Aman raj

23116009 (ECE)

**Elaborate Report on Data Cleaning and Exploratory Data Analysis (EDA)**

Introduction: Data analysis is a crucial step in deriving insights from raw data. This report covers **data cleaning** and **exploratory data analysis (EDA)** performed on a cricket dataset. The key objectives include handling missing values, removing duplicates, detecting outliers, and performing univariate, bivariate, and multivariate analysis to understand relationships within the dataset.

1. **2. Data Cleaning**

**2.1 Loading and Inspecting the Dataset**

The dataset was loaded and inspected using df.info() and df.head().

Data types, missing values, and inconsistencies were identified.

**2.2 Handling Missing Values**

1. Missing values were treated using **imputation (mean, mode)** or removed if necessary.
2. Categorical missing values were replaced with the mode, while numerical columns were imputed using the mean.
3. **2.3 Removing Duplicates**
4. Duplicate records were detected and removed using df.drop\_duplicates() to avoid redundancy.
5. **2.4 Outlier Detection and Treatment**
6. Outliers were identified using box plots and the **interquartile range (IQR) method**.
7. Extreme values were either **winsorized** (capping at a threshold) or removed.
8. **2.5 Standardizing Categorical Values**
9. Inconsistent categorical values (e.g., "Bat first" vs. "bat first") were standardized using df['column'].str.lower().str.strip().
10. Typos were corrected to ensure uniformity in analysis.
11. **3. Exploratory Data Analysis (EDA)**
12. **3.1 Univariate Analysis**
13. **Summary Statistics**
14. Measures such as **mean, median, mode, variance, and skewness** were computed using df.describe().
15. This helped identify distributions and detect skewness in variables like Total\_Runs and Run\_Rate.
16. **Frequency Distributions**
17. The most frequent categories in columns like Venue, Toss\_Decision, and Winner were analyzed.
18. **Visualizations**
19. **Histograms** were plotted to visualize the distribution of numerical variables.
20. **Box plots** were used to detect outliers and assess data spread.
21. **3.2 Bivariate Analysis**
22. **Correlation Analysis**
23. A **correlation matrix** was plotted using a heatmap to understand relationships between numerical variables.
24. Strong correlations included Total\_Runs vs. Run\_Rate (0.79) and Wickets\_Lost vs. Win\_Margin (-0.58).
25. **Scatter Plots**
26. **Total Runs vs. Run Rate**: Showed a strong positive trend.
27. **Wickets Lost vs. Win Margin**: Showed a negative correlation, confirming that losing more wickets reduces winning margins.
28. **Comparisons Between Categorical and Numerical Variables**
29. **Bar plots** showed the impact of Toss\_Decision on total runs scored.
30. **Violin plots** displayed the distribution of Run\_Rate for different winning teams.
31. **Box plots** analyzed Win\_Margin based on Toss\_Decision.
32. **3.3 Multivariate Analysis**
33. **Pair Plots**
34. Pair plots were generated to analyze multiple numerical relationships simultaneously.
35. Strong associations between **Total Runs, Wickets Lost, and Win Margin** were observed.
36. **Heatmaps**
37. A **heatmap visualization** of correlations reinforced numerical relationships.
38. **Grouped Comparisons**
39. **Box plots comparing Total Runs by Winner** helped analyze team performances.
40. **Violin plots of Run Rate by Toss Decision** showed distribution variations based on game strategies.
41. **4. Key Insights and Conclusions**
42. **4.1 Data Cleaning Insights**
43. Missing values were handled efficiently without losing essential data.
44. Duplicate records were removed to improve dataset integrity.
45. Outliers were treated to prevent biased statistical analysis.
46. Categorical inconsistencies were resolved, improving accuracy in categorical analysis.
47. **4.2 EDA Insights**
48. **Univariate analysis** revealed skewed distributions in Total\_Runs and Run\_Rate.
49. **Bivariate analysis** confirmed strong relationships, particularly Total\_Runs vs. Run\_Rate and Wickets\_Lost vs. Win\_Margin.
50. **Multivariate analysis** highlighted complex interactions, like how different teams perform based on toss decisions.
51. **4.3 Recommendations**
52. Further statistical modeling can be performed on the cleaned dataset.
53. Advanced machine learning techniques (e.g., regression or classification) can be applied.
54. Understanding team strategies based on match conditions can improve performance analysis.

**5. Conclusion**

* This report successfully demonstrates the importance of **data cleaning and exploratory data analysis (EDA)**. The insights drawn from this dataset can be useful for deeper cricket analytics, such as match predictions and team performance evaluations. Future studies can incorporate more advanced techniques, such as **predictive modeling and feature engineering**, to gain further insights.