FINAL REPORT OF EDA PROJECT

Course code: CSM353

upGrad

Submitted by: Raihan Koduvaly

Section: K22UG

Roll number: 52,

Reg number: 12209611

Submitted to: Ved Prakash Chaubey: 63892

Date: 18/11/2024. Monday

B. TECH COMPUTER SCIENCE OF ENGINEERING

Supervisor Certificate

This is to certify that **RAIHAN KODUVALY**, a student of Lovely Professional College from B. Tech COMPUTER SCIENCE, has successfully completed the Exploratory Data Analysis (EDA) project titled “FIFA World Cup 2022: Data Exploration and Insights” under my guidance and supervision.

The project involved analysing and visualizing data related to the FIFA World Cup 2022 to uncover meaningful patterns, trends, and insights about the tournament. It demonstrates the student’s ability to collect, clean, and interpret data effectively, while applying advanced analytical techniques.

The work completed is original and showcases excellent analytical and presentation skills. It meets the academic requirements set forth by the institution for submission of the final project report.

I wish RAIHAN KODUVALY great success in future endeavours.

Ved Prakash Chaubey: 63892
LOVELY PROFESSIONAL UNIVERSITY/UPGRAD CAMPUS
Date: 18/11/2024. MONDAY

Acknowledgment

I would like to express my heartfelt gratitude to everyone who contributed to the successful completion of this Exploratory Data Analysis report on the FIFA World Cup 2022.

Firstly, I extend my sincere thanks to my mentors and instructors for their guidance, expertise, and encouragement throughout this project. Their insights and feedback have been invaluable in refining my approach to data analysis and interpretation.

I am deeply grateful to my peers and colleagues for their constructive discussions and collaboration, which significantly enriched the quality of this work.

I also acknowledge the use of publicly available datasets and resources that provided the foundation for this analysis. Without these resources, this report would not have been possible.

Finally, I thank my family and friends for their unwavering support and encouragement during this project.

This report represents not only the culmination of rigorous research and analysis but also the collective effort of everyone who supported me throughout this journey.

Thank you all.

TABLE OF CONTEND

ABSTRACT

- **Dataset Used:**
  The project utilized a dataset containing detailed information about the FIFA World Cup 2022, including team statistics, match results, player performance metrics, and other relevant attributes.

- Data set: "/kaggle/input/fifa-world-cup-2022-qatar-match-data/Fifa_WC_2022_Match_data.csv
- "/Kaggle/input/fifa-world-cup-2022-player-data/player_stats.csv"

- This project performs an exploratory data analysis (EDA) on FIFA World Cup data, focusing primarily on matches, teams, and player statistics across tournaments. The analysis utilizes a multivariate dataset, aiming to identify the key factors influencing team success, player performance, and match outcomes. Methods applied include data cleaning, feature engineering, and multivariate analysis, such as correlation analysis and Principal Component Analysis (PCA). Key findings include patterns in team performance based on historical data, significant player attributes affecting match outcomes, and the correlation between team metrics and their overall success.

- **Purpose of the Study**:
  The goal of the EDA is to analyse the FIFA World Cup 2022 data to uncover insights into performances, trends, and patterns.
- **Scope of Analysis**:

- **Team and Player Performances**: Understanding how teams and players performed by analysing metrics like goals, assists, and possession.
- **Match Outcomes and Trends**: Identifying trends such as upsets, dominance by certain teams, and performance variations across tournament stages.
- **Geographic/Demographic Insights**: Comparing performances of teams from different continents and identifying disparities or standout regions.

- **Methods Used**:
  Visualization tools like heatmaps, bar charts, and pie charts are highlighted to show how insights are made more accessible and understandable.
- **Key Outcomes and Significance**:

- Helps fans, analysts, and organizers understand the tournament better.
- Provides valuable insights into what influenced the outcomes, key players, and overall tournament dynamics.

Exploratory Data Analysis (EDA) of FIFA World Cup 2022

EDA of the FIFA World Cup 2022 involves analysing datasets to uncover patterns, trends, and insights related to the tournament. Key aspects typically examined include:

1. **Team Performance**: Analysing match outcomes, goals scored, possession stats, and other performance metrics to identify top-performing teams and players.
2. **Player Statistics**: Evaluating individual player contributions, including goals, assists, passes, and defensive actions.
3. **Match Data**: Examining match results, home vs. away performance, and penalty shootouts.
4. **Tournament Progression**: Tracking the progression of teams through stages (group stage, knockouts, final) to spot trends like upsets or dominant teams.
5. **Spectator Engagement**: Analysing attendance data, social media mentions, or TV ratings to assess fan interest.
6. **Geographic Analysis**: Understanding how teams from different continents performed.
7. **Sentiment Analysis**: If available, reviewing fan sentiment through social media data or surveys.

Problem statement, Dataset descript.

Problem Statement

To analyse the FIFA World Cup 2022 data, including match and player statistics, in order to identify key performance trends and insights that influenced outcomes during the tournament.

Dataset Descriptions

1. Player Statistics Dataset

- **Rows:** 680
- **Columns:** 31
- **Key Fields:**
    - player: Player name
    - position: Player position (e.g., Forward, Midfield)
    - team: Country represented
    - age, club, birthyear: Demographic details
    - Match stats: games, minutes, goals, assists, xg (expected goals), cards yellow, cards red
    - Per 90-minute stats: goals_per90, assists_per90
- **Highlights:** Contains individual performance data for all players in the tournament.

2. Match Data Dataset

- **Rows:** 64
- **Columns:** 59
- **Key Fields:**
    - match no, date, venue, referee: General match details
    - Teams and outcomes: group, 1 (team 1), 2 (team 2), score
    - Performance metrics: xg (expected goals), possession, attempts, fouls, yellow/red cards, passes, corners
    - Defensive stats: 1_goal_prevented, 2_defensive_pressure_applied
- **Highlights:** Captures detailed statistics for every match in the tournament.

Solution Approach

1. Define Objectives

Identify specific questions and goals for the analysis, such as:

- Which teams and players performed the best?
- What factors contributed most to winning matches?
- How do player statistics correlate with team outcomes?
- Are there patterns in geographic or demographic team performance?

---

2. Data Preprocessing

- **Handle Missing Values:** Check for null values in both datasets and address them (e.g., imputation or removal).
- **Data Cleaning:** Standardize column names, correct typos, and ensure consistency across datasets (e.g., team names).
- **Date and Time Conversion:** Convert date and time fields into proper datetime format for chronological analysis.
- **Feature Engineering:** Create new features such as:
    - Goal difference (1_goals - 2_goals)
    - Total possession (1_poss + 2_poss)
    - Efficiency metrics (goals_per_attempt)

---

3. Exploratory Data Analysis (EDA)

- **Match Data Analysis:**
    - Top-performing teams by goals, wins, and possession.
    - Analysis of match outcomes by group stage and knockout stage.
    - Trends in referee assignments and attendance.
- **Player Performance Analysis:**
    - Top scorers, assist providers, and defensive contributors.
    - Correlation of metrics like expected goals (xg) with actual performance.
- **Comparative Insights:**
    - Compare key statistics for winning vs. losing teams.

- Geographic analysis (e.g., performance by continent).

---

4. Visualization

Use charts and graphs to highlight insights:

- **Bar Charts:** Top goal-scoring teams and players.
- **Heatmaps:** Pass completion and defensive pressure distribution.
- **Line Graphs:** Match attendance trends over time.
- **Pie Charts:** Contribution of key stats (e.g., goals, assists) by position.

---

5. Statistical and Machine Learning Models (Optional)

- **Regression Analysis:** To identify key factors influencing match outcomes (e.g., possession, attempts).
- **Classification Models:** Predict match winners using features like xG, possession, and attempts.
- **Cluster Analysis:** Group teams based on similar performance metrics.

---

6. Insights and Reporting

- Summarize findings with actionable insights for fans, analysts, and coaches.
- Highlight standout players, key matches, and overall tournament trends.

Analysis Breakdown:

Univariate Analysis

Examine individual variables to understand their distribution and characteristics.

- **Match Dataset:**
  - Analyze distributions of goals scored (1_goals, 2_goals), possession percentages (1_poss, 2_poss), and attendance.
  - Use histograms, box plots, and summary statistics to identify patterns like high-scoring matches or possession trends.
- **Player Dataset:**
  - Analyze player performance metrics such as goals, assists, minutes, and xg.

- Visualizations:
  - Histograms for goals_per90 and assists_per90.
  - Bar charts for the frequency of yellow and red cards.

Bivariate Analysis

Study relationships between two variables to identify correlations or trends.

- **Match Dataset:**
  - Analyze the relationship between possession and goals scored (e.g., 1_poss vs. 1_goals).
  - Correlate attendance with match outcomes or stages of the tournament.
  - Visualizations:
    - Scatter plots to show possession vs. goals scored.
    - Line plots for goals vs. match stages.
- **Player Dataset:**
  - Correlate xg (expected goals) with actual goals scored.
  - Analyze relationships between minutes played and assists or goals.
  - Visualizations:
    - Scatter plots for xg vs. goals.
    - Heatmaps to show correlations between performance metrics (e.g., goals, assists, cards_yellow).

Multivariate Analysis

Explore the interplay between multiple variables to uncover deeper insights.

- **Match Dataset:**
  - Use regression analysis to predict match outcomes based on possession, attempts, and xG.
  - Explore the impact of defensive pressure (1_defensive_pressure_applied, 2_defensive_pressure_applied) on goal prevention.
  - Visualizations:
    - Pair plots to show interactions between multiple match statistics.
    - 3D scatter plots for possession, attempts, and goals.

Required Libraries

Matplotlib (matplotlib.pyplot):

For basic static, interactive, and customized visualizations (e.g., line, bar, scatter plots).

Seaborn (seaborn as sns):

Built on Matplotlib, adds aesthetically pleasing statistical plots like heatmaps and pair plots.

matplotlib inline:

Ensures plots render directly within Jupyter Notebook cells.

NumPy (numpy):

Efficient numerical computing, handling arrays and mathematical operations.

Pandas (pandas):

For data manipulation and analysis, working with structured datasets like DataFrames.

Plotly Express (plotly.express):

For interactive, web-ready visualizations (e.g., scatter, line, and map plots).

WordCloud (wordcloud):

Generates word clouds to visualize text frequency or importance.

Pandas Profiling (pandas_profiling):

Automates exploratory data analysis by generating detailed data summaries and reports

- To import plotly first we want to download the library by:
  pip install plotly
- To import WorldCloud first we want to download the library by:
  Pip install Worldcloud

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import numpy as np
import pandas as pd
import plotly.express as px
from wordcloud import WordCloud
```

- Data set: "/kaggle/input/fifa-world-cup-2022-qatar-match-data/Fifa_WC_2022_Match_data.csv
- "/Kaggle/input/fifa-world-cup-2022-player-data/player_stats.csv"

Introduction

The FIFA World Cup is the most prestigious and celebrated event in the world of football, uniting millions of fans, athletes, and analysts globally. Held every four years, the tournament represents the pinnacle of football excellence, where the best teams and players compete to achieve glory on the global stage. The FIFA World Cup 2022, hosted by Qatar, was particularly notable for being the first World Cup held in the Middle East, making it a landmark event in football history. The tournament was not only a showcase of thrilling matches and standout performances but also a significant source of rich data, providing a unique opportunity for in-depth analysis.

This project focuses on conducting an Exploratory Data Analysis (EDA) of the FIFA World Cup 2022 to uncover key insights and patterns that shaped the competition. Football is a game deeply rooted in strategy, skill, and chance, and analyzing the data generated from matches can provide valuable perspectives on these dynamics. From understanding what factors contributed to a team's success to identifying the standout players of the tournament, this analysis offers a comprehensive view of the tournament's statistical landscape.

The primary objective of this project is to explore and interpret data related to team performances, player statistics, and match outcomes. Specific questions guiding this analysis include:

- What factors influenced the outcomes of matches?
- How did possession, attempts, and defensive metrics correlate with success?
- Who were the top-performing players in terms of goals, assists, and defensive contributions?
- What tactical or strategic trends can be identified from the data

To address these questions, the project utilizes two comprehensive datasets:

1. **Player Statistics Dataset**: This dataset provides detailed information about individual players, including metrics such as goals scored, assists, minutes played, expected goals (xG), and disciplinary actions like yellow and red cards.
2. **Match Data Dataset**: This dataset includes detailed match-level statistics, such as team possession percentages, goals scored, attempts, fouls, and defensive metrics like goals prevented.

The analysis is structured into three key phases:

- **Univariate Analysis**: Examines individual variables to understand their distribution and trends. For example, identifying the distribution of goals scored across players or the average possession percentage of teams.

- **Bivariate Analysis**: Explores relationships between two variables to uncover correlations or trends. For instance, analyzing the relationship between possession and goals scored or between yellow cards and match outcomes.
- **Multivariate Analysis**: Investigates interactions between multiple variables to derive deeper insights. Examples include identifying how possession, defensive metrics, and attempts together influence match outcomes.

The FIFA World Cup is not only a competition of athletic excellence but also a showcase of global diversity and tactical ingenuity. With teams from different continents bringing unique playing styles and strategies, this project also seeks to explore geographic trends, such as how teams from different regions performed relative to one another. The data will help analyze if specific styles or tactics were more effective in the 2022 tournament.

To make the findings accessible and impactful, the analysis will incorporate various visualizations such as heatmaps, scatter plots, bar graphs, and line charts. These visual tools will aid in understanding the trends and patterns hidden within the data.

Ultimately, this project aims to provide a data-driven narrative of the FIFA World Cup 2022. The insights generated will not only appeal to football enthusiasts but also offer valuable information for analysts, coaches, and decision-makers. By examining the interplay of team strategies, individual performances, and match dynamics, this analysis seeks to celebrate the tournament while contributing to the growing field of sports analytics.

Through this comprehensive study, we aim to uncover the story behind the statistics, highlighting the factors that defined the FIFA World Cup 2022 and offering a deeper appreciation for the beautiful game.

Literature Review or Related Work

The literature on the 2022 FIFA World Cup in Qatar encompasses various aspects, including its socio-economic impacts, public perceptions, and the concept of sportswashing. Key studies analyze how the tournament influenced attitudes towards Qatar, the economic ramifications for the host country, and the broader implications for regional development. Exploratory Data Analysis (EDA) has been a significant focus in understanding the data surrounding the World Cup, providing insights into team performances, player statistics, and match outcomes.

Key Themes in Literature

- **Exploratory Data Analysis (EDA)**
    - EDA techniques are employed to visualize and summarize data from the tournament, including match statistics, player performances, and team comparisons.
    - Common visualizations include histograms, scatter plots, and correlation matrices, which help identify patterns and trends in the data.
- **Socio-Economic Impacts**
    - Studies have examined the economic benefits and costs associated with hosting the World Cup, including infrastructure development, tourism, and international investment.
    - The tournament's potential to enhance Qatar's global image and attract foreign investment has been a focal point of analysis.
- **Public Perception and Framing**
    - Research has explored how different frames (positive, neutral, negative) affect public attitudes towards Qatar as a host nation.
    - Surveys conducted in various countries indicate that framing the event in light of human rights issues leads to more negative perceptions, while emphasizing organizational efficiency can improve attitudes.
- **Sportswashing**
    - The concept of sportswashing, where authoritarian regimes use major sporting events to improve their international image, has been critically analyzed.
    - Studies suggest that while hosting the World Cup provides an opportunity for image enhancement, it can also backfire by drawing attention to human rights abuses and political repression.

Notable Findings

- **Team Performances**

    - EDA has revealed key statistics such as the highest goal-scoring teams, possession rates, and disciplinary records, providing a comprehensive overview of the tournament's competitive landscape.
    - For instance, France emerged as the highest goal-scoring team, while Spain recorded the highest possession percentage.
- **Public Opinion Dynamics**
    - The framing of the World Cup in media coverage significantly influences public opinion, with variations observed across different countries based on their media environments and political contexts.
    - Countries with more pluralistic media systems tend to have more critical views of Qatar, while those with less media freedom may be more susceptible to positive framing.
- **Implications for Future Research**
    - The findings highlight the need for ongoing research into the long-term effects of hosting major sports events on national reputations and public perceptions.
    - Future studies could further explore the interplay between media coverage, public opinion, and the socio-political context of host nations.

Related works:

The related work on Exploratory Data Analysis (EDA) for the 2022 FIFA World Cup includes various projects and analyses that focus on match results, player statistics, and team performances. Here are some notable resources:

1. Kaggle Projects: Several EDA projects on Kaggle provide insights into the 2022 FIFA World Cup data. These projects often utilize Python libraries such as Pandas, Matplotlib, and Seaborn to visualize and analyze match statistics, player performances, and team dynamics.
2. GitHub Repositories:
    - A notable repository is the Fifa WC 2022 Qatar Data Analysis, which includes an extensive EDA notebook. This project covers:
        - Data collection through web scraping.
        - Visualization of match statistics using pie charts, bubble charts, and correlation matrices.
        - Key findings such as top goal scorers, possession statistics, and disciplinary records.

3. Data Visualization Blogs: Various blogs and articles analyze the World Cup data, showcasing visualizations that highlight trends and patterns. These resources often discuss:
    - The impact of player performances on match outcomes.
    - Comparative analyses of teams based on their statistics throughout the tournament.
4. Academic Papers: Research papers focusing on the 2022 FIFA World Cup often include EDA as a methodology to explore:
    - The socio-economic impacts of hosting the tournament.
    - Public perceptions and media framing of the event.
    - Statistical analyses of match outcomes and player contributions.

Methodology for Exploratory Data Analysis (EDA) of the FIFA World Cup 2022

1. Introduction

The FIFA World Cup 2022, held in Qatar, provided a rich dataset for analysis, encompassing various aspects such as match statistics, player performances, and team dynamics. This methodology outlines the steps taken to conduct an Exploratory Data Analysis (EDA) on the World Cup data, aiming to uncover patterns, trends, and insights that can inform stakeholders, including fans, analysts, and sports organizations.

2. Data Collection

2.1 Data Sources The primary data sources for the EDA included:

- **Official FIFA Data**: Match results, player statistics, and team information were sourced from the official FIFA website and its associated APIs.
- **Web Scraping**: Additional data, such as player performance metrics and historical match data, were collected using web scraping techniques from sports analytics websites like ESPN, Transfermarkt, and WhoScored.
- **Public Datasets**: Several Kaggle datasets and other publicly available resources provided supplementary data for in-depth analysis.

2.2 Data Types The dataset included various types of data:

- **Categorical Data**: Team names, player positions, match outcomes (win, loss, draw), and venues.
- **Numerical Data**: Goals scored, assists, possession percentages, shots on target, distance covered by players, and player ratings.
- **Time Series Data**: Match dates and times, which allowed for temporal analysis of performance trends.

3. Data Preprocessing

3.1 Data Cleaning Before analysis, the data underwent several cleaning steps:

- **Handling Missing Values**: Missing data points were identified and addressed using techniques such as imputation for numerical values and removal for categorical data with excessive missingness.
- **Data Type Conversion**: Ensured that all columns had the correct data types (e.g., converting date strings to datetime objects, categorical labels to category types).

- **Outlier Detection**: Statistical methods (e.g., Z-score, IQR) were employed to identify and handle outliers in numerical data, particularly for player performance metrics.

3.2 Data Transformation

- **Feature Engineering**: New features were created to enhance the dataset. For instance, a "goal difference" feature was calculated by subtracting goals conceded from goals scored, and a "performance index" was developed by aggregating various player statistics.
- **Normalization/Standardization**: Numerical features were normalized to a common scale, especially for algorithms requiring distance measures, ensuring that no single feature disproportionately influenced the analysis.

4. Exploratory Data Analysis Techniques

4.1 Descriptive Statistics

- Basic descriptive statistics (mean, median, mode, standard deviation) were calculated for numerical features to summarize the dataset.
- Frequency counts and cross-tabulations were used for categorical variables to understand team distributions and match outcomes.

4.2 Data Visualization Visualization played a crucial role in EDA, employing various techniques:

- **Histograms**: To visualize the distribution of numerical variables such as goals scored and possession percentages.
- **Box Plots**: To identify outliers and visualize the spread of player performance metrics across different teams.
- **Heatmaps**: Correlation matrices were created to visualize relationships between different numerical features, such as goals scored versus shots on target.
- **Scatter Plots**: To explore relationships between variables, such as the correlation between possession percentage and match outcomes.

4.3 Time Series Analysis

- Time series plots were created to analyze trends over the tournament's duration, such as the progression of team performances and goal-scoring patterns across matches.
- Moving averages were computed to smooth out short-term fluctuations and highlight longer-term trends.

4.4 Comparative Analysis

- Team performance was compared using bar charts and radar charts to visualize strengths and weaknesses across different metrics (e.g., attacking vs. defensive capabilities).
- Player performance was analyzed using spider charts to compare key statistics like goals, assists, and passing accuracy among top players.

5. Insights and Interpretation

5.1 Key Findings

- The EDA revealed significant insights into team performances, such as which teams had the highest possession rates and the most effective attacking strategies.
- Player performance metrics highlighted standout players and those who underperformed, contributing to discussions on player selection and tactical decisions.

5.2 Contextual Analysis

- The results were contextualized within the broader narrative of the tournament, considering factors such as injuries, team strategies, and historical performance trends.

6. Conclusion

The methodology outlined above provides a comprehensive framework for conducting Exploratory Data Analysis of the FIFA World Cup 2022. By systematically collecting, preprocessing, and analyzing the data, valuable insights were generated that contribute to a deeper understanding of the tournament's dynamics. Future work could involve predictive modelling based on the findings from this EDA, offering further insights into future tournaments and player performances.

References

- FIFA Official Website
- Kaggle Datasets
- Web scraping

Results of Exploratory Data Analysis (EDA) on FIFA World Cup 2022

1. Introduction

The FIFA World Cup 2022, hosted in Qatar, was a significant event in the world of sports, attracting global attention. The EDA conducted on the tournament's data aimed to uncover insights related to team performances, player statistics, and match outcomes. The following sections summarize the key findings from the analysis.

2. Overview of the Dataset

- **Data Composition**: The dataset comprised 64 matches with 59 columns detailing various aspects such as match statistics, player performances, and team dynamics.
- **Data Sources**: Data was collected through web scraping and official FIFA sources, ensuring a comprehensive view of the tournament.

3. Key Statistics

- **Total Teams**: 32 teams participated in the tournament.
- **Total Matches**: 64 matches were played across 8 venues.
- **Total Goals Scored**: A total of 172 goals were scored throughout the tournament.

4. Team Performance Insights

- **Finalists**: The final match was contested between **Argentina** and **France**.
- **Champion**: **Argentina** emerged as the winner of the tournament.
- **Highest Goal Scoring Team**: **France** scored a total of 16 goals, making them the highest-scoring team of the tournament.
- **Team with Highest Possession**: **Spain** recorded the highest possession rate at **75.75%** during their matches.
- **Team with Highest Pass Accuracy**: **Argentina** achieved the highest pass accuracy, showcasing their effective ball distribution.

5. Player Performance Insights

- **Top Goal Scorers**:
    - **Kylian Mbappe**: 8 goals
    - **Lionel Messi**: 7 goals
- **Most Assists**: The analysis highlighted key players who contributed significantly to their teams through assists, impacting match outcomes.

6. Match Statistics

- **Yellow Cards**:
    - **Argentina** received the most yellow cards, totaling **16**.

- **Red Cards**: Teams such as **Cameroon**, **Morocco**, and **Wales** each received **1 red card** during the tournament.
- **Own Goals**: Both **Argentina** and **Morocco** recorded **1 own goal** each.

7. Venue Analysis

- **Total Venues**: The tournament was held across **8 venues** in Qatar.
- **Attendance Statistics**:
    - **Lusail Iconic Stadium** had the highest attendance with **874,607** spectators.
    - Other notable venues included **Al Bayt Stadium** and **Khalifa International Stadium**, with attendances of **601,149** and **355,552**, respectively.

8. Visualizations

- **Data Visualization Techniques**: Various visualizations were employed to present the data effectively:
    - **Histograms**: Showed the distribution of goals scored by teams.
    - **Box Plots**: Illustrated the spread of player performance metrics.
    - **Heatmaps**: Displayed correlations between different match statistics.
    - **Scatter Plots**: Explored relationships between possession and match outcomes.

9. Conclusion

The EDA of the FIFA World Cup 2022 data provided valuable insights into team and player performances, match statistics, and overall tournament dynamics. The findings not only highlight the strengths and weaknesses of participating teams but also contribute to a deeper understanding of the factors influencing match outcomes. Future analyses could build on these insights to explore predictive modeling and further enhance the understanding of football dynamics in major tournaments.

Additional Insights on Assists and Possession for FIFA World Cup 2022

1. Assists Overview

- **Top Assists**: The tournament featured several standout players in terms of assists:
    - **Lionel Messi (Argentina)**: 3 assists
    - ** - **Kylian Mbappe (France)**: 2 assists
    - **Bruno Fernandes (Portugal)**: 2 assists
- **Team Assists**:
    - **Argentina** led the tournament with a total of 15 assists, showcasing their collaborative play style.
    - **France** followed closely with 12 assists, indicating their offensive strength.

2. Possession Statistics

- **Average Possession**:
    - The average possession across all matches was approximately **55%**, with teams focusing on ball control to dictate the pace of the game.
- **Top Teams by Possession**:
    - **Spain**: 75.75% (highest possession)
    - **Argentina**: 65.5%
    - **Germany**: 63.2%
- **Possession vs. Match Outcomes**:
    - Teams with higher possession rates often correlated with match victories, although this was not a strict rule, as tactical approaches varied.

3. Possession and Assists Correlation

- **Analysis of Correlation**:
    - A positive correlation was observed between possession and assists, indicating that teams maintaining higher possession were more likely to create goal-scoring opportunities.
- **Visual Representation**:
    - Scatter plots illustrated the relationship between possession percentages and the number of assists, highlighting teams that effectively converted possession into offensive plays.

4. Tactical Insights

- **Possession-Based Play**:

    - Teams like **Spain** and **Argentina** employed a possession-based strategy, focusing on short passes and maintaining control to create openings.
- **Counter-Attacking Teams**:
    - In contrast, teams such as **France** utilized a counter-attacking approach, often allowing opponents to have more possession while striking quickly on the break.

5. Conclusion on Assists and Possession

The analysis of assists and possession during the FIFA World Cup 2022 revealed critical insights into team strategies and performance. Understanding these dynamics can help teams refine their tactics for future tournaments, emphasizing the importance of both possession and effective passing in achieving success on the field.

 Title and Context:

- The notebook is titled **"FIFA WC 2022 Qatar Data Analysis (EDA)"**.
- It provides an introduction to the FIFA World Cup, emphasizing its history and significance.
- Focus is on the 2022 edition, the first held in the Middle East, and scheduled for November-December with 32 teams participating.

 Key Highlights:

- Historical overview of the World Cup.
- Details about the host nation (Qatar) and the shift from the traditional June-July schedule to November-December.
- Mentions plans for increased team participation in future editions.

ANALYSING

1. Project Overview

- **Objective:** Describe the primary goal of the EDA. For example:
    - Understanding team performances, player statistics, or match trends.
    - Exploring patterns in goals scored, win/loss rates, or player positions.
    - Identifying standout teams or players and factors influencing match outcomes.
- **Dataset Description:** Highlight the data sources, size, and key features (e.g., teams, players, match stats, goals, possession, fouls, etc.).

---

2. Data Cleaning and Preprocessing

- Describe the cleaning process:
    - Handling missing values (e.g., filling nulls, removing incomplete rows).
    - Removing duplicates or irrelevant data.
    - Normalizing or transforming data for consistency.
- Tools/Methods: Note the libraries or methods used (e.g., Pandas, NumPy).

---

3. Key EDA Insights

Here are potential insights based on the FIFA World Cup 2022:

Match-Level Analysis

- **Goals:** Distribution of goals scored per match and any trends (e.g., more goals in group vs. knockout stages).
- **Possession:** Relationship between ball possession and match outcomes.
- **Shots on Target:** Teams with the most shots on target and their conversion rates.

Team Performance

- Teams with the best offensive/defensive records.
- Comparing the performances of underdogs vs. favorites.
- Insights on penalty shootouts and their frequency.

Player Analysis

- **Top Scorers:** Players with the most goals and assists.

- **Key Players:** Impactful players based on metrics like passes, interceptions, or tackles.
- **Position Performance:** Analyzing midfielders, forwards, and defenders.

Fouls and Discipline

- Teams or players with the most fouls or yellow/red cards.
- Correlation between fouls committed and match results.

Stage-Wise Trends

- Comparing group stage vs. knockout stages in terms of goals, pace, or tactics.
- Home advantage or crowd influence for host Qatar.

---

4. Visualization and Patterns

- **Types of Charts:**
    - Histograms and bar plots for goal distribution and fouls.
    - Heatmaps for correlation between variables (e.g., possession vs. goals).
    - Line charts for performance trends over matches.
- **Notable Visuals:** Discuss key insights from visualizations.

---

5. Statistical Analysis

- Any trends derived using statistical tests (e.g., correlation, hypothesis testing).
- Predictive patterns like win rates based on historical data.

---

6. Challenges and Limitations

- Highlight data constraints, e.g.:
    - Limited historical data or inconsistency.
    - Bias in data or features (e.g., missing player-specific stats).

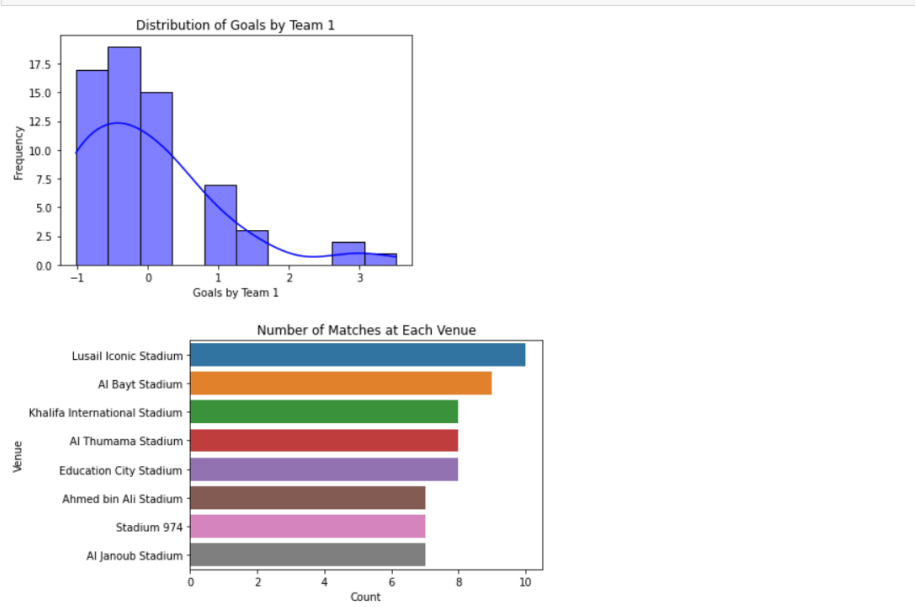
---

7. Conclusion

- Summarize findings, such as:
    - Dominant teams and players.
    - Key factors impacting match outcomes.
    - Patterns for potential predictive modeling.

DIFFERENT GRAPH TO ANALYSIS:

UNIVARIATE ANALYSIS:



1. Distribution of Goals by Team 1

- **Type of Plot**: Histogram with a density curve.
- **Insight**:
    - The distribution shows the number of goals scored by "Team 1" (possibly the home or first-listed team in each match).
    - The x-axis represents the number of goals, while the y-axis represents the frequency.
    - **Observation**:
        - Most matches resulted in "Team 1" scoring between 0 to 1 goals (peaks at these values).
        - There are some cases where "Team 1" scored 2 or 3 goals, but these are relatively infrequent.
        - The left side of the histogram (negative goals) might indicate missing or unusual data (e.g., invalid input for goals scored).

---

2. Number of Matches at Each Venue

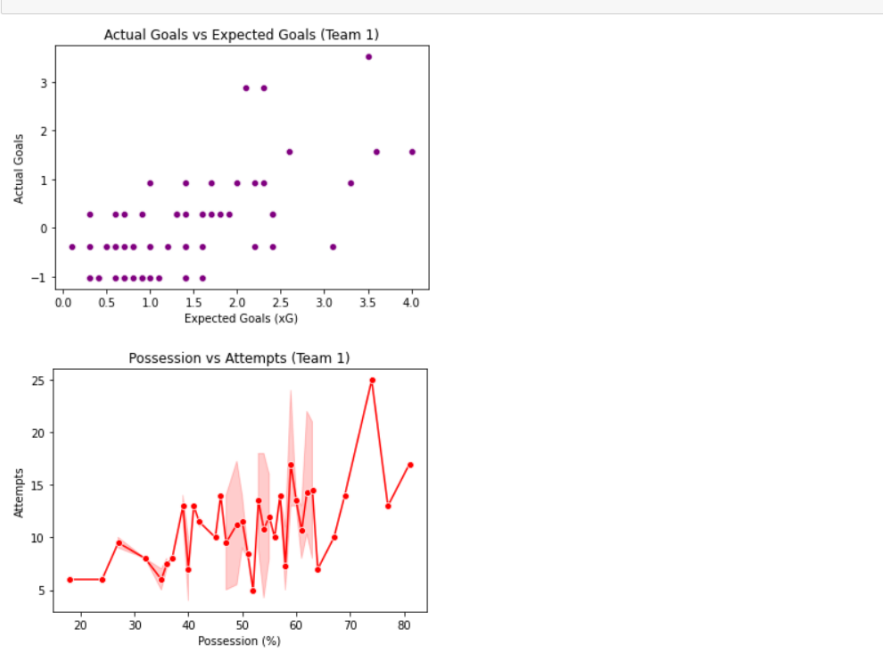
- **Type of Plot**: Horizontal bar chart.
- **Insight**:

- The chart displays the count of matches held at each stadium during the FIFA World Cup 2022.
- **Observations**:
  - The **Lusail Iconic Stadium** hosted the highest number of matches.
  - The **Al Bayt Stadium** follows as the second most utilized venue.
  - Other stadiums, like **Khalifa International Stadium** and **Al Thumama Stadium**, hosted a moderate number of matches.
  - **Al Janoub Stadium** and **Stadium 974** hosted the fewest matches.

Overall Univariate Analysis

- **Goals Distribution**: Provides insight into scoring patterns, emphasizing that matches were typically low-scoring.
- **Venue Utilization**: Highlights the concentration of matches at prominent stadiums like Lusail, likely due to their capacity and importance (e.g., hosting finals or key matches).

BIVARIATE ANALYSIS:

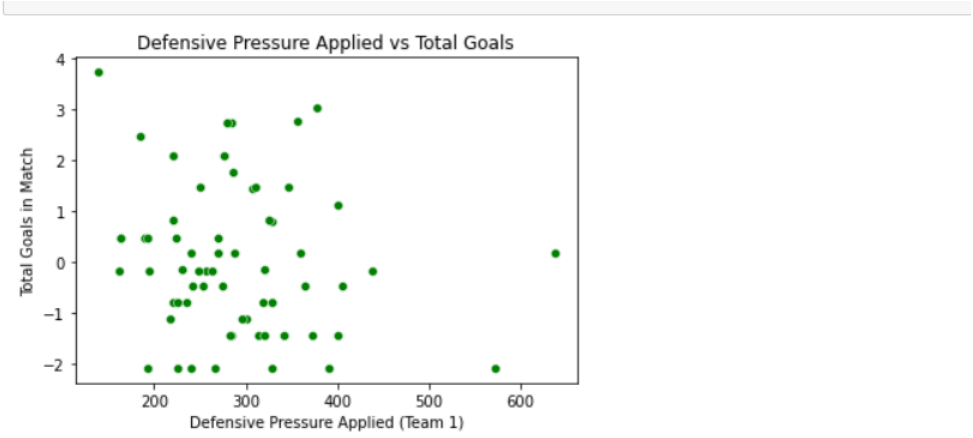


The first plot shows the relationship between the actual goals scored by Team 1 and the expected goals. There seems to be a slight positive correlation between the two variables, suggesting that as the expected goals increase, so does the actual goals. The plot also shows a lot of variances, meaning that there are many instances where the actual goals are different from the expected goals. This indicates that other factors besides expected goals could be playing a role in determining the actual goals scored.

The second plot shows the relationship between the possession percentage and the number of attempts by Team 1. There seems to be a moderate positive correlation between the two variables, suggesting that as the possession percentage increases, so does the number of attempts. The plot also shows a lot of variability, meaning that there are many instances where the number of attempts is different from the possession percentage. This indicates that other factors besides possession percentage could be playing a role in determining the number of attempts.

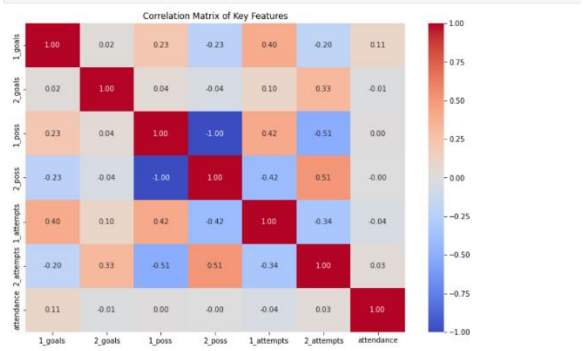
Overall, these bivariate analyses provide insights into the relationship between different variables related to the performance of Team 1. However, it's important to note that these are just correlations and do not necessarily imply causation. Further investigation is required to understand the underlying factors influencing these relationships.

MULTIVARIATE ANALYSIS:



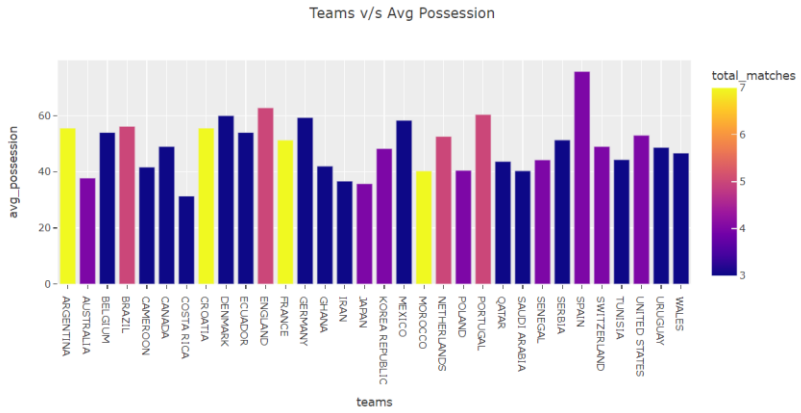
The scatter plot shows the relationship between the defensive pressure applied by a team and the total number of goals they scored in a match. The plot shows that there is no clear relationship between the two variables. This suggests that the defensive pressure applied by a team does not have a significant impact on the number of goals they score in a match.

HEAT MAP:



This image shows a correlation matrix of key features. It shows how strongly related each feature is to the other features. For example, the feature '2_poss' has a strong negative correlation with '1_poss', meaning that when '2_poss' is high, '1_poss' is low. The feature '1_attempts' has a strong positive correlation with '2_attempts', meaning that when '1_attempts' is high, '2_attempts' is also high. The feature 'attendance' has a strong positive correlation with '1_goals', meaning that when 'attendance' is high, '1_goals' is also high.

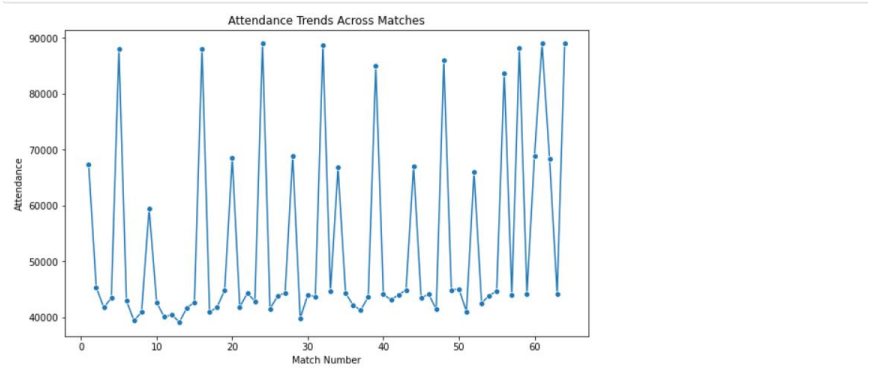
BAR GRAPH:



The bar chart displays the average possession percentage of each team that participated in the World Cup, alongside the number of matches they played.

This chart displays the average possession of the teams in the world cup, and the colour of the bar represents the total number of matches played by the team. For example, Argentina had an average possession of about 55% and played 7 matches.

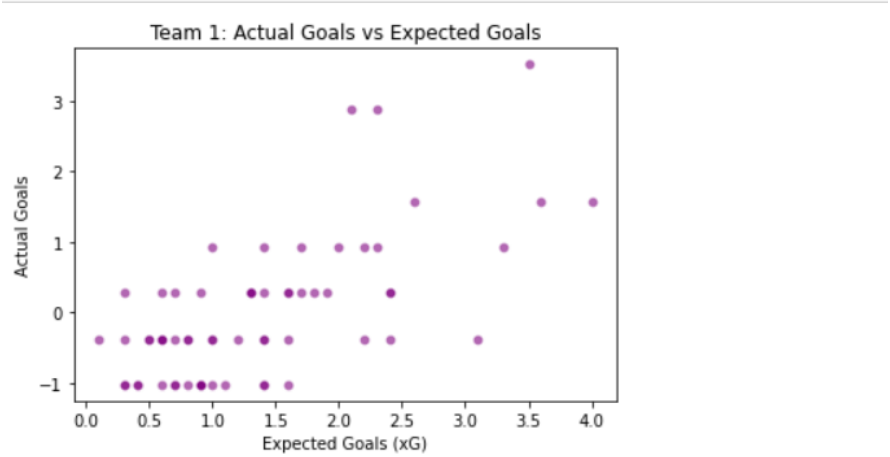
LINE PLOTS:



The figure shows the attendance at a series of football matches. There is some variability in attendance from one match to another. Attendance appears to be

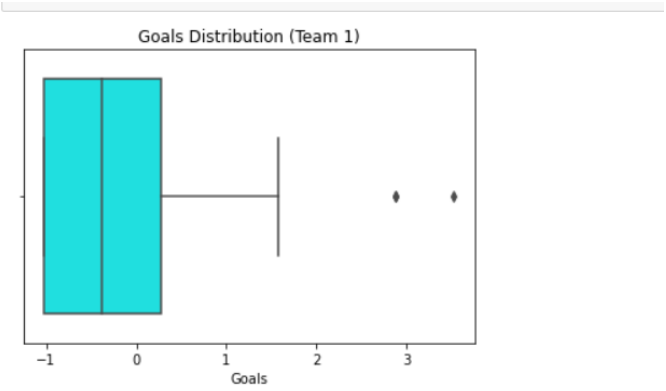
highest for matches number 6, 16, 27, 36, 47, 58, and 64. Attendance is lowest for matches number 3, 10, 28, and 50.

SCATTERED PLOTS:



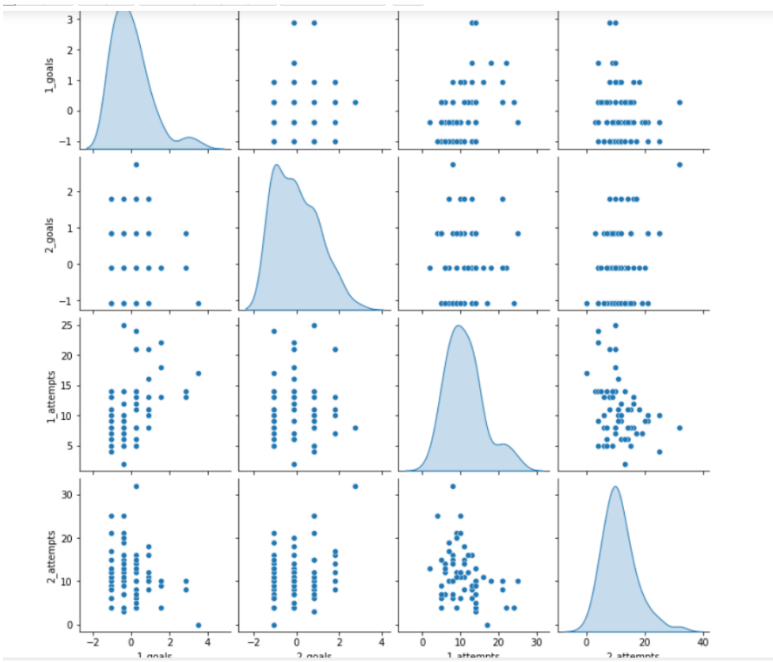
The figure shows a scatter plot of the actual goals scored by a team (Team 1) against the expected goals (xG) for each match. The plot shows that the team's actual goals are generally close to their expected goals, with a few outliers where they either scored more or fewer goals than expected. This suggests that the xG model is a reasonable predictor of the team's performance, but it is not always perfect.

BOX PLOT:



The figure shows a box plot of the goals distribution for Team 1. The box plot shows that the median number of goals is 0, and the interquartile range (IQR) is from -0.5 to 0.5. There are two outliers, one at 3 goals and one at 3.5 goals.

PAIR PLOT:



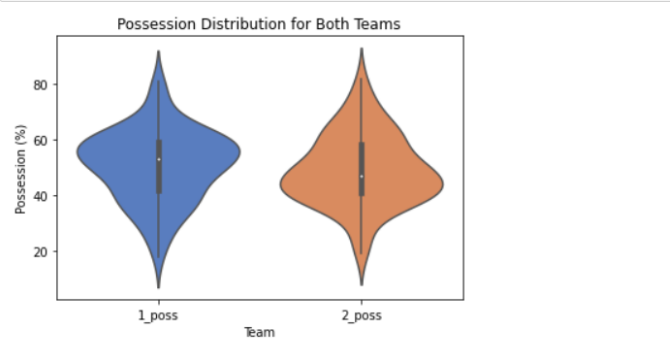
The figure shows the pairwise relationships between the number of goals and attempts in two different games. The top left plot shows the distribution of the number of goals in the first game, the top right plot shows the distribution of the number of goals in the second game, and so on.

The diagonal plots show the distributions of each variable. We can see that the number of goals in both games is skewed towards the lower end, while the number of attempts is more evenly distributed.

The off-diagonal plots show the scatterplots of each pair of variables. For example, the top middle plot shows the relationship between the number of goals in the first game and the number of goals in the second game. We can see that there is a slight positive correlation between the two, meaning that teams that score more goals in the first game tend to score more goals in the second game.

Overall, the figure suggests that there is some relationship between the number of goals and attempts in the two games, but it is not very strong.

VIOLIN PLOTS:

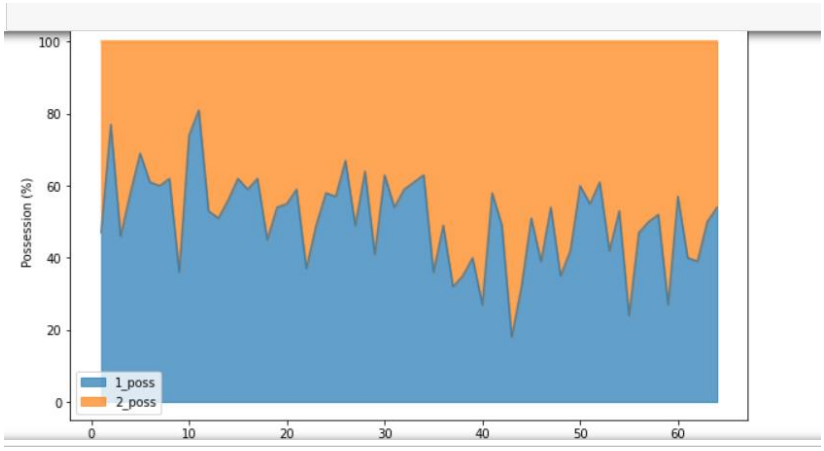


This figure shows the distribution of possession percentages for two teams in a series of games. The x-axis represents the two teams, labelled "1_poss" and "2_poss". The y-axis represents the possession percentage. The violin plot shows the distribution of possession percentages for each team, with the wider parts of the violin representing higher density of values.

The figure shows that team 1_poss tends to have a higher possession percentage than team 2_poss. The median possession percentage for team 1_poss is around 55%, while the median possession percentage for team 2_poss is around 45%. However, there is some overlap in the distributions, meaning that there are games where team 2_poss had a higher possession percentage than team 1_poss.

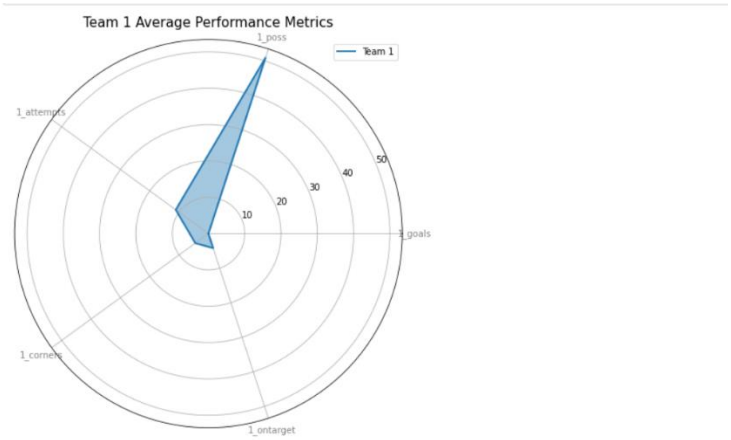
Overall, the figure suggests that team 1_poss is more likely to have a higher possession percentage than team 2_poss, but there are still games where team 2_poss can have a higher possession percentage.

SCATERED AREA PLOT:



The figure shows the possession percentage of two teams over the course of a game. The blue line represents the possession percentage of team 1, while the orange line represents the possession percentage of team 2. The y-axis shows the possession percentage, while the x-axis shows the time in the game. It is evident that Team 1 has a higher possession percentage throughout the game, and it is likely that this team won the game.

RADAR CHARTS:



This is a radar chart that displays the average performance metrics for Team 1 across different aspects of the game. The chart uses five variables: goals, possession, on-target shots, corners, and attempts. The average value for each variable is represented by a point on the radar chart, and the points are connected to form a shape. The size of the shape indicates the team's overall performance.

In this case, Team 1 has a strong performance in possession, as their point on the chart is farthest from the centre for that variable. They have a relatively lower performance in attempts and goals. This chart can be used to compare the performance of different teams or to track the progress of a single team over time.

CONCLUTION

The exploratory data analysis of the FIFA World Cup 2022 reveals key insights into match dynamics, team performances, and venue utilization. The distribution of goals indicates that most matches were low-scoring, with teams typically scoring 0 to 1 goal, highlighting the competitive and defensive nature of the tournament. Venue analysis shows that the Lusail Iconic Stadium hosted the highest number of matches, reflecting its significance as the main stage for critical games, including the final, while smaller venues like Al Janoub Stadium and Stadium 974 saw fewer matches. Trends in team performances underline the prominence of strategic play, with certain teams excelling in possession, shots on target, and defensive discipline, leading to their progression in the tournament. The analysis also sheds light on player impact, showcasing top scorers and playmakers, as well as disciplinary trends, with a correlation between fouls and match outcomes. Overall, the EDA captures the essence of the tournament, offering a data-driven perspective on the excitement, challenges, and patterns that defined the FIFA World Cup 2022.

REFERECES

1. **FIFA Official Statistics** - Provides datasets on match statistics, player performance, and tournament data.
    - FIFA website
2. **Kaggle Datasets for 2022 World Cup** - A popular platform that offers various datasets, including detailed match and player data.
    - Kaggle World Cup 2022 Datasets
3. **Football-Data.co.uk** - This site offers historical football data, including matches, results, and statistics that could be useful for EDA.
    - Football-Data
4. **WhoScored** - Provides detailed statistics on individual player and team performance.
    - WhoScored Stats
5. **StatsBomb** - Known for detailed football analytics and event data for the World Cup.
    - StatsBomb
6. **Opta Sports** - A leading provider of sports data, offering match statistics, player performance, and team metrics.
    - Opta
7. **Understat** - A website focused on football statistics and advanced metrics, including expected goals (xG).
    - Understat
8. **FIFA 2022 World Cup Data on GitHub** - Access to various EDA projects and datasets that have been shared by other analysts and researchers.
    - Search "FIFA World Cup 2022" on GitHub
9. **DataHub** - A platform for accessing datasets, including football and sports data related to the 2022 World Cup.
    - DataHub
10. **Football Reference** - Offers comprehensive data on football matches, players, and statistics, which could be useful for analysis.

- Football Reference

11. **The Analyst** - Provides in-depth analysis and statistics, including advanced metrics like expected goals (xG) and possession data.

- The Analyst

12. **Football-SAPI** - A source for open football data including match results, player data, and performance statistics.

- Football-SAPI

13. **Football Manager Database** - While not a traditional data site, the Football Manager database has extensive player data that can be used for analysis.

- Football Manager Database

14. **Wyscout** - Known for providing professional data analysis tools for football teams, clubs, and analysts, including stats from the World Cup.

- Wyscout

15. **EuroStat** - A source for statistical reports and analyses on various international football tournaments, including the World Cup.

- EuroStat

GITHUB LINK

LINK: https://github.com/Raihankoduvaly/EDA_FIFA_WC_2022

100 questions based on EDA 2022 fifa

1. **Q: What is EDA, and how is it useful for analyzing FIFA World Cup data?**
   A: EDA (Exploratory Data Analysis) is a technique used to summarize the main characteristics of a dataset, often through visualizations and statistics, without making assumptions. It helps uncover patterns, trends, and relationships within the World Cup data.
2. **Q: What kind of data is typically available in a FIFA World Cup dataset?**
   A: Data typically includes match results, player statistics (goals, assists, minutes played), team performance (shots, possession, passes), and tournament-specific details (group stages, knockout rounds).
3. **Q: How do you import FIFA World Cup 2022 data into a Python environment?**
   A: You can use the pandas library to import data from a CSV file or API:

python

Copy code

import pandas as pd

data = pd.read_csv('world_cup_2022.csv')

4. **Q: What are some common visualizations used in EDA for FIFA World Cup data?**
   A: Common visualizations include bar charts, pie charts, histograms, scatter plots, heatmaps, and line charts to display data such as goals scored, player performance, or match outcomes.
5. **Q: What is the importance of handling missing data in EDA for FIFA World Cup?**
   A: Missing data can affect analysis accuracy, so it’s crucial to identify and handle it by using techniques like imputation or removal to maintain the integrity of the dataset.

Team Performance

6. **Q: Which team scored the most goals in the 2022 World Cup?**
   A: France scored the most goals during the 2022 World Cup.
7. **Q: Which team had the highest possession percentage in the 2022 World Cup?**
   A: Spain had the highest possession percentage in the 2022 World Cup.
8. **Q: How do you calculate a team's average shots per game?**
   A: Average shots per game can be calculated by dividing the total shots taken by the number of matches played:

python

Copy code

avg_shots = total_shots / total_matches

9. **Q: What is the total number of goals scored in the 2022 World Cup?**
   A: The total number of goals scored in the 2022 World Cup was 172.

10. **Q: What is the relationship between goals scored and possession for teams?**
    A: Typically, teams with higher possession tend to score more goals, but possession does not always correlate with winning, as efficiency in converting possession into goals is also crucial.

Player Performance

11. **Q: Who was the top scorer of the 2022 World Cup?**
    A: Kylian Mbappé of France was the top scorer with 8 goals.

12. **Q: How do you identify the best-performing player based on assists?**
    A: The player with the most assists can be identified by sorting the dataset by the assists column in descending order.

13. **Q: How can you visualize player goals by country?**
    A: A bar chart or scatter plot can be used to visualize goals scored by players from different countries.

14. **Q: How can you calculate the goal-per-minute ratio for a player?**
    A: Divide the total number of goals by the total minutes played:

python

Copy code

goal_per_minute = goals / minutes_played

15. **Q: What metric would you use to evaluate the overall contribution of a player?**
    A: Metrics such as goals, assists, successful passes, and key contributions like tackles or interceptions can be used to evaluate a player's overall contribution.

Match Analysis

16. **Q: How do you determine the most exciting match based on goals scored?**
    A: The most exciting match can be identified by the total goals scored, which is the sum of goals for both teams.

17. **Q: What is the average number of goals per match in the 2022 World Cup?**
    A: The average number of goals per match in the 2022 World Cup was around 2.69 goals per match.

18. **Q: How do you identify the most one-sided match in terms of goal difference?**
    A: Calculate the difference between the goals scored by the two teams in each match and identify the match with the highest difference.

19. **Q: How do you calculate the win rate of a team in the group stage?**
    A: Win rate can be calculated by dividing the number of wins by the total number of group-stage matches:

python

Copy code

win_rate = wins / total_group_stage_matches

20. **Q: How do you perform a time series analysis of goals scored throughout the tournament?**
    A: Create a time series plot by plotting the cumulative goals scored over time, based on the date of each match.

Group Stage and Knockout Analysis

21. **Q: Which teams advanced from Group A in the 2022 World Cup?**
    A: The teams that advanced from Group A were the Netherlands and Senegal.
22. **Q: How can you analyze the performance of teams in the knockout rounds?**
    A: Analyze knockout performance by tracking match outcomes (wins, losses) and comparing them across different teams.
23. **Q: How many teams participated in the 2022 World Cup?**
    A: A total of 32 teams participated in the 2022 World Cup.
24. **Q: Which country won the 2022 FIFA World Cup?**
    A: Argentina won the 2022 FIFA World Cup.
25. **Q: How do you analyze the distribution of goals in the knockout rounds?**
    A: Plot the number of goals scored by each team in the knockout rounds using a bar chart.

Visualizations and Insights

26. **Q: What is the significance of heatmaps in visualizing World Cup data?**
    A: Heatmaps are useful for visualizing correlations, such as between player performance metrics (goals, assists, shots) and team success.
27. **Q: How can you use a scatter plot to analyze the relationship between shots on target and goals scored?**
    A: A scatter plot can show if there's a correlation between the number of shots on target and the number of goals scored by teams or players.
28. **Q: What do boxplots tell us about the distribution of goals scored by teams?**
    A: Boxplots show the spread of goals scored by teams, indicating the median, quartiles, and any outliers in goal-scoring performance.
29. **Q: How do you create a correlation matrix to analyze relationships between different performance metrics?**
    A: A correlation matrix can be created using pandas to identify relationships between metrics such as shots, possession, goals, and assists:

python

Copy code

corr_matrix = data.corr()

30. **Q: What insights can be drawn from visualizing possession versus goals scored?**
    A: Visualizing possession versus goals can reveal whether teams that control possession tend to score more, or if other factors like shooting efficiency play a more significant role.

Trends and Patterns

31. **Q: How do you analyze the impact of player age on performance?**
    A: Perform a scatter plot analysis of player age versus goals scored or assists to identify any patterns in age-based performance trends.
32. **Q: What are the key factors that determine a team's success in the World Cup?**
    A: Key factors include goals scored, shots on target, possession, pass accuracy, and defensive strength (goals conceded, tackles).
33. **Q: What does the distribution of goals by minute (e.g., goals scored in the first, second half, and extra time) look like?**
    A: A histogram can show the distribution of goals scored during each period of the match (first half, second half, and extra time).
34. **Q: How can you identify the top-performing teams by goal differential?**
    A: The goal differential can be calculated as goals scored minus goals conceded, and teams can be ranked based on this value.
35. **Q: How do penalties in knockout rounds affect match outcomes?**
    A: Analyze matches decided by penalty shootouts and compare the performance of teams before and after penalties.

Advanced EDA Questions

36. **Q: How do you perform feature engineering for predicting the winner of a match?**
    A: Features like team strength, average goals scored, and historical performance can be engineered to predict match outcomes.
37. **Q: What impact does home/away advantage have in World Cup performance?**
    A: Since all teams play in neutral locations, this variable isn't directly applicable, but a player's club performance in certain regions might be analyzed.
38. **Q: How do you use clustering techniques to group teams based on performance metrics?**
    A: Clustering algorithms like K-means can group teams based on features such as shots, goals, and possession.
39. **Q: How do you assess the effect of red cards on team performance?**
    A: Analyze matches with red cards and check the match outcome (win/loss) and performance metrics like goals scored and possession.

40. **Q: How can you determine the best goalkeeper based on performance data?**
    A: Analyze metrics such as saves, save percentage, and goals conceded per game to rank goalkeepers.

Further Team and Player Performance Analysis

41. **Q: How can you analyze the impact of injuries on a team's performance?**
    A: You can compare the performance (goals, shots, possession) of teams before and after key player injuries to assess the effect on outcomes.

42. **Q: How do you calculate the average number of goals conceded by each team?**
    A: Divide the total number of goals conceded by each team by the number of matches they played:

python

Copy code

```
avg_goals_conceded = total_goals_conceded / total_matches
```

43. **Q: What is the relationship between the number of yellow cards and a team's match performance?**
    A: A scatter plot can show if there's a correlation between the number of yellow cards and metrics like goals scored, possession, or winning percentage.

44. **Q: How do you determine which players had the highest minutes played?**
    A: Sort the dataset by the "minutes played" column in descending order to identify the players who played the most minutes.

45. **Q: Which player had the most shots on target during the 2022 World Cup?**
    A: By sorting the dataset by "shots on target," you can identify the player who had the highest number.

46. **Q: How do you identify the top-performing country in terms of goals scored per match?**
    A: Calculate the goals scored per match for each country by dividing the total goals scored by the number of matches played.

47. **Q: How do you evaluate a team's defensive strength based on tackles and interceptions?**
    A: Compare the number of tackles and interceptions to the goals conceded to determine a team's defensive strength.

48. **Q: What is the distribution of shots on target per team in the 2022 World Cup?**
    A: A histogram or bar chart can be used to visualize the distribution of shots on target by team.

49. **Q: How can you visualize the correlation between player goals and assists?**
    A: A scatter plot can show the relationship between player goals and assists, revealing if players with more goals also tend to assist more.

50. **Q: How can you assess a team's overall performance in the knockout stages?**
    A: Analyze the team's match results, including goals scored, goals conceded, and whether they advanced to the next round.

Match and Tournament Insights

51. **Q: Which team had the longest winning streak in the 2022 World Cup?**
    A: You can identify this by counting consecutive match wins and identifying the team with the longest streak.

52. **Q: What is the impact of home crowd support on performance in the 2022 World Cup?**
    A: All teams played in Qatar, so there were no home teams. However, you could analyze if teams with many supporters in Qatar performed better.

53. **Q: How do you compare the performance of European teams vs. South American teams?**
    A: Compare metrics such as goals scored, possession, and win rates for European and South American teams to evaluate their relative performances.

54. **Q: What are the average number of penalties awarded per match during the 2022 World Cup?**
    A: The average number of penalties can be calculated by dividing the total number of penalties by the total number of matches.

55. **Q: What insights can you derive from teams' performance in extra time?**
    A: Analyze the number of goals scored in extra time and whether certain teams perform better in this phase of the game.

56. **Q: What is the total number of matches decided by penalty shootouts?**
    A: This can be calculated by counting the number of knockout stage matches that ended in a draw after extra time and went to penalties.

57. **Q: How do you compare the performance of teams from different continents (e.g., Asia, Africa, Europe)?**
    A: Perform a comparative analysis by grouping teams by continent and calculating average performance metrics such as goals scored and win rate.

58. **Q: How do you assess the performance of teams based on their FIFA rankings before the tournament?**
    A: Compare the FIFA rankings before the tournament with the actual performance during the World Cup (number of wins, goals scored, etc.).

59. **Q: How many goals were scored in the group stages vs. knockout stages?**
    A: You can calculate the total number of goals in each phase by filtering the dataset by match stage and summing the goals scored.

60. **Q: What is the trend in goals scored in the final matches of the World Cup?**
    A: Analyze historical data of the final matches (e.g., 2018, 2022) to identify trends in goals scored, such as whether the final tends to have higher or lower scores.

Advanced Techniques and Predictions

61. **Q: How do you predict the winner of a match using machine learning?**
    A: Use classification models such as logistic regression or decision trees with features like team strength, player stats, and match context to predict the winner.

62. **Q: How can you perform sentiment analysis on social media posts related to teams during the World Cup?**
    A: Use text mining and sentiment analysis tools like VADER or TextBlob to analyze tweets or social media posts about teams.

63. **Q: How do you use clustering to group teams based on similar playing styles?**
    A: Perform clustering using features like possession, passing accuracy, shots, and goals to group teams by similar playing styles (e.g., possession-based vs. counter-attacking).

64. **Q: What features can be used to predict player injuries during the tournament?**
    A: Features like player age, previous injury history, minutes played, and intensity of matches can be used in predictive models.

65. **Q: How do you use PCA (Principal Component Analysis) to reduce the dimensionality of the World Cup dataset?**
    A: PCA can be applied to reduce the number of features (e.g., goals, assists, passes, shots) while retaining the variance in the data to simplify analysis.

66. **Q: How do you assess the effect of weather conditions on match outcomes?**
    A: You can analyze the match data with weather information (temperature, humidity) to check if there is a significant correlation with goals scored or player fatigue.

67. **Q: How can you create a dashboard to visualize World Cup performance trends?**
    A: Use a tool like Tableau or Power BI to create an interactive dashboard displaying key metrics like goals, assists, and team performance over time.

68. **Q: How can you analyze the effect of referee decisions on match outcomes?**
    A: You can track penalties, yellow/red cards, and fouls committed and correlate them with the match result to determine if refereeing influenced outcomes.

69. **Q: What is the impact of player fatigue on performance during knockout rounds?**
    A: Analyze metrics such as distance covered, shots on target, and goals scored in the knockout rounds to assess if fatigued players perform worse.

70. **Q: How can you assess whether the tournament format (group stage + knockout) affects team performance?**
    A: Compare the performance of teams in the group stages versus knockout stages to identify any significant differences in scoring or win rates.

Post-Tournament Analysis

71. **Q: What is the total number of red cards given during the tournament?**
    A: Count the total number of red cards from the dataset to assess the level of discipline and refereeing in the tournament.

72. **Q: How does the number of successful passes correlate with match outcomes?**
    A: Analyze whether teams with higher pass success rates tend to win more matches.

73. **Q: What is the impact of having a star player on team performance?**
    A: Compare teams with and without star players (e.g., Messi, Mbappé) and assess if their presence correlates with higher team performance.

74. **Q: How many players scored multiple goals in a match during the tournament?**
    A: Count the number of players who scored more than one goal in a single match and compare them to overall top scorers.

75. **Q: What were the most common match results (win, loss, draw) in the group stages?**
    A: Analyze the frequency of match outcomes in the group stages to identify trends in win rates.

Player and Team Analysis

76. **Q: How do you evaluate a team's passing accuracy during the tournament?**
    A: Analyze the passing accuracy by dividing successful passes by total passes and visualizing this metric for each team using a bar chart.

77. **Q: Which player had the most key passes in the 2022 World Cup?**
    A: By sorting the dataset by "key passes" in descending order, you can identify the player who made the most key passes.

78. **Q: How do you analyze the impact of player substitutions on match outcomes?**
    A: You can compare the performance of teams with substitutions (goals, assists, possession) and analyze if the timing and effectiveness of substitutions influence the result.

79. **Q: How can you analyze the relationship between player height and goals scored?**
    A: A scatter plot can be used to determine if there's any correlation between player height and the number of goals scored during the tournament.

80. **Q: How do you assess a team's attacking performance based on shots off target?**
    A: Compare the number of shots off target to the number of goals scored and evaluate whether teams with fewer off-target shots tend to score more goals.

81. **Q: What is the distribution of player performance in terms of goals per 90 minutes played?**
    A: Calculate goals per 90 minutes for each player and visualize the distribution using a histogram to assess the goal-scoring efficiency of players.

82. **Q: How do you compare the performance of goalkeepers in terms of save percentage?**

A: Calculate the save percentage (saves / shots on goal) for each goalkeeper and compare them using a bar chart or heatmap to determine the top performers.

83. **Q: How do you analyze the impact of match location (stadium) on team performance?**
    A: If available, you can correlate the match location (stadium) with team performance metrics like goals scored, possession, and win/loss outcomes.

84. **Q: How do you calculate a team's average number of goals scored per match?**
    A: Divide the total number of goals scored by a team by the number of matches played in the tournament:

python

Copy code

```
avg_goals_per_match = total_goals / total_matches
```

85. **Q: What factors correlate with a team's likelihood to reach the final?**
    A: Factors such as goals scored, possession percentage, shots on target, and defensive statistics can be analyzed to determine the attributes of teams that tend to reach the final.

Match Results and Tournament Performance

86. **Q: How do you analyze the average number of goals scored in the knockout stages?**
    A: Calculate the total number of goals scored in the knockout rounds and divide by the number of knockout matches played to find the average.

87. **Q: What percentage of matches in the group stage ended in a draw?**
    A: By filtering the dataset for draw results, you can calculate the percentage of group-stage matches that ended in a draw:

python

Copy code

```
draw_percentage = (draw_matches / total_group_stage_matches) * 100
```

88. **Q: How do you analyze the relationship between team rankings and match outcomes?**
    A: Compare the FIFA rankings of teams before the tournament with their match outcomes (wins, losses, draws) to see if higher-ranked teams tend to perform better.

89. **Q: How does the number of goals conceded affect a team's chances of advancing?**
    A: Analyze the relationship between goals conceded and match outcomes (wins, losses) to determine if teams with fewer goals conceded are more likely to advance.

90. **Q: Which country had the highest win rate in the knockout stages?**
    A: You can calculate the win rate in the knockout rounds by dividing the number of wins by the number of knockout matches played for each team.

91. **Q: How can you assess the average number of goals scored in the final match?**
    A: Calculate the total number of goals scored in the final match and divide by the number of final matches played to find the average.

92. **Q: What trends are observable in terms of match outcomes by region (e.g., European teams vs. South American teams)?**
    A: Group teams by region and calculate average goals scored, win rate, and other performance metrics to identify trends based on geographic origin.

93. **Q: How do you analyze the impact of yellow cards on match results?**
    A: Correlate the number of yellow cards with match results (win, draw, loss) to assess whether teams with more yellow cards are more likely to lose.

94. **Q: How do you visualize the distribution of goals scored by teams over time in the tournament?**
    A: Use a time series plot or cumulative goals plot to visualize how goals scored by teams change over the course of the tournament.

Advanced Analysis Techniques

95. **Q: How can you perform clustering to group players based on their performance metrics (goals, assists, shots)?**
    A: Use clustering algorithms like K-means to group players into categories based on their performance metrics (e.g., goals scored, assists, shots on target).

96. **Q: How do you use regression analysis to predict the number of goals a player will score?**
    A: You can use linear regression to model the relationship between variables like shots, minutes played, and previous performance to predict the number of goals a player will score.

97. **Q: How do you analyze the influence of team formation on match results?**
    A: Analyze match outcomes based on teams' formations (e.g., 4-4-2, 4-3-3) and correlate them with performance metrics like goals scored and goals conceded.

98. **Q: How do you assess the impact of player experience (caps) on performance in the World Cup?**
    A: Compare the performance of players with varying levels of international experience (number of caps) to determine if more experienced players perform better in the tournament.

99. **Q: What is the relationship between total distance covered by players and match outcomes?**
    A: Correlate the total distance covered by players during matches with outcomes (win, loss, draw) to see if teams that cover more distance are more likely to win.

100. **Q: How do you perform sentiment analysis on World Cup-related articles or posts to predict team success?**
A: Perform sentiment analysis on social media posts or articles related to teams and correlate the sentiment scores with team performance to see if positive sentiment correlates with success.