pyspark-ipl-dataset

November 10, 2024

1 Steps to Mount S3 Bucket in Databricks

[0]:
```
# Define AWS Access Key and Secret Key
ACCESS_KEY = "AKIA6G75DSLTO5DTGHFD"
SECRET_KEY = "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"

# Configure SparkContext with S3 credentials
sc._jsc.hadoopConfiguration().set("fs.s3a.access.key", ACCESS_KEY)
sc._jsc.hadoopConfiguration().set("fs.s3a.secret.key", SECRET_KEY)

# Set the S3 endpoint for the region (assuming eu-north-1 for Stockholm)
aws_region = "eu-north-1"
sc._jsc.hadoopConfiguration().set("fs.s3a.endpoint", f"s3.{aws_region}.
 ↪amazonaws.com")

# Paths for each file in the S3 bucket
ball_by_ball_path = "s3a://ipldatasetpyspark/Ball_By_Ball.csv"
match_path = "s3a://ipldatasetpyspark/Match.csv"
player_match_path = "s3a://ipldatasetpyspark/Player_match.csv"
player_path = "s3a://ipldatasetpyspark/Player.csv"
team_path = "s3a://ipldatasetpyspark/Team.csv"

# Read each file and save in separate DataFrames
ball_by_ball_df = spark.read.csv(ball_by_ball_path, inferSchema=True,␣
 ↪header=True)
match_df = spark.read.csv(match_path, inferSchema=True, header=True)
player_match_df = spark.read.csv(player_match_path, inferSchema=True,␣
 ↪header=True)
player_df = spark.read.csv(player_path, inferSchema=True, header=True)
team_df = spark.read.csv(team_path, inferSchema=True, header=True)

# Display the first few rows of each DataFrame to confirm they loaded correctly
print("Ball By Ball DataFrame:")
ball_by_ball_df.show(5)

print("Match DataFrame:")
match_df.show(5)
```

```
print("Player Match DataFrame:")
player_match_df.show(5)

print("Player DataFrame:")
player_df.show(5)

print("Team DataFrame:")
team_df.show(5)
```

```
Ball By Ball DataFrame:
+--------+-------+-------+----------+-----------+-----------+----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
----------+-------------+------+------+-------+---+-----------+-------+-------
----------+----------+---------------+-------------+----------+------+-------+-
----------+------+----------+--------+----------------+---------+--------------
----+-------------+---------------+----------+---------------+---------+-------
-----------+--------------+--------------+------------+-------------+-----------
+
|MatcH_id|Over_id|Ball_id|Innings_No|Team_Batting|Team_Bowling|Striker_Batting_P
osition|Extra_Type|Runs_Scored|Extra_runs|Wides|Legbyes|Byes|Noballs|Penalty|Bow
ler_Extras|      Out_type|Caught|Bowled|Run_out|LBW|Retired_hurt|Stumped|caught_
and_bowled|hit_wicket|ObstructingFeild|Bowler_Wicket|Match_Date|Season|Striker|N
on_Striker|Bowler|Player_Out|Fielders|Striker_match_SK|StrikerSK|NonStriker_matc
h_SK|NONStriker_SK|Fielder_match_SK|Fielder_SK|Bowler_match_SK|BOWLER_SK|PlayerO
ut_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|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|    277|
104|    83|      null|    null|           20336|      276|              20333|
103|              -1|        -1|          20343|       82|                -1|
4|             1|           0|            0|   20130420|
|  598028|     14|      1|         1|           5|           2|
5| No Extras|          1|         0|    0|      0|   0|      0|      0|
0|Not Applicable|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|    104|
6|   346|      null|    null|           20333|      103|              20328|
```

```
5|              -1|        -1|          20348|      345|                -1|
4|             1|           0|            0|   20130420|
|  598028|     14|      2|         1|           5|          2|
3| No Extras|          1|         0|    0|      0|   0|      0|      0|
0|Not Applicable|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|      6|
104|   346|      null|    null|           20328|        5|               20333|
103|              -1|        -1|          20348|      345|                -1|
4|             1|           0|            0|   20130420|
|  598028|     14|      3|         1|           5|          2|
5| No Extras|          1|         0|    0|      0|   0|      0|      0|
0|Not Applicable|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|    104|
6|   346|      null|    null|           20333|      103|               20328|
5|              -1|        -1|          20348|      345|                -1|
4|             1|           0|            0|   20130420|
|  598028|     14|      4|         1|           5|          2|
3| No Extras|          0|         0|    0|      0|   0|      0|      0|
0|Not Applicable|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|      6|
104|   346|      null|    null|           20328|        5|               20333|
103|              -1|        -1|          20348|      345|                -1|
4|             1|           0|            0|   20130420|
+--------+-------+-------+----------+-----------+-----------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
----------+--------------+------+------+-------+---+------------+-------+-------
----------+----------+---------------+-------------+----------+------+-------+-
----------+------+----------+--------+----------------+---------+--------------
----+-------------+---------------+----------+---------------+---------+-------
-----------+-------------+--------------+-----------+------------+-----------
+
only showing top 5 rows

Match DataFrame:
+--------+--------+-------------------+-------------------+----------+--------
---+--------------------+----------+-----------+-------------------+----------
----------+---------+--------+------------+-----------+----------+----------+
|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…|2008-04-18|
2008|M Chinnaswamy Sta…| Bangalore|       India|Royal Challengers…|Kolkata
Knight Ri…|    field|    runs|      Result|BB McCullum|       140|         1|
|       1|  335988|     Kings XI Punjab| Chennai Super Kings|2008-04-19|
```

```
2008|Punjab Cricket As…|Chandigarh|       India| Chennai Super Kings| Chennai
Super Kings|      bat|    runs|      Result| MEK Hussey|        33|         1|
|       2|  335989|    Delhi Daredevils|    Rajasthan Royals|2008-04-19|
2008|    Feroz Shah Kotla|     Delhi|       India|    Rajasthan Royals|    Delhi
Daredevils|      bat| wickets|      Result|MF Maharoof|         9|         1|
|       3|  335990|      Mumbai Indians|Royal Challengers…|2008-04-20|
2008|    Wankhede Stadium|    Mumbai|       India|      Mumbai Indians|Royal
Challengers…|      bat| wickets|      Result| MV Boucher|         5|
1|
|       4|  335991|Kolkata Knight Ri…|     Deccan Chargers|2008-04-20|
2008|        Eden Gardens|   Kolkata|       India|     Deccan Chargers|Kolkata
Knight Ri…|      bat| wickets|      Result|  DJ Hussey|         5|         1|
+--------+--------+--------------------+--------------------+----------+--------
---+--------------------+----------+------------+--------------------+----------
----------+---------+--------+------------+-----------+----------+-----------+
only showing top 5 rows

Player Match DataFrame:
+---------------+---------------+--------+---------+-----------+----------+-----
---------+--------------------+------------+---------+--------------------+-----
--------------+-----------+----------------+---------------+------------------+
--------------+--------------+--------------+---------------+-------------+-----
---------+
|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_manofThematch|Age_As_on_match|IsPlayers_Team_won|Bat
ting_Status|Bowling_Status|Player_Captain|Opposit_captain|Player_keeper|Opposit_
keeper|
+---------------+---------------+--------+---------+-----------+----------+-----
---------+--------------------+------------+---------+--------------------+-----
--------------+-----------+----------------+---------------+------------------+
--------------+--------------+--------------+---------------+-------------+-----
---------+
|             -1|           -1.0|      -1|       -1|        N/A|      null|
null|                null|        null|     null|               null|
null|       null|            null|           null|              null|
null|          null|          null|           null|         null|          null|
|          12694|3.3598700006E10|  335987|        6|   R
Dravid|1973-01-11|Right-hand bat|  Right-arm offbreak|       India|
Captain|Royal Challengers…|Kolkata Knight Ri…|       2008|              0|
35|                0|          null|          null|      R Dravid|     SC
Ganguly|   MV Boucher|       WP Saha|
|          12695|3.3598700007E10|  335987|        7|   W
Jaffer|1978-02-16|Right-hand bat|  Right-arm offbreak|       India|
Player|Royal Challengers…|Kolkata Knight Ri…|       2008|              0|
30|                0|          null|          null|      R Dravid|     SC
Ganguly|   MV Boucher|       WP Saha|
|          12696|3.3598700008E10|  335987|        8|    V
```

```
Kohli|1988-11-05|Right-hand bat|    Right-arm medium|       India|
Player|Royal Challengers…|Kolkata Knight Ri…|       2008|               0|
20|                0|          null|          null|      R Dravid|     SC
Ganguly|   MV Boucher|       WP Saha|
|          12697|3.3598700009E10|  335987|        9|  JH
Kallis|1975-10-16|Right-hand bat|Right-arm fast-me…|South Africa|
Player|Royal Challengers…|Kolkata Knight Ri…|       2008|               0|
33|                0|          null|          null|      R Dravid|     SC
Ganguly|   MV Boucher|       WP Saha|
+--------------+--------------+--------+---------+-----------+----------+-----
---------+-------------------+-----------+---------+-------------------+-----
---------------+-----------+---------------+---------------+-----------------+
-------------+-------------+-------------+--------------+------------+-----
---------+
only showing top 5 rows

Player DataFrame:
+---------+---------+--------------+----------+--------------+----------------
-+------------+
|PLAYER_SK|Player_Id|    Player_Name|       DOB|  Batting_hand|
Bowling_skill|Country_Name|
+---------+---------+--------------+----------+--------------+----------------
-+------------+
|        0|        1|     SC Ganguly|1972-07-08| Left-hand bat|  Right-arm
medium|       India|
|        1|        2|    BB McCullum|1981-09-27|Right-hand bat|  Right-arm
medium| New Zealand|
|        2|        3|     RT Ponting|1974-12-19|Right-hand bat|  Right-arm
medium|   Australia|
|        3|        4|      DJ Hussey|1977-07-15|Right-hand bat|Right-arm
offbreak|   Australia|
|        4|        5|Mohammad Hafeez|1980-10-17|Right-hand bat|Right-arm
offbreak|    Pakistan|
+---------+---------+--------------+----------+--------------+----------------
-+------------+
only showing top 5 rows

Team DataFrame:
+-------+-------+--------------------+
|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|
+-------+-------+--------------------+
only showing top 5 rows
```

2 import necessory functions

[0]:
```
from pyspark.sql.functions import col, count, avg, when, expr , sum as spark_sum
```

3 Data Cleaning Process

3.1 1 Drop Duplicate recods

[0]:
```
# Drop duplicates based on all columns
ball_by_ball_df = ball_by_ball_df.dropDuplicates()
player_match_df = player_match_df.dropDuplicates()
player_df = player_df.dropDuplicates()
match_df = match_df.dropDuplicates()
team_df = team_df.dropDuplicates()
```

3.2 2 Handle Missing values

[0]:
```
#null values checks

for df, name in zip([ball_by_ball_df, player_match_df, player_df, match_df,␣
 ↪team_df],
                    ['ball_by_ball_df', 'player_match_df', 'player_df',␣
 ↪'match_df', 'team_df']):
    print(f"Null values in {name}:")
    df.select([count(when(col(c).isNull(), c)).alias(c) for c in df.columns]).
 ↪show()
```

```
Null values in ball_by_ball_df:
+--------+-------+-------+----------+------------+------------+----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
----------+--------+------+------+-------+---+------------+-------+-------------
----+----------+----------------+------------+----------+------+-------+-------
----+------+----------+--------+----------------+---------+--------------------+-
------------+----------------+----------+--------------+---------+-------------
-----+--------------+-------------+-----------+-------------+-----------+
|MatcH_id|Over_id|Ball_id|Innings_No|Team_Batting|Team_Bowling|Striker_Batting_P
osition|Extra_Type|Runs_Scored|Extra_runs|Wides|Legbyes|Byes|Noballs|Penalty|Bow
ler_Extras|Out_type|Caught|Bowled|Run_out|LBW|Retired_hurt|Stumped|caught_and_bo
wled|hit_wicket|ObstructingFeild|Bowler_Wicket|Match_Date|Season|Striker|Non_Str
iker|Bowler|Player_Out|Fielders|Striker_match_SK|StrikerSK|NonStriker_match_SK|N
ONStriker_SK|Fielder_match_SK|Fielder_SK|Bowler_match_SK|BOWLER_SK|PlayerOut_mat
ch_SK|BattingTeam_SK|BowlingTeam_SK|Keeper_Catch|Player_out_sk|MatchDateSK|
+--------+-------+-------+----------+------------+------------+----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
```

```
----------+--------+------+------+-------+---+------------+-------+-------------
----+----------+----------------+-------------+----------+------+-------+-------
----+------+----------+--------+---------------+---------+-------------------+-
-----------+----------------+----------+---------------+---------+-------------
-----+--------------+--------------+------------+-------------+-----------+
|       0|      0|      0|         0|           0|           0|
13861|         0|          0|         0|    0|      0|   0|      0|      0|
0|       0|     0|     0|      0|  0|           0|      0|                 0|
0|               0|            0|         0|     0|      0|          0|     0|
143013|  145100|               0|        0|                  0|            0|
0|         0|              0|        0|                  0|             0|
0|           0|            0|          0|
+--------+------+-------+----------+-----------+------------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
----------+--------+------+------+-------+---+------------+-------+-------------
----+----------+----------------+-------------+----------+------+-------+-------
----+------+----------+--------+---------------+---------+-------------------+-
-----------+----------------+----------+---------------+---------+-------------
-----+--------------+--------------+------------+-------------+-----------+
```

Null values in player_match_df:

```
+--------------+---------------+--------+---------+-----------+---+------------
+-------------+------------+---------+-----------+------------+-----------+-----
----------+--------------+-----------------+-------------+--------------+---
----------+--------------+-------------+--------------+
|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_ma
nofThematch|Age_As_on_match|IsPlayers_Team_won|Batting_Status|Bowling_Status|Pla
yer_Captain|Opposit_captain|Player_keeper|Opposit_keeper|
+--------------+---------------+--------+---------+-----------+---+------------
+-------------+------------+---------+-----------+------------+-----------+-----
----------+--------------+-----------------+-------------+--------------+---
----------+--------------+-------------+--------------+
|              0|              0|       0|        0|          0|  1|
1|            1|           1|        1|          1|           1|          1|
1|              1|                 1|         13993|         13993|
1|              1|            1|             1|
+--------------+---------------+--------+---------+-----------+---+------------
+-------------+------------+---------+-----------+------------+-----------+-----
----------+--------------+-----------------+-------------+--------------+---
----------+--------------+-------------+--------------+
```

Null values in player_df:

```
+---------+---------+-----------+---+------------+-------------+------------+
|PLAYER_SK|Player_Id|Player_Name|DOB|Batting_hand|Bowling_skill|Country_Name|
+---------+---------+-----------+---+------------+-------------+------------+
|        0|        0|          0|  0|           0|            0|           0|
+---------+---------+-----------+---+------------+-------------+------------+
```

```
Null values in match_df:
+--------+--------+-----+-----+----------+-----------+----------+---------+-----
-------+-----------+------------+---------+--------+-----------+---------+-----
-----+----------+
|Match_SK|match_id|Team1|Team2|match_date|Season_Year|Venue_Name|City_Name|Count
ry_Name|Toss_Winner|match_winner|Toss_Name|Win_Type|Outcome_Type|ManOfMach|Win_M
argin|Country_id|
+--------+--------+-----+-----+----------+-----------+----------+---------+-----
-------+-----------+------------+---------+--------+-----------+---------+-----
-----+----------+
|       0|       0|    0|    0|         0|          0|         0|        0|
0|          1|           0|        0|       1|           0|        0|         0|
0|
+--------+--------+-----+-----+----------+-----------+----------+---------+-----
-------+-----------+------------+---------+--------+-----------+---------+-----
-----+----------+

Null values in team_df:
+-------+-------+---------+
|Team_SK|Team_Id|Team_Name|
+-------+-------+---------+
|      0|      0|        0|
+-------+-------+---------+
```

3.3 3 Fill Null Values

[0]:
```
# Fill missing values with default values
ball_by_ball_df = ball_by_ball_df.fillna({"Extra_Type": "None", "Out_type":␣
 ↪"Not Out", "Bowler_Wicket": 0})
player_match_df = player_match_df.fillna({"Bowling_skill": "Unknown",␣
 ↪"Role_Desc": "Unknown"})
```

3.4 4 Drop Rows

[0]:
```
# Drop rows where important columns have null values
ball_by_ball_df = ball_by_ball_df.dropna(subset=["Match_id", "Over_id",␣
 ↪"Ball_id", "Team_Batting", "Team_Bowling"])
player_match_df = player_match_df.dropna(subset=["Player_Name", "Match_Id"])
```

3.5 5 Convert Data Types

[0]:
```
from pyspark.sql.types import IntegerType, DateType, StringType

# Convert numerical columns to IntegerType
```

```
ball_by_ball_df = ball_by_ball_df.withColumn("Runs_Scored",␣
 ↪ball_by_ball_df["Runs_Scored"].cast(IntegerType()))
player_match_df = player_match_df.withColumn("Age_As_on_match",␣
 ↪player_match_df["Age_As_on_match"].cast(IntegerType()))

# Convert date columns to DateType
match_df = match_df.withColumn("Match_Date", match_df["Match_Date"].
 ↪cast(DateType()))
player_df = player_df.withColumn("DOB", player_df["DOB"].cast(DateType()))
```

3.6 6 Standardize Sting Columns

[0]:
```
from pyspark.sql.functions import trim, lower

# Standardize team names and player names
ball_by_ball_df = ball_by_ball_df.withColumn("Team_Batting",␣
 ↪trim(lower(ball_by_ball_df["Team_Batting"])))
player_df = player_df.withColumn("Player_Name", trim(player_df["Player_Name"]))
```

3.7 7 Handle Outliers

[0]:
```
# Check for extreme values in Runs_Scored
ball_by_ball_df.describe(["Runs_Scored"]).show()
```

```
+-------+------------------+
|summary|       Runs_Scored|
+-------+------------------+
|  count|            150451|
|   mean|1.2221985895740142|
| stddev|1.5943108833469897|
|    min|                 0|
|    max|                 6|
+-------+------------------+
```

3.8 8 Verify Data Integrity with key Relationships

- check if match_id in ball_by_ball_df has matching match_id valeus in match_df

[0]:
```
missing_matches = ball_by_ball_df.join(match_df, on="match_id" ,␣
 ↪how="left_anti")
print("records with missing match_id in Match_df:")
missing_matches.show()
```

```
records with missing match_id in Match_df:
+--------+-------+-------+----------+------------+------------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
```

```
---------+--------+------+------+-------+---+-----------+-------+------------
----+----------+----------------+-------------+----------+------+-------+-------
----+------+----------+--------+---------------+---------+------------------+-
-----------+----------------+----------+--------------+---------+------------
-----+--------------+--------------+-----------+-------------+-----------+
|MatcH_id|Over_id|Ball_id|Innings_No|Team_Batting|Team_Bowling|Striker_Batting_P
osition|Extra_Type|Runs_Scored|Extra_runs|Wides|Legbyes|Byes|Noballs|Penalty|Bow
ler_Extras|Out_type|Caught|Bowled|Run_out|LBW|Retired_hurt|Stumped|caught_and_bo
wled|hit_wicket|ObstructingFeild|Bowler_Wicket|Match_Date|Season|Striker|Non_Str
iker|Bowler|Player_Out|Fielders|Striker_match_SK|StrikerSK|NonStriker_match_SK|N
ONStriker_SK|Fielder_match_SK|Fielder_SK|Bowler_match_SK|BOWLER_SK|PlayerOut_mat
ch_SK|BattingTeam_SK|BowlingTeam_SK|Keeper_Catch|Player_out_sk|MatchDateSK|
+--------+-------+-------+----------+-----------+-----------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
---------+--------+------+------+-------+---+-----------+-------+------------
----+----------+----------------+-------------+----------+------+-------+-------
----+------+----------+--------+---------------+---------+------------------+-
-----------+----------------+----------+--------------+---------+------------
-----+--------------+--------------+-----------+-------------+-----------+
+--------+-------+-------+----------+-----------+-----------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
---------+--------+------+------+-------+---+-----------+-------+------------
----+----------+----------------+-------------+----------+------+-------+-------
----+------+----------+--------+---------------+---------+------------------+-
-----------+----------------+----------+--------------+---------+------------
-----+--------------+--------------+-----------+-------------+-----------+
```

3.9 Final CLeaned Data Check

[0]:
```
# Display schema and some data to verify
ball_by_ball_df.printSchema()
ball_by_ball_df.show(2)

player_match_df.printSchema()
player_match_df.show(2)
```

```
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 = false)
 |-- Runs_Scored: integer (nullable = false)
 |-- 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 = false)
 |-- Caught: integer (nullable = true)
 |-- Bowled: integer (nullable = true)
 |-- Run_out: integer (nullable = true)
 |-- LBW: integer (nullable = true)
 |-- Retired_hurt: integer (nullable = true)
 |-- Stumped: integer (nullable = true)
 |-- caught_and_bowled: integer (nullable = true)
 |-- hit_wicket: integer (nullable = true)
 |-- ObstructingFeild: integer (nullable = true)
 |-- Bowler_Wicket: integer (nullable = false)
 |-- Match_Date: date (nullable = true)
 |-- Season: integer (nullable = true)
 |-- Striker: integer (nullable = true)
 |-- Non_Striker: integer (nullable = true)
 |-- Bowler: integer (nullable = true)
 |-- Player_Out: integer (nullable = true)
 |-- Fielders: integer (nullable = true)
 |-- Striker_match_SK: integer (nullable = true)
 |-- StrikerSK: integer (nullable = true)
 |-- NonStriker_match_SK: integer (nullable = true)
 |-- NONStriker_SK: integer (nullable = true)
 |-- Fielder_match_SK: integer (nullable = true)
 |-- Fielder_SK: integer (nullable = true)
 |-- Bowler_match_SK: integer (nullable = true)
 |-- BOWLER_SK: integer (nullable = true)
 |-- PlayerOut_match_SK: integer (nullable = true)
 |-- BattingTeam_SK: integer (nullable = true)
 |-- BowlingTeam_SK: integer (nullable = true)
 |-- Keeper_Catch: integer (nullable = true)
 |-- Player_out_sk: integer (nullable = true)
 |-- MatchDateSK: integer (nullable = true)

+--------+-------+-------+----------+------------+------------+-----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
----------+-------------+------+------+-------+---+------------+-------+-------
----------+----------+----------------+-------------+----------+------+-------+-
----------+------+----------+--------+----------------+---------+--------------
----+-------------+----------------+----------+---------------+---------+-------
-----------+-------------+-------------+------------+-------------+-----------
+
|MatcH_id|Over_id|Ball_id|Innings_No|Team_Batting|Team_Bowling|Striker_Batting_P
```

```
osition|Extra_Type|Runs_Scored|Extra_runs|Wides|Legbyes|Byes|Noballs|Penalty|Bow
ler_Extras|      Out_type|Caught|Bowled|Run_out|LBW|Retired_hurt|Stumped|caught_
and_bowled|hit_wicket|ObstructingFeild|Bowler_Wicket|Match_Date|Season|Striker|N
on_Striker|Bowler|Player_Out|Fielders|Striker_match_SK|StrikerSK|NonStriker_matc
h_SK|NONStriker_SK|Fielder_match_SK|Fielder_SK|Bowler_match_SK|BOWLER_SK|PlayerO
ut_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|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|    277|
104|    83|      null|    null|           20336|      276|              20333|
103|              -1|        -1|          20343|       82|                -1|
4|             1|           0|            0|   20130420|
|  598028|     14|      1|         1|           5|           2|
5| No Extras|          1|         0|    0|      0|   0|      0|      0|
0|Not Applicable|     0|     0|      0|  0|           0|      0|
0|         0|               0|            0|2013-04-20|  2013|    104|
6|   346|      null|    null|           20333|      103|              20328|
5|              -1|        -1|          20348|      345|                -1|
4|             1|           0|            0|   20130420|
+--------+------+-------+----------+-----------+-----------+----------------
-------+----------+-----------+----------+-----+-------+----+-------+-------+---
---------+-------------+------+------+-------+---+-----------+-------+-------
---------+----------+---------------+-------------+----------+------+-------+-
----------+------+----------+--------+----------------+---------+--------------
----+-------------+---------------+----------+--------------+---------+-------
-----------+-------------+-------------+-----------+-------------+-----------
+
only showing top 2 rows

root
 |-- Player_match_SK: integer (nullable = true)
 |-- PlayerMatch_key: double (nullable = true)
 |-- Match_Id: integer (nullable = true)
 |-- Player_Id: integer (nullable = true)
 |-- Player_Name: string (nullable = true)
 |-- DOB: date (nullable = true)
 |-- Batting_hand: string (nullable = true)
 |-- Bowling_skill: string (nullable = false)
```

```
 |-- Country_Name: string (nullable = true)
 |-- Role_Desc: string (nullable = false)
 |-- Player_team: string (nullable = true)
 |-- Opposit_Team: string (nullable = true)
 |-- Season_year: integer (nullable = true)
 |-- is_manofThematch: integer (nullable = true)
 |-- Age_As_on_match: integer (nullable = true)
 |-- IsPlayers_Team_won: integer (nullable = true)
 |-- Batting_Status: string (nullable = true)
 |-- Bowling_Status: string (nullable = true)
 |-- Player_Captain: string (nullable = true)
 |-- Opposit_captain: string (nullable = true)
 |-- Player_keeper: string (nullable = true)
 |-- Opposit_keeper: string (nullable = true)

+---------------+---------------+--------+---------+-----------+----------+-----
---------+-----------------+------------+---------+--------------------+-------
------------+-----------+----------------+--------------+------------------+--
------------+--------------+--------------+--------------+-------------+-------
-------+
|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_manofThematch|Age_As_on_match|IsPlayers_Team_won|Bat
ting_Status|Bowling_Status|Player_Captain|Opposit_captain|Player_keeper|Opposit_
keeper|
+---------------+---------------+--------+---------+-----------+----------+-----
---------+-----------------+------------+---------+--------------------+-------
------------+-----------+----------------+--------------+------------------+--
------------+--------------+--------------+--------------+-------------+-------
-------+
|             -1|           -1.0|      -1|       -1|        N/A|      null|
null|           Unknown|        null|  Unknown|               null|
null|       null|            null|          null|             null|
null|          null|          null|           null|         null|          null|
|          12695|3.3598700007E10|  335987|        7|   W
Jaffer|1978-02-16|Right-hand bat|Right-arm offbreak|       India|   Player|Royal
Challengers…|Kolkata Knight Ri…|       2008|               0|
30|                 0|          null|          null|      R Dravid|     SC
Ganguly|   MV Boucher|       WP Saha|
+---------------+---------------+--------+---------+-----------+----------+-----
---------+-----------------+------------+---------+--------------------+-------
------------+-----------+----------------+--------------+------------------+--
------------+--------------+--------------+--------------+-------------+-------
-------+
only showing top 2 rows
```

4 Most Common PySpark used functions

4.0.1 Select

[0]:
```
# Selecting specific columns
player_match_df.select("Player_Name", "Batting_hand", "Bowling_skill",␣
  ↪"Country_Name").show(5)
```

```
+-----------+--------------+--------------------+------------+
|Player_Name|  Batting_hand|       Bowling_skill|Country_Name|
+-----------+--------------+--------------------+------------+
|        N/A|          null|             Unknown|        null|
|    V Kohli|Right-hand bat|    Right-arm medium|       India|
|  JH Kallis|Right-hand bat|Right-arm fast-me…|South Africa|
|   CL White|Right-hand bat|     Legbreak googly|   Australia|
|   W Jaffer|Right-hand bat|  Right-arm offbreak|       India|
+-----------+--------------+--------------------+------------+
only showing top 5 rows
```

4.0.2 filter or where

[0]:
```
player_match_df.filter((col("Role_desc") == "Captain")  & (col("Country_Name")␣
  ↪== "India")).show(5)
```

```
+---------------+---------------+--------+---------+---------------+----------+-
------------+--------------------+-----------+---------+--------------------+-
------------------+-----------+----------------+---------------+--------------
---+--------------+--------------+--------------+---------------+-------------+
--------------+
|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_manofThematch|Age_As_on_match|IsPlayers_Team_won|Bat
ting_Status|Bowling_Status|
Player_Captain|Opposit_captain|Player_keeper|Opposit_keeper|
+---------------+---------------+--------+---------+---------------+----------+-
------------+--------------------+-----------+---------+--------------------+-
------------------+-----------+----------------+---------------+--------------
---+--------------+--------------+--------------+---------------+-------------+
--------------+
|          12767| 3.359900005E10|  335990|       50|Harbhajan
Singh|1980-07-03|Right-hand bat|  Right-arm offbreak|       India|  Captain|
Mumbai Indians|Royal Challengers…|       2008|                0|
28|                 0|          null|          null|Harbhajan Singh|       R
Dravid|     L Ronchi|    MV Boucher|
|          12719|3.3598800027E10|  335988|       27|   Yuvraj Singh|1981-12-12|
Left-hand bat|Slow left-arm ort…|       India|  Captain|     Kings XI Punjab|
Chennai Super Kings|       2008|                0|             27|
```

```
0|          null|          null|   Yuvraj Singh|       MS Dhoni|KC Sangakkara|
MS Dhoni|
|          12705|3.3598700001E10|  335987|        1|     SC Ganguly|1972-07-08|
Left-hand bat|    Right-arm medium|       India|  Captain|Kolkata Knight
Ri…|Royal Challengers…|       2008|               0|             36|
1|          null|          null|     SC Ganguly|       R Dravid|      WP Saha|
MV Boucher|
|          12771|3.3599000006E10|  335990|        6|       R
Dravid|1973-01-11|Right-hand bat|  Right-arm offbreak|       India|
Captain|Royal Challengers…|      Mumbai Indians|       2008|               0|
35|                1|          null|          null|       R Dravid|Harbhajan
Singh|   MV Boucher|      L Ronchi|
|          12739|3.3598900041E10|  335989|       41|       V
Sehwag|1978-10-20|Right-hand bat|  Right-arm offbreak|       India|  Captain|
Delhi Daredevils|    Rajasthan Royals|       2008|               0|
30|                1|          null|          null|       V Sehwag|       SK
Warne|   KD Karthik|       M Rawat|
+--------------+---------------+--------+---------+---------------+----------+-
------------+--------------------+-----------+---------+--------------------+-
------------------+-----------+---------------+---------------+--------------
---+--------------+--------------+---------------+---------------+-------------+
--------------+
only showing top 5 rows
```

4.0.3 Group By and Aggregation

[0]:
```
player_match_df.groupBy("Player_Name").count().orderBy("count" ,␣
  ↪ascending=False).show(5)
```

```
+-----------+-----+
|Player_Name|count|
+-----------+-----+
|   SK Raina|  160|
|  RG Sharma|  159|
|   MS Dhoni|  158|
| KD Karthik|  152|
| RV Uthappa|  149|
+-----------+-----+
only showing top 5 rows
```

[0]:
```
player_match_df.groupBy("Country_Name").agg(avg("Age_As_on_match").
  ↪alias("Avg_age")).show(5)
```

```
+------------+------------------+
|Country_Name|           Avg_age|
+------------+------------------+
```

```
| Afghanistan|21.294117647058822|
|   Sri Lanka|31.812280701754386|
|   Zimbabwea|              27.5|
|        null|              null|
|       India|27.313101160862356|
+------------+------------------+
only showing top 5 rows
```

4.0.4 Sorting

[0]:
```
# Order players by age
player_match_df.orderBy(col("Age_As_on_match").desc()).select("Player_Name",␣
 ↪"Age_As_on_match").distinct().show(10)
```

```
+-----------+---------------+
|Player_Name|Age_As_on_match|
+-----------+---------------+
|   R Dravid|             35|
|    V Kohli|             20|
|  AA Noffke|             31|
|    P Kumar|             22|
| MV Boucher|             32|
|   CL White|             25|
|  JH Kallis|             33|
|    B Akhil|             31|
|        N/A|           null|
|     Z Khan|             30|
+-----------+---------------+
only showing top 10 rows
```

4.0.5 Sql Query Using Pyspark Sql

[0]:
```
# Register the DataFrame as a global temporary view
player_match_df.createOrReplaceGlobalTempView("Player_match")

# Write a different SQL query
sql_query = spark.sql("""
    SELECT Player_Name, COUNT(DISTINCT Match_Id) AS total_matches_played
    FROM global_temp.Player_match
    GROUP BY Player_Name
    ORDER BY total_matches_played DESC
""")

# Show the top 5 results
sql_query.show(5)
```

```
+-----------+--------------------+
|Player_Name|total_matches_played|
+-----------+--------------------+
|   SK Raina|                 160|
|  RG Sharma|                 159|
|   MS Dhoni|                 158|
| KD Karthik|                 152|
| RV Uthappa|                 149|
+-----------+--------------------+
only showing top 5 rows
```

5 EDA

5.1 1. count the number of matches played in each season

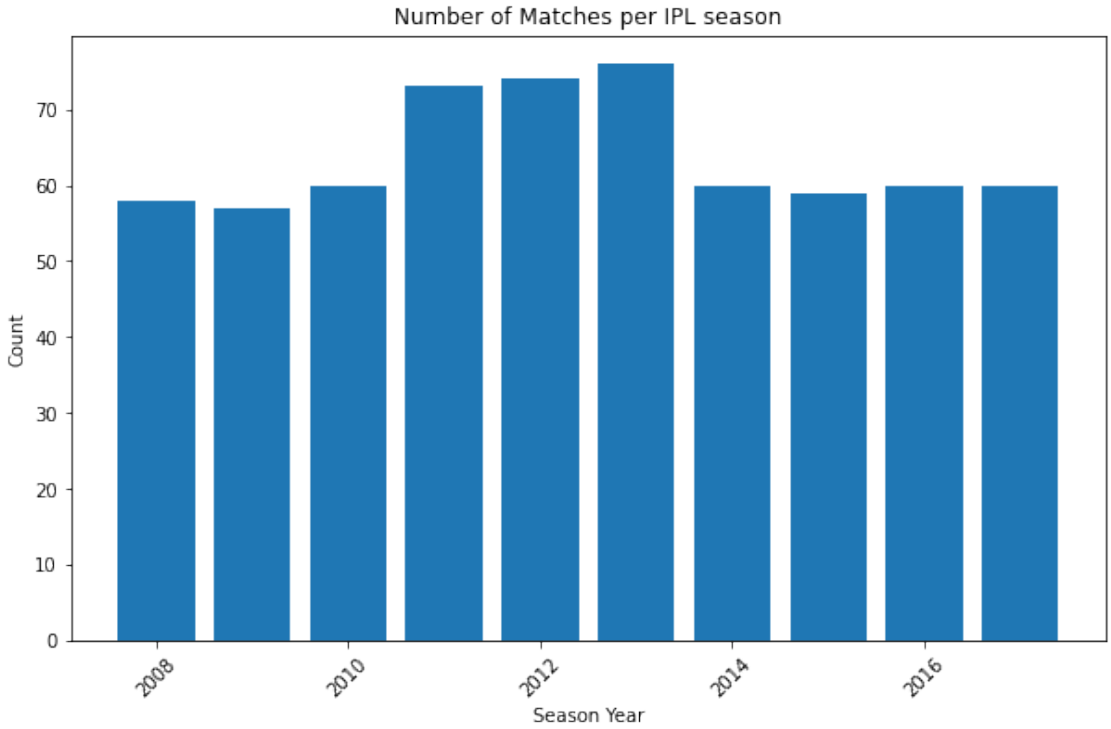
[0]:
```
season_count_df = match_df.groupBy("Season_Year").count().orderBy('Season_Year')
season_count_df.show()
```

```
+-----------+-----+
|Season_Year|count|
+-----------+-----+
|       2008|   58|
|       2009|   57|
|       2010|   60|
|       2011|   73|
|       2012|   74|
|       2013|   76|
|       2014|   60|
|       2015|   59|
|       2016|   60|
|       2017|   60|
+-----------+-----+
```

5.1.1 Plotting

[0]:
```
import matplotlib.pyplot as plt
season_count_pd = season_count_df.toPandas()

# plotting
plt.figure(figsize=(10,6))
plt.bar(season_count_pd["Season_Year"], season_count_pd["count"])
plt.xlabel("Season Year")
plt.ylabel("Count")
plt.title("Number of Matches per IPL season")
plt.xticks(rotation=45)
```

```
plt.show()
```



5.2 2. Top 5 Bowlers by wicket taken

[0]:
```
top_bowler_df = ball_by_ball_df.filter(col("Bowler_Wicket") ==1 ).
  ↪groupBy("Bowler").agg(count("Bowler_Wicket").alias("Wicket_taken")).
  ↪orderBy(col("Wicket_taken").desc()).limit(5)

top_bowler_df.show()
```

```
+------+------------+
|Bowler|Wicket_taken|
+------+------------+
|   194|         154|
|   136|         134|
|    50|         127|
|    67|         126|
|    71|         122|
+------+------------+
```

[0]:
```
#create unique list of players
unique_player_df = player_match_df.select("Player_Id", "Player_Name").distinct()
```

```
# Step 2: Join `top_bowlers_df` with `unique_players_df` to get bowler names
top_bowlers_with_names_df = top_bowler_df \
    .join(unique_player_df, top_bowler_df["bowler"] ==␣
 ↪unique_player_df["player_id"], "inner") \
    .select("player_name", "Wicket_taken") \
    .distinct() \
    .orderBy("Wicket_taken", ascending=False)

top_bowlers_with_names_df.show()
```

```
+---------------+------------+
|    player_name|Wicket_taken|
+---------------+------------+
|     SL Malinga|         154|
|       A Mishra|         134|
|Harbhajan Singh|         127|
|      PP Chawla|         126|
|       DJ Bravo|         122|
+---------------+------------+
```

6 Univariate Analysis

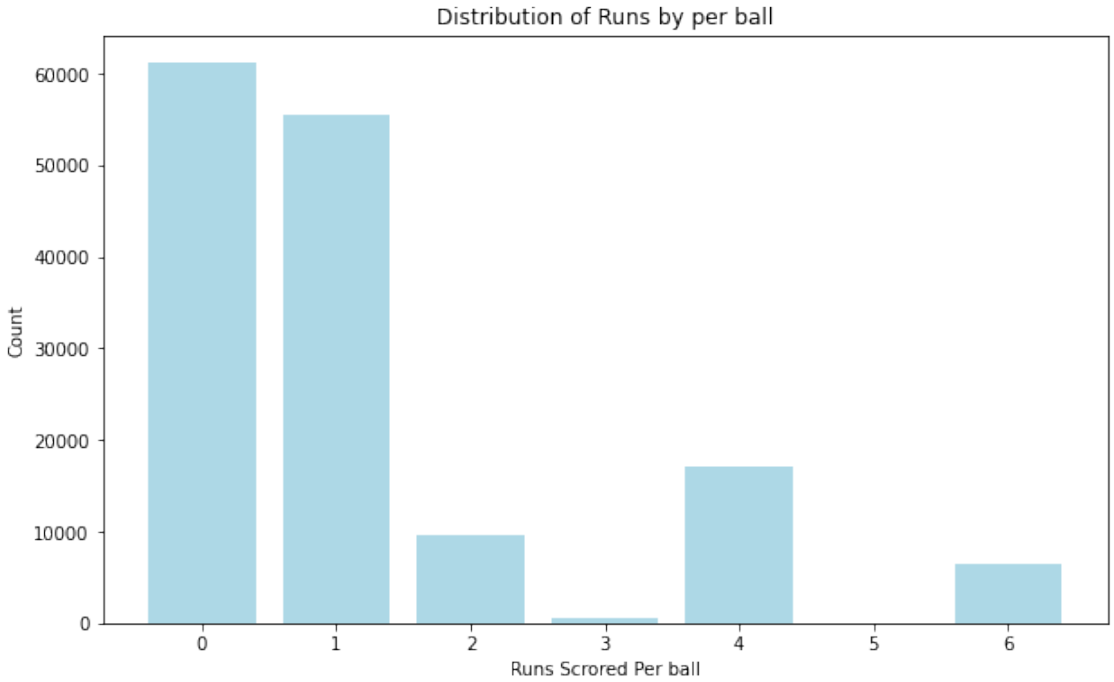
- What is the distribution of runs Scored per ball ?
- what are the differenent dismissal types and how frequenctly do they occur ?

[0]:
```
# Distribution of runs scored per ball
runs_distribution_df = ball_by_ball_df.groupBy("Runs_Scored").count().
 ↪orderBy("Runs_scored")
runs_distribution_df.show()
```

```
+-----------+-----+
|Runs_Scored|count|
+-----------+-----+
|          0|61151|
|          1|55495|
|          2| 9705|
|          3|  509|
|          4|17026|
|          5|   45|
|          6| 6520|
+-----------+-----+
```

[0]:
```
#plotting
runs_distribution_pd = runs_distribution_df.toPandas()
```

```
plt.figure(figsize=(10,6))
plt.bar(runs_distribution_pd["Runs_Scored"], runs_distribution_pd["count"] ,␣
 ↪color="lightblue" )
plt.title("Distribution of Runs by per ball")
plt.xlabel("Runs Scrored Per ball")
plt.ylabel("Count")
plt.show()
```

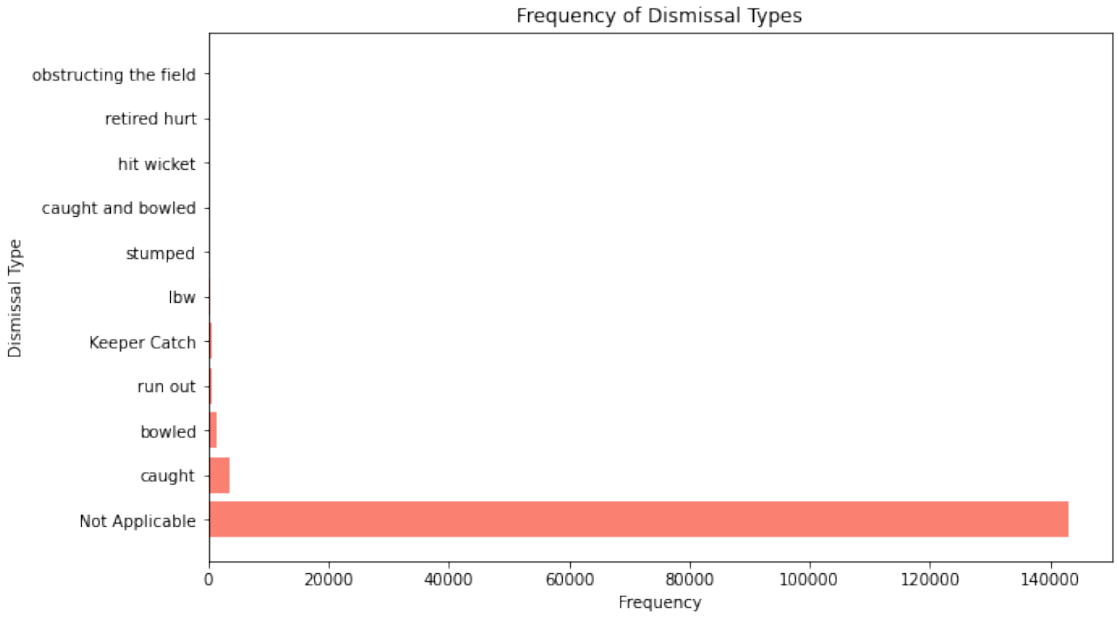


[0]:
```
# question 2 dismissal type frequency
dismissal_type_df= ball_by_ball_df.groupby("Out_type").count().
 ↪orderBy(col("count").desc())
dismissal_type_df.show()
```

```
+--------------------+------+
|            Out_type| count|
+--------------------+------+
|      Not Applicable|143013|
|              caught|  3678|
|              bowled|  1382|
|             run out|   755|
|        Keeper Catch|   695|
|                 lbw|   455|
|             stumped|   243|
|   caught and bowled|   211|
|          hit wicket|     9|
```

```
|        retired hurt|     9|
|obstructing the f…|     1|
+--------------------+------+
```

[0]:
```
# Plotting
dismissal_type_pd = dismissal_type_df.toPandas()
plt.figure(figsize=(10, 6))
plt.barh(dismissal_type_pd["Out_type"], dismissal_type_pd["count"],␣
 ↪color="salmon")
plt.xlabel("Frequency")
plt.ylabel("Dismissal Type")
plt.title("Frequency of Dismissal Types")
plt.show()
```



7 Bivariate Analysis

- How does the players Batting positon affect the runs scored ?
- What is the distribution of wickets taken by different bowlers ?

[0]:
```
# runs scored by batting position
run_by_position_df = ball_by_ball_df.groupby("Striker_Batting_Position").
 ↪agg(spark_sum("Runs_Scored").alias("Total_Runs")).
 ↪orderBy("Striker_Batting_Position")
run_by_position_df.show()
```

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

```
|Striker_Batting_Position|Total_Runs|
+------------------------+----------+
|                    null|     17920|
|                       1|     31004|
|                       2|     31254|
|                       3|     28775|
|                       4|     24958|
|                       5|     20640|
|                       6|     13334|
|                       7|      8098|
|                       8|      4425|
|                       9|      2159|
|                      10|      1032|
|                      11|       282|
+------------------------+----------+
```

8 3. Multivariate Analysis

- what is the relationship between runs scored , player role , and dismissal type ?
- how do different venues copare in terms of scoreing and wicket-taking patterns ?

[0]:
```
# Question 1: Runs, Role, and Dismissal Type
# Joining `ball_by_ball_df` and `player_match_df` to get player role for each␣
 ↪dismissal
runs_role_dismissal_df = ball_by_ball_df \
    .join(player_match_df, ball_by_ball_df["Striker"] ==␣
 ↪player_match_df["Player_Id"]) \
    .groupBy("Role_Desc", "Out_type") \
    .agg(spark_sum("Runs_Scored").alias("Total_Runs")) \
    .orderBy(col("Total_Runs").desc())

runs_role_dismissal_df.show()
```

```
+-------------+-----------------+----------+
|    Role_Desc|         Out_type|Total_Runs|
+-------------+-----------------+----------+
|       Player|   Not Applicable|  10442512|
|      Captain|   Not Applicable|   2534390|
|       Keeper|   Not Applicable|   1824879|
|CaptainKeeper|   Not Applicable|    712781|
|       Player|          run out|     14241|
|      Captain|          run out|      2670|
|       Keeper|          run out|      1518|
|CaptainKeeper|          run out|       794|
|      Captain|     retired hurt|       373|
|       Player|     retired hurt|       283|
```

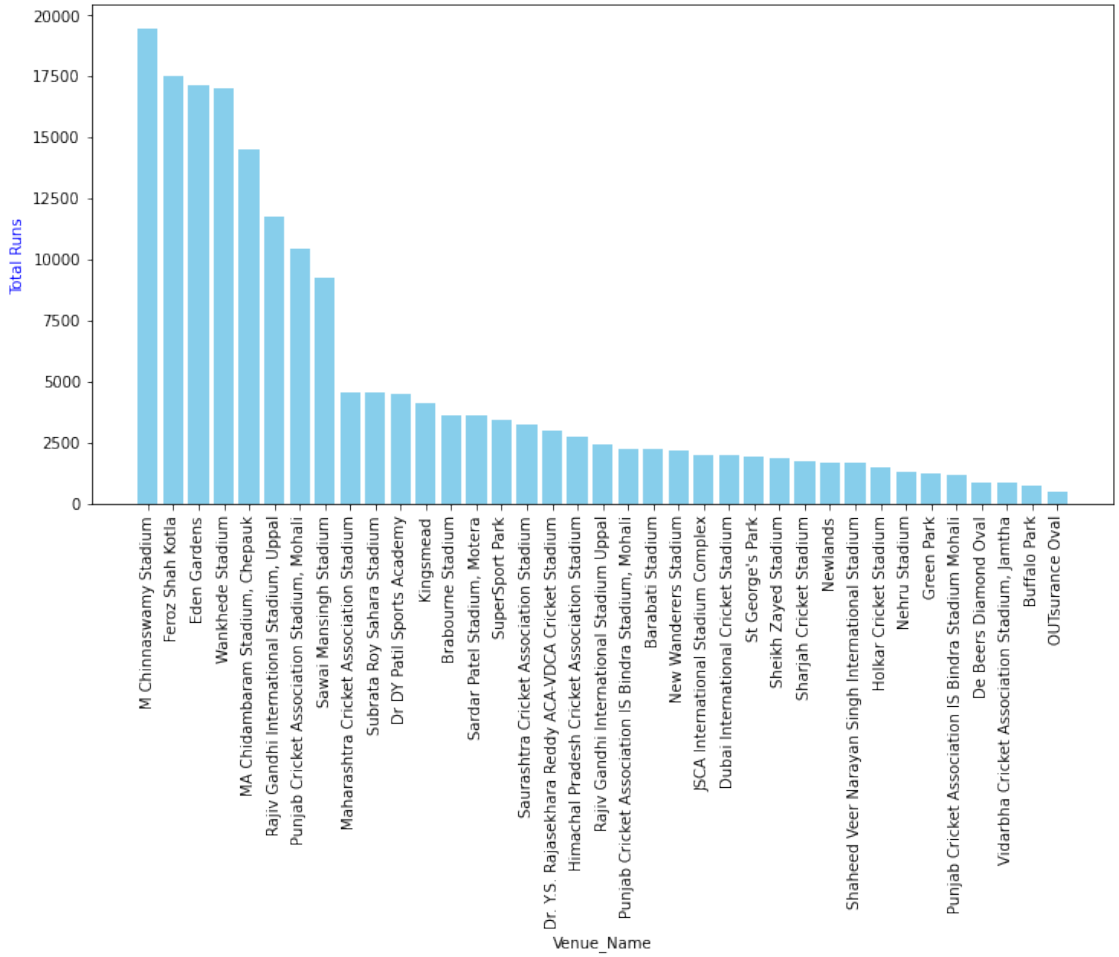
```
|       Player|           caught|         0|
|       Keeper|          stumped|         0|
|       Player|       hit wicket|         0|
|       Player|caught and bowled|         0|
|CaptainKeeper|          stumped|         0|
|       Keeper|              lbw|         0|
|      Captain|              lbw|         0|
|       Keeper|     Keeper Catch|         0|
|       Player|     Keeper Catch|         0|
|      Captain|          stumped|         0|
+-------------+-----------------+----------+
only showing top 20 rows
```

[0]:
```
from pyspark.sql.functions import sum as spark_sum, col

# Ensure numeric data in Runs_Scored and Bowler_Wicket columns
ball_by_ball_df = ball_by_ball_df.withColumn("Runs_Scored", col("Runs_Scored").
 ↪cast("int")) \
                                 .withColumn("Bowler_Wicket",␣
 ↪col("Bowler_Wicket").cast("int"))

# Replace nulls in Runs_Scored and Bowler_Wicket with 0 to avoid aggregation␣
 ↪errors
ball_by_ball_df = ball_by_ball_df.na.fill({"Runs_Scored": 0, "Bowler_Wicket":␣
 ↪0})

# Group by venue and aggregate runs and wickets
venue_performance_df = ball_by_ball_df \
    .join(match_df, "Match_id") \
    .groupBy("Venue_Name") \
    .agg(spark_sum("Runs_Scored").alias("Total_Runs"),␣
 ↪spark_sum("Bowler_Wicket").alias("Total_Wickets")) \
    .orderBy(col("Total_Runs").desc())

venue_performance_df.show()
```

```
+--------------------+----------+-------------+
|          Venue_Name|Total_Runs|Total_Wickets|
+--------------------+----------+-------------+
|M Chinnaswamy Sta…|     19423|          687|
|    Feroz Shah Kotla|     17491|          619|
|        Eden Gardens|     17105|          609|
|    Wankhede Stadium|     16996|          637|
|MA Chidambaram St…|     14471|          507|
|Rajiv Gandhi Inte…|     11748|          415|
|Punjab Cricket As…|     10421|          375|
```

```
|Sawai Mansingh St…|      9231|          320|
|Maharashtra Crick…|      4538|          153|
|Subrata Roy Sahar…|      4518|          178|
|Dr DY Patil Sport…|      4490|          199|
|           Kingsmead|      4076|          178|
|   Brabourne Stadium|      3595|          114|
|Sardar Patel Stad…|      3589|          120|
|     SuperSport Park|      3440|          117|
|Saurashtra Cricke…|      3198|           99|
|Dr. Y.S. Rajasekh…|      2995|          127|
|Himachal Pradesh …|      2716|          102|
|Rajiv Gandhi Inte…|      2419|           87|
|Punjab Cricket As…|      2232|           67|
+--------------------+----------+-------------+
only showing top 20 rows
```

[0]:
```
# Plotting Venue Analysis
venue_performance_pd = venue_performance_df.toPandas()
fig, ax1 = plt.subplots(figsize=(12, 6))

# Bar chart for total runs
ax1.bar(venue_performance_pd["Venue_Name"], venue_performance_pd["Total_Runs"],␣
 ↪color="skyblue", label="Total Runs")
ax1.set_xlabel("Venue_Name")
ax1.set_ylabel("Total Runs", color="blue")
ax1.tick_params(axis="x", rotation=90)
```



[0]:
```
import matplotlib.pyplot as plt

# Assuming 'venue_performance_pd' is a Pandas DataFrame
fig, ax1 = plt.subplots(figsize=(12, 7))

# Bar chart for Total Runs
ax1.bar(venue_performance_pd["Venue_Name"], venue_performance_pd["Total_Runs"],␣
 ↪color="blue", label="Total Runs")
ax1.set_xlabel("Venue Name")
ax1.set_ylabel("Total Runs", color="blue")
ax1.tick_params(axis="y", labelcolor="blue")

# Line chart for Total Wickets
ax2 = ax1.twinx()  # Set up the secondary y-axis
ax2.plot(venue_performance_pd["Venue_Name"],␣
 ↪venue_performance_pd["Total_Wickets"], color="red", marker="o", label="Total␣
 ↪Wickets")
```

```
ax2.set_ylabel("Total Wickets", color="red")
ax2.tick_params(axis="y", labelcolor="red")

# Rotate x-axis labels for better readability
plt.xticks(rotation=45, ha="right")

plt.title("Venue Analysis: Total Runs vs. Total Wickets")
fig.tight_layout()  # Adjust layout to prevent overlap
plt.show()
```