databricks IPL-Data-Analysis

(https://databricks.com)

End-to-End Data Engineering Project on IPL Data

Business Case for IPL Data Analysis Project

Project Overview:

The IPL Data Analysis Project aims to extract, transform, and analyze data from the Indian Premier League (IPL) to provide actionable insights for stakeholders, including team managers, coaches, analysts, and fans. Leveraging data engineering and analytical techniques, this project seeks to improve decision-making in player selection, match strategies, and overall team performance.

Objectives:

Performance Analysis: Assess player performances through various metrics like runs scored, wickets taken, and overall contributions to match outcomes. Match Insights: Analyze match outcomes in relation to toss results, venue conditions, and player performance to identify patterns that influence match results. Player Development: Identify strengths and weaknesses of players based on historical data to guide training and development efforts. Fan Engagement: Provide fans with engaging visualizations and statistics that enhance their viewing experience and understanding of the game.

Methodology:

Data Extraction: Use AWS S3 to store raw data files, including player statistics, match details, and ball-by-ball commentary. Data Transformation: Load data into Databricks, clean and normalize datasets, and create relevant schemas for analysis. Analysis and Reporting: Calculate total and average runs scored by players and teams. Identify economical bowlers and high-impact balls in matches. Analyze the impact of winning the toss on match outcomes. Visualize insights through charts and graphs for better comprehension. Expected Outcomes: Enhanced Decision-Making: Provide teams with data-driven insights to enhance player selection and match strategies. Performance Benchmarks: Establish benchmarks for player performance that can be tracked over seasons. Engagement Metrics: Deliver compelling data visualizations that enhance fan engagement and interest in the league.

3

```
### importing packages
from pyspark.sql import SparkSession

## create session
spark = SparkSession.builder.appName("IPL Data Analysis").getOrCreate()
```

4

```
from pyspark.sql.types import StructField,StructType,IntegerType,StringType,BooleanType,DateType,DecimalType
from pyspark.sql.functions import *
from pyspark.sql.window import Window
```

```
5
```

```
### connecting aws to databricks vis pyspark
spark.conf.set("fs.s3a.access.key", "AKIAZANQMQDEJ5KLCAXG")
spark.conf.set("fs.s3a.secret.key", "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")
spark.conf.set("fs.s3a.endpoint", "s3.eu-west-1.amazonaws.com")
```

Creating manual schemas for accurate datatype of columns

```
7
```

```
ball_by_ball_schema = StructType([
    StructField("match_id", IntegerType(), True),
    StructField("over_id", IntegerType(), True),
    StructField("ball_id", IntegerType(), True),
    StructField("innings_no", IntegerType(), True),
    StructField("team_batting", StringType(), True),
    StructField("team_bowling", StringType(), True),
    StructField("striker_batting_position", IntegerType(), True),
    StructField("extra_type", StringType(), True),
    StructField("runs_scored", IntegerType(), True),
    StructField("extra_runs", IntegerType(), True),
    StructField("wides", IntegerType(), True),
    StructField("legbyes", IntegerType(), True),
    StructField("byes", IntegerType(), True),
    StructField("noballs", IntegerType(), True),
    StructField("penalty", IntegerType(), True),
    StructField("bowler_extras", IntegerType(), True),
    StructField("out_type", StringType(), True),
    StructField("caught", BooleanType(), True),
    StructField("bowled", BooleanType(), True),
    StructField("run_out", BooleanType(), True),
    StructField("lbw", BooleanType(), True),
    StructField("retired_hurt", BooleanType(), True),
    StructField("stumped", BooleanType(), True),
    StructField("caught_and_bowled", BooleanType(), True),
    StructField("hit_wicket", BooleanType(), True),
    StructField("obstructingfeild", BooleanType(), True),
    StructField("bowler_wicket", BooleanType(), True),
    StructField("match_date", DateType(), True),
    StructField("season", IntegerType(), True),
    StructField("striker", IntegerType(), True),
    StructField("non_striker", IntegerType(), True),
    StructField("bowler", IntegerType(), True),
    StructField("player_out", IntegerType(), True),
    StructField("fielders", IntegerType(), True),
    StructField("striker_match_sk", IntegerType(), True),
    StructField("strikersk", IntegerType(), True),
    StructField("nonstriker_match_sk", IntegerType(), True),
    StructField("nonstriker_sk", IntegerType(), True),
    StructField("fielder_match_sk", IntegerType(), True),
    StructField("fielder_sk", IntegerType(), True),
    StructField("bowler_match_sk", IntegerType(), True),
    StructField("bowler_sk", IntegerType(), True),
    StructField("playerout_match_sk", IntegerType(), True)
```

```
    StructField("playerout_match_sk", IntegerType(), True),
    StructField("battingteam_sk", IntegerType(), True),
    StructField("bowlingteam_sk", IntegerType(), True),
    StructField("keeper_catch", BooleanType(), True),
    StructField("player_out_sk", IntegerType(), True),
    StructField("matchdatesk", DateType(), True)
])
```

8

```
ball_by_ball_df=spark.read.schema(ball_by_ball_schema).format("csv")\
    .option("header","true")\
    .option("InferSchema","true")\
    .load("s3://ipl-data-analysis-bucket1/Ball_By_Ball.csv")
```

9

```
ball_by_ball_df.show()
```

```
+--------+-------+-------+----------+------------+------------+-----------------------+----------+-----------+----------+-----+-------+----+-------+-------+-------------+--------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+-------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+--------------+--------------+------------+-------------+-----------+
|match_id|over_id|ball_id|innings_no|team_batting|team_bowling|striker_batting_position|extra_type|runs_scored|extra_runs|wides|legbyes|byes|noballs|penalty|bowler_extras|      out_type|caught|bowle
d|run_out| lbw|retired_hurt|stumped|caught_and_bowled|hit_wicket|obstructingfeild|bowler_wicket|match_date|season|striker|non_striker|bowler|player_out|fielders|striker_match_sk|strikersk|nonstriker
_match_sk|nonstriker_sk|fielder_match_sk|fielder_sk|bowler_match_sk|bowler_sk|playerout_match_sk|battingteam_sk|bowlingteam_sk|keeper_catch|player_out_sk|matchdatesk|
+--------+-------+-------+----------+------------+------------+-----------------------+----------+-----------+----------+-----+-------+----+-------+-------+-------------+--------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+-------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+--------------+--------------+------------+-------------+-----------+
|  598028|     15|      6|         1|           5|           2|                      6| No Extras|          4|         0|    0|      0|   0|      0|      0|            0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2013|    277|        104|    83|      null|    null|          20336|      276|
20333|          103|             -1|        -1|          20343|       82|               -1|            4|            1|        null|            0|       null|
|  598028|     14|      1|         1|           5|           2|                      5| No Extras|          1|         0|    0|      0|   0|      0|      0|            0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2013|    104|          6|   346|      null|    null|          20333|      103|
20328|            5|             -1|        -1|          20348|      345|               -1|            4|            1|        null|            0|       null|
|  598028|     14|      2|         1|           5|           2|                      3| No Extras|          1|         0|    0|      0|   0|      0|      0|            0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2013|      6|        104|   346|      null|    null|          20328|        5|
20333|          103|             -1|        -1|          20348|      345|               -1|            4|            1|        null|            0|       null|
|  598028|     14|      3|         1|           5|           2|                      5| No Extras|          1|         0|    0|      0|   0|      0|      0|            0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2013|    104|          6|   346|      null|    null|          20333|      103|
20328|            5|             -1|        -1|          20348|      345|               -1|            4|            1|        null|            0|       null|
```

10

```
ball_by_ball_df.printSchema()
```

```
root
 |-- match_id: integer (nullable = true)
 |-- over_id: integer (nullable = true)
 |-- ball_id: integer (nullable = true)
 |-- innings_no: integer (nullable = true)
 |-- team_batting: string (nullable = true)
 |-- team_bowling: string (nullable = true)
 |-- striker_batting_position: integer (nullable = true)
 |-- extra_type: string (nullable = true)
 |-- runs_scored: integer (nullable = true)
 |-- extra_runs: integer (nullable = true)
 |-- wides: integer (nullable = true)
 |-- legbyes: integer (nullable = true)
 |-- byes: integer (nullable = true)
 |-- noballs: integer (nullable = true)
 |-- penalty: integer (nullable = true)
 |-- bowler_extras: integer (nullable = true)
 |-- out_type: string (nullable = true)
 |-- caught: boolean (nullable = true)
 |-- bowled: boolean (nullable = true)
 |-- run_out: boolean (nullable = true)
```

11

```
match_schema = StructType([
    StructField("match_sk", IntegerType(), True),
    StructField("match_id", IntegerType(), True),
    StructField("team1", StringType(), True),
    StructField("team2", StringType(), True),
    StructField("match_date", DateType(), True),
    StructField("season_year", IntegerType(), True),
    StructField("venue_name", StringType(), True),
    StructField("city_name", StringType(), True),
    StructField("country_name", StringType(), True),
    StructField("toss_winner", StringType(), True),
    StructField("match_winner", StringType(), True),
    StructField("toss_name", StringType(), True),
    StructField("win_type", StringType(), True),
    StructField("outcome_type", StringType(), True),
    StructField("manofmach", StringType(), True),
    StructField("win_margin", IntegerType(), True),
    StructField("country_id", IntegerType(), True)
])
match_df = spark.read.schema(match_schema).format("csv").option("header","true").load("s3://ipl-data-analysis-bucket1/Match.csv")
```

12

```
match_df.show()
```

```
+--------+--------+--------------------+--------------------+----------+-----------+--------------------+----------+------------+--------------------+--------------------+---------+--------+--------
----+-------------+----------+----------+
|match_sk|match_id|               team1|               team2|match_date|season_year|          venue_name| city_name|country_name|         toss_winner|        match_winner|toss_name|win_type|outcome_
type|    manofmach|win_margin|country_id|
+--------+--------+--------------------+--------------------+----------+-----------+--------------------+----------+------------+--------------------+--------------------+---------+--------+--------
----+-------------+----------+----------+
|       0|  335987|Royal Challengers...|Kolkata Knight Ri...|      null|       2008|M Chinnaswamy Sta...| Bangalore|       India|Royal Challengers...|Kolkata Knight Ri...|    field|    runs|      Re
sult|  BB McCullum|       140|         1|
|       1|  335988|     Kings XI Punjab| Chennai Super Kings|      null|       2008|Punjab Cricket As...|Chandigarh|       India| Chennai Super Kings| Chennai Super Kings|      bat|    runs|      Re
sult|   MEK Hussey|        33|         1|
|       2|  335989|    Delhi Daredevils|    Rajasthan Royals|      null|       2008|    Feroz Shah Kotla|     Delhi|       India|    Rajasthan Royals|    Delhi Daredevils|      bat| wickets|      Re
sult|  MF Maharoof|         9|         1|
|       3|  335990|      Mumbai Indians|Royal Challengers...|      null|       2008|    Wankhede Stadium|    Mumbai|       India|      Mumbai Indians|Royal Challengers...|      bat| wickets|      Re
sult|   MV Boucher|         5|         1|
|       4|  335991|Kolkata Knight Ri...|     Deccan Chargers|      null|       2008|        Eden Gardens|   Kolkata|       India|     Deccan Chargers|Kolkata Knight Ri...|      bat| wickets|      Re
sult|    DJ Hussey|         5|         1|
|       5|  335992|    Rajasthan Royals|     Kings XI Punjab|      null|       2008|Sawai Mansingh St...|    Jaipur|       India|     Kings XI Punjab|    Rajasthan Royals|      bat| wickets|      Re
sult|    SR Watson|         6|         1|
|       6|  335993|     Deccan Chargers|    Delhi Daredevils|      null|       2008|Rajiv Gandhi Inte...| Hyderabad|       India|     Deccan Chargers|    Delhi Daredevils|      bat| wickets|      Re
```

```
sult|     V Sehwag|         9|         1|
|       7|  335994| Chennai Super Kings|      Mumbai Indians|      null|       2008|MA Chidambaram St...|   Chennai|       India|      Mumbai Indians| Chennai Super Kings|    field|    runs|      Re
```

13

```
player_schema = StructType([
    StructField("player_sk", IntegerType(), True),
    StructField("player_id", IntegerType(), True),
    StructField("player_name", StringType(), True),
    StructField("dob", DateType(), True),
    StructField("batting_hand", StringType(), True),
    StructField("bowling_skill", StringType(), True),
    StructField("country_name", StringType(), True)
])

player_df = spark.read.schema(player_schema).format("csv").option("header","true").load("s3://ipl-data-analysis-bucket1/Player.csv")
```

14

```
player_df.show()
```

```
+---------+---------+---------------+----+--------------+--------------------+------------+
|player_sk|player_id|    player_name| dob|  batting_hand|       bowling_skill|country_name|
+---------+---------+---------------+----+--------------+--------------------+------------+
|        0|        1|     SC Ganguly|null| Left-hand bat|    Right-arm medium|       India|
|        1|        2|    BB McCullum|null|Right-hand bat|    Right-arm medium| New Zealand|
|        2|        3|     RT Ponting|null|Right-hand bat|    Right-arm medium|   Australia|
|        3|        4|      DJ Hussey|null|Right-hand bat|  Right-arm offbreak|   Australia|
|        4|        5|Mohammad Hafeez|null|Right-hand bat|  Right-arm offbreak|    Pakistan|
|        5|        6|       R Dravid|null|Right-hand bat|  Right-arm offbreak|       India|
|        6|        7|       W Jaffer|null|Right-hand bat|  Right-arm offbreak|       India|
|        7|        8|        V Kohli|null|Right-hand bat|    Right-arm medium|       India|
|        8|        9|      JH Kallis|null|Right-hand bat|Right-arm fast-me...|South Africa|
|        9|       10|       CL White|null|Right-hand bat|     Legbreak googly|   Australia|
|       10|       11|     MV Boucher|null|Right-hand bat|    Right-arm medium|South Africa|
|       11|       12|        B Akhil|null|Right-hand bat|Right-arm medium-...|       India|
|       12|       13|      AA Noffke|null|Right-hand bat|Right-arm fast-me...|   Australia|
|       13|       14|        P Kumar|null|Right-hand bat|    Right-arm medium|       India|
|       14|       15|         Z Khan|null|Right-hand bat|Left-arm fast-medium|       India|
|       15|       16|       SB Joshi|null| Left-hand bat|Slow left-arm ort...|       India|
|       16|       17|       PA Patel|null| Left-hand bat|                 N/A|       India|
|       17|       18|      ML Hayden|null| Left-hand bat|    Right-arm medium|   Australia|
```

15

```
player_match_schema = StructType([
    StructField("player_match_sk", IntegerType(), True),
    StructField("playermatch_key", DecimalType(), True),
    StructField("match_id", IntegerType(), True),
    StructField("player_id", IntegerType(), True),
    StructField("player_name", StringType(), True),
    StructField("dob", DateType(), True),
    StructField("batting_hand", StringType(), True),
    StructField("bowling_skill", StringType(), True),
    StructField("country_name", StringType(), True),
    StructField("role_desc", StringType(), True),
    StructField("player_team", StringType(), True),
    StructField("opposit_team", StringType(), True),
    StructField("season_year", IntegerType(), True),
    StructField("is_manofthematch", BooleanType(), True),
    StructField("age_as_on_match", IntegerType(), True),
    StructField("isplayers_team_won", BooleanType(), True),
    StructField("batting_status", StringType(), True),
    StructField("bowling_status", StringType(), True),
    StructField("player_captain", StringType(), True),
    StructField("opposit_captain", StringType(), True),
    StructField("player_keeper", StringType(), True),
    StructField("opposit_keeper", StringType(), True)
])

player_match_df = spark.read.schema(player_match_schema).format("csv").option("header","true").load("s3://ipl-data-analysis-bucket1/Player_match.csv")
```

16

```
player_match_df.show()
```

```
+---------------+--------------+--------+---------+--------------+----+-------------+--------------------+------------+---------+--------------------+--------------------+-----------+------------
----+--------------+-----------------+-------------+-------------+-------------+--------------+-------------+--------------+
|player_match_sk|playermatch_key|match_id|player_id|    player_name| dob|  batting_hand|       bowling_skill|country_name|role_desc|         player_team|        opposit_team|season_year|is_manofthem
atch|age_as_on_match|isplayers_team_won|batting_status|bowling_status|player_captain|opposit_captain|player_keeper|opposit_keeper|
+---------------+--------------+--------+---------+--------------+----+-------------+--------------------+------------+---------+--------------------+--------------------+-----------+------------
----+--------------+-----------------+-------------+-------------+-------------+--------------+-------------+--------------+
|             -1|            -1|      -1|       -1|           N/A|null|         null|                null|        null|     null|                null|                null|       null|
null|          null|             null|         null|         null|         null|          null|         null|          null|
|          12694|          null|  335987|        6|       R Dravid|null|Right-hand bat|  Right-arm offbreak|       India|  Captain|Royal Challengers...|Kolkata Knight Ri...|       2008|
```

```
null|             35|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
|          12695|           null|  335987|        7|       W Jaffer|null|Right-hand bat|  Right-arm offbreak|       India|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             30|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
|          12696|           null|  335987|        8|        V Kohli|null|Right-hand bat|    Right-arm medium|       India|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             20|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
|          12697|           null|  335987|        9|      JH Kallis|null|Right-hand bat|Right-arm fast-me...|South Africa|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             33|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
|          12698|           null|  335987|       10|       CL White|null|Right-hand bat|     Legbreak googly|   Australia|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             25|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
|          12699|           null|  335987|       11|     MV Boucher|null|Right-hand bat|    Right-arm medium|South Africa|   Keeper|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             32|              null|          null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|
```

17

```
team_schema = StructType([
    StructField("team_sk", IntegerType(), True),
    StructField("team_id", IntegerType(), True),
    StructField("team_name", StringType(), True)
])

team_df = spark.read.schema(team_schema).format("csv").option("header","true").load("s3://ipl-data-analysis-bucket1/Team.csv")
```

18

```
team_df.show()
```

```
+-------+-------+--------------------+
|team_sk|team_id|           team_name|
+-------+-------+--------------------+
|      0|      1|Kolkata Knight Ri...|
|      1|      2|Royal Challengers...|
|      2|      3| Chennai Super Kings|
|      3|      4|     Kings XI Punjab|
|      4|      5|    Rajasthan Royals|
|      5|      6|    Delhi Daredevils|
|      6|      7|      Mumbai Indians|
|      7|      8|     Deccan Chargers|
|      8|      9|Kochi Tuskers Kerala|
|      9|     10|       Pune Warriors|
|     10|     11| Sunrisers Hyderabad|
|     11|     12|Rising Pune Super...|
|     12|     13|       Gujarat Lions|
```

```
+-------+-------+--------------------+
```

19

```
# Filtering data to include only valid deliveries (excluding extras like wides and no balls for specific analyses)
ball_by_ball_df = ball_by_ball_df.filter((col("wides") == 0) & (col("noballs")==0))

# Aggregation: Calculating the total and average runs scored in each match and inning
total_and_avg_runs = ball_by_ball_df.groupBy("match_id", "innings_no").agg(
    sum("runs_scored").alias("total_runs"),
    round(avg("runs_scored"), 2).alias("average_runs")
)
```

20

```
total_and_avg_runs.show()
```

```
+--------+----------+----------+------------+
|match_id|innings_no|total_runs|average_runs|
+--------+----------+----------+------------+
|  980940|         1|       138|        1.15|
|  419132|         1|       162|        1.35|
| 1082632|         2|       202|        1.92|
|  335993|         2|       131|        1.68|
| 1082617|         1|       123|        1.05|
|  980910|         2|       156|        1.47|
|  598057|         2|       141|        1.24|
|  980966|         1|       143|        1.19|
|  980982|         2|       169|        1.41|
|  419114|         2|       130|        1.12|
|  734042|         2|       146|         1.7|
|  829772|         2|       111|        1.34|
|  980934|         1|       130|        1.08|
|  733976|         1|       142|        1.39|
|  734018|         1|       132|         1.1|
|  501257|         2|       143|         1.4|
|  548328|         2|       123|        1.03|
|  733994|         1|       122|        1.02|
```

Calculating running total of runs in each match for each over

22

```
# Window Functions use:
window=Window.partitionBy("match_id","innings_no").orderBy("over_id")

ball_by_ball_df=ball_by_ball_df.withColumn("running_total_runs",sum("runs_scored").over(window))
```

23

```
ball_by_ball_df.show()
```

```
+--------+-------+-------+----------+------------+------------+----------------------+----------+-----------+----------+-----+-------+----+-------+-------+------------+-------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+-------------+-------------+-----------+-------------+----------+-----------------+
|match_id|over_id|ball_id|innings_no|team_batting|team_bowling|striker_batting_position|extra_type|runs_scored|extra_runs|wides|legbyes|byes|noballs|penalty|bowler_extras|      out_type|caught|bowle
d|run_out| lbw|retired_hurt|stumped|caught_and_bowled|hit_wicket|obstructingfeild|bowler_wicket|match_date|season|striker|non_striker|bowler|player_out|fielders|striker_match_sk|strikersk|nonstriker
_match_sk|nonstriker_sk|fielder_match_sk|fielder_sk|bowler_match_sk|bowler_sk|playerout_match_sk|battingteam_sk|bowlingteam_sk|keeper_catch|player_out_sk|matchdatesk|running_total_runs|
+--------+-------+-------+----------+------------+------------+----------------------+----------+-----------+----------+-----+-------+----+-------+-------+------------+-------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+-------------+-------------+-----------+-------------+----------+-----------------+
|  335987|      1|      1|         1|           1|           2|                     1|   legbyes|          0|         1|    0|      1|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|        null|      null|  2008|      1|          2|    14|      null|    null|          12705|        0|
12706|            1|             -1|        -1|          12702|       13|               -1|            0|             1|       null|           0|       null|                0|
|  335987|      1|      2|         1|           1|           2|                     2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|        null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|            0|             -1|        -1|          12702|       13|               -1|            0|             1|       null|           0|       null|                0|
|  335987|      1|      4|         1|           1|           2|                     2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|        null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|            0|             -1|        -1|          12702|       13|               -1|            0|             1|       null|           0|       null|                0|
|  335987|      1|      5|         1|           1|           2|                     2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|        null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|            0|             -1|        -1|          12702|       13|               -1|            0|             1|       null|           0|       null|                0|
```

24

```
# Conditional Column: Flag for high impact balls (either a wicket or more than 6 runs including extras)
ball_by_ball_df=ball_by_ball_df.withColumn(
    "high_impact",
    when((col("runs_scored") + col("extra_runs") > 6) | (col("bowler_wicket") == True),True).otherwise(False)
)
```

25

```
ball_by_ball_df.show()
```

```
+--------+-------+-------+----------+------------+------------+-----------------------+----------+-----------+----------+-----+-------+----+-------+-------+------------+-------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+-------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+-------------+-------------+-----------+-------------+----------+-----------------+-----------+
|match_id|over_id|ball_id|innings_no|team_batting|team_bowling|striker_batting_position|extra_type|runs_scored|extra_runs|wides|legbyes|byes|noballs|penalty|bowler_extras|      out_type|caught|bowle
d|run_out| lbw|retired_hurt|stumped|caught_and_bowled|hit_wicket|obstructingfeild|bowler_wicket|match_date|season|striker|non_striker|bowler|player_out|fielders|striker_match_sk|strikersk|nonstriker
_match_sk|nonstriker_sk|fielder_match_sk|fielder_sk|bowler_match_sk|bowler_sk|playerout_match_sk|battingteam_sk|bowlingteam_sk|keeper_catch|player_out_sk|matchdatesk|running_total_runs|high_impact|
+--------+-------+-------+----------+------------+------------+-----------------------+----------+-----------+----------+-----+-------+----+-------+-------+------------+-------------+------+-----
-+-------+----+------------+-------+-----------------+----------+---------------+-------------+----------+------+-------+-----------+------+----------+--------+---------------+---------+----------
---------+-------------+----------------+----------+--------------+---------+-----------------+-------------+-------------+-----------+-------------+----------+-----------------+-----------+
|  335987|      1|      1|         1|           1|           2|                      1|   legbyes|          0|         1|    0|      1|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2008|      1|          2|    14|      null|    null|          12705|        0|
12706|           1|             -1|        -1|         12702|      13|               -1|            0|           1|       null|           0|       null|                0|      false|
|  335987|      1|      2|         1|           1|           2|                      2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|           0|             -1|        -1|         12702|      13|               -1|            0|           1|       null|           0|       null|                0|      false|
|  335987|      1|      4|         1|           1|           2|                      2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|           0|             -1|        -1|         12702|      13|               -1|            0|           1|       null|           0|       null|                0|      false|
|  335987|      1|      5|         1|           1|           2|                      2| No Extras|          0|         0|    0|      0|   0|      0|      0|           0|Not Applicable|  null|  nul
l|   null|null|        null|   null|             null|      null|           null|         null|      null|  2008|      2|          1|    14|      null|    null|          12706|        1|
12705|           0|             -1|        -1|         12702|      13|               -1|            0|           1|       null|           0|       null|                0|      false|
```

Starting with data cleaning and transformation

```
# Extracting year, month, and day from the match date for more detailed time-based analysis
match_df = match_df.withColumn("year", year("match_date"))
match_df = match_df.withColumn("month", month("match_date"))
match_df = match_df.withColumn("day", dayofmonth("match_date"))

# High margin win: categorizing win margins into 'high', 'medium', and 'low'
match_df = match_df.withColumn(
    "win_margin_category",
    when(col("win_margin") >= 100, "High")
    .when((col("win_margin") >= 50) & (col("win_margin") < 100), "Medium")
    .otherwise("Low")
)

# Analyze the impact of the toss: who wins the toss and the match
match_df = match_df.withColumn(
    "toss_match_winner",
    when(col("toss_winner") == col("match_winner"), "Yes").otherwise("No")
)

# Show the enhanced match DataFrame
match_df.show()
```

```
+--------+--------+--------------------+--------------------+----------+-----------+--------------------+----------+------------+--------------------+--------------------+---------+--------+--------
----+-------------+----------+----------+----+-----+----+-------------------+-----------------+
|match_sk|match_id|               team1|               team2|match_date|season_year|          venue_name| city_name|country_name|        toss_winner|        match_winner|toss_name|win_type|outcome_
type|    manofmach|win_margin|country_id|year|month| day|win_margin_category|toss_match_winner|
+--------+--------+--------------------+--------------------+----------+-----------+--------------------+----------+------------+--------------------+--------------------+---------+--------+--------
----+-------------+----------+----------+----+-----+----+-------------------+-----------------+
|       0|  335987|Royal Challengers...|Kolkata Knight Ri...|      null|       2008|M Chinnaswamy Sta...| Bangalore|       India|Royal Challengers...|Kolkata Knight Ri...|    field|    runs|      Re
sult|  BB McCullum|       140|         1|null| null|null|               High|               No|
|       1|  335988|     Kings XI Punjab| Chennai Super Kings|      null|       2008|Punjab Cricket As...|Chandigarh|       India| Chennai Super Kings| Chennai Super Kings|      bat|    runs|      Re
sult|   MEK Hussey|        33|         1|null| null|null|                Low|              Yes|
|       2|  335989|    Delhi Daredevils|    Rajasthan Royals|      null|       2008|    Feroz Shah Kotla|     Delhi|       India|    Rajasthan Royals|    Delhi Daredevils|      bat| wickets|      Re
sult|  MF Maharoof|         9|         1|null| null|null|                Low|               No|
|       3|  335990|      Mumbai Indians|Royal Challengers...|      null|       2008|    Wankhede Stadium|    Mumbai|       India|      Mumbai Indians|Royal Challengers...|      bat| wickets|      Re
sult|   MV Boucher|         5|         1|null| null|null|                Low|               No|
|       4|  335991|Kolkata Knight Ri...|     Deccan Chargers|      null|       2008|        Eden Gardens|   Kolkata|       India|     Deccan Chargers|Kolkata Knight Ri...|      bat| wickets|      Re
sult|    DJ Hussey|         5|         1|null| null|null|                Low|               No|
|       5|  335992|    Rajasthan Royals|     Kings XI Punjab|      null|       2008|Sawai Mansingh St...|    Jaipur|       India|     Kings XI Punjab|    Rajasthan Royals|      bat| wickets|      Re
sult|    SR Watson|         6|         1|null| null|null|                Low|               No|
|       6|  335993|     Deccan Chargers|    Delhi Daredevils|      null|       2008|Rajiv Gandhi Inte...| Hyderabad|       India|     Deccan Chargers|    Delhi Daredevils|      bat| wickets|      Re
sult|     V Sehwag|         9|         1|null| null|null|                Low|               No|
|       7|  335994| Chennai Super Kings|      Mumbai Indians|      null|       2008|MA Chidambaram St...|   Chennai|       India|      Mumbai Indians| Chennai Super Kings|    field|    runs|      Re
```

```
# Normalizing and cleaning player names
player_df = player_df.withColumn("player_name", lower(regexp_replace("player_name", "[^a-zA-Z0-9 ]", "")))

# Handle missing values in 'batting_hand' and 'bowling_skill' with a default 'unknown'
player_df = player_df.na.fill({"batting_hand": "unknown", "bowling_skill": "unknown"})

# Categorizing players based on batting hand
player_df = player_df.withColumn(
    "batting_style",
    when(col("batting_hand").contains("left"), "Left-Handed").otherwise("Right-Handed")
)

# Show the modified player DataFrame
player_df.show()
```

```
+---------+---------+---------------+----+-------------+--------------------+------------+-------------+
|player_sk|player_id|    player_name| dob|  batting_hand|       bowling_skill|country_name|batting_style|
+---------+---------+---------------+----+-------------+--------------------+------------+-------------+
|        0|        1|     sc ganguly|null| Left-hand bat|    Right-arm medium|       India| Right-Handed|
|        1|        2|    bb mccullum|null|Right-hand bat|    Right-arm medium| New Zealand| Right-Handed|
|        2|        3|     rt ponting|null|Right-hand bat|    Right-arm medium|   Australia| Right-Handed|
|        3|        4|      dj hussey|null|Right-hand bat|  Right-arm offbreak|   Australia| Right-Handed|
|        4|        5|mohammad hafeez|null|Right-hand bat|  Right-arm offbreak|    Pakistan| Right-Handed|
|        5|        6|       r dravid|null|Right-hand bat|  Right-arm offbreak|       India| Right-Handed|
|        6|        7|       w jaffer|null|Right-hand bat|  Right-arm offbreak|       India| Right-Handed|
|        7|        8|        v kohli|null|Right-hand bat|    Right-arm medium|       India| Right-Handed|
|        8|        9|      jh kallis|null|Right-hand bat|Right-arm fast-me...|South Africa| Right-Handed|
|        9|       10|       cl white|null|Right-hand bat|     Legbreak googly|   Australia| Right-Handed|
|       10|       11|     mv boucher|null|Right-hand bat|    Right-arm medium|South Africa| Right-Handed|
|       11|       12|        b akhil|null|Right-hand bat|Right-arm medium-...|       India| Right-Handed|
|       12|       13|      aa noffke|null|Right-hand bat|Right-arm fast-me...|   Australia| Right-Handed|
|       13|       14|        p kumar|null|Right-hand bat|    Right-arm medium|       India| Right-Handed|
|       14|       15|         z khan|null|Right-hand bat|Left-arm fast-medium|       India| Right-Handed|
|       15|       16|       sb joshi|null| Left-hand bat|Slow left-arm ort...|       India| Right-Handed|
|       16|       17|       pa patel|null| Left-hand bat|                 N/A|       India| Right-Handed|
|       17|       18|      ml hayden|null| Left-hand bat|    Right-arm medium|   Australia| Right-Handed|
```

```
# Adding a 'veteran_status' column based on player age
player_match_df = player_match_df.withColumn(
    "veteran_status",
    when(col("age_as_on_match") >= 35, "Veteran").otherwise("Non-Veteran")
)

# Dynamic column to calculate years since debut
player_match_df = player_match_df.withColumn(
    "years_since_debut",
    (year(current_date()) - col("season_year"))
)

# Show the enriched DataFrame
player_match_df.show()
```

```
+---------------+---------------+--------+---------+---------------+----+-------------+-------------------+------------+---------+-------------------+-------------------+-----------+------------
----+---------------+-----------------+-------------+-------------+-------------+---------------+-------------+-------------+-------------+-----------------+
|player_match_sk|playermatch_key|match_id|player_id|    player_name| dob|  batting_hand|       bowling_skill|country_name|role_desc|        player_team|       opposit_team|season_year|is_manofthem
atch|age_as_on_match|isplayers_team_won|batting_status|bowling_status|player_captain|opposit_captain|player_keeper|opposit_keeper|veteran_status|years_since_debut|
+---------------+---------------+--------+---------+---------------+----+-------------+-------------------+------------+---------+-------------------+-------------------+-----------+------------
----+---------------+-----------------+-------------+-------------+-------------+---------------+-------------+-------------+-------------+-----------------+
|             -1|             -1|      -1|       -1|            N/A|null|         null|               null|        null|     null|               null|               null|       null|
null|           null|             null|         null|          null|         null|           null|         null|          null|   Non-Veteran|             null|
|          12694|           null|  335987|        6|       R Dravid|null|Right-hand bat|  Right-arm offbreak|       India|  Captain|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             35|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|       Veteran|               16|
|          12695|           null|  335987|        7|       W Jaffer|null|Right-hand bat|  Right-arm offbreak|       India|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             30|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|   Non-Veteran|               16|
|          12696|           null|  335987|        8|        V Kohli|null|Right-hand bat|    Right-arm medium|       India|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             20|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|   Non-Veteran|               16|
|          12697|           null|  335987|        9|      JH Kallis|null|Right-hand bat|Right-arm fast-me...|South Africa|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             33|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|   Non-Veteran|               16|
|          12698|           null|  335987|       10|       CL White|null|Right-hand bat|     Legbreak googly|   Australia|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             25|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|   Non-Veteran|               16|
|          12699|           null|  335987|       11|     MV Boucher|null|Right-hand bat|    Right-arm medium|South Africa|   Keeper|Royal Challengers...|Kolkata Knight Ri...|       2008|
null|             32|             null|         null|          null|      R Dravid|     SC Ganguly|   MV Boucher|       WP Saha|   Non-Veteran|               16|
|          12700|           null|  335987|       12|        B Akhil|null|Right-hand bat|Right-arm medium-...|       India|   Player|Royal Challengers...|Kolkata Knight Ri...|       2008|
```

```
### creating temporary views to further write sql queries for analysis
ball_by_ball_df.createOrReplaceTempView("ball_by_ball")
match_df.createOrReplaceTempView("match")
player_df.createOrReplaceTempView("player")
player_match_df.createOrReplaceTempView("player_match")
team_df.createOrReplaceTempView("team")
```

31

```
ball_by_ball_df.columns
```

```
Out[73]: ['match_id',
 'over_id',
 'ball_id',
 'innings_no',
 'team_batting',
 'team_bowling',
 'striker_batting_position',
 'extra_type',
 'runs_scored',
 'extra_runs',
 'wides',
 'legbyes',
 'byes',
 'noballs',
 'penalty',
 'bowler_extras',
 'out_type',
 'caught',
 'bowled',
 'run_out',
 'lbw',
```

Identifying top scoring batsman for particular seasons

33

```
##Using  spark sql to query and retreive data
top_scoring_batsmen_per_season = spark.sql("""
SELECT
p.player_name,
m.season_year,
SUM(b.runs_scored) AS total_runs
FROM ball_by_ball b
JOIN match m ON b.match_id = m.match_id
JOIN player_match pm ON m.match_id = pm.match_id AND b.striker = pm.player_id
JOIN player p ON p.player_id = pm.player_id
GROUP BY p.player_name, m.season_year
ORDER BY m.season_year, total_runs DESC
""")
```



```
top_scoring_batsmen_per_season.show(30)
```

```
+---------------+-----------+----------+
|    player_name|season_year|total_runs|
+---------------+-----------+----------+
|       se marsh|       2008|       614|
|      g gambhir|       2008|       532|
|  st jayasuriya|       2008|       508|
|      sr watson|       2008|       463|
|       gc smith|       2008|       437|
|   ac gilchrist|       2008|       431|
|      yk pathan|       2008|       430|
|       sk raina|       2008|       420|
|       ms dhoni|       2008|       414|
|       v sehwag|       2008|       399|
|      rg sharma|       2008|       399|
|       r dravid|       2008|       370|
|     sc ganguly|       2008|       349|
|       s dhawan|       2008|       340|
|  kc sangakkara|       2008|       319|
|      dj hussey|       2008|       318|
|     rv uthappa|       2008|       316|
|    sa asnodkar|       2008|       311|
```

Identifying Economical Bowlers During Powerplay Over

36

```
economical_bowlers_powerplay = spark.sql("""
SELECT
p.player_name,
AVG(b.runs_scored) AS avg_runs_per_ball,
COUNT(b.bowler_wicket) AS total_wickets
FROM ball_by_ball b
JOIN player_match pm ON b.match_id = pm.match_id AND b.bowler = pm.player_id
JOIN player p ON pm.player_id = p.player_id
WHERE b.over_id <= 6
GROUP BY p.player_name
HAVING COUNT(*) >= 1
ORDER BY avg_runs_per_ball, total_wickets DESC
""")
economical_bowlers_powerplay.show()
```

```
+---------------+------------------+-------------+
|    player_name| avg_runs_per_ball|total_wickets|
+---------------+------------------+-------------+
|     sm harwood|0.3333333333333333|            0|
|     ankit soni|               0.5|            0|
|      gr napier|               0.5|            0|
|       aj finch|               0.5|            0|
|        a zampa|               0.5|            0|
|     avesh khan|               0.5|            0|
|       nb singh|0.5833333333333334|            0|
|     ag murtaza|0.6538461538461539|            0|
|      sb bangar|0.6666666666666666|            0|
|     d du preez|0.6666666666666666|            0|
|        s gopal|0.6666666666666666|            0|
|     fh edwards|0.6923076923076923|            0|
|       a kumble|0.7685185185185185|            0|
|j syed mohammad|0.7777777777777778|            0|
|   kp pietersen|0.7777777777777778|            0|
|       umar gul|0.7777777777777778|            0|
|  la carseldine|0.8333333333333334|            0|
|    rj peterson|0.8333333333333334|            0|
```

Analyzing Toss Impact on Individual Match Outcomes

38

```
toss_impact_individual_matches = spark.sql("""
SELECT m.match_id, m.toss_winner, m.toss_name, m.match_winner,
       CASE WHEN m.toss_winner = m.match_winner THEN 'Won' ELSE 'Lost' END AS match_outcome
FROM match m
WHERE m.toss_name IS NOT NULL
ORDER BY m.match_id
""")
toss_impact_individual_matches.show()
```

```
+--------+--------------------+---------+--------------------+-------------+
|match_id|         toss_winner|toss_name|        match_winner|match_outcome|
+--------+--------------------+---------+--------------------+-------------+
|  335987|Royal Challengers...|    field|Kolkata Knight Ri...|         Lost|
|  335988| Chennai Super Kings|      bat| Chennai Super Kings|          Won|
|  335989|    Rajasthan Royals|      bat|    Delhi Daredevils|         Lost|
|  335990|      Mumbai Indians|      bat|Royal Challengers...|         Lost|
|  335991|     Deccan Chargers|      bat|Kolkata Knight Ri...|         Lost|
|  335992|     Kings XI Punjab|      bat|    Rajasthan Royals|         Lost|
|  335993|     Deccan Chargers|      bat|    Delhi Daredevils|         Lost|
|  335994|      Mumbai Indians|    field| Chennai Super Kings|         Lost|
|  335995|    Rajasthan Royals|    field|    Rajasthan Royals|          Won|
|  335996|      Mumbai Indians|    field|     Kings XI Punjab|         Lost|
|  335997|    Rajasthan Royals|    field|    Rajasthan Royals|          Won|
|  335998|Kolkata Knight Ri...|      bat| Chennai Super Kings|         Lost|
|  335999|     Deccan Chargers|    field|     Deccan Chargers|          Won|
|  336000|    Delhi Daredevils|      bat|     Kings XI Punjab|         Lost|
|  336001| Chennai Super Kings|      bat| Chennai Super Kings|          Won|
|  336002|Kolkata Knight Ri...|      bat|      Mumbai Indians|         Lost|
|  336003|Royal Challengers...|    field|    Delhi Daredevils|         Lost|
|  336004|     Kings XI Punjab|    field|     Kings XI Punjab|          Won|
```

Calculating Average Runs Scored by Players in Matches They Won

```
average_runs_in_wins = spark.sql("""
SELECT p.player_name, AVG(b.runs_scored) AS avg_runs_in_wins, COUNT(*) AS innings_played
FROM ball_by_ball b
JOIN player_match pm ON b.match_id = pm.match_id AND b.striker = pm.player_id
JOIN player p ON pm.player_id = p.player_id
JOIN match m ON pm.match_id = m.match_id
WHERE m.match_winner = pm.player_team
GROUP BY p.player_name
ORDER BY avg_runs_in_wins ASC
""")
average_runs_in_wins.show()
```

```
+--------------+-------------------+--------------+
|   player_name|   avg_runs_in_wins|innings_played|
+--------------+-------------------+--------------+
|       a nehra|                0.0|             2|
|    kp appanna|                0.0|             1|
|     jj bumrah|                0.0|             2|
|      i sharma|                0.0|             1|
|      ts mills|                0.0|             3|
|      j theron|                0.0|             1|
|      vr aaron|                0.0|             5|
|     sn thakur|                0.0|             2|
| anirudh singh|                0.0|             1|
|    t thushara|                0.2|             5|
|     sa abbott|               0.25|             4|
| yashpal singh| 0.3157894736842105|            19|
|   s sreesanth| 0.3333333333333333|             3|
|   kc cariappa| 0.3333333333333333|             3|
|    jd unadkat|                0.4|             5|
|    sm harwood|0.42857142857142855|             7|
|      r shukla|                0.5|             2|
|       jm kemp|                0.5|             8|
```

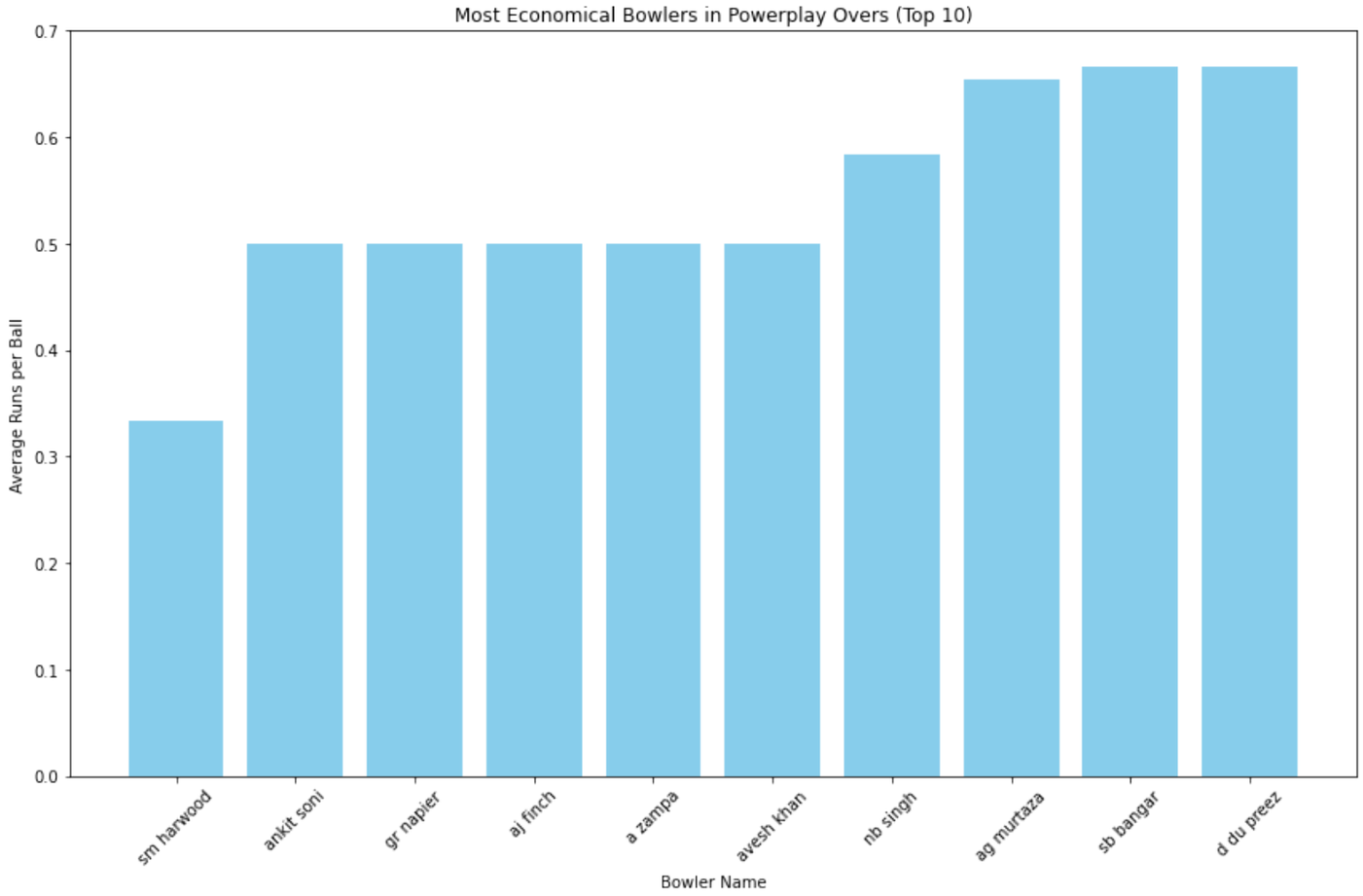
Starting with Data Visulization


```
## importing libraries
import matplotlib.pyplot as plt
import seaborn as sns
```

Visualizing the Most Economical Bowlers in Powerplay Overs

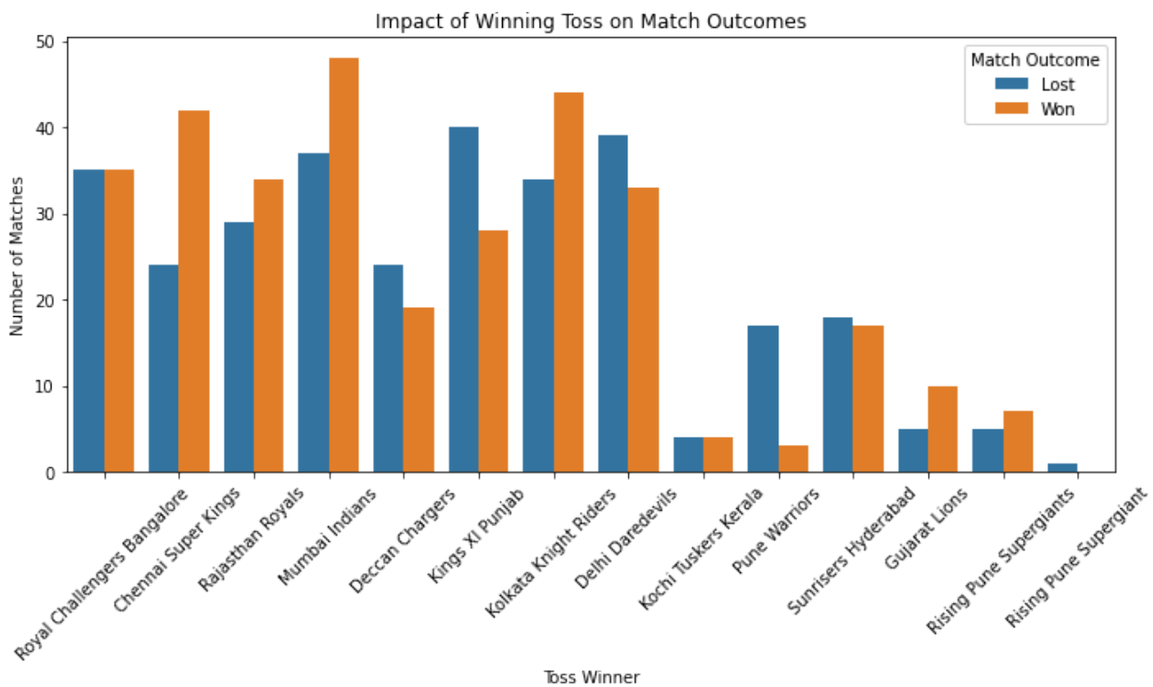
44

```
# Assuming 'economical_bowlers_powerplay' is already executed and available as a Spark DataFrame
economical_bowlers_pd = economical_bowlers_powerplay.toPandas()

# Visualizing using Matplotlib
plt.figure(figsize=(12, 8))
# Limiting to top 10 for clarity in the plot
top_economical_bowlers = economical_bowlers_pd.nsmallest(10, 'avg_runs_per_ball')
plt.bar(top_economical_bowlers['player_name'], top_economical_bowlers['avg_runs_per_ball'], color='skyblue')
plt.xlabel('Bowler Name')
plt.ylabel('Average Runs per Ball')
plt.title('Most Economical Bowlers in Powerplay Overs (Top 10)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



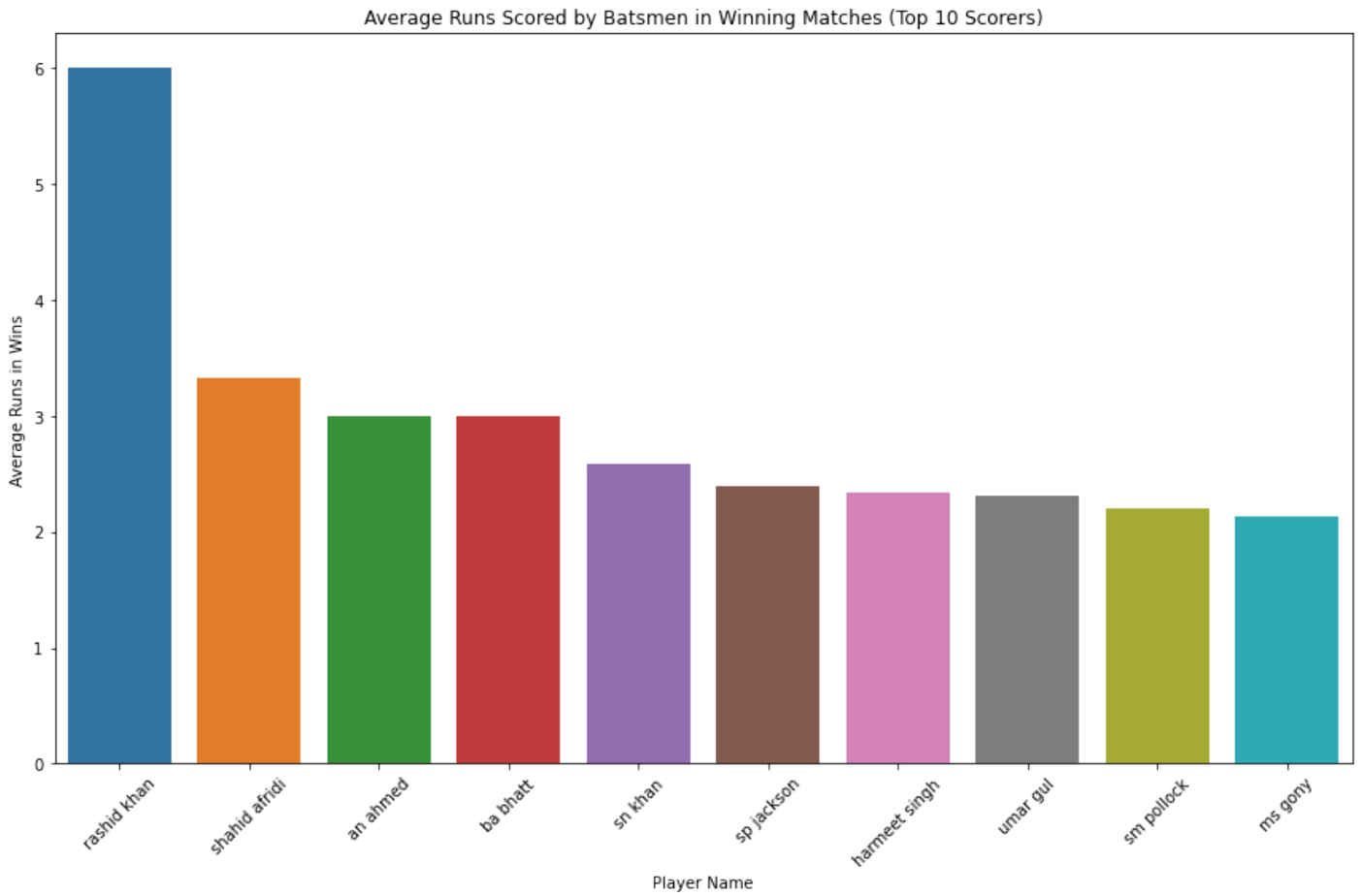
Visualizing the Impact of Winning the Toss on Match Outcomes

```
##Converting spark dataframe to pandas
toss_impact_pd = toss_impact_individual_matches.toPandas()

# Creating a countplot to show win/loss after winning toss
plt.figure(figsize=(10, 6))
sns.countplot(x='toss_winner', hue='match_outcome', data=toss_impact_pd)
plt.title('Impact of Winning Toss on Match Outcomes')
plt.xlabel('Toss Winner')
plt.ylabel('Number of Matches')
plt.legend(title='Match Outcome')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Visualizing Average Runs Scored by Batsmen in Winning Matches

```
## converting average_runs_pd dataframe to pandas
average_runs_pd = average_runs_in_wins.toPandas()

# Using seaborn to plot average runs in winning matches
plt.figure(figsize=(12, 8))
top_scorers = average_runs_pd.nlargest(10, 'avg_runs_in_wins')
sns.barplot(x='player_name', y='avg_runs_in_wins', data=top_scorers)
plt.title('Average Runs Scored by Batsmen in Winning Matches (Top 10 Scorers)')
plt.xlabel('Player Name')
plt.ylabel('Average Runs in Wins')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Analyzing Average and Highest Scores by Venue

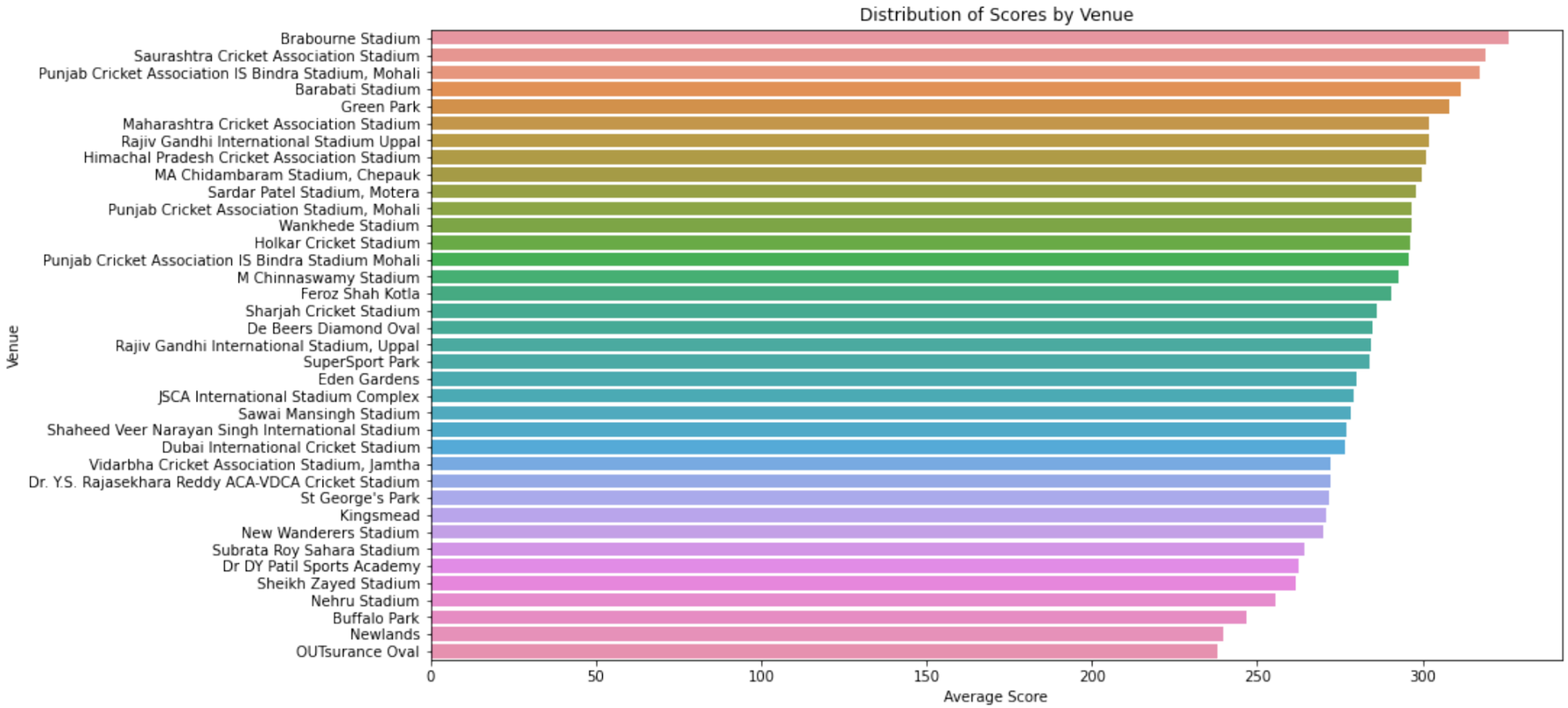
50

```
# Executing Spark SQL Query
scores_by_venue = spark.sql("""
SELECT venue_name, AVG(total_runs) AS average_score, MAX(total_runs) AS highest_score
FROM (
    SELECT ball_by_ball.match_id, match.venue_name, SUM(runs_scored) AS total_runs
    FROM ball_by_ball
    JOIN match ON ball_by_ball.match_id = match.match_id
    GROUP BY ball_by_ball.match_id, match.venue_name
)
GROUP BY venue_name
ORDER BY average_score DESC
""")
```

Visualizing the Distribution of Scores by Venue

52

```
# Converting to Pandas DataFrame
scores_by_venue_pd = scores_by_venue.toPandas()

# Plotting the average scores by venue
plt.figure(figsize=(14, 8))
sns.barplot(x='average_score', y='venue_name', data=scores_by_venue_pd)
plt.title('Distribution of Scores by Venue')
plt.xlabel('Average Score')
plt.ylabel('Venue')
plt.show()
```

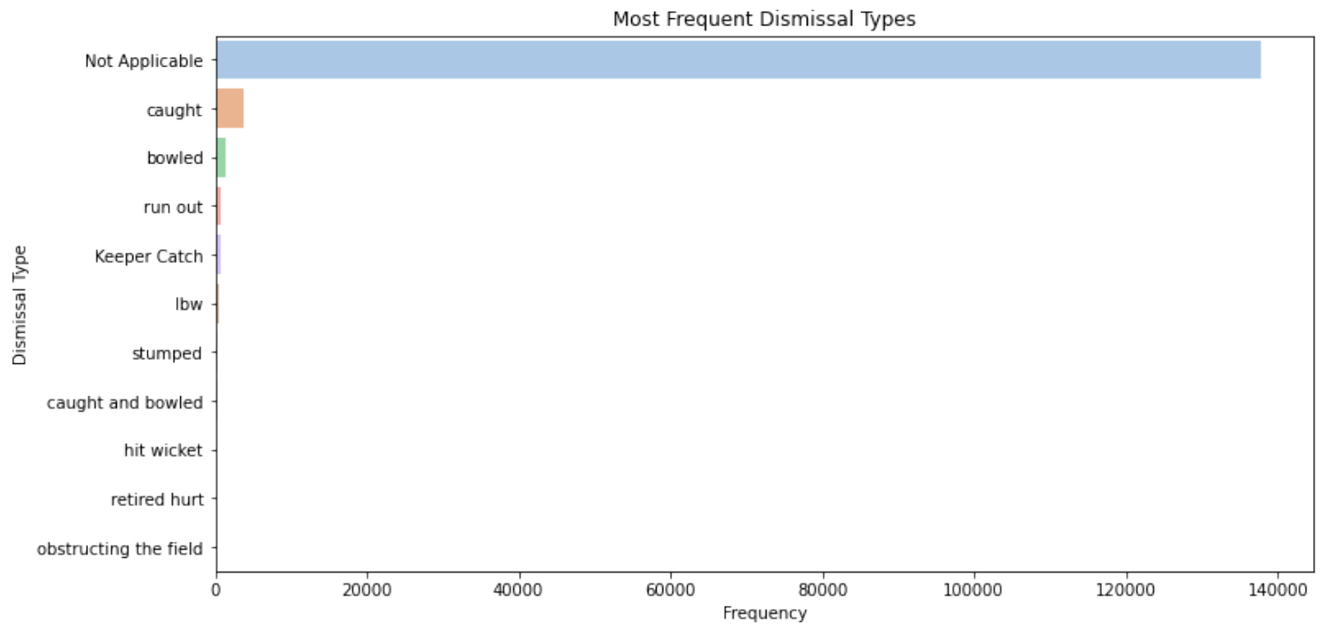


Analysis of Dismissal Types in IPL Matches


```
# Execute SQL Query
dismissal_types = spark.sql("""
SELECT out_type, COUNT(*) AS frequency
FROM ball_by_ball
WHERE out_type IS NOT NULL
GROUP BY out_type
ORDER BY frequency DESC
""")
```

55

```
# Convert to Pandas DataFrame
dismissal_types_pd = dismissal_types.toPandas()

# Plot
plt.figure(figsize=(12, 6))
sns.barplot(x='frequency', y='out_type', data=dismissal_types_pd, palette='pastel')
plt.title('Most Frequent Dismissal Types')
plt.xlabel('Frequency')
plt.ylabel('Dismissal Type')
plt.show()
```



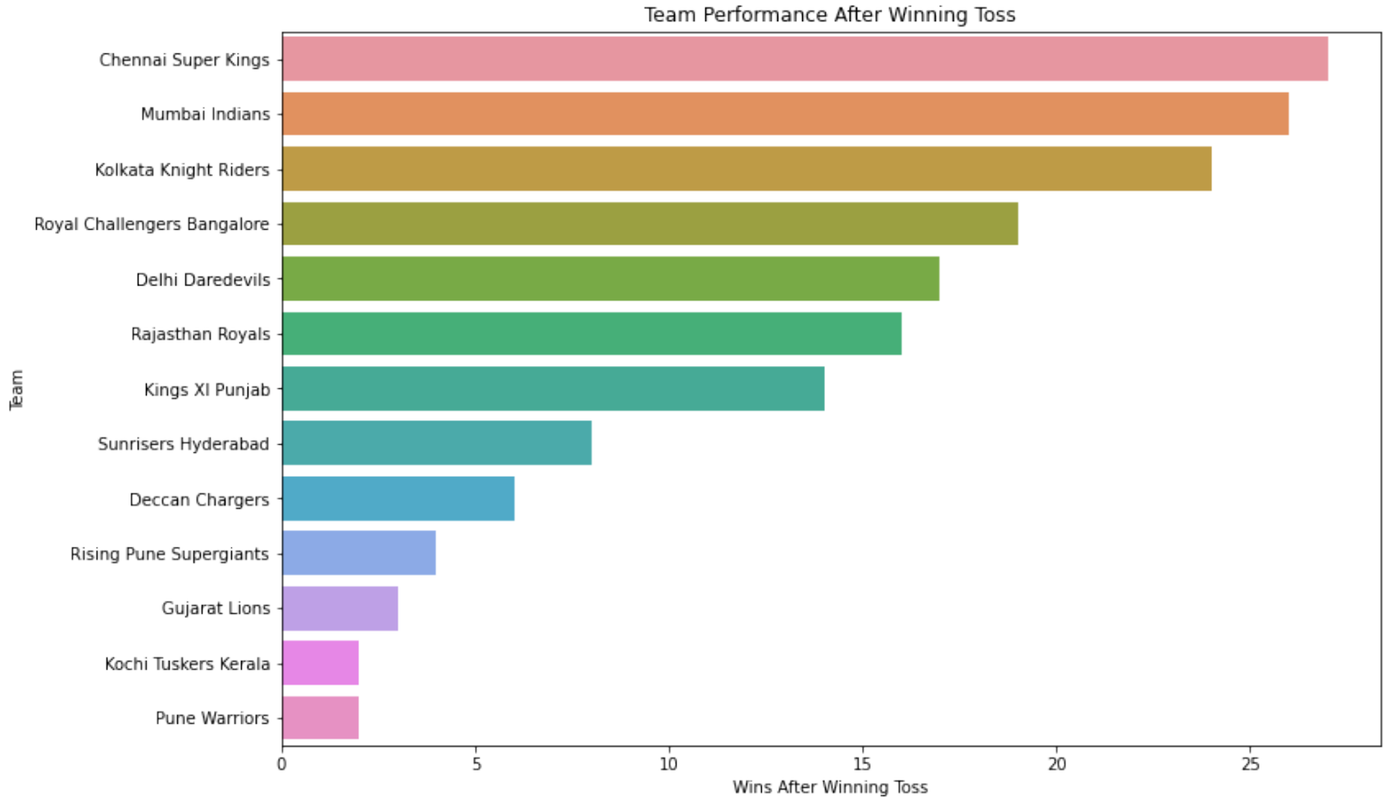
Team Performance After Winning Toss

57

```
# Execute SQL Query
team_toss_win_performance = spark.sql("""
SELECT team1, COUNT(*) AS matches_played, SUM(CASE WHEN toss_winner = match_winner THEN 1 ELSE 0 END) AS wins_after_toss
FROM match
WHERE toss_winner = team1
GROUP BY team1
ORDER BY wins_after_toss DESC
""")
```



```
# Convert to Pandas DataFrame
team_toss_win_pd = team_toss_win_performance.toPandas()

# Plot
plt.figure(figsize=(12, 8))
sns.barplot(x='wins_after_toss', y='team1', data=team_toss_win_pd)
plt.title('Team Performance After Winning Toss')
plt.xlabel('Wins After Winning Toss')
plt.ylabel('Team')
plt.show()
```



59

```
# SQL Query to identify clutch performances
clutch_performances = spark.sql("""
SELECT
    p.player_name,
    COUNT(*) AS clutch_innings
FROM ball_by_ball b
JOIN player_match pm ON b.match_id = pm.match_id AND b.striker = pm.player_id
JOIN player p ON pm.player_id = p.player_id
WHERE b.over_id > 18 -- Last 2 overs considered clutch
AND b.runs_scored >= 6 -- Assuming scoring 6 or more in clutch situations
GROUP BY p.player_name
ORDER BY clutch_innings DESC
""")
clutch_performances.show()
```

```
+--------------+-------------+
|    player_name|clutch_innings|
+--------------+-------------+
|       ms dhoni|           54|
|      rg sharma|           36|
|     ka pollard|           31|
| ab de villiers|           29|
|        v kohli|           22|
|   yuvraj singh|           21|
|       dj bravo|           21|
|      ja morkel|           20|
|harbhajan singh|           20|
|      jp duminy|           18|
|      da miller|           17|
|      ik pathan|           17|
|       bj hodge|           16|
|      hh pandya|           15|
|      at rayudu|           14|
|      ra jadeja|           14|
|    jp faulkner|           14|
|      km jadhav|           13|
```

60

```
+--------------------+-------------+----------+----------+
|          venue_name|matches_played|wins_team1|wins_team2|
+--------------------+-------------+----------+----------+
|M Chinnaswamy Sta...|           66|        32|        32|
|        Eden Gardens|           61|        38|        23|
|    Feroz Shah Kotla|           60|        26|        33|
|    Wankhede Stadium|           57|        34|        23|
|MA Chidambaram St...|           48|        34|        14|
|Rajiv Gandhi Inte...|           41|        17|        24|
|Punjab Cricket As...|           35|        18|        17|
|Sawai Mansingh St...|           33|        24|         9|
|Dr DY Patil Sport...|           17|         7|        10|
|Subrata Roy Sahar...|           17|         4|        13|
|Maharashtra Crick...|           15|         6|         9|
|           Kingsmead|           15|         8|         7|
|Sardar Patel Stad...|           12|         7|         5|
|     SuperSport Park|           12|         6|         6|
|   Brabourne Stadium|           11|         9|         2|
|Dr. Y.S. Rajasekh...|           11|         4|         7|
|Saurashtra Cricke...|           10|         3|         6|
```

| Himachal Pradesh | | 9 | 5 | 4 |