Usman Institute of Technology

V1 BS CS

BIG DATA ANALYTICS ,CSC-316

#### 

Topic: Analysis of Cricket Match Outcomes

#### 

#### 

#### 

#### **Group Details**

* + Member 1: Syed Saad Akhtar 22sp-029-cs
  + Member 2: Kinza Fatima 22sp-032-cs
  + Member 3: Syeda Akfa Feroz 22sp-046-cs

### 

### **Description about code:**

### **Data Cleaning Steps:**

### **Dropped the column umpire3 as it contained no meaningful data (all null values).**

### **Identified columns with missing data (city, winner, player\_of\_match, umpire1, and umpire2) and filled their missing values with the placeholder "unknown".**

### **Converted the date column from string format to a proper datetime format for easier date manipulation and analysis.**

### **Removed duplicate rows from the dataset to ensure uniqueness and data integrity.**

### **Standardized text in specific columns (team1, team2, toss\_winner, etc.) by stripping extra spaces and capitalizing words for consistent formatting.**

### **Handling Outliers:**

### **Identified potential outliers in the win\_by\_runs and win\_by\_wickets columns (values less than 0) and printed them for review.**

### **Feature Engineering:**

### **Added a new derived column, match\_outcome, to indicate how a match was won:**

### **"Target is defended" if the win was by runs.**

### **"Target is chased" if the win was by wickets.**

### **"Tie/No Result" if neither condition was met.**

### **Logical Consistency:**

### **Ensured that the dl\_applied column only contained binary values (1 for true, 0 for false).**

### **Output:**

### **Saved the cleaned dataset to a user-specified file path in .csv format.**

### 

### 

### **Cricket Match Analysis: Target Variables and Key Features:**

### **1. Match Outcome Analysis**

* **Toss Impact**:
  + Target Variable: winner (whether the toss-winning team wins the match).
  + Features to analyze: toss\_winner, toss\_decision, and match\_outcome.

### **2. Team Performance**

* **Top Performing Teams**:
  + Target Variable: winner (to identify the team with the most wins).
  + Features to analyze: season, team1, and team2.

### **3. Player Performance**

* **Seasonal Standouts**:
  + Target Variable: player\_of\_match (to find standout players in specific seasons).
  + Features to analyze: season and player\_of\_match.

### **4. Venue and Location Insights**

* **City and Venue Trends**:
  + Target Variable: match\_outcome (to analyze if a venue favors batting or bowling).
  + Features to analyze: venue, city, win\_by\_runs, and win\_by\_wickets.
* **Home Advantage**:
  + Target Variable: winner (to check if home teams perform better).
  + Features to analyze: city, team1, team2, and toss\_winner.