10Alytics

CASE STUDY: PANDAS

CHU NGWOKE

FIFA WORLD CUP: Introduction

You've been hired as a data engineer by an international sports organization responsible for organizing major sporting events, including the FIFA World Cup. As the organization gears up for the upcoming World Cup tournament, they are facing challenges in analyzing historical match data to derive insights. Your task is to analyze a dataset containing information about FIFA World Cup matches since its inception, clean and manipulate the data, and provide actionable insights to inform tournament planning and strategy.

FIFA WORLD CUP: The Challenge

The organization needs to understand historical trends and patterns in FIFA World Cup matches to optimize tournament logistics, enhance fan engagement, and improve team performance. However, the dataset containing match data is large and unstructured, making it challenging to extract meaningful insights. Additionally, inconsistencies and missing values in the data require thorough cleaning and manipulation to ensure accuracy and reliability in the analysis.

FIFA WORLD CUP: The Data

The dataset can be downloaded here. The dataset contains the following information

- ID: Unique identifier for each match
- Year: Year in which the match took place
- Date: Date of the match
- Stage: Stage of the tournament (group stage, knockout stage, final, etc.)
- Home Team: Home team in the match
- Home Goals: Number of goals scored by the home team
- Away Goals: Number of goals scored by the away team
- Away Team: Away team in the match
- Win Conditions: Conditions under which the match was won (if applicable)
- Host Team: Boolean indicating if the match was hosted by the home team

TASKS: Data Exploration

1. How many rows and columns are in the dataset?
2. Are there missing values in the dataset?
3. Are there duplicates in the dataset?
4. Check the datatypes of all columns in the dataset
5. Check for Outliers in all numerical columns
6. What is the earliest tournament year recorded in the dataset
7. What is the latest tournament year recorded in the dataset

TASKS: Data Cleaning

1. Convert the date column to datetime datatype.
2. Standardize the column names.
3. Handle missing values appropriately, if they exist. Justify your chosen method for handling the missing values.
4. Are there duplicates in the dataset? If yes, record the number of duplicates and handle appropriately.
5. Handle outliers in the dataset, if they exist. Justify your method of handling the outliers.

TASKS: Data Transformation

1. Create a new column named 'Total Goals' that contains the sum of home and away goals scored in each match.
2. Create a conditional column ‘Score Category’ that contains over ‘High Scoring’ for matches with more than 2 goals and ‘Low Scoring’ otherwise
3. Create a new dataset that holds only the group stage matches
4. Create a new dataset that holds only matches from 2010 and above.
5. Create a pivot table to summarize the total goals scored by each team in each year of the tournament.

TASKS: Generating Insights with Pandas

1. Which team has played the most number of FIFA World Cup matches.
2. Find the top 5 highest scoring tournament years? Which countries hosted the Tournament in these years.
3. How many goals has each team scored in the FIFA World Cup? Which team has scored the highest number of goals? How many goals is it?
4. Which team has hosted the most number of FIFA World Cup tournaments
5. What ratio of goal per game for each team.
6. Which tournament stage are most goals scored in? (Find the goal per game ratio for each tournament stage)

TASKS: Generating Insights with Pandas

7. Calculate the win rate for all teams. Which team has the highest win rate
8. Which team has played the most Final? (Stage == Final)
9. Which team has won the most FIFA world Cup (Which team has won the most Final)
10. What percentage of matches went to Extra time in the knockout stages in 2010 (Find same metric for the last 3 tournaments)
11. What percentage of matches went to Penalties in the knockout stages in 2010 (Find same metric for the last 3 tournaments)
12. How many times has the Host Nation (Host Team) won the tournament?
13. Which other insights can you think of? Get them for your new employer!


10Alytics
Returns
Happy Engineering!