```
from pyspark.sql import SparkSession
import pyspark.sql.functions
from pyspark.sql import functions as f
from pyspark.sql.types import StructType, StructField, StringType, IntegerType


schema_city = StructType([
                         StructField("city_id", IntegerType(),True),
                          StructField("City", StringType(),True),
                          StructField("Venue", StringType(),True),
                          StructField("Neutral_venue", StringType(),True),
                         ])

df_cityDimension = spark.read.csv("/FileStore/tables/city_dimension-2.csv", header = True, schema = schema_city )
# df_cityDimension = spark.read.csv("/FileStore/tables/city_dimension-2.csv", header = True, schema = 'city_id INT,
City String, Venue String, Neutral_venue String' )
df_dateDimension = spark.read.csv("/FileStore/tables/date_dimension-2.csv", header = True, schema = 'date_id INT, Date
Date, Day String, Year INT')
df_matchDimension = spark.read.csv("/FileStore/tables/match_dimension-2.csv", header = True, schema = 'match_id INT,
id INT, Team1 String, Team2 String, toss_winner String, Toss_decision String, Winner String, Player_of_the_match
String, Eliminator String')
df_umpireDimension = spark.read.csv("/FileStore/tables/umpire_dimension-2.csv", header = True, schema = 'Umpire_id
INT, id INT, Umpire1 String, Umpire2 String')
df_factTable = spark.read.csv("/FileStore/tables/fact_table.csv", header = True, schema = 'id INT, city_id INT,
date_id INT, match_id INT, umpire_id INT, team1 String, team2 String, result_margin INT, result String')
```

```
# df_cityDimension.show()
df_dateDimension.show()
# df_matchDimension.show()
# df_umpireDimension.show()
# df_factTable.show()
```

```
+-------+----------+---------+----+
|date_id|      Date|      Day|Year|
+-------+----------+---------+----+
|      1|2008-04-18|   Friday|2008|
|      2|2008-04-19| Saturday|2008|
|      3|2008-04-19| Saturday|2008|
|      4|2008-04-20|   Sunday|2008|
|      5|2008-04-20|   Sunday|2008|
|      6|2008-04-21|   Monday|2008|
|      7|2008-04-22|  Tuesday|2008|
|      8|2008-04-23|Wednesday|2008|
|      9|2008-04-24| Thursday|2008|
|     10|2008-04-25|   Friday|2008|
|     11|2008-04-26| Saturday|2008|
|     12|2008-04-26| Saturday|2008|
|     13|2008-04-27|   Sunday|2008|
|     14|2008-04-27|   Sunday|2008|
|     15|2008-04-28|   Monday|2008|
|     16|2008-04-29|  Tuesday|2008|
|     17|2008-04-30|Wednesday|2008|
|     18|2008-05-01| Thursday|2008|
```

```
# df_cityDimension.printSchema()
# df_dateDimension.printSchema()
# df_matchDimension.printSchema()
# df_umpireDimension.printSchema()
# df_factTable.printSchema()
```

```
df_cityDimension.filter(df_cityDimension.City == "Bangalore").show()
```

```
+-------+---------+--------------------+-------------+
|city_id|     City|               Venue|Neutral_venue|
+-------+---------+--------------------+-------------+
|      1|Bangalore|M Chinnaswamy Sta...|            0|
|     11|Bangalore|M Chinnaswamy Sta...|            0|
|     15|Bangalore|M Chinnaswamy Sta...|            0|
|     25|Bangalore|M Chinnaswamy Sta...|            0|
|     31|Bangalore|M Chinnaswamy Sta...|            0|
|     45|Bangalore|M Chinnaswamy Sta...|            0|
|     52|Bangalore|M Chinnaswamy Sta...|            0|
|    121|Bangalore|M Chinnaswamy Sta...|            0|
|    124|Bangalore|M Chinnaswamy Sta...|            0|
|    131|Bangalore|M Chinnaswamy Sta...|            0|
|    136|Bangalore|M Chinnaswamy Sta...|            0|
|    153|Bangalore|M Chinnaswamy Sta...|            0|
|    156|Bangalore|M Chinnaswamy Sta...|            0|
|    165|Bangalore|M Chinnaswamy Sta...|            0|
|    181|Bangalore|M Chinnaswamy Sta...|            0|
|    207|Bangalore|M Chinnaswamy Sta...|            0|
|    219|Bangalore|M Chinnaswamy Sta...|            0|
|    222|Bangalore|M Chinnaswamy Sta...|            0|
```

```
# Q1. Find the match data in which winning margin is more than 100 runs

display(
    df_matchDimension.join(df_factTable, df_matchDimension.id == df_factTable.id,
"inner").filter(df_factTable.result_margin>100 )
       )
```

	match_id	id	Team1	Team2	toss_winner	Toss_decisi
1	1	335982	Royal Challengers Bangalore	Kolkata Knight Riders	Royal Challengers Bangalore	field
2	56	336038	Delhi Daredevils	Rajasthan Royals	Delhi Daredevils	field
3	235	501260	Kings XI Punjab	Royal Challengers Bangalore	Kings XI Punjab	bat
4	347	598027	Royal Challengers Bangalore	Pune Warriors	Pune Warriors	field
5	490	829785	Royal Challengers Bangalore	Kings XI Punjab	Kings XI Punjab	field
6	552	980987	Royal Challengers Bangalore	Gujarat Lions	Gujarat Lions	field
7	611	1082635	Delhi Daredevils	Mumbai Indians	Delhi Daredevils	field

Showing all 9 rows.

```
# Q2. Find players with most 'player of the match' award

display(
    df_matchDimension.groupby('Player_of_the_match').count().sort(f.col('count').desc())
)
```

	Player_of_the_match	count
1	CH Gayle	22
2	AB de Villiers	22
3	RG Sharma	18
4	MS Dhoni	17

5	DA Warner	17
6	SR Watson	16
7	YK Pathan	15

Showing all 232 rows.

```
# Q3. How many match MI has won at wankhede stadium while chose to bat first.

display(
    df_cityDimension.join(df_matchDimension, df_cityDimension.city_id == df_matchDimension.match_id,"inner").
    filter(   (df_cityDimension.Venue == 'Wankhede Stadium') & (df_matchDimension.Winner == 'Mumbai Indians') &
(df_matchDimension.Toss_decision == 'bat')   ).
    groupby('Winner').count()
)
```

	Winner	count
1	Mumbai Indians	13

Showing all 1 rows.

```
# Q4. Top match winners in Eden Gardens

display(
    df_cityDimension.join(df_matchDimension, df_cityDimension.city_id == df_matchDimension.match_id,"inner").
    filter(df_cityDimension.Venue == 'Eden Gardens').
    groupby('Winner').count().sort(f.col('count').desc())
)
```

	Winner	count
1	Kolkata Knight Riders	45

2	Mumbai Indians	10
3	Chennai Super Kings	5
4	Royal Challengers Bangalore	4
5	Kings XI Punjab	3
6	Gujarat Lions	2
7	Rajasthan Royals	2

Showing all 12 rows.

```
# Q5. Find out the venue where RCB has beaten MI
```

```
display(
    df_cityDimension.join(df_matchDimension, df_cityDimension.city_id == df_matchDimension.match_id,"inner").
    where(((f.col('Team1') == 'Mumbai Indians') & (f.col('Team2') == 'Royal Challengers Bangalore')) |
((f.col('Team2') == 'Mumbai Indians') |
    (f.col('Team1') == 'Royal Challengers Bangalore')) & (f.col('Winner') == 'Royal Challengers Bangalore'))
    .select('Venue').distinct()
    )
```

	Venue
1	Dubai International Cricket Stadium
2	Maharashtra Cricket Association Stadium
3	M Chinnaswamy Stadium
4	Brabourne Stadium
5	M.Chinnaswamy Stadium
6	Sharjah Cricket Stadium

7	Wankhede Stadium

Showing all 13 rows.

```
# Q6. Find out how many matches did RCB played in Bangalore and won.

display(
    df_cityDimension.join(df_matchDimension, df_cityDimension.city_id == df_matchDimension.match_id,"inner").
    where(    ((f.col('Team1') == 'Royal Challengers Bangalore')  | (f.col('Team2') == 'Royal Challengers Bangalore'))
& (f.col('Winner') == 'Royal Challengers Bangalore')          & (f.col('City') == 'Bangalore')    )
    .groupby('City','Winner').count()
    )
```

	City	Winner	count
1	Bangalore	Royal Challengers Bangalore	28

Showing all 1 rows.

```
# Q7. Top teams which played most no. of matches on weekends

df_team1 =  (df_factTable.join(df_dateDimension, df_factTable.date_id == df_dateDimension.date_id,"inner").
    where(   (f.col('Day') == 'Saturday') | (f.col('Day') == 'Sunday')
    ).
    groupby('Team1').count().withColumnRenamed("count","count1")
       )
df_team2 =  (df_factTable.join(df_dateDimension, df_factTable.date_id == df_dateDimension.date_id,"inner").
    where(   (f.col('Day') == 'Saturday') | (f.col('Day') == 'Sunday')
    ).
    groupby('Team2').count().withColumnRenamed("count","count2")
       )

display(
    df_team1.join(df_team2, df_team1.Team1 == df_team2.Team2,"inner").
    withColumn('Total_match_played', (f.col('count1') + f.col('count2'))).
    withColumnRenamed("Team1","Teams").
    select('Teams', 'Total_match_played').sort(f.col('Total_match_played').desc())
)
```

	Teams	Total_match_played
1	Mumbai Indians	80
2	Royal Challengers Bangalore	76
3	Kings XI Punjab	76
4	Kolkata Knight Riders	76
5	Chennai Super Kings	71
6	Rajasthan Royals	64
7	Delhi Daredevils	61

Showing all 15 rows.

```
# Q8. Who were the umpire1 most of the times when RR has won the match?

(
    df_matchDimension.join(df_umpireDimension, df_matchDimension.id == df_umpireDimension.id,"inner").
    where(  (f.col('Winner') == 'Rajasthan Royals')).
    groupby('Umpire1').count().sort(f.desc('count'))

).show()
```

```
+--------------------+-----+
|             Umpire1|count|
+--------------------+-----+
|           BF Bowden|    8|
|           Aleem Dar|    7|
|     HDPK Dharmasena|    6|
|         RE Koertzen|    4|
|           Asad Rauf|    4|
|              S Ravi|    4|
|         BR Doctrove|    3|
|             YC Barde|    3|
|           M Erasmus|    3|
|       C Shamshuddin|    3|
|           SS Hazare|    3|
|        AK Chaudhary|    2|
|           MR Benson|    2|
|           JD Cloete|    2|
|KN Ananthapadmana...|    2|
|         Nitin Menon|    2|
|         AY Dandekar|    2|
|         CB Gaffaney|    2|
```

```
#Q9. Find the team which won maximum no. of tosses.

display(
    df_matchDimension.groupby('toss_winner').count().sort(f.desc('count'))
)
```

	toss_winner	count
1	Mumbai Indians	103
2	Kolkata Knight Riders	97
3	Chennai Super Kings	96
4	Royal Challengers Bangalore	84
5	Rajasthan Royals	84
6	Kings XI Punjab	83
7	Delhi Daredevils	79

Showing all 15 rows.

```
+--------------+-----------+------------------+
|Matches_Played|Matches_Won|      Probability|
+--------------+-----------+------------------+
|           199|        118|0.592964824120603|
+--------------+-----------+------------------+
```