NAME: ABHISHEK KUMAR SINGH

REG. NO: 12113396

ROLL NO: RK21URA02

INT353 CA-2

UEFA CHAMPIONS LEAGUE

Q1

Domain knowledge:

Domain: UEFA Champions League

1. **UEFA Champions League Overview:** The UEFA Champions League is an annual club football competition organized by the Union of European Football Associations (UEFA). It is considered one of the most prestigious tournaments in the world, featuring top-tier football clubs from various European countries. The tournament consists of several stages, including qualifying rounds, group stages, knockout rounds, and ultimately the final.

2. Key Concepts and Terminology:

- **Goals:** Goals scored by players during matches. Goals are a primary indicator of a team's offensive performance.
- **Assists:** Assists are passes or plays that directly lead to goals. They provide insights into a player's ability to create scoring opportunities.
- **Attack:** The offensive aspect of a team's performance, including metrics like goals, shots on target, and chances created.
- **Defence:** The defensive aspect of a team's performance, including metrics like clean sheets, interceptions, and tackles.

- **Passing:** Metrics related to successful and accurate passes, pass completion rate, and key passes that create goal-scoring chances.
- **Field Control:** Refers to a team's ability to maintain possession and control the play in various areas of the field.
- **GK Data:** Goalkeeper-specific metrics such as saves, clean sheets, and distribution accuracy.

3. Challenges and Considerations:

- **Data Quality:** Ensure that the dataset is accurate, complete, and free from errors. Missing or inconsistent data can lead to skewed analyses.
- **Data Granularity:** Consider the granularity of the data. Are you analysing data at the player level, team level, or match level? This affects the insights you can draw.
- **Contextual Understanding:** It's important to have a deep understanding of football and the Champions League format to interpret the data effectively.
- **Comparative Analysis:** We might want to compare teams or players from different leagues or countries, which could introduce challenges due to varying playing styles and levels of competition.
- **Time Series Analysis:** Given that the Champions League progresses through different stages over time, we might perform time series analysis to understand how team performance evolves across the tournament.
- **Feature Engineering:** Creating derived features such as goal difference, points per match, or conversion rates can provide additional insights.

4. Potential Insights:

- Identify top goal scorers, assist providers, and players with exceptional passing accuracy.
- Analyse team performance in terms of goals scored, conceded, and clean sheets to assess their offensive and defensive strengths.

- Investigate correlations between passing accuracy and possession control to understand teams' playing styles.
- Study goalkeepers' performance in terms of saves made, clean sheets, and distribution accuracy.

Conclusion:

Understanding the domain of the UEFA Champions League is essential before diving into Exploratory Data Analysis. Gaining familiarity with key concepts, terminology, challenges, and potential insights will enable me to extract meaningful information from the dataset and provide valuable insights into the performance of teams and players during the 2021-2022 season.

Data Understanding:

Below is a summary of the data understanding for the UEFA Champions League dataset chosen, broken down by each CSV file:

attacking.csv:

- This dataset contains information related to attacking statistics of players in the UEFA Champions League for the 2021-2022 season.
- Key columns include Player Name, Club, Position, Assists, Corners Taken, Offsides, Dribbles, and Matches Played.
- The dataset allows for the analysis of player performance in terms of assists, corner taking, dribbling skills, and offside situations.

attempts.csv:

- This dataset provides details on players' attempts at goal during UEFA Champions League matches in the 2021-2022 season.

- Columns include Playing Position, Total Attempts, On Target Attempts, Off Target Attempts, Blocked Attempts, and Matches Played.
- It offers insights into players' shooting accuracy, their ability to place shots on target, and how often their attempts are blocked by opposing players.

defending.csv:

- The dataset focuses on defensive statistics of players in the UEFA Champions League for the 2021-2022 season.
- Important columns include Balls Recovered, Tackles, Tackles Won, Tackles Lost, Clearance Attempted, and Matches Played.
- It allows for the analysis of players' defensive contributions, including ball recoveries, tackling efficiency, and clearance attempts.

disciplinary.csv:

- This dataset contains information related to disciplinary actions taken against players during UEFA Champions League matches in the 2021-2022 season.
- Key columns include Fouls Committed, Fouls Suffered, Red Cards, Yellow Cards, Minutes Played, and Matches Played.
- It provides insights into players' disciplinary records, including the number of fouls committed, cards received, and minutes played.

distribution.csv:

- This dataset focuses on players' distribution and passing statistics in UEFA Champions League matches for the 2021-2022 season.
- Columns include Pass Accuracy, Passes Attempted, Passes Completed, Cross Accuracy, Crosses Attempted, Crosses Completed, Free Kicks Taken, and Matches Played.
- It enables analysis of players' passing accuracy, cross effectiveness, and their role in free-kick situations.

goalkeeping.csv:

- This dataset provides goalkeeping statistics for players in the UEFA Champions League during the 2021-2022 season.
- Important columns include Position, Saves, Goals Conceded, Saved Penalties, Clean Sheets, Punches Made, and Matches Played.
- It allows for the assessment of goalkeepers' performance in terms of saves, goals conceded, penalty saves, and clean sheets.

goals.csv:

- This dataset contains information on goals scored by players in the UEFA Champions League for the 2021-2022 season.
- Key columns include Player Name, Club, Position, Goals, Goals with Right Foot, Goals with Left Foot, Header Goals, Goals from Other Body Parts, Goals Inside the Penalty Area, Goals Outside the Penalty Area, Penalty Goals, and Matches Played.
- It provides insights into players' goal-scoring patterns, including the types of goals (e.g., headers, penalties) and where they score from.

key_stats.csv:

- This dataset offers key statistics related to player performance in UEFA Champions League matches in the 2021-2022 season.
- Columns include Player Name, Club, Position, Minutes Played, Matches Played, Goals, Assists, and Distance Covered.
- It allows for an overview of player contributions in terms of goals, assists, playing time, and distance covered during matches.

Understanding these datasets is essential for conducting meaningful exploratory data analysis (EDA) and extracting insights about player and team performance in the UEFA Champions League for the specified

season. Further analysis can now be conducted based on research questions and objectives.

Reason for choosing dataset:

Subject: Selection of the UEFA Champions League Dataset - A Tribute to Cristiano Ronaldo

As a football enthusiast and a devoted fan of Cristiano Ronaldo, I wanted to take a moment to share my thought process behind my choice of the UEFA Champions League dataset for upcoming project.

Given my passion for football and my admiration for Cristiano Ronaldo, I believe I'll resonate with the reasons that led me to select this dataset. Here's why I find the UEFA Champions League dataset to be a fitting choice:

1. The Grandest Stage in Football: The UEFA Champions League represents the pinnacle of club football, bringing together the best teams from across Europe to compete for glory. For fans like us, it's a platform where dreams are realized, history is made, and unforgettable moments are etched in the annals of football history.

2. Celebrating Ronaldo's Journey: Cristiano Ronaldo, my favourite footballer, has left an indelible mark on the Champions League. His incredible performances, stunning goals, and unmatched dedication have made him a true icon of the tournament. By exploring this dataset, I not only honour his legacy but also gain insights into his impact on the competition over the years.

3. Insights into Excellence: As fans, I admire the skill, teamwork, and strategies that go into each match. This dataset allows me to delve into the nuances of team and player performance, uncover patterns, and analyse the factors that contribute to success on the grand stage.

4. The Joy of Discovery: Exploratory data analysis of the UEFA Champions League dataset presents me with an opportunity to uncover hidden gems of information.

Whether it's discovering rising talents, identifying trends, or revisiting historic matchups, the process promises to be an exciting journey.

5. Bridging Passion and Analysis: By working with this dataset, I'll be able to merge our love for football with the analytical skills I've developed. It's a chance to combine my fandom with a rigorous approach to data analysis, creating a unique blend of excitement and expertise.

6. A Homage to Ronaldo's Impact: Cristiano Ronaldo's journey through the Champions League, from his days at Manchester United to his triumphs with Real Madrid and beyond, is a testament to his dedication and excellence. This dataset gives me a chance to retrace his steps and quantify his influence on the tournament statistically.

In a nutshell, my choice of the UEFA Champions League dataset is a tribute to passion for football and admiration for Cristiano Ronaldo. It's an opportunity to immerse ourselves in the world of football data, celebrate the sport I love, and honour the accomplishments of a true legend.

I'm excited about the insights I'll uncover and the knowledge I'll gain through my exploratory data analysis. Let's embark on this journey and make my project a fitting ode to the beautiful game and to the remarkable athlete who has touched our hearts.

Looking forward to diving into the dataset and creating something truly special.

With enthusiasm and football fervour.

Analysis Questions:

General Player Performance:

1. Who was the top goal scorer in the UEFA Champions League for the 2021-2022 season?
2. Which player provided the most assists in the tournament?
3. Who had the highest pass accuracy among all players?
4. Which player attempted the most dribbles?

5. Who received the most yellow cards in the tournament?
6. Which goalkeeper kept the highest number of clean sheets?
7. Who had the most minutes played in the UEFA Champions League?
8. Which player covered the most distance during matches?

Scoring Patterns:

9. What percentage of goals were scored with headers in the tournament?
10. Which player scored the most goals with their left foot?
11. How many goals were scored from outside the penalty area?
12. Who was the most effective penalty taker in terms of conversion rate?
13. What was the average number of goals scored per match in the tournament?

Positional Analysis:

14. Which position (e.g., forward, midfielder) had the highest average number of goals scored?
15. Do players in certain positions tend to have higher pass accuracy than others?
16. Are defenders more likely to receive yellow cards compared to midfielders or forwards?
17. Which position had the highest average number of tackles made per match?

Team Performance:

18. Which club scored the most goals in the UEFA Champions League?
19. Which team had the best defensive record in terms of goals conceded?
20. Did teams with higher possession percentages tend to win more matches?
21. Which club had the highest pass completion rate?

Discipline and Cards:

22. Is there a correlation between the number of fouls committed and the number of yellow cards received?
23. Which player had the highest ratio of red cards to matches played?
24. Did players who received red cards tend to have fewer minutes played in subsequent matches?
25. How many players received multiple red cards during the tournament?

1. These questions can serve as a starting point for your exploratory data analysis (EDA) and can help you uncover interesting insights about player and team performance in the UEFA Champions League for the 2021-2022 season.

Q2)

Libraries used:

NumPy (np):

- NumPy is a fundamental library for numerical computing in Python.
- It provides support for arrays and matrices, which are essential for numerical operations.
- Commonly used for mathematical calculations, data manipulation, and working with large datasets.

2. **Pandas (pd)**:
    - Pandas is a data manipulation library that provides data structures like DataFrames and Series.
    - It is used for data cleaning, exploration, and transformation.
    - Great for importing, exporting, and working with structured data.
3. **Matplotlib (plt)**:

- Matplotlib is a popular plotting library for creating static, interactive, and animated visualizations.
- It provides fine-grained control over plot elements.
- Useful for creating various types of charts, graphs, and plots.

4. **Matplotlib.ticker (ticker)**:
    - Matplotlib's ticker module is used to control the formatting of tick locations on plot axes.
    - It allows you to customize the appearance of ticks, tick labels, and tick positions on the axes.
5. **Plotly Express (px)**:
    - Plotly Express is a high-level interface to create interactive plots and dashboards.
    - It simplifies the creation of complex visualizations and interactive web-based dashboards.
    - Often used for data exploration and sharing insights with stakeholders.
6. **Seaborn (sns)**:
    - Seaborn is a statistical data visualization library based on Matplotlib.
    - It provides a high-level interface for creating aesthetically pleasing and informative statistical graphics.
    - Frequently used for creating attractive, informative, and complex visualizations with minimal code.
7. **Warnings**:
    - The 'warnings' library is used to control how warning messages are displayed or handled in Python.
    - It can be used to suppress or customize warning messages to improve code readability and debugging.

1. These libraries play critical roles in data analysis and visualization, enabling data professionals to perform tasks such as data exploration, data cleaning, statistical analysis, and creating compelling visual representations of data. They provide a powerful toolkit for working with data in various formats and for different analytical purposes.

Approaches:

Data Filtering:

The practice of choosing and extracting a subset of data from a larger dataset according to particular standards or requirements is known as data filtering. It enables you to filter out useless or superfluous information and concentrate on pertinent sections of the material. Conditions can be used to filter data. For example, they can be used to pick rows where a certain column satisfies a set of requirements (e.g., all rows where "Genre" equals "Strategy").

Aggregation:

Data is combined and summarized during the aggregation process in order to provide a single result or a reduced group of outcomes. It is frequently used to compute or provide summary statistics for data sets. The functions sum, average, count, minimum, and maximum are frequently used for aggregation. For instance, you may compile data to determine the average revenue for a genre or the overall revenue for a particular game.

Statistical Analysis:

The process of applying statistical tools and procedures to data in order to derive findings, infer conclusions, and get new insights is known as statistical analysis. It entails looking at the data's correlations, trends, and variances. Regression analysis, correlation analysis, and descriptive statistics (such as mean, median, and standard deviation) are a few examples of statistical

analysis. It is employed to identify statistically important results, trends, and linkages in the data.

Q3)

Steps of EDA:

1 Data Summary:

- Start by obtaining a summary of your dataset, including the number of rows and columns, data types, and basic statistics like mean, median, standard deviation, and quartiles.
- Use Pandas functions like **info()**, **describe()**, and **head()** to get an initial overview.

2. **Data Cleaning**:
    - Identify and handle missing data by either imputing missing values or removing rows/columns with missing data.
    - Check for and handle duplicate records if applicable.
    - Address outliers and anomalies that may skew the analysis.
3. **Data Visualization**:
    - Create visualizations such as histograms, bar charts, box plots, scatter plots, and heatmaps to explore the distribution and relationships between variables.
    - Use libraries like Matplotlib, Seaborn, and Plotly for visualization.
4. **Univariate Analysis**:
    - Analyze individual variables one at a time to understand their distributions and characteristics.
    - Explore categorical variables with frequency counts, bar charts, or pie charts.

- For continuous variables, use histograms, density plots, or summary statistics.

5. **Bivariate Analysis**:
    - Examine relationships between pairs of variables to uncover correlations or associations.
    - Create scatter plots, box plots, or violin plots to visualize how variables interact with each other.
    - Calculate correlation coefficients to quantify relationships.
6. **Multivariate Analysis**:
    - Extend the analysis to multiple variables simultaneously.
    - Use techniques like pair plots, heatmaps, or parallel coordinates to visualize complex interactions between variables.
7. **Hypothesis Testing**:(to be done in future)
    - Formulate hypotheses about relationships in the data and conduct statistical tests to confirm or refute them.
    - Common tests include t-tests, chi-square tests, ANOVA, and correlation tests.
8. **Documentation and Reporting**:
    - Keeping detailed records of your EDA process, including code, visualizations, and observations.
    - Summarize key findings and insights in a clear and concise report or presentation.
9. **Iterative Process**:
    - EDA is often iterative, meaning you may revisit and refine your analysis as you gain more insights or encounter new questions.

10. **Domain Knowledge**:
    - Leverage domain knowledge to interpret findings and contextualize results within the relevant field or industry.

Visualizing the answers to the questions and providing insights and conclusions for each can be a comprehensive process. Below, I'll outline the visualization techniques and key findings for each question:

Q3)

Visualisation of all the Questions for Analysis:

General Player Performance:

1. **Top Goal Scorer**:
    - Create a bar chart with players on the x-axis and their goal counts on the y-axis.
    - Insight: Identify the player with the highest goal count.
    - Conclusion: The top goal scorer for the UEFA Champions League in the 2021-2022 season is **Benzema** with **15** goals.
2. **Most Assists**:
    - Create a bar chart with players on the x-axis and their assist counts on the y-axis.
    - Insight: Identify the player with the most assists.
    - Conclusion: The player with the most assists in the tournament is **Bruno Fernandez** with **7** assists.
3. **Highest Pass Accuracy**:
    - Create a bar chart with players on the x-axis and their pass accuracy percentages on the y-axis.
    - Insight: Identify the player with the highest pass accuracy.

- Conclusion: **Erokhin** has the highest pass accuracy among all players, with a pass accuracy of **98.0%**

4. **Most Dribbles Attempted**:
    - Create a bar chart with players on the x-axis and the number of dribbles attempted on the y-axis.
    - Insight: Identify the player who attempted the most dribbles.
    - Conclusion: **Vinicius Junior** attempted the most dribbles in the tournament, with **83** dribbles.
5. **Most Yellow Cards**:
    - Create a bar chart with players on the x-axis and the number of yellow cards on the y-axis.
    - Insight: Identify the player with the most yellow cards.
    - Conclusion: **Felipe** received the most yellow cards in the tournament, with **2** yellow cards.
6. **Most Clean Sheets (Goalkeepers)**:
    - Create a bar chart with goalkeepers on the x-axis and the number of clean sheets on the y-axis.
    - Insight: Identify the goalkeeper with the highest number of clean sheets.
    - Conclusion: **Courtois** kept the highest number of clean sheets in the tournament, with **5** clean sheets.
7. **Most Minutes Played**:
    - Create a bar chart with players on the x-axis and the minutes played on the y-axis.
    - Insight: Identify the player who played the most minutes.
    - Conclusion: **Courtois** played the most minutes in the UEFA Champions League, with **1230** minutes.

8. **Most Distance Covered**:
    - Create a bar chart with players on the x-axis and the distance covered on the y-axis.
    - Insight: Identify the player who covered the most distance.
    - Conclusion: **Lewandoski** covered the most distance during matches, covering a total of **99**.**7** kilometers.

Scoring Patterns:

9. **Percentage of Goals Scored with Headers**:
    - Create a pie chart to show the percentage of goals scored with headers.
    - Insight: Determine the percentage of header goals.
    - Conclusion: **16.22%** of goals in the tournament were scored with headers.
10. **Most Goals with Left Foot**:
    - Create a bar chart with players on the x-axis and the number of left-footed goals on the y-axis.
    - Insight: Identify the player with the most left-footed goals.
    - Conclusion: **Salah** scored the most goals with their left foot, with **8** goals.
11. **Goals from Outside Penalty Area**:
    - Create a bar chart to show the number of goals scored from outside the penalty area.
    - Insight: Count the number of goals from outside the penalty area.
    - Conclusion: A total of **38** goals were scored from outside the penalty area.
12. **Effective Penalty Taker**:

- Create a bar chart with players on the x-axis and their penalty conversion rates on the y-axis.
- Insight: Identify the player with the highest penalty conversion rate.
- Conclusion: **Benzema** was the most effective penalty taker, with a conversion rate of **8.33%.**

13. **Average Goals per Match**:
    - Calculate the average number of goals per match.
    - Insight: Determine the average goal-scoring rate.
    - Conclusion: The average number of goals scored per match in the tournament was **0.33**.

Positional Analysis:

14. **Position with Highest Average Goals**:
    - Create a bar chart with player positions on the x-axis and the average number of goals on the y-axis.
    - Insight: Identify the position with the highest average goals.
    - Conclusion: Forward had the highest average number of goals scored per player.
15. **Pass Accuracy by Position**:
    - Create a box plot or violin plot to visualize pass accuracy by position.
    - Insight: Compare pass accuracy across different positions.
    - Conclusion: Medfield players tend to have higher pass accuracy compared to other positions.
16. **Yellow Cards by Position**:
    - Create a bar chart or box plot to show the distribution of yellow cards by player position.
    - Insight: Determine if certain positions receive more yellow cards.

- Conclusion: Defenders players are more likely to receive yellow cards compared to Forwards and midfielders.

17. **Average Tackles by Position**:
    - Create a bar chart with player positions on the x-axis and the average number of tackles on the y-axis.
    - Insight: Identify the position with the highest average number of tackles.
    - Conclusion: Midfield players had the highest average number of tackles per match.

Team Performance:

18. **Club with Most Goals**:
    - Bar chart showing clubs with the most goals scored.
    - Insight: Identifying the club that scored the most goals.
    - Bayern scored the most goals.
19. **Best Defensive Team**:
    - Bar chart showing clubs with the fewest goals conceded.
    - Insight: Determining the team with the best defensive record.
20. **Possession Percentage and Match Wins**:
    - Scatter plot or correlation matrix showing the relationship between possession percentage and match wins.
    - Insight: Analyzing the impact of possession on match outcomes.
21. **Club with Highest Pass Completion Rate**:
    - Bar chart showing clubs with the highest pass completion rates.
    - Insight: Identifying the club with the best passing accuracy.

Discipline and Cards:

22. **Correlation between Fouls and Yellow Cards**:
    - Scatter plot or correlation coefficient showing the relationship between fouls committed and yellow cards.
    - Insight: Examining the relationship between fouls and disciplinary actions.
23. **Player with Highest Red Card Ratio**:
    - Bar chart or table showing players with the highest ratio of red cards to matches played.
    - Insight: Identifying players with a high red card-to-match ratio.
24. **Minutes Played after Red Cards**:
    - Box plots or histograms showing minutes played by players who received red cards.
    - Insight: Understanding how red cards affect players' playing time.
25. **Players with Multiple Red Cards**:
    - Bar chart or table showing the number of red cards received by each player.
    - Insight: Identifying players who received multiple red cards during the tournament.