**Exploratory Data Analysis (EDA) Report**

**1. Introduction**

This report summarizes the Exploratory Data Analysis (EDA) performed on the dataset. The goal of EDA is to clean the data, handle missing values, detect outliers, and analyze feature relationships to prepare the dataset for machine learning.

**2. Dataset Overview**

The dataset was loaded and the following initial checks were performed:

* **Data Types:** Verified data types to ensure correct formats (numerical, categorical).
* **Missing Values:** Identified columns with missing values.
* **Duplicate Records:** Checked for duplicate rows.
* **Basic Statistics:** Used .describe() to understand data distribution.

**3. Data Cleaning Steps**

**Handling Missing Values**

* **Categorical Features:** Mode imputation was applied (e.g., Day.of.Week filled with mode value).
* **Numerical Features:** Median imputation was used for missing numerical values (MPC, RSI, PPC, EB, OR, MS, LA, NES, OS).
* **Date Column Handling:** Extracted day, month, and year from Date and then dropped the original column.

**Duplicate Removal**

* Checked for duplicate rows using df.duplicated().sum(), but no significant duplicates were found.

**Dropping Irrelevant Columns**

* Removed unnecessary columns (Date, and columns containing Athlete\_ if present) to focus on relevant data.

**4. Feature Engineering**

**Encoding Categorical Variables**

* Used **One-Hot Encoding (OHE)** to convert categorical features into numerical form (pd.get\_dummies(drop\_first=True)).

**Correlation Analysis**

* **Heatmap Visualization:** Generated a correlation matrix to understand feature relationships.
* **Feature Selection:**
  + Dropped highly correlated features (correlation > 0.9) to reduce multicollinearity.
  + Used np.triu() to extract the upper triangle of the correlation matrix.

**5. Outlier Detection and Treatment**

**Outlier Capping (Winsorization)**

* Used the **Interquartile Range (IQR) Method** to cap extreme outliers instead of removing them.
* Formula used:
  + **Lower Bound:** Q1 - 1.5 \* IQR
  + **Upper Bound:** Q3 + 1.5 \* IQR
* Applied to numerical columns: PPC, MPC, EB, OR, MS, LA, NES, OS, RSI.

**6. Data Visualization**

**Heatmap**

* Correlation heatmaps were plotted using **Seaborn** to visualize feature relationships.
* Helped identify redundant variables.

|  |  |
| --- | --- |
| **7. Why These Methods Were Used ?** |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **METHODS USED AS PER THE DATASET** | |  |  |
| ***TECHNIQUES*** | ***REASON FOR USE*** |  |  |
| **Mode Imputation** | Best for categorical data to maintain mode frequency |  |  |
| **Median Imputation** | Best for numerical data with skewed distributions |  |  |
| **One-Hot Encoding** | Prevents introducing ordinal relationships in categorical features |  |  |
| **IQR Method** | Caps outliers while preserving data instead of outright removal |  |  |
| **Feature Selection via Correlation** | Reduces multicollinearity, improving ML model performance |  |  |

**8. Conclusion**

This EDA process prepared the dataset for machine learning by ensuring data cleanliness, reducing redundancy, and handling missing values and outliers effectively. Next steps would involve feature scaling and model selection.

**9. Next Week's Targets**

* **Feature Scaling:** Standardizing numerical features using MinMaxScaler or StandardScaler.
* **Model Training:** Implementing initial ML models such as Logistic Regression, Decision Trees, or Random Forest.
* **Hyperparameter Tuning:** Optimizing model parameters using GridSearchCV.
* **Model Evaluation:** Assessing model performance using accuracy, precision, recall, and F1-score.
* **Feature Importance Analysis:** Understanding which features contribute most to predictions.