|  |  |
| --- | --- |
| **Project Code** | Project 1 : Mandatory |
| **Project Name** | Load and Process IPL Match Data |

|  |  |  |
| --- | --- | --- |
| **Prepared/Modified by** | **Role** | **Date of Preparation** |
| Charudatta Joshi | AWS Trainer | 31-05-2023 |
| **Reviewed by** | **Role** | **Date of Review** |
|  |  |  |
| **Approved by** | **Role** | **Date of Approval** |
|  |  |  |
| **Circulation List** |  | **Version Number of the template:** |
| **Version Number** | 1.0 | 1.0 |

**Dataset Details:**

**About Dataset**

**This data has ball by ball information for every match played between 2008 and 2022 seasons.**

**Context**:

The Indian Premier League (IPL) is a professional men's Twenty20 cricket league, contested by ten teams based out of ten Indian cities. The league was founded by the Board of Control for Cricket in India (BCCI) in 2007. It is usually held between March and May of every year and has an exclusive window in the ICC Future Tours Programme.

The IPL is the most-attended cricket league in the world and in 2014 was ranked sixth by average attendance among all sports leagues. In 2010, the IPL became the first sporting event in the world to be broadcast live on YouTube. The brand value of the IPL in 2019 was ₹47,500 crores (6.3 billion US dollars), according to Duff & Phelps. According to BCCI, the 2015 IPL season contributed ₹1,150 crores (150 million US dollars) to the GDP of the Indian economy. The 2020 IPL season set a massive viewership record with 31.57 million average impressions and with an overall consumption increase of 23 percent from the 2019 season.

**Tables in Dataset: -**

Table 1: Match Details (File Name : IPL\_Matches\_2008\_2022.csv)



Table 2: Ball\_By\_Ball\_Details (File Name : IPL\_Ball\_by\_Ball\_2008\_2022)



**Abbreviations:**

1. BF: Balls Faced
2. SR: Strike Rate (Batsman: Total Runs/Balls Faced)
3. POS: Position
4. Mat: Matches
5. Inns Innings
6. NO: Not Out
7. HS: Highest Score
8. AVG: Average ( Batsman: Total Runs / Innings)
9. Ov: Overs
10. Econ: Economy Rate of Bowler (Total runs conceded/ Overs balled, Round the half completed over to next whole number)
11. Wkts: Wickets
12. Maid: Maiden Over

**Infrastructure Setup:**

S3 for data store and Glue for data processing

**Mandatory Requirements**

**Requirement No 1: Fastest Centuries**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Runs | BF | 4s | 6s | Against | Venue | Match Date |
| 1 | Adam Gilchrist | 109 | 42 | 9 | 10 | MI | DY Patil Stadium | 27-Apr-08 |
| 2 | Sanath Jayasuriya | 114 | 45 | 9 | 11 | CSK | Wankhede Stadium | 14-May-08 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list players with fastest century on the top.
3. List should position players in ascending order of number of balls required to score first 100 Runs.
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 2: Most Fours Innings**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Runs | BF | SR | 4s | 6s | Against | Venue | Match Date |
| 1 | Ajinkya Rahane | 98 | 66 | 148.48 | 16 | 1 | PBKS | Sawai Mansingh Stadium | 06-Apr-12 |
| 2 | Murali Vijay | 113 | 58 | 194.82 | 15 | 4 | DC | Chidambaram | 25-May-12 |
| 3 | Virender Sehwag | 73 | 38 | 192.1 | 12 | 1 | RR | Sawai Mansingh Stadium | 01-May-12 |
| 4 | Rohit Sharma | 109 | 60 | 181.66 | 12 | 5 | KKR | Eden Gardens | 12-May-12 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list players with the greatest number of fours on top
3. File should position players in descending order of 4s in a single inning
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 3: Most Runs**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Mat | Inns | NO | Runs | HS | Avg | BF | SR | 100 | 50 | 4s | 6s |
| 1 | Sachin Tendulkar | 15 | 15 | 2 | 618 | 89 | 47.53 | 466 | 132.61 | 0 | 5 | 86 | 3 |
| 2 | Jacques Kallis | 16 | 16 | 4 | 572 | 89 | 47.66 | 494 | 115.78 | 0 | 6 | 67 | 9 |
| 3 | Suresh Raina | 16 | 16 | 5 | 520 | 83 | 47.27 | 364 | 142.85 | 0 | 4 | 45 | 22 |
| 4 | Sourav Ganguly | 14 | 14 | 1 | 493 | 88 | 37.92 | 419 | 117.66 | 0 | 4 | 58 | 15 |
| 5 | Murali Vijay | 15 | 15 | 2 | 458 | 127 | 35.23 | 292 | 156.84 | 1 | 2 | 36 | 26 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list statics of players as shown above
3. List should position players in descending order of 4s in a single inning
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 4: Most Wickets**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Mat | Inns | Ov | Runs | Wkts | Econ | 4w | 5w |
| 1 | Dwayne Bravo | 18 | 18 | 62.3 | 497 | 32 | 7.95 | 1 | 0 |
| 2 | James Faulkner | 16 | 16 | 63.1 | 427 | 28 | 6.75 | 0 | 2 |
| 3 | Harbhajan Singh | 19 | 19 | 70 | 456 | 24 | 6.51 | 0 | 0 |
| 4 | Mitchell Johnson | 17 | 17 | 64 | 459 | 24 | 7.17 | 0 | 0 |
| 5 | Vinay Kumar | 16 | 16 | 60.1 | 493 | 23 | 8.19 | 0 | 0 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list statics of bowler as shown above
3. File should position players in descending order of wickets taken in all innings
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 5: Most Runs Conceded In Innings**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Ov | Runs | Wkts | SR | Against | Venue | Match Date |
| 1 | RP Singh | 4 | 59 | 1 | 24 | KKR | Rajiv Gandhi Intl. Cricket Stadium | 11-May-08 |
| 2 | Palani Amarnath | 4 | 57 | 1 | 24 | MI | Chidambaram | 23-Apr-08 |
| 3 | Vikram Singh | 4 | 56 | 0 | 0 | MI | Wankhede Stadium | 21-May-08 |
| 4 | Munaf Patel | 4 | 55 | 0 | 0 | DEC | Rajiv Gandhi Intl. Cricket Stadium | 24-Apr-08 |
| 5 | Munaf Patel | 4 | 55 | 1 | 24 | CSK | Chidambaram | 24-May-08 |
| 6 | Jacob Oram | 4 | 53 | 0 | 0 | MI | Chidambaram | 23-Apr-08 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list statics of bowler as shown above
3. File should position bowlers in descending order of runs given in single innings
4. List should be in descending order of wickets
5. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 6: Fastest Fifties**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Runs | BF | 4s | 6s | Against | Venue | Match Date |
| 1 | Yusuf Pathan | 61 | 21 | 4 | 6 | DEC | Rajiv Gandhi Intl. Cricket Stadium | 24-Apr-08 |
| 2 | Kumar Sangakkara | 94 | 23 | 13 | 1 | MI | IS Bindra Stadium | 25-Apr-08 |
| 3 | Virender Sehwag | 51 | 23 | 2 | 5 | PBKS | Arun Jaitley Stadium | 17-May-08 |
| 4 | Kumar Sangakkara | 50 | 23 | 7 | 2 | DEC | IS Bindra Stadium | 23-May-08 |
| 5 | Virender Sehwag | 71 | 24 | 7 | 4 | CSK | Chidambaram | 02-May-08 |
| 6 | James Hopes | 71 | 24 | 10 | 3 | CSK | IS Bindra Stadium | 19-Apr-08 |
| 7 | MS Dhoni | 65 | 24 | 9 | 3 | RCB | M. Chinnaswamy Stadium | 28-Apr-08 |
| 8 | David Hussey | 57 | 24 | 4 | 3 | DEC | Rajiv Gandhi Intl. Cricket Stadium | 11-May-08 |
| 9 | Sanath Jayasuriya | 114 | 25 | 9 | 11 | CSK | Wankhede Stadium | 14-May-08 |
| 10 | Rahul Dravid | 75 | 25 | 6 | 6 | RR | Sawai Mansingh Stadium | 17-May-08 |
| 11 | Albie Morkel | 71 | 26 | 6 | 4 | RR | Chidambaram | 24-May-08 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list statics of players as shown above
3. File should position players in ascending order of balls required to score first 50 runs.
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”

**Requirement No 7: Most Sixes Innings**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| POS | Player | Runs | BF | SR | 4s | 6s | Against | Venue | Match Date |
| 1 | Brendon McCullum | 158 | 73 | 216.43 | 10 | 13 | RCB | M. Chinnaswamy Stadium | 18-Apr-08 |
| 2 | Sanath Jayasuriya | 114 | 48 | 237.5 | 9 | 11 | CSK | Wankhede Stadium | 14-May-08 |
| 3 | Adam Gilchrist | 109 | 47 | 231.91 | 9 | 10 | MI | DY Patil Stadium | 27-Apr-08 |
| 4 | Michael Hussey | 116 | 54 | 214.81 | 8 | 9 | PBKS | IS Bindra Stadium | 19-Apr-08 |
| 5 | Andrew Symonds | 117 | 53 | 220.75 | 11 | 7 | RR | Rajiv Gandhi Intl. Cricket Stadium | 24-Apr-08 |
| 6 | Shaun Marsh | 115 | 69 | 166.66 | 11 | 7 | RR | IS Bindra Stadium | 28-May-08 |
| 7 | Yuvraj Singh | 49 | 16 | 306.25 | 2 | 6 | RR | IS Bindra Stadium | 28-May-08 |
| 8 | Virender Sehwag | 94 | 41 | 229.26 | 10 | 6 | DEC | Rajiv Gandhi Intl. Cricket Stadium | 22-Apr-08 |
| 9 | Yusuf Pathan | 61 | 28 | 217.85 | 4 | 6 | DEC | Rajiv Gandhi Intl. Cricket Stadium | 24-Apr-08 |
| 10 | Rahul Dravid | 75 | 36 | 208.33 | 6 | 6 | RR | Sawai Mansingh Stadium | 17-May-08 |

**Acceptance Criteria:**

1. Separate CSV file for every season should be created
2. File should list statics of players as shown above
3. File should position players in ascending order of sixes hit in single innings
4. Create a output file with the same name as of title (Example: if title is “Good Aws Batch” file name should be “good\_aws\_batch.csv”