CYART

inquiry@cyart.io

www.cyart.io

Data Cleaning Challenges and Solutions

Problem 1 : Missing Values

Problem:

Some columns in the dataset contained missing (NaN/null) values, especially in critical fields like scores or match outcomes.

Solution:

- Used df.isnull().sum() to detect missing values.
- Decided to drop rows with missing values in essential columns using df.dropna().
- This ensured only complete match records were analyzed.

Problem 2: Outliers

Problem:

The venue_id column had extremely large or small values that did not align with expected ranges (possibly invalid encodings or rare cases).

Solution:

- Used Interquartile Range (IQR) method:
- This helped filter out extreme outlier values and retained normal distribution.

Problem 3: Incorrect Data Types

Problem:

Several numeric columns such as scores or runs were stored as strings (object type), preventing mathematical operations.

Solution:

- Converted those columns using astype():
- Ensured proper numeric formatting for statistical and visualization tasks.

Problem 4: Unparsed Date Columns

Problem:

Date columns like start_date and end_date were in string format, preventing date-based filtering or sorting.

Solution:

- Converted using Pandas' date parser:

```
df['start_date'] = pd.to_datetime(df['start_date'])
```

- Enabled time series analysis and seasonal comparisons.

Regression Modeling Summary

- Goal: Predict total Runs scored by a player using match performance metrics
- Model Used: Linear Regression via statsmodels
- Steps:
  - Cleaned and engineered numeric features
  - Added constant term for intercept
  - Fitted model using OLS from statsmodels.api
- Assumptions checked: Linearity, residuals distribution

Key Regression Results

Model Summary Highlights:

- R^2 Score: ~0.61 — model explains 61% variance in player runs

Significant Predictors ($p < 0.05$):

  - Innings
  - Average
  - Strike Rate

- Insignificant Predictor:
  - Matches ($p > 0.05$ — low impact on total runs)

Interpretation:

- More innings and higher averages/strike rates strongly predict total runs
- Number of matches alone isn't a reliable performance indicator