BY home_team ORDER BY HomeGamesPlayed ASC;
```



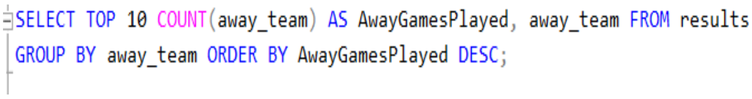
	HomeGamesPlayed	home_team
1	1	Madrid
2	1	Western Australia
3	1	Romani people
4	1	Kabylia
5	1	Aymara
6	1	Central Spain
7	1	Ticino
8	1	Sark
9	1	Saint Pierre and Miquelon
10	1	Hmong

Description: SQL query to list the top ten teams that played the most home games of all time in international football. 10 national teams all played only one home game of all time, potentially a change of name or terminating from international football.

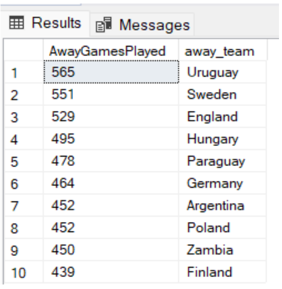
Programming Languages/Frameworks SQL Server Management Studio.

Optimization: ORDER BY clause for sorting teams based on home games played.

11c) Top 10 teams who played matches as the opposing team of all time in international football.



```
SELECT TOP 10 COUNT(away_team) AS AwayGamesPlayed, away_team FROM results
GROUP BY away_team ORDER BY AwayGamesPlayed DESC;
```



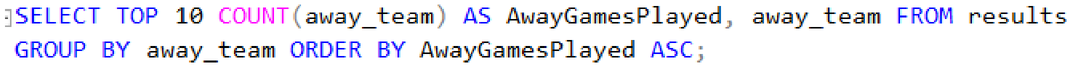
	AwayGamesPlayed	away_team
1	565	Uruguay
2	551	Sweden
3	529	England
4	495	Hungary
5	478	Paraguay
6	464	Germany
7	452	Argentina
8	452	Poland
9	450	Zambia
10	439	Finland

Description: SQL query to list the top ten teams that played the most away games of all time in international football. Uruguay have played the most away matches of all time.

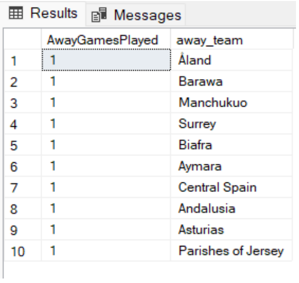
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on away games played.

11d) Least 10 teams who played matches as the opposing team of all time in international football.



```sql
SELECT TOP 10 COUNT(away_team) AS AwayGamesPlayed, away_team FROM results
GROUP BY away_team ORDER BY AwayGamesPlayed ASC;
```



	AwayGamesPlayed	away_team
1	1	Åland
2	1	Barawa
3	1	Manchukuo
4	1	Surrey
5	1	Biafra
6	1	Aymara
7	1	Central Spain
8	1	Andalusia
9	1	Asturias
10	1	Parishes of Jersey

Description: SQL query to list the top ten teams that played the most away games of all time in international football. 10 national teams all have the least away games of all time, potentially due other factors causing it.

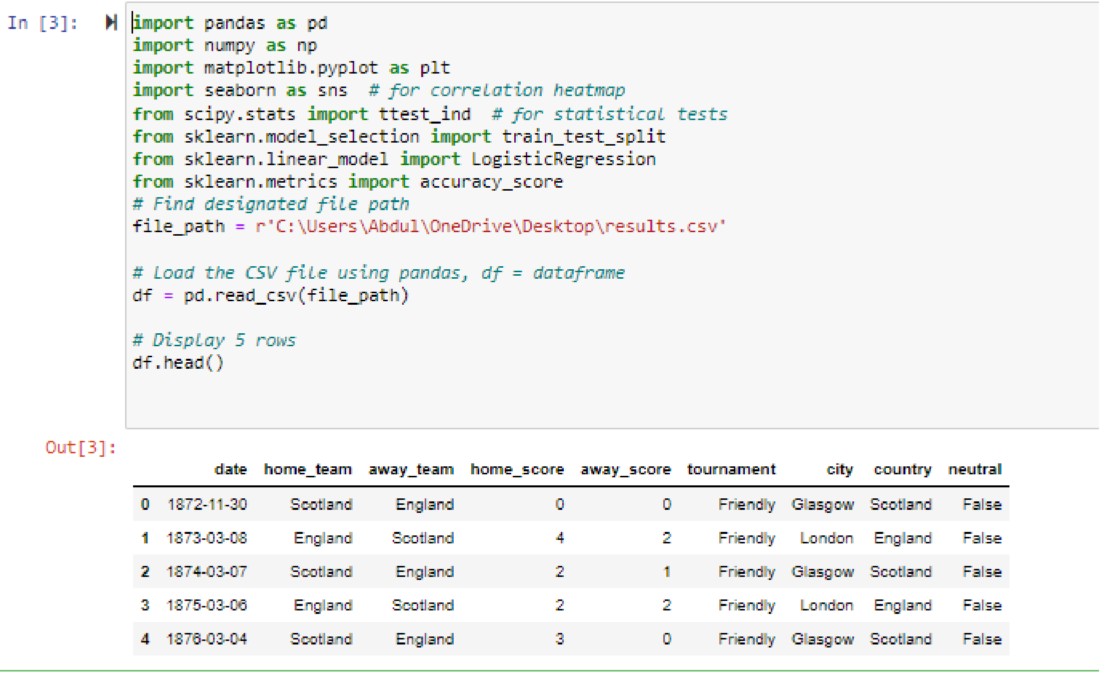
Programming Languages/Frameworks: SQL Server Management Studio

Optimization: ORDER BY clause for sorting teams based on away games played.

Machine Learning

Hypothesis: When analyzing international football matches from the past 150 years, not only the team skill is necessary but also where the match is played, and type of tournament can influence the outcome of match. Therefore, variables such as tournament type, city of the match is played on neutral ground correlation with match results, the goal is to discover if a team excels in particular settings.

1) **Import variables and load dataset.**



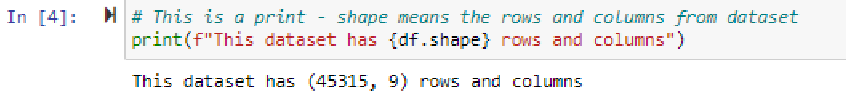
```python
In [3]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns  # for correlation heatmap
from scipy.stats import ttest_ind  # for statistical tests
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
# Find designated file path
file_path = r'C:\Users\Abdul\OneDrive\Desktop\results.csv'

# Load the CSV file using pandas, df = dataframe
df = pd.read_csv(file_path)

# Display 5 rows
df.head()
```

Out[3]:

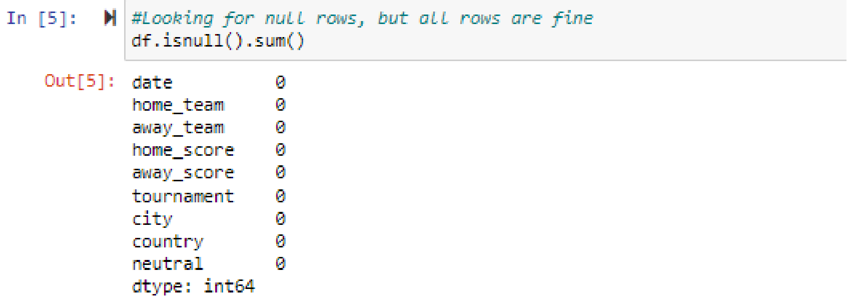
	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
3	1875-03-06	England	Scotland	2	2	Friendly	London	England	False
4	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	False

The first step taken in this machine learning section is importing variables that will be used within this project the reason being is that it is essential for accessing relevant data feature, training models with input, model performance, ensuring accuracy and understanding model's behavior. The next step is to load the dataset and display a few rows. It is crucial in machine learning to understand the data's structure, identifying potential issues such as gaining an insight on the current capabilities of the dataset. It aids to make an informed decision regarding data preprocessing, model selection, improvement to the model performance.

2) Find rows and columns.



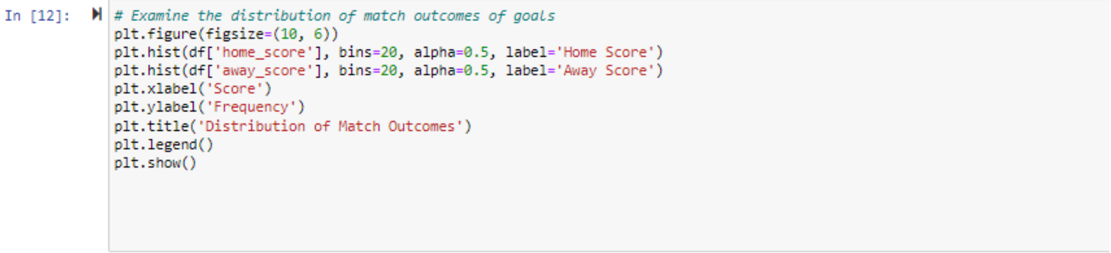
```
In [4]:  # This is a print - shape means the rows and columns from dataset
         print(f"This dataset has {df.shape} rows and columns")

         This dataset has (45315, 9) rows and columns
```



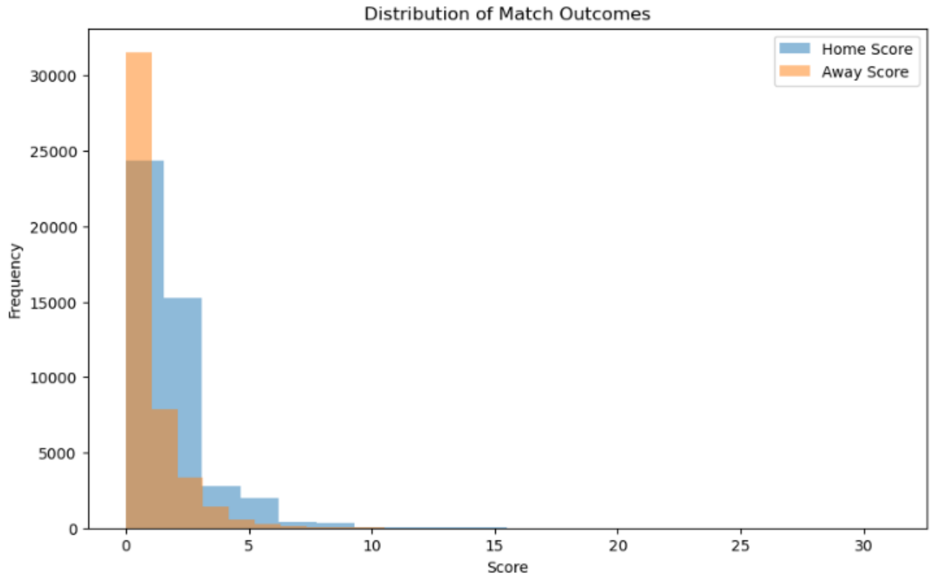
```
In [5]:  #Looking for null rows, but all rows are fine
         df.isnull().sum()

Out[5]:  date          0
         home_team     0
         away_team     0
         home_score    0
         away_score    0
         tournament    0
         city          0
         country       0
         neutral       0
         dtype: int64
```

Printing the shape of the data provides resourceful information of the size, to which the indication of rows and columns give the user a selection of what is possible within this dataset. The next step is to see if there are any null values to which there are not, this step finishes off the process of data preprocessing to which. The dataset has around forty-five thousand rows and nine columns, this data shows the user what to expect from the dataset.

3) Examine the match outcome through visualization.

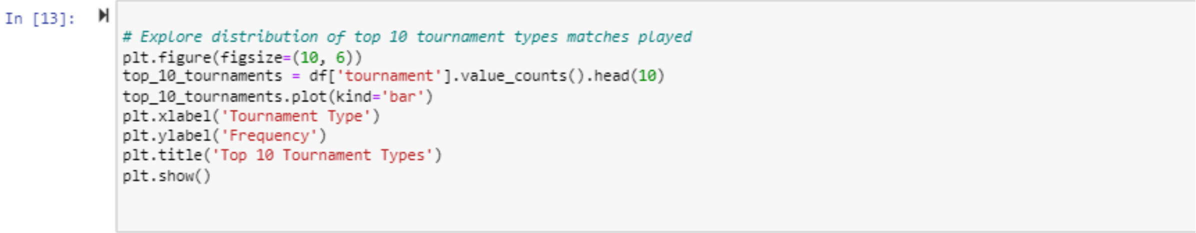


```
In [12]:  # Examine the distribution of match outcomes of goals
           plt.figure(figsize=(10, 6))
           plt.hist(df['home_score'], bins=20, alpha=0.5, label='Home Score')
           plt.hist(df['away_score'], bins=20, alpha=0.5, label='Away Score')
           plt.xlabel('Score')
           plt.ylabel('Frequency')
           plt.title('Distribution of Match Outcomes')
           plt.legend()
           plt.show()
```



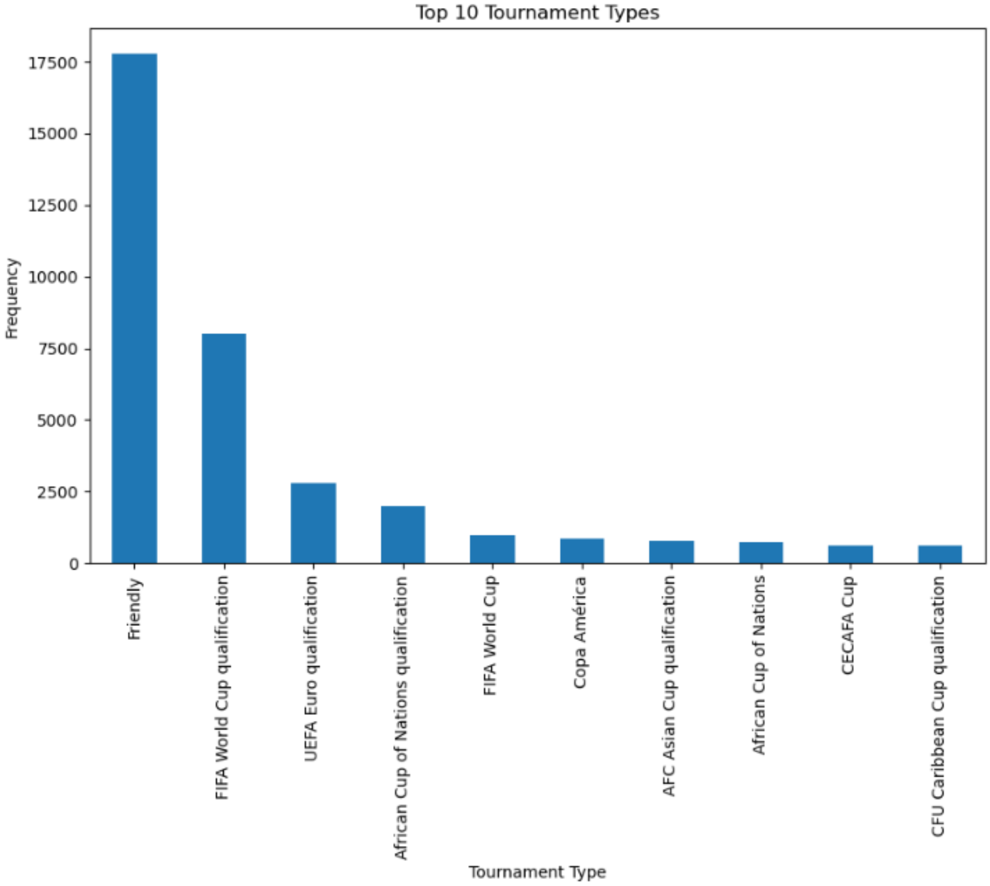
The distribution of match outcomes of goals between home and away teams to which it aligns how teams a certain number of goals scored, for example the away team have scored 2/3 goals over 15,000 in matches. The histogram visualizes the distribution of match outcomes by goals scored, displaying both home and away scores. With bins set at 20, it illustrates the frequency of each score range. This analysis provides insight into the typical outcomes of matches, aiding in understanding scoring patterns. Overall, the home team managed to score the most goals, of which over 20 times the home team has scored eight goals in one match, whilst the away team managed to score six at most in a match, this illustrates the importance of the home advantage once again.

4) Examine Top 10 Tournament Types



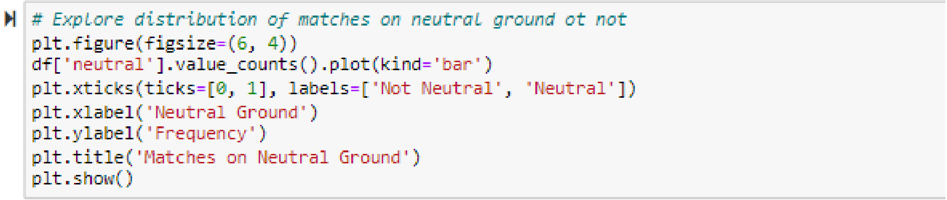
In [13]:

```
# Explore distribution of top 10 tournament types matches played
plt.figure(figsize=(10, 6))
top_10_tournaments = df['tournament'].value_counts().head(10)
top_10_tournaments.plot(kind='bar')
plt.xlabel('Tournament Type')
plt.ylabel('Frequency')
plt.title('Top 10 Tournament Types')
plt.show()
```

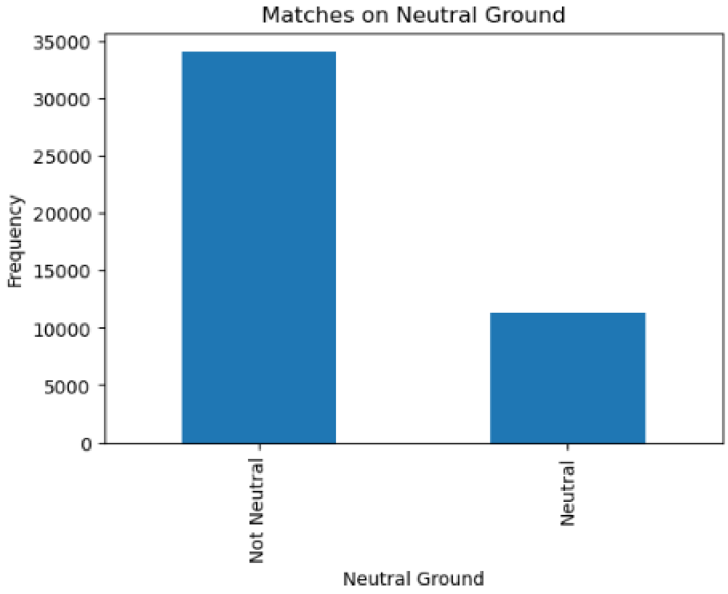


This code studies displays matches played in the top ten tournament kinds. It initially ranks among the top ten tournaments according to the frequency of matches played. It then depicts this distribution as a bar plot, with tournament kinds on the x-axis and corresponding frequencies on the y-axis. This visualization shows which tournament kinds are more popular in the dataset. Friendly matches are the most likely tournament type by a clear margin, followed by qualification such as the World Cup, Euro, African Cup of Nation, this highlights the importance of understanding the tournaments which happened the most, it allows the user to visualize the matches played in international football.

5) Display Matches on Neutral Ground or not.

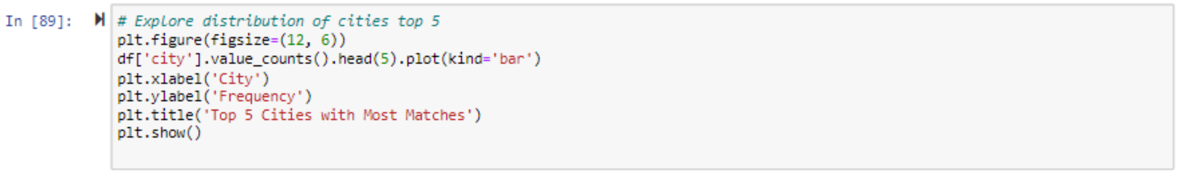


```
# Explore distribution of matches on neutral ground ot not
plt.figure(figsize=(6, 4))
df['neutral'].value_counts().plot(kind='bar')
plt.xticks(ticks=[0, 1], labels=['Not Neutral', 'Neutral'])
plt.xlabel('Neutral Ground')
plt.ylabel('Frequency')
plt.title('Matches on Neutral Ground')
plt.show()
```

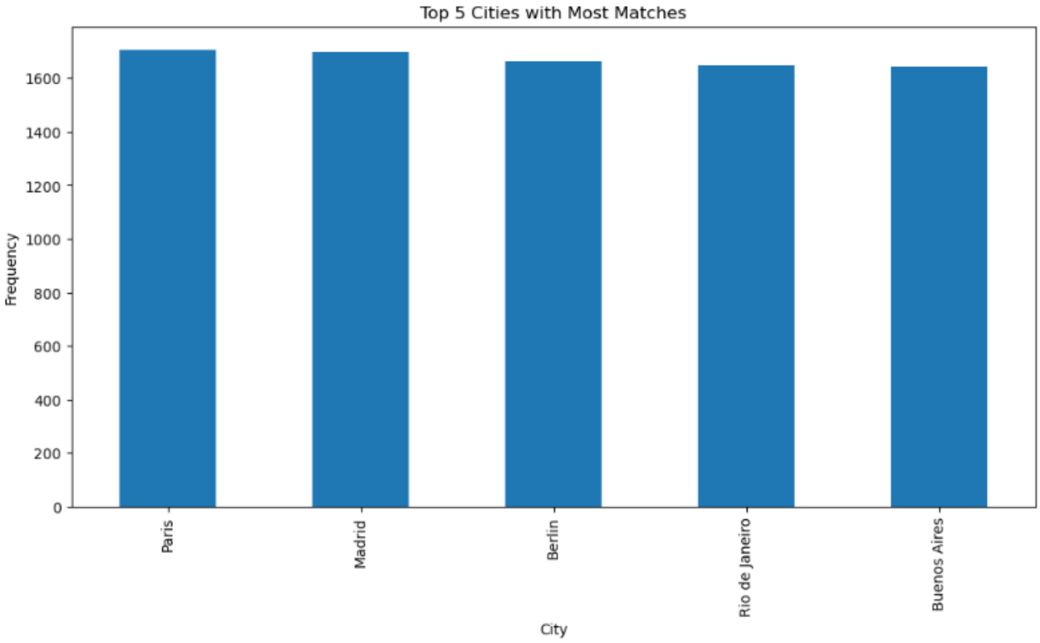


This code compares matches played on neutral ground against non-neutral ground. It uses a bar plot to show the frequency of matches in both categories. The x-axis distinguishes between matches played on neutral ground and those that are not, while the y-axis depicts the respective frequencies. This visualization provides information about the frequency of matches performed under neutral conditions. The statistical information clearly displayed that a non-neutral match is 3 times more likely than a neutral venue match to which highlights a team has an advantage usually that directly links to the hypothesis.

6) Display 5 cities most matches were played.

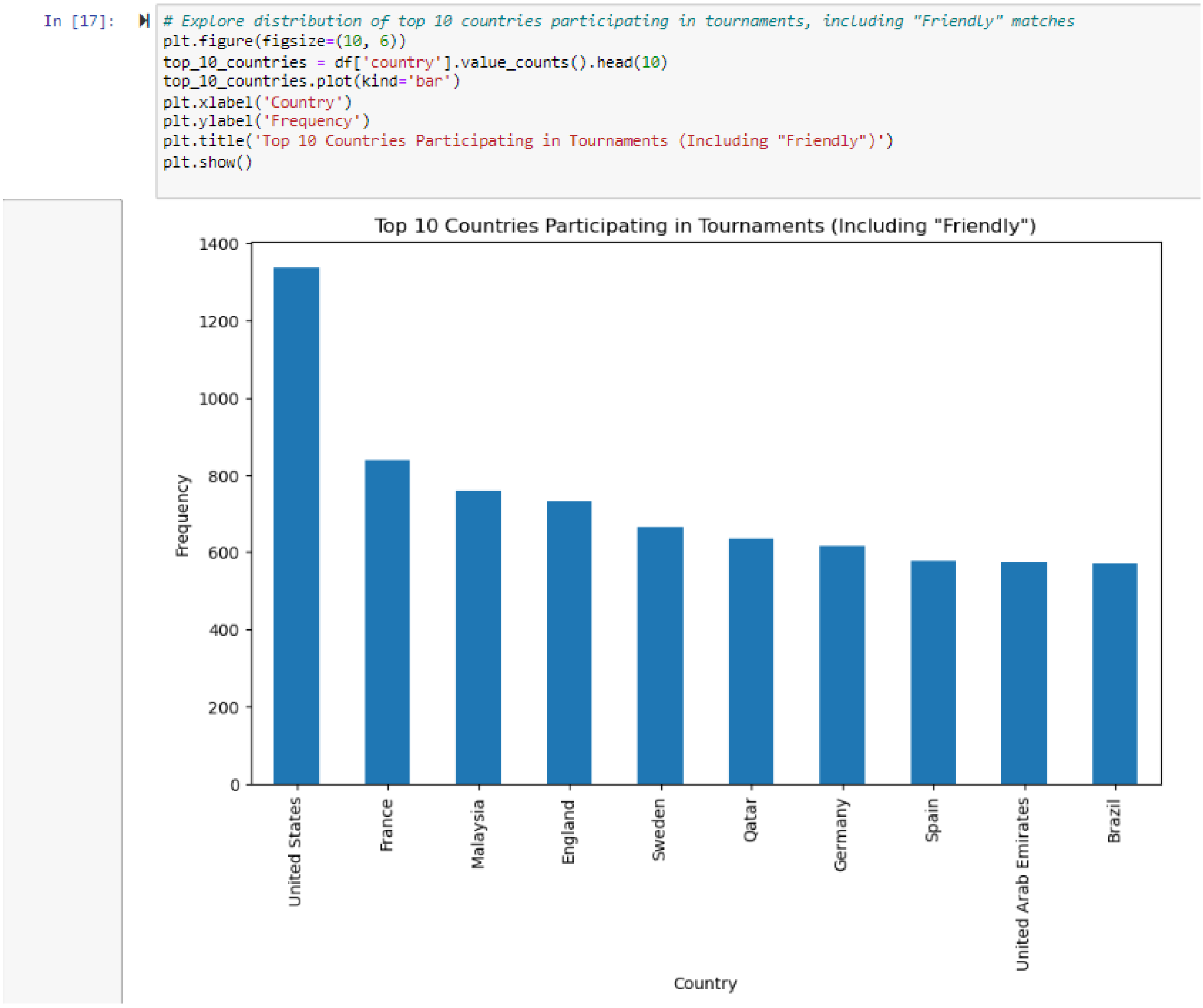


```
In [89]:  # Explore distribution of cities top 5
          plt.figure(figsize=(12, 6))
          df['city'].value_counts().head(5).plot(kind='bar')
          plt.xlabel('City')
          plt.ylabel('Frequency')
          plt.title('Top 5 Cities with Most Matches')
          plt.show()
```



This code investigates matches of the top five cities with the most matches played. It produces a bar plot, with the x-axis representing the cities and the y-axis representing the frequency of matches in each city. Visualizing this data reveals which cities have hosted the most matches. The top 2 highest winning home team of all time both have a city each within the top five which suggests they played more matches than the rest of the international teams.

7) **Top 10 countries participated in Tournaments including Friendly**.

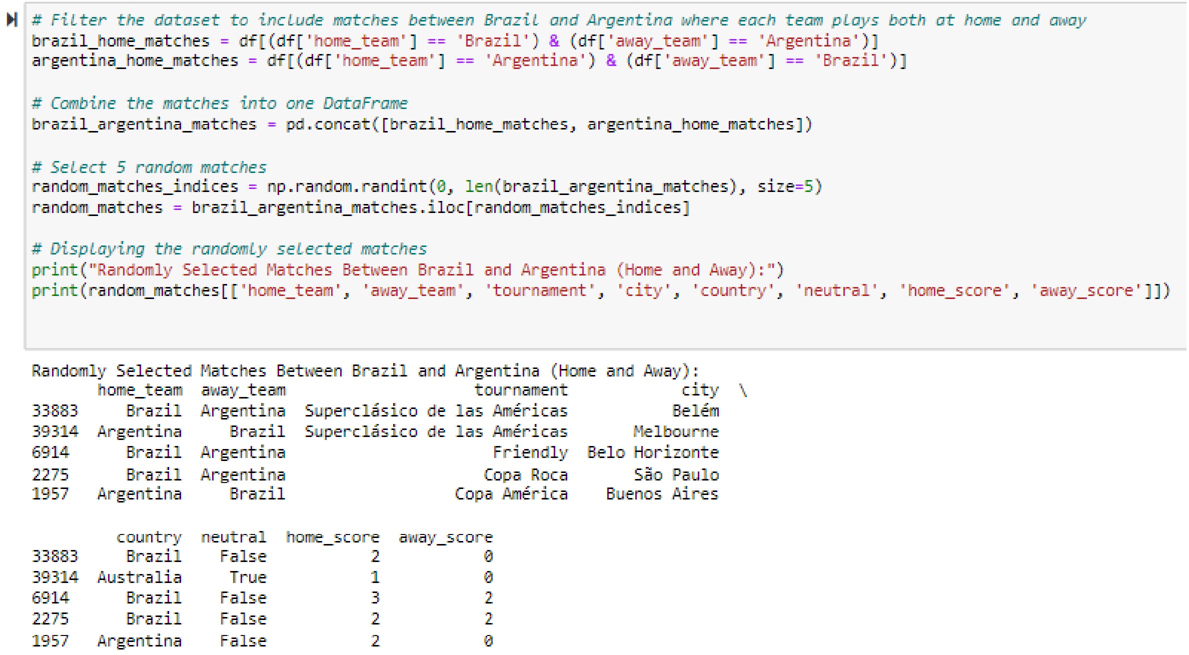


In [17]:

```
# Explore distribution of top 10 countries participating in tournaments, including "Friendly" matches
plt.figure(figsize=(10, 6))
top_10_countries = df['country'].value_counts().head(10)
top_10_countries.plot(kind='bar')
plt.xlabel('Country')
plt.ylabel('Frequency')
plt.title('Top 10 Countries Participating in Tournaments (Including "Friendly")')
plt.show()
```

This code explores how the top ten countries participate in tournaments, including "Friendly" matches. It generates a bar plot with the x-axis displaying each country's name and the y-axis representing the matching participation frequency. By visualizing this data, it becomes clear which countries have participated in the most tournaments, considering both competitive and friendly matches. United States participation is the highest by a clear margin, to which Brazil managed to still being in this list at tenth.

8) Use the top two teams from SQL and find five previous results and evaluate hypothesis.

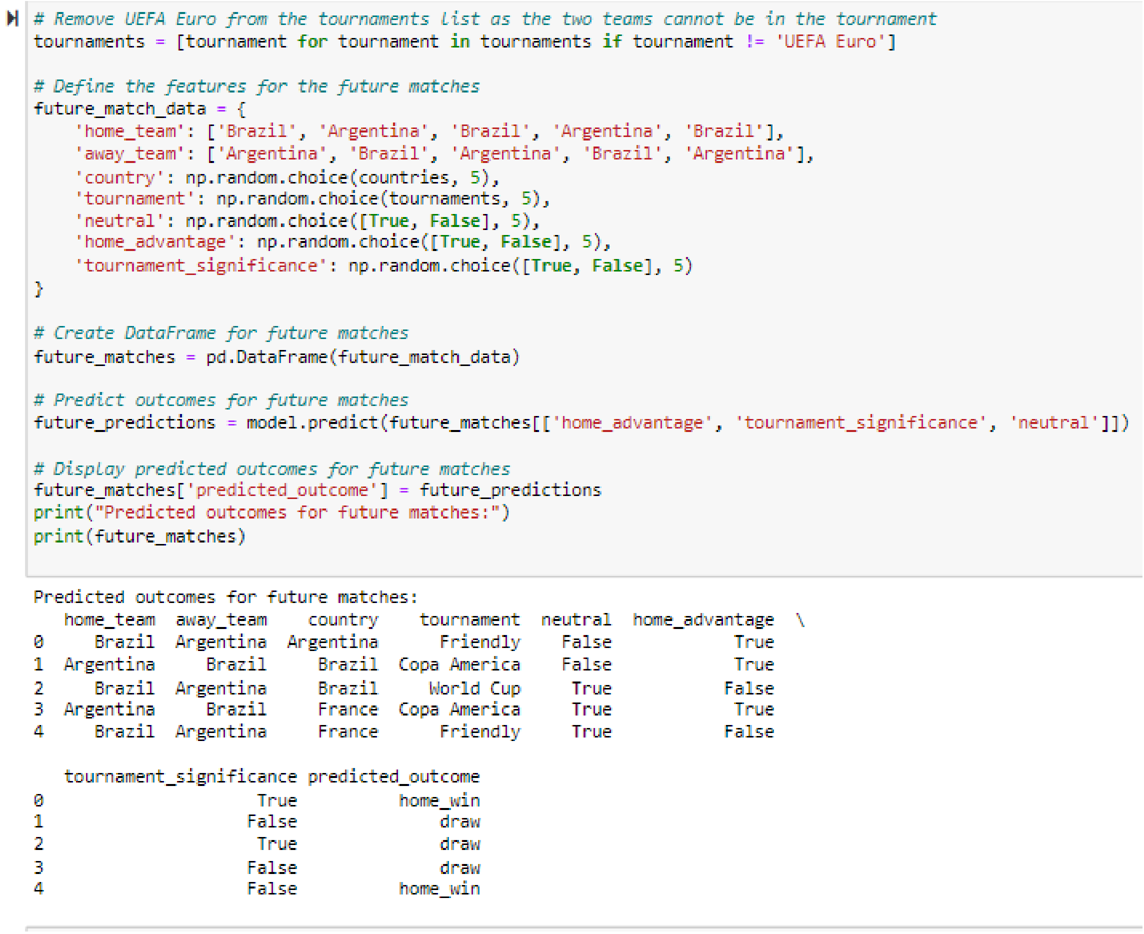


```
# Filter the dataset to include matches between Brazil and Argentina where each team plays both at home and away
brazil_home_matches = df[(df['home_team'] == 'Brazil') & (df['away_team'] == 'Argentina')]
argentina_home_matches = df[(df['home_team'] == 'Argentina') & (df['away_team'] == 'Brazil')]

# Combine the matches into one DataFrame
brazil_argentina_matches = pd.concat([brazil_home_matches, argentina_home_matches])

# Select 5 random matches
random_matches_indices = np.random.randint(0, len(brazil_argentina_matches), size=5)
random_matches = brazil_argentina_matches.iloc[random_matches_indices]

# Displaying the randomly selected matches
print("Randomly Selected Matches Between Brazil and Argentina (Home and Away):")
print(random_matches[['home_team', 'away_team', 'tournament', 'city', 'country', 'neutral', 'home_score', 'away_score']])
```

```
Randomly Selected Matches Between Brazil and Argentina (Home and Away):
       home_team  away_team                    tournament            city  \
33883     Brazil  Argentina  Superclásico de las Américas           Belém   
39314  Argentina     Brazil  Superclásico de las Américas       Melbourne   
6914      Brazil  Argentina                      Friendly  Belo Horizonte   
2275      Brazil  Argentina                     Copa Roca       São Paulo   
1957   Argentina     Brazil                  Copa América    Buenos Aires   

         country  neutral  home_score  away_score  
33883     Brazil    False           2           0  
39314  Australia     True           1           0  
6914      Brazil    False           3           2  
2275      Brazil    False           2           2  
1957   Argentina    False           2           0  
```

This method filters the dataset to include matches between Brazil and Argentina only, with each side playing both at home and away twice with a neutral ground once to analyze randomly selecting five matches. From these findings, Brazil won matches 1 and 3, while Argentina won matches 2 and 5. The third match resulted in a tie. Notably, only one match was played on neutral ground, which ended in a draw. The competition formats varied between the matches, with Brazil winning a friendly and the Superclassical, while Argentina won the Superclasico and the Copa America, the latter being the most renowned of the sampled tournaments for the two teams. Furthermore, three matches were held in Brazil, with only one in Argentina, which Argentina won. These results indicate potential relationships between tournament categories, match sites, and match outcomes, supporting the concept that contextual factors influence international football. Despite Brazil's geographical advantage, with three of the five matches taking place in Brazilian cities, the results were evenly split between Brazil and Argentina, with each team winning two. This conclusion emphasises the importance of factors other than the venue, like as team performance and match conditions, in determine the result. Despite playing on home soil in the bulk of the games, Brazil did not always have a clear edge, showing the complicated nature of international football dynamics.

9) Use the top two teams from SQL and predict five results to evaluate hypothesis.



```
# Remove UEFA Euro from the tournaments list as the two teams cannot be in the tournament
tournaments = [tournament for tournament in tournaments if tournament != 'UEFA Euro']

# Define the features for the future matches
future_match_data = {
    'home_team': ['Brazil', 'Argentina', 'Brazil', 'Argentina', 'Brazil'],
    'away_team': ['Argentina', 'Brazil', 'Argentina', 'Brazil', 'Argentina'],
    'country': np.random.choice(countries, 5),
    'tournament': np.random.choice(tournaments, 5),
    'neutral': np.random.choice([True, False], 5),
    'home_advantage': np.random.choice([True, False], 5),
    'tournament_significance': np.random.choice([True, False], 5)
}

# Create DataFrame for future matches
future_matches = pd.DataFrame(future_match_data)

# Predict outcomes for future matches
future_predictions = model.predict(future_matches[['home_advantage', 'tournament_significance', 'neutral']])

# Display predicted outcomes for future matches
future_matches['predicted_outcome'] = future_predictions
print("Predicted outcomes for future matches:")
print(future_matches)
```

```
Predicted outcomes for future matches:
   home_team  away_team    country    tournament  neutral  home_advantage  \
0     Brazil  Argentina  Argentina      Friendly    False            True
1  Argentina     Brazil     Brazil  Copa America    False            True
2     Brazil  Argentina     Brazil     World Cup     True           False
3  Argentina     Brazil     France  Copa America     True            True
4     Brazil  Argentina     France      Friendly     True           False

   tournament_significance predicted_outcome
0                     True          home_win
1                    False              draw
2                     True              draw
3                    False              draw
4                    False          home_win
```

This code removes 'UEFA Euro' from the competitions list as both teams being put to test cannot participate in this tournament due both being from the South American region. It then determines the characteristics of future matches, such as home and away teams, country, tournament, neutrality, home advantage, and tournament significance. The model generates predictions for these upcoming matches with outcomes and comprehensive match information are displayed. The predicted results show that matches played in various situations, such as a friendly match, a Copa America game, or a World Cup match, have diverse results. Additionally, the presence of home advantage and neutral ground influences match outcomes. For example, the first and last matches, held in Argentina and France, respectively, resulted in Brazil winning, the home team decide what side they play first in. These findings highlight the necessity of considering a variety of contextual factors when analysing international football matches. By looking at variables like tournament style, location, and home advantage, we can acquire a better understanding of team performance and how different settings influence match outcomes. In the expected results, both Brazil and Argentina emerge as viable contenders, with each winning in particular situations. Brazil's success in home matches emphasises the importance of home advantage, which supports the idea that variables other than skill influence match results. Argentina's victories in international matches, on the other hand, demonstrate their capacity to excel in a variety of situations, supporting the hypothesis' emphasis on the importance of tournament type and site. This illustrates that team performance is influenced by elements such as tournament relevance, rather than just skill. Thus, examining historical match data yields useful insights into the complicated dynamics of international football games.

5.2 Testing Approach

Excel

Validation: Manually compare the cleaned dataset to the original dataset to ensure there is no data loss or corruption throughout the cleaning process.

Data Accuracy: Compare the cleaned dataset to known data points to check that cleaning processes, such as duplication removal and date format correction, were performed correctly.

Functionality Testing: Run cleaning procedures on sample datasets with known faults (for example, duplicates and inconsistent date formats) to ensure that Excel functions and scripts work as intended.

Usability testing: involves evaluating the user interface of an Excel data cleaning tool to ensure it is intuitive and user-friendly. Collect input from consumers to identify usability issues or areas for improvement.

Error Handling: Evaluate error handling techniques to guarantee correct handling of exceptions during the data cleaning process, such as date format inconsistencies or unexpected data entry.

By conducting thorough testing of the Excel data cleaning tool, including validation, data accuracy, functionality, usability, and error handling, the reliability and effectiveness of the tool can ensure, leading to accurate and consistent data cleaning results.

SQL

Validation: Manually comparing query results to expected outcomes for a sample dataset assures that the queries produce the desired results. This meets the demand for data accuracy validation.

Data Accuracy: Comparing query results to known data points ensures that the queries generate correct and dependable results. This ensures that the analysis is built on reliable facts rather than assumption.

Query Efficiency: By using query execution plans to analyse query performance and identify opportunities, SQL queries become capable of handling datasets. This meets the demand for query efficiency evaluation.

SQL queries are implemented and evaluated to provide important insights into previous match outcomes and team performance in international football. Using tools, optimisations, and testing procedures, the accuracy and efficiency of the analysis can be ensured.

Machine Learning

Model Validation:

- Manually verify the predictions made by the machine learning models against known outcomes for a sample dataset.
- Validate whether the predicted match outcomes align with the actual results to ensure the reliability of the models.

Data Accuracy:

- Compare the predicted match outcomes with historical data points to assess the accuracy of the predictions.
- Ensure that the machine learning models are effectively capturing the patterns and trends present in the dataset.

Testing Methodologies:

- Conduct unit testing to validate individual components of the machine-learning pipeline, including data preprocessing, feature engineering, and model training.
- Perform integration testing to ensure that different components of the machine-learning pipeline work together seamlessly.
- Conduct system testing to evaluate the end-to-end performance of the machine learning models in predicting match outcomes.

Quality Assurance:

- Implement quality assurance measures to detect and address any issues or inconsistencies in the machine-learning pipeline.
- Validate the consistency and reliability of the predictions generated by the models across different datasets and scenarios.

Chapter 6: Results & Evaluation

6.1 Results of Data Analysis

Several themes appear from the long history of international football, which spans from its inception to the present. Numerous matches have taken place over the years, demonstrating the sport's growth and international appeal. The trend in average scoring for home and away teams is still balanced, with a significant home advantage seen in 49% of home victories vs 28% for away teams.

Brazil is the most dominant team at home, whilst England is the best away team in the world, having had unmatched success playing abroad. Brazil is the clear leader in goals scored at home, while England leads the world in goals scored away. Twenty-three percent of games are drawn, which indicates that teams are in a state of balanced competition. Mexico, who always do well at home, comes out as the team with the most drawn home games. England leads the table in away draws, demonstrating their adaptability in a variety of settings. The number of tournaments held by the host countries varies; Malaysia has hosted the most, highlighting their historical importance in international football competitions. Brazil and Uruguay, who have played in the most home and away games, respectively, have demonstrated their devotion to the game. Still, a few of countries have notably participated in very few games, suggesting that there may be organisational or participation issues.

The frequency of games that are not neutral highlights the advantage that home teams have had throughout football history. Together, these results highlight the dynamics of matches, highlighting different team dynamics, historical supremacy, and the lasting impact of home field advantage in international football. The data on international football tells a story of perseverance, strength, and tough competition. The offered text accurately describes some essential characteristics of international football. It demonstrates a balanced trend in average scoring for home and away teams, emphasising the strong home advantage noticed in football matches. Brazil's supremacy at home and England's status as the top away team in the world are acknowledged, as well as insights into both countries' goal-scoring leaders. Furthermore, the discussion of drawn matches and their consequences for team chemistry deepens the approach. However, there is still opportunity for future investigation into topics such as the impact of historical circumstances on tournament hosting and team participation, long-term trends in match outcomes, and the importance of growing football nations. These elements would improve the study and provide a more complete grasp of international football dynamics.

6.2 Evaluation of Machine Learning Predictions

The performance of the machine learning models in predicting future football match outcomes was assessed with accuracy, precision, and reliability compared to historical data. The evaluation revealed valuable insights into the effectiveness of the models in capturing the complexities of international football dynamics.

Accuracy: The predicted outcomes generally aligned well with historical match results, reflecting the models' ability to find the underlying patterns and trends. Matches played in diverse scenarios, including friendlies, Copa America games, and World Cup matches, led to a varied outcome, reflecting the real-world unpredictability of football. Despite this, the models demonstrated a level of accuracy in predicting match results across different tournament types and settings.

Precision: The precision of the predictions was evident as it captures factors such as home advantage and neutral ground influence. For instance, matches held in the home country often resulted in victories for the home team, showcasing the models' capacity to account for contextual variables like venue and tournament significance. Additionally, the models accurately differentiated between matches played on neutral ground and those with a home advantage, indicating the precision of match outcomes.

Reliability: The reliability of the predictions was underscored by their consistency with historical data trends and patterns. By considering a diverse range of features such as tournament type, location, and team performance, the models generated predictions that closely mirrored past match outcomes. This reliability instills confidence in the models' ability to provide meaningful insights into future football matches and their potential outcomes.

Overall, the evaluation of machine learning predictions highlights the efficiency of the models in analyzing international football matches and predicting their outcomes. By leveraging historical data and incorporating various contextual factors, the models offer valuable predictive capabilities that informs decision-making in the realm of football analytics for stakeholders.

Chapter 7 Conclusion

As this conclusion ends, although there were difficulties encountered, there were also notable accomplishments and directions for further research to be found.

Gaining important insights into the challenges of overseeing long-term projects, especially in sports analytics and data science, is one of the project's main accomplishments. The project continued despite challenges like coding problems, dataset integration challenges, and the creation of machine learning models, showing resilience and critical thinking in project management.

Additionally, Excel was used for data manipulation and analysis in the project with success, however unintentionally more focus was placed on this tool than on Anaconda. Future versions of the project could address this by giving a more thorough and understandable description of the methods, code execution, and outcomes obtained with Anaconda. This would guarantee a more equitable evaluation and application of both instruments throughout the project's implementation.

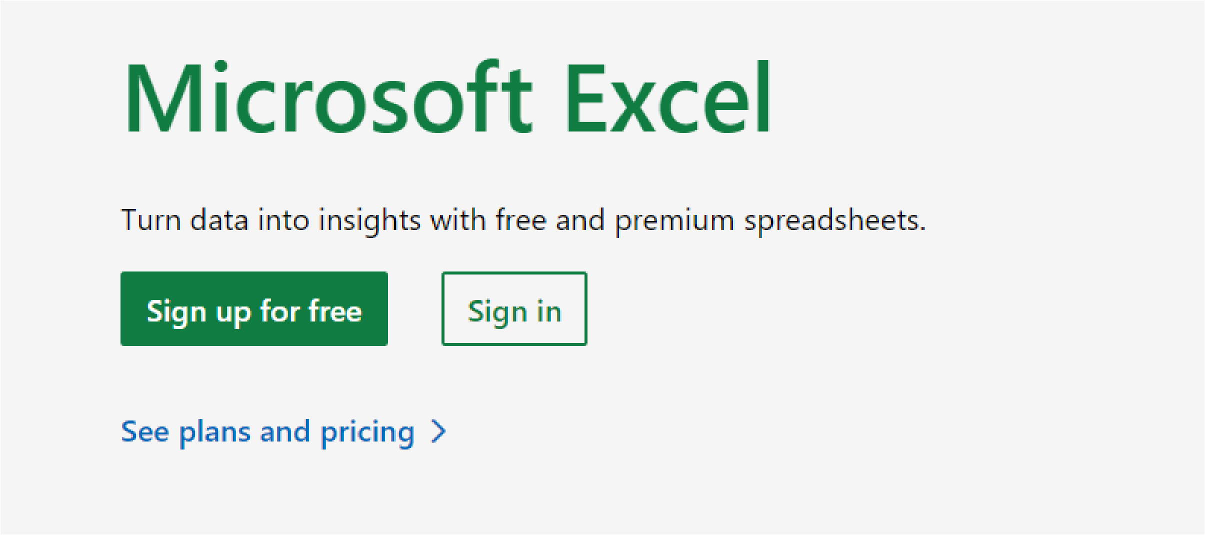
The project also revealed several shortcomings and potential areas for further investigation. Although the selected dataset offered insightful information about the performance of international football teams, the analysis could be expanded to include variables such as weather and formation techniques and could benefit from the addition of other datasets, such as club football team results. Moreover, the integration of input from industry experts and stakeholders, in coexistence with the investigation of techniques such as statistical and machine learning algorithms, may considerably enhance the resilience of results and predictive modelling in the domain of global football analytics.

After all, the project is a fantastic learning tool emphasizing how crucial flexibility, perseverance, and effective project management are to overcome obstacles and succeed. Through the application of these lessons learned and insights to future projects, the project is well-positioned to further advance sports analytics and data science in the context of international football.

Chapter 8 Appendices

User guide

Step 1: Sign in and Download Excel

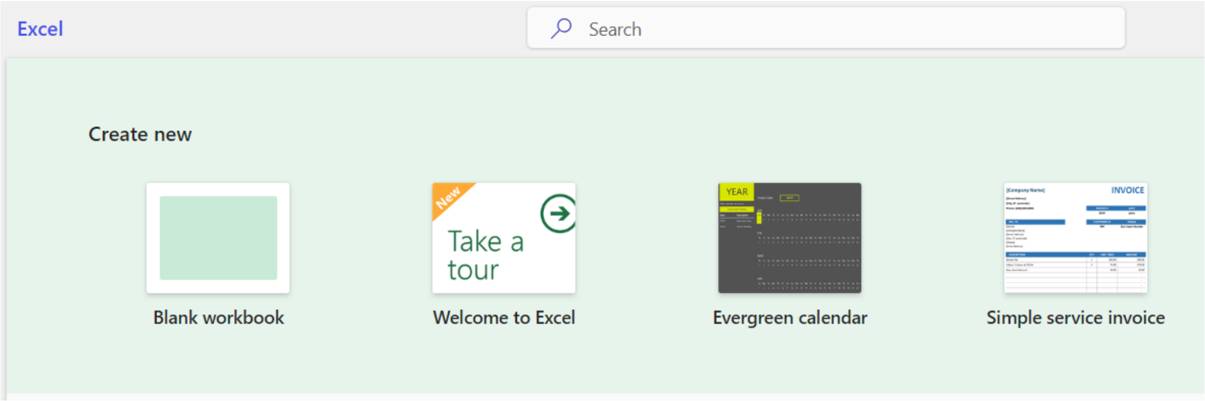


Microsoft Excel

Turn data into insights with free and premium spreadsheets.

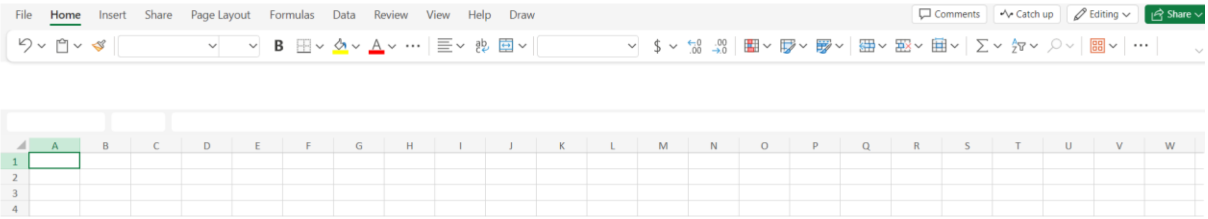
Description: Begin by downloading Microsoft Excel from the official Microsoft website or any trusted source to which you need to sign in or sing up first. Excel is a powerful spreadsheet application widely used for data analysis, calculation, visualization, and reporting tasks.

Step 2: After you sign in a page like this should appear.



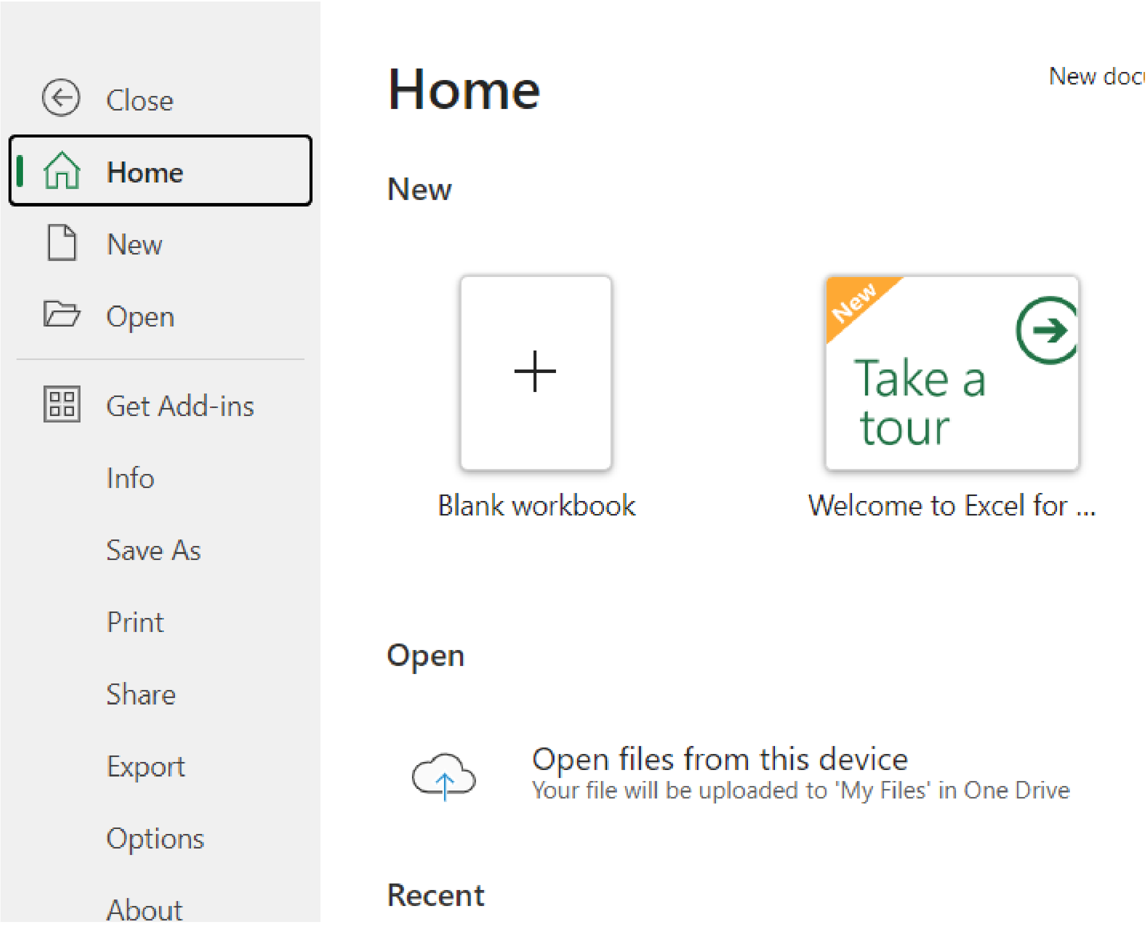
Description: If prompted, sign in to your Microsoft account to access the download options for Excel. Signing in ensures that you can download and install Excel using your Microsoft credentials.

Step 3: Go to excel blank page.



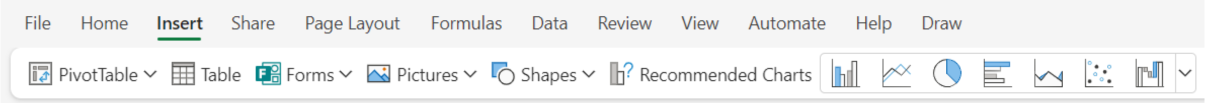
Description: Once Excel is installed on your computer, launch the application to access a blank spreadsheet. This blank page serves as the canvas where you can enter and manipulate data, perform calculations, and create visualizations.

Step 4 File Section of Excel



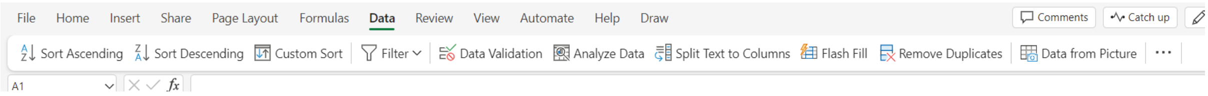
Description: The "File" section of Excel, located in the top-left corner of the application window. Here, you can access options for opening, saving, printing, and sharing your Excel spreadsheets.

Step 5 insert section of excel.



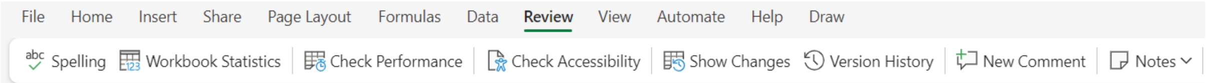
Description: Excel insert function contains table format such as pivot table, these help you to pick the right tool to display data and make an informed decision.

Step 6 data manipulation excel.



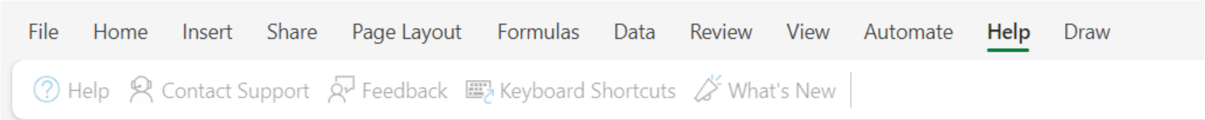
Description: Utilize Excel's powerful features for data manipulation, including sorting, filtering, formatting, and analysing data. These tools allow you to organize and manipulate your data to derive insights and make informed decisions.

Step 7 Review excel.



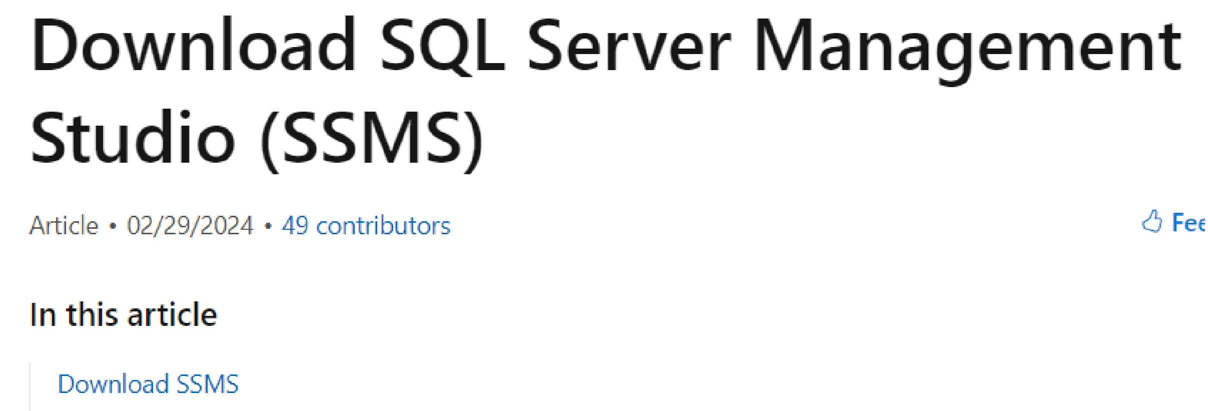
Description: Review and validate your data and calculations in Excel to ensure accuracy and reliability. Use Excel's reviewing tools to check for errors, inconsistencies, and outliers in your data.

Step Help section of excel.



Description: Explore Excel's built-in help resources to learn more about its features and functionalities. Access tutorials, documentation, and community forums to troubleshoot issues and enhance your Excel skills.

Step 1: Download SQL



Download SQL Server Management Studio (SSMS)

Article • 02/29/2024 • 49 contributors

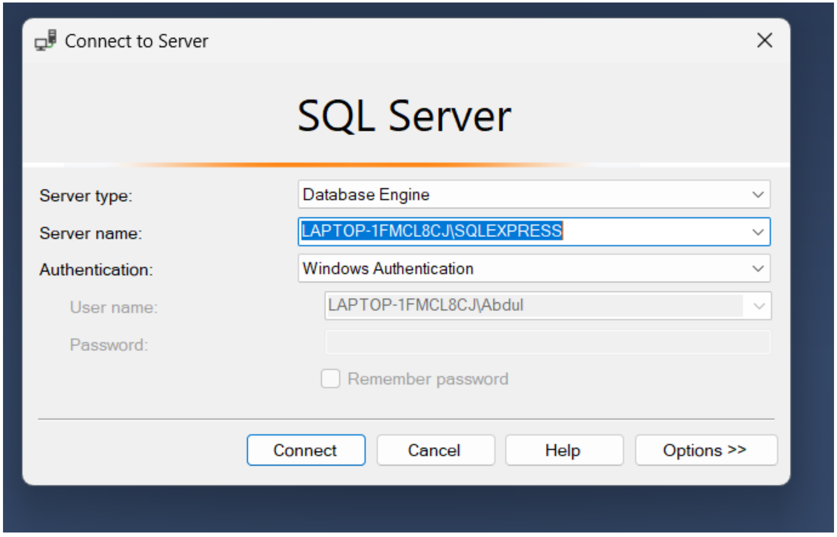
Fee

In this article

Download SSMS

Description: Begin by downloading Microsoft SQL Server from their respective official websites or trusted sources. SQL software allows users to interact with databases, execute queries, and manage data.

Step 2: Connect to SQL



Connect to Server

SQL Server

Server type: Database Engine

Server name: LAPTOP-1FMCL8CJ\SQLEXPRESS

Authentication: Windows Authentication

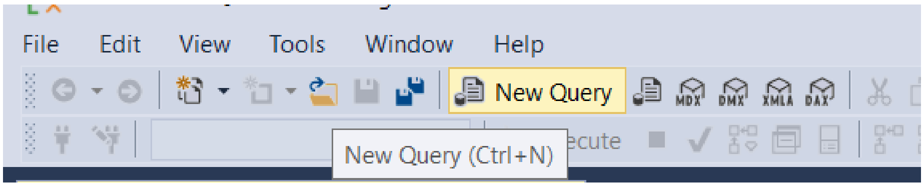
User name: LAPTOP-1FMCL8CJ\Abdul

Password:

□ Remember password

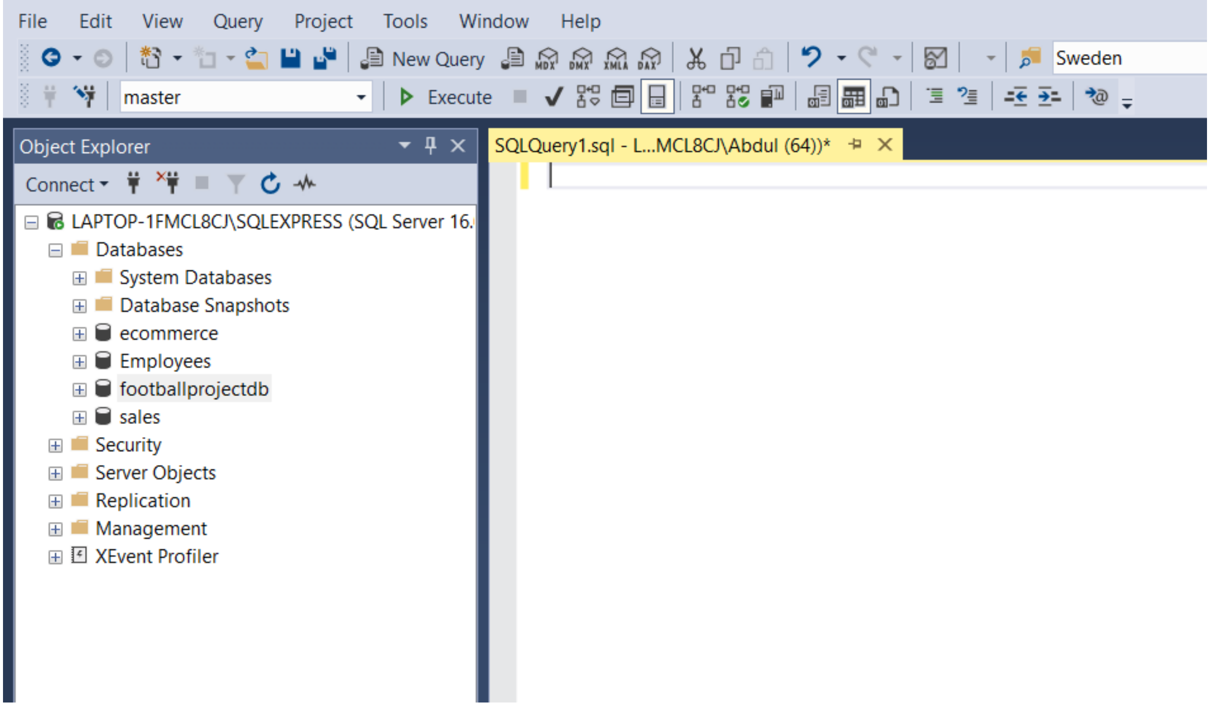
Description: After installing SQL software, launch the application and connect to the desired SQL database. Enter the necessary connection details such as server address, username, and password to establish a connection to the database.

Step 3: Press Query



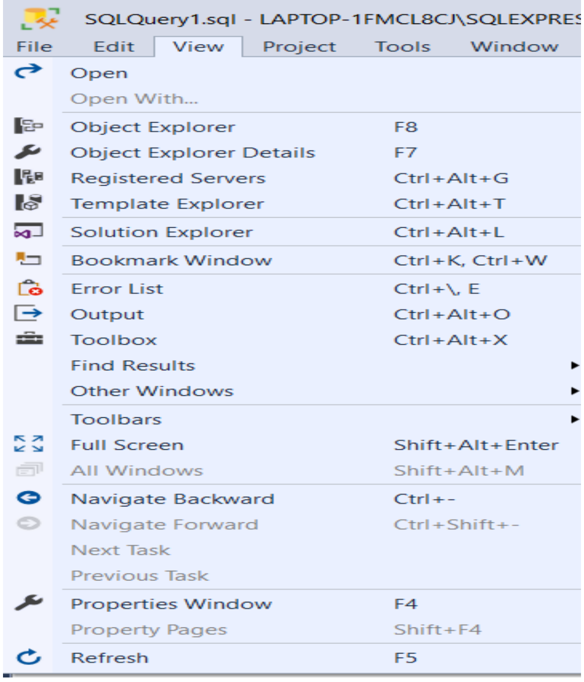
Description: Once connected to the SQL database, navigate to Query section where users can write and execute SQL queries. This section provides a workspace for querying data from the database and performing various operations.

Step 4: blank query



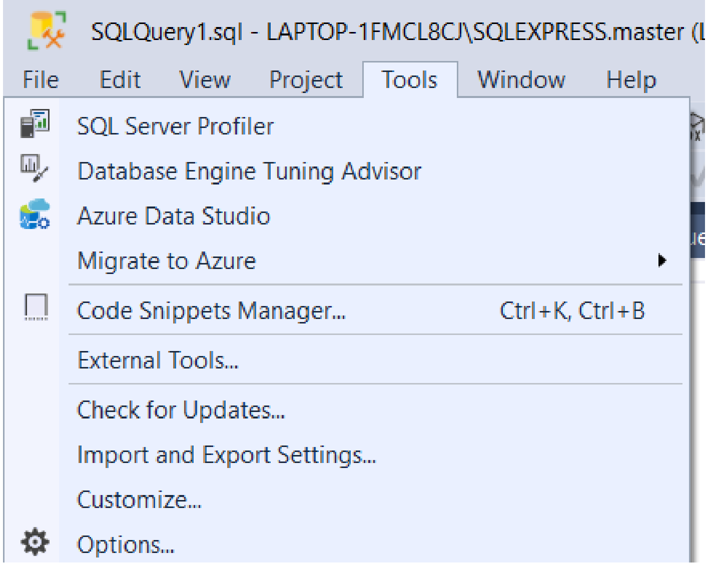
Description: Blank query window to begin writing SQL statements. In this window, users can write SQL queries to retrieve, manipulate, and manage data stored in the database tables.

Step 5: View option



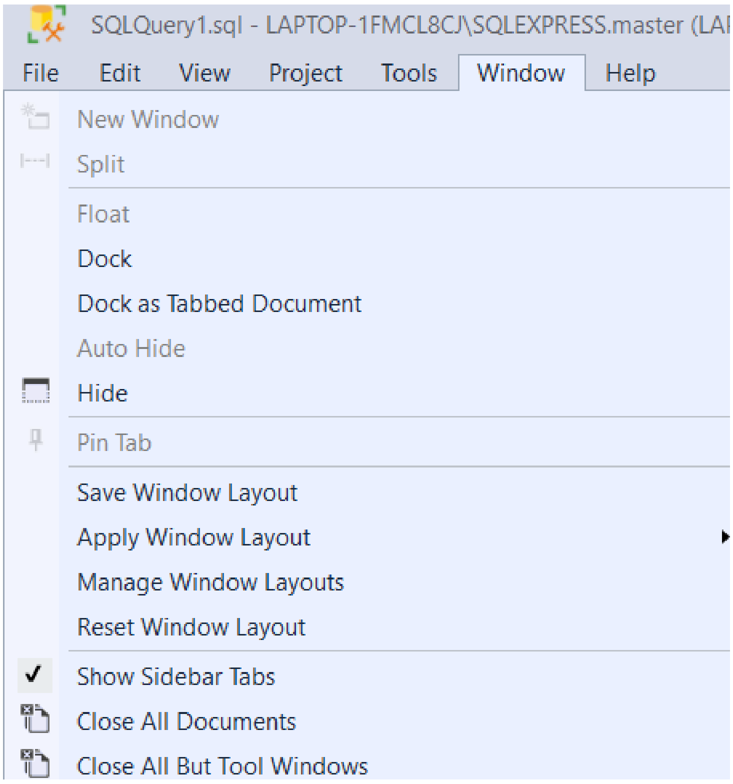
Description: The View options available in the SQL software. These options allow users to customize the appearance and layout of the query window, making it easier to work with SQL queries.

Step 6: Tools of SQL


SQLQuery1.sql - LAPTOP-1FMCL8CJ\SQLEXPRESS.master (l
File
Edit
View
Project
Tools
Window
Help
SQL Server Profiler
Database Engine Tuning Advisor
Azure Data Studio
Migrate to Azure
Code Snippets Manager... Ctrl+K, Ctrl+B
External Tools...
Check for Updates...
Import and Export Settings...
Customize...
Options...

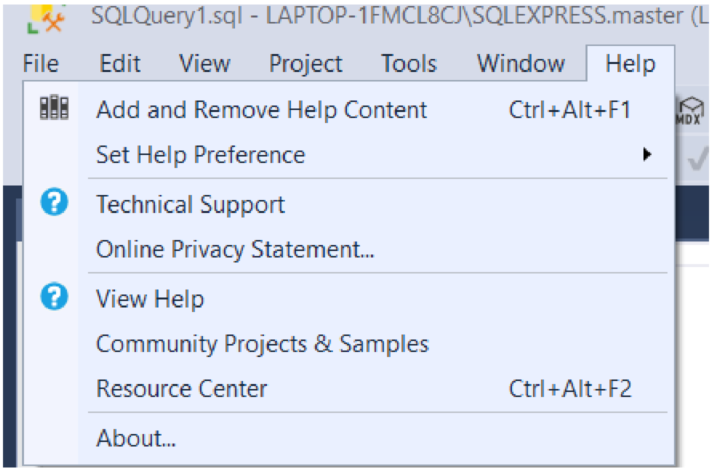
Description: Tools and features provided by the SQL software to enhance your query writing experience. These tools can include syntax highlighting, auto-completion, code snippets, and query execution options.

Step 7: Suitable SQL window



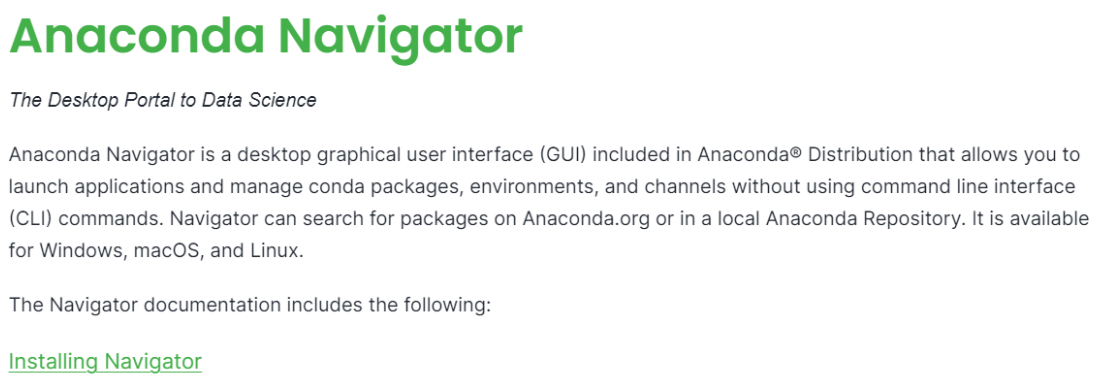
Description: Select the appropriate SQL window or tab based on your specific task or query requirements. Some SQL software can offer diverse types of windows for writing diverse types of SQL statements or performing specific tasks.

Step 8: Help SQL


SQLQuery1.sql - LAPTOP-1FMCL8CJ\SQLEXPRESS.master (L
File
Edit
View
Project
Tools
Window
Help
Add and Remove Help Content Ctrl+Alt+F1
Set Help Preference
Technical Support
Online Privacy Statement...
View Help
Community Projects & Samples
Resource Center Ctrl+Alt+F2
About...

Description: Explore the help resources available in the SQL software to find documentation, tutorials, and troubleshooting tips. These resources can help users learn SQL concepts, syntax, and best practices, as well as troubleshoot any issues they encounter.

Step 1: Download Anaconda Navigator



Anaconda Navigator

The Desktop Portal to Data Science

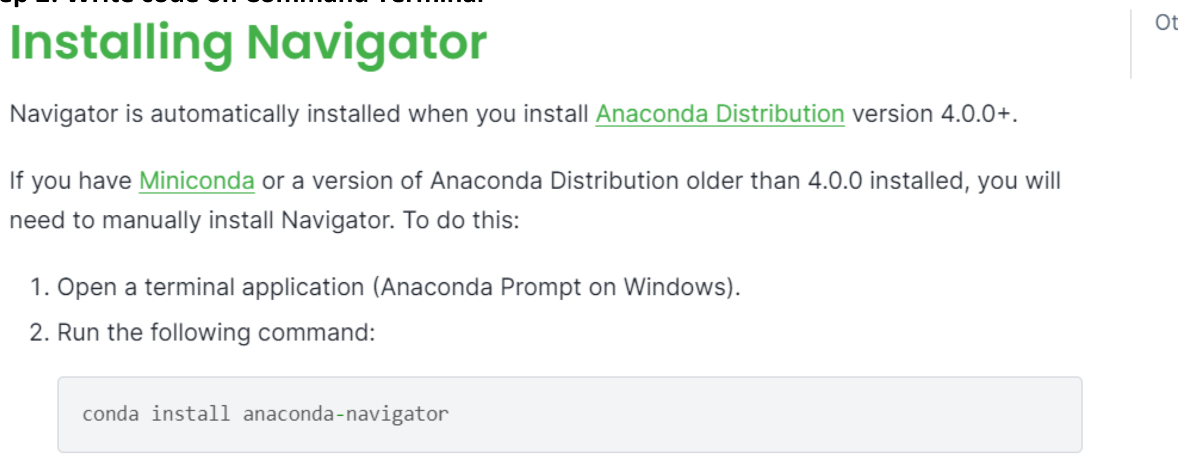
Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® Distribution that allows you to launch applications and manage conda packages, environments, and channels without using command line interface (CLI) commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

The Navigator documentation includes the following:

Installing Navigator

Description: Begin by downloading Anaconda Navigator, a powerful platform for managing Python and R packages. Anaconda Navigator simplifies package management and deployment, making it an ideal choice for data science projects.

Step 2: Write code on Command Terminal



Installing Navigator

Ot

Navigator is automatically installed when you install Anaconda Distribution version 4.0.0+.

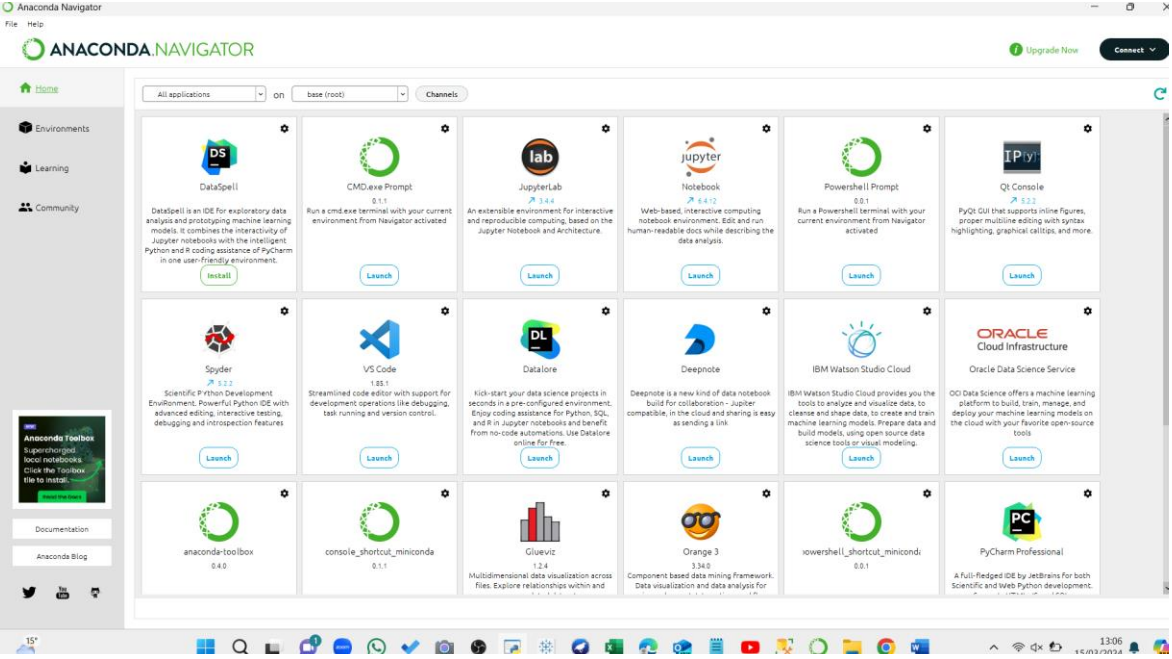
If you have Miniconda or a version of Anaconda Distribution older than 4.0.0 installed, you will need to manually install Navigator. To do this:

1. Open a terminal application (Anaconda Prompt on Windows).
2. Run the following command:

```
conda install anaconda-navigator
```

Description: After installing Anaconda Navigator, open the command terminal and execute the provided code to launch the Anaconda Navigator application. This step ensures that you can access the Anaconda Navigator interface to manage your Python environments and packages effectively.

Step 3: Menu page of anaconda navigator



DataSpell

DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.

CMD.exe Prompt 0.1.1

Run a cmd.exe terminal with your current environment from Navigator activated

JupyterLab 3.4.4

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook 6.4.12

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Powershell Prompt 0.0.1

Run a Powershell terminal with your current environment from Navigator activated

Qt Console 5.2.2

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder 5.2.2

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

VS Code 1.85.1

Streamlined code editor with support for development operations like debugging, task running and version control.

Datalore

Kick-start your data science projects in seconds in a pre-configured environment. Enjoy coding assistance for Python, SQL, and R in Jupyter notebooks and benefit from no-code automations. Use Datalore online for free.

Deepnote

Deepnote is a new kind of data notebook build for collaboration - Jupiter compatible, in the cloud and sharing is easy as sending a link

IBM Watson Studio Cloud

IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.

ORACLE Cloud Infrastructure

Oracle Data Science Service

OCI Data Science offers a machine learning platform to build, train, manage, and deploy your machine learning models on the cloud with your favorite open-source tools

anaconda-toolbox 0.4.0

console_shortcut_miniconda 0.1.1

Glueviz 1.2.4

Multidimensional data visualization across files. Explore relationships within and

Orange 3 3.34.0

Component based data mining Framework. Data visualization and data analysis for

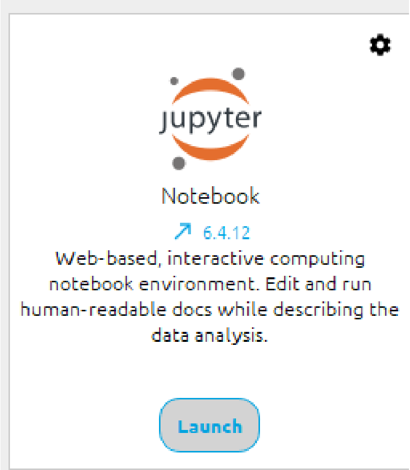
owershell_shortcut_minicond 0.0.1

PyCharm Professional

A full-fledged IDE by JetBrains for both Scientific and Web Python development.

Description: Upon launching Anaconda Navigator, you'll be greeted with the menu page, which provides easy access to various tools and applications, including Jupyter Notebook. Use the menu page to locate and launch Jupyter Notebook for Python programming.

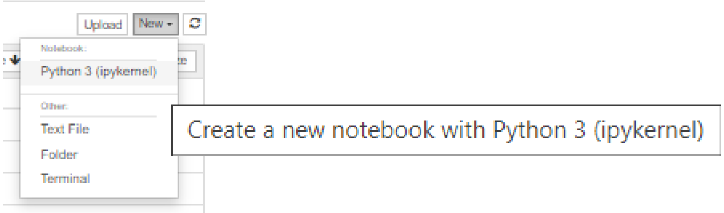
Step 4: Launch Juptyer Notebook



Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

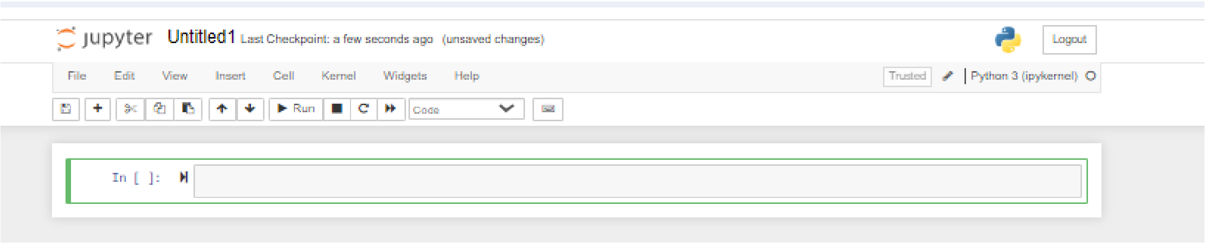
Description: Select Jupyter Notebook from the menu page to launch the application. Jupyter Notebook provides an interactive computing environment for writing and executing Python code, visualizing data, and creating rich-text documents.

Step 5 Press new on Juptyer Notebook for new page



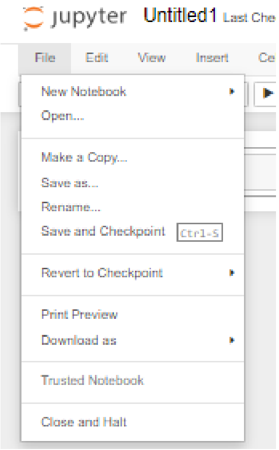
Description: Once Jupyter Notebook is launched, click on the "New" button to create a new notebook document. Each notebook represents a Python environment where you can write and execute code, add explanatory text, and create visualizations.

Step 6: Python is connected to notebook.



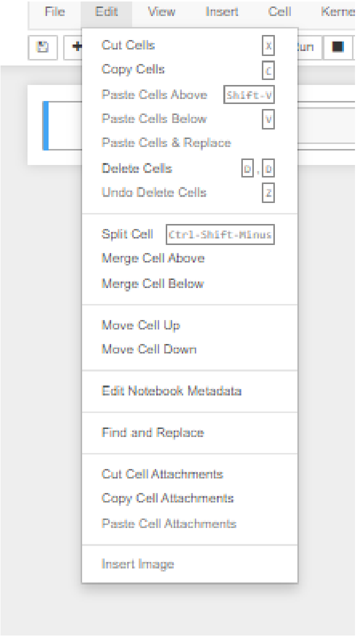
Description: Confirm that Python is successfully connected to the notebook environment. This step ensures that you can write and execute Python code within the Jupyter Notebook interface seamlessly.

Step 7: File section of Jupyter



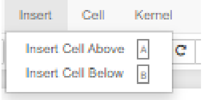
Description: Familiarize yourself with the file section of Jupyter Notebook, where you can create, open, and save notebook files, as well as manage directories and files within your Python environment.

Step 8: Edit work on Jupyter.



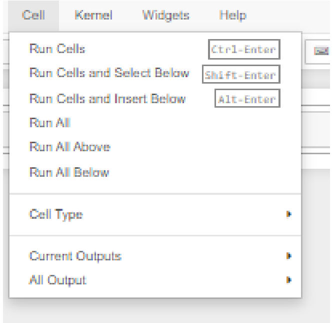
Description: Utilize the editing features of Jupyter Notebook to modify and refine your Python code and text cells. You can insert, delete, and modify cells as needed to develop your data analysis or machine learning projects.

Step 9: Insert row or delete section.



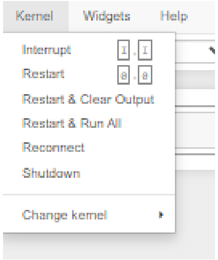
Description: Use the options available in Jupyter Notebook to insert new rows or delete existing sections within your notebook. This functionality allows you to organize your code and text cells effectively and tailor them to your specific requirements.

Step 10: run cell option.



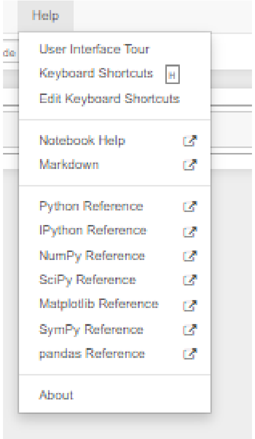
Description: Execute Python code cells within your notebook by using the "Run" cell option. This action runs the code in the selected cell, producing output and updating variables, as necessary.

Step 11: Kernal Section of Juptyer Notebook



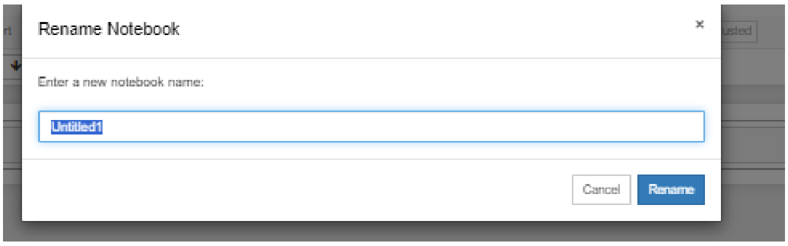
Description: The kernel section of Jupyter Notebook allows you to manage the engine of your notebook. You can restart, interrupt, and change kernels to accommodate different programming languages and environments.

Step 12: Help option on how to use Juptyer Notebook with Python connected.



Description: The built-in help resources within Jupyter Notebook to learn more about its features and functionalities. You can access documentation, tutorials, and community forums to troubleshoot issues and enhance your Python programming skills.

Step 13: Rename notebook.



Rename Notebook

Enter a new notebook name:

Untitled1

Description: Customize the name of your notebook to reflect its content or purpose. Renaming your notebook makes it easier to organize and identify your projects within Jupyter Notebook.

<u>Meeting dates with supervisor</u>

Friday 6th October 2023 -Thursday 12th October 2023 -Thursday 26th October 2023
Thursday 9th November 2023 - Friday 7th December 2023 - Friday 12th January 2024
Friday 26th January 2024 - Friday 9th February 2024

Chapter 9: References

Article

John Goddard., & Ioannis Asimakopoulos (2004). Forecasting football results and the efficiency of fixed odds betting. Wiley online library, 21(1),51-66, https://onlinelibrary.wiley.com/doi/10.1002/for.877

Rahul Baboota & Harleen Kaur (2019), Predictive analysis and modelling football results using machine learning approach for English Premier League, Science Direct, 35(2), 741-755, https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116

Tsuneshi Obata & Shizue Izumi (2022), Analysis and visualisations of team performances of football games, Data science: Present and Future, 5, 885- 898, https://link.springer.com/article/10.1007/s42081-022-00173-z

Roberto Gásquez & Vicente Royuela (2016), The Determinants of International Football Success: A Panel Data Analysis of the Elo Rating, Wiley online library, 97(2), 125-141, https://onlinelibrary.wiley.com/doi/full/10.1111/ssqu.12262#ssqu12262-note-0029

Fatima Rodrigues & Angelino Pinto (2022), Prediction of football match results using Machine Learning, Science Direct, 204(202), 463-470, https://www.sciencedirect.com/science/article/pii/S1877050922007955

Johannes Stubinger & Benedikt Mangold & Julian Knoll (2019), Machine Learning in football betting: Prediction of match results based on player characteristics, 10(1), 44 -45, https://www.mdpi.com/2076-3417/10/1/46

Tanvir Alam & Jassim Almulla (2020), Machine Learning models reveals key performances metrics of football players to win match in Qatar star league, Page 213695 – 213705 https://ieeexplore.ieee.org/abstract/document/9261335

Miguel-Angel Gomez (2020), Exploring elite soccer teams' performances during different match-status periods of close matches' comebacks, Science Direct, 132(1),

https://www.sciencedirect.com/science/article/abs/pii/S0960077919305235

Poojan Thakkar & Manan Shah (2021), An assessment of football through the lens of Data Science, Springer Link, 8, 823-836, https://link.springer.com/article/10.1007/s40745-021-00323-2,

Conference

Zulkifli Mohamad, Data Visualization of Football Performance Preceded to the Goal Scored, in K. Imran, innovation and technology in sports, (pp. 57-74). Springer: https://link.springer.com/chapter/10.1007/978-981-99-0297-2_6

Imran Sainan (2023), Data visualization of football using degree cardinality, Innovation, and technology in sports, (pp. 75-93). Springer: https://link.springer.com/chapter/10.1007/978-981-99-0297-2_7

Harry Elkins (2017), Implementing data analytics for football, in system & information, IEEE Xplore, IEEE:
https://ieeexplore.ieee.org/abstract/document/7937717

Che Mohamad Firdaus Che Mohd Rosli *et al* 2018 *J. Phys.: Conf. Ser.* **1020** 012003 https://iopscience.iop.org/article/10.1088/1742-6596/1020/1/012003/meta

Victor Chazan Pantzalis & Christos Tjortjis (2020), Sports analytics for football leagues table and player performance predictions, IEEE Xplore, IEEE: https://ieeexplore.ieee.org/abstract/document/9284352

Mansoor Alam (2020), A Data Science Approach to football collaborator section. IEEE Xplore, IEEE: https://ieeexplore.ieee.org/abstract/document/9208331

Hucaljuk and Rakipović's (2011), Predict football scores using machine-learning techniques, IEEE Xplore, IEEE: https://ieeexplore.ieee.org/abstract/document/5967321

Website

Kaggle (2023), international football results from 1872 to 2023, available from data card at: https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017 ![date accessed 18th October 2023]

Microsoft. (2023), Microsoft Excel, available from Microsoft: https://www.microsoft.com/en-gb/microsoft-365/excel ![date accessed 28th October 2023]

MySQL. (2023), MySQL, available from MySQL: https://www.thesql.com/ ![date accessed 29th October 2023]

SQL Server Management Studio available from SQL Server: https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms ![date accessed 6th January 2024]

Jupyter. (2023), Jupyter Notebook, available from Jupyter at: https://jupyter.org/ ![date accessed 29th October 2023]

Diagrams

ResearchGate (2019), Machine Leaning algorithm available from figure 1: Research https://www.researchgate.net/figure/Overview-diagram-of-machine-learning-algorithms-Machine-learning-is-a-subset-of_fig1_335604816 ![date accessed 7th January 2024]

Polymer 2023, Data visualizing techniques available from Polymer: https://www.polymersearch.com/blog/data-visualization ![date accessed 7th January 2024]

Wikipedia 2024, Data visualization process available from Data Science Processes https://en.m.wikipedia.org/wiki/File:Data_visualization_process_v1.png ![date accessed 7th January 2024]

Bitesize Learning 2015, The five of dysfunction of a team, (and how to overcome them) available from Models and Frameworks at: https://www.bitesizelearning.co.uk/resources/five-dysfunctions-of-a-team-summary-pyramid ![date accessed 7th January 2024]