Student1: Sarthak Singhal(20171091)

Student2: Sajal Asati(20171183)

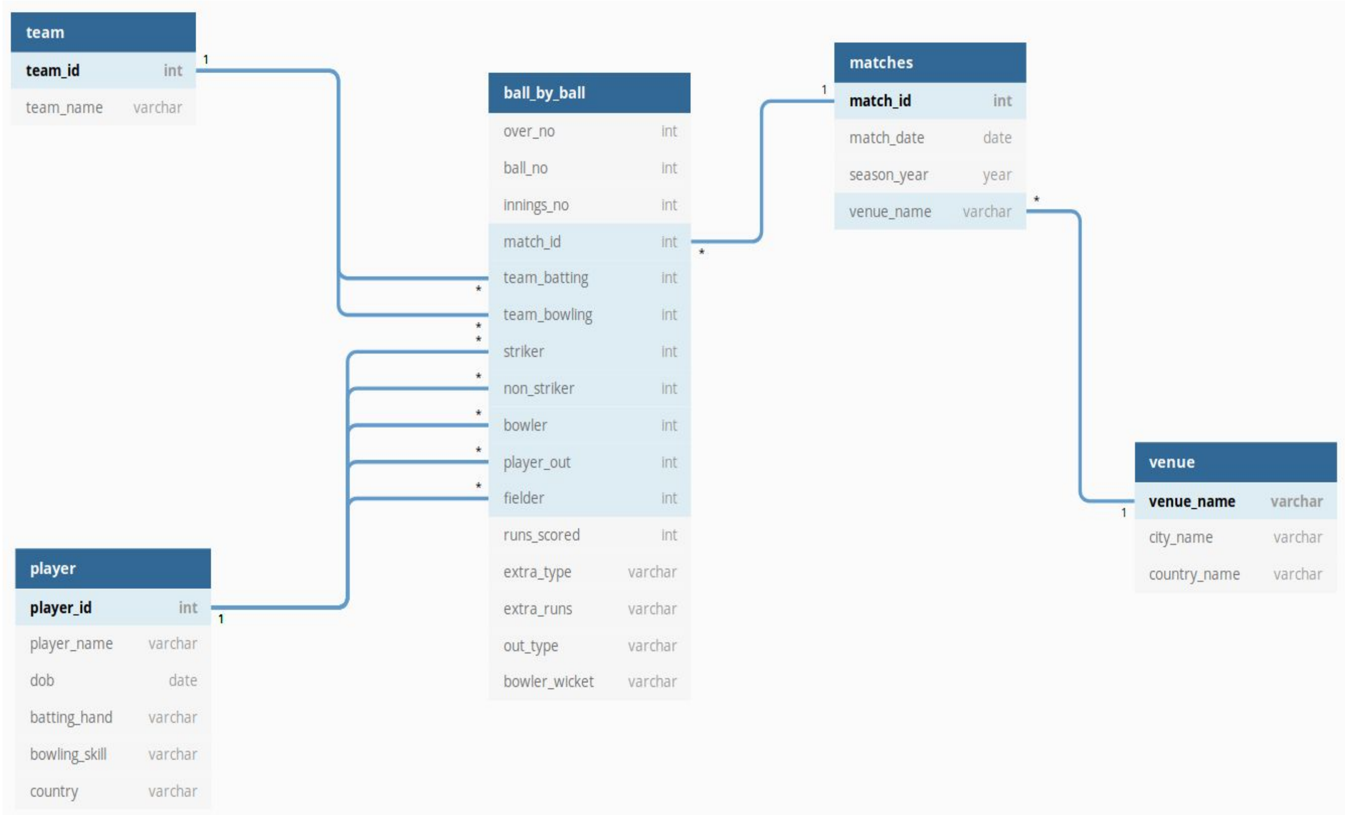
Data set used: IPL dataset

No. of pages: 8

Rationale for picking the data set

- ❏ So basically to do any sort of analysis one should have domain specific knowledge which helps a data analyst to analyse the data better.
- ❏ This is why we have chosen the dataset on IPL matches as we closely follow up with this league and thus can find important stats regarding the matches.
- ❏ This prerequisite knowledge allows us to explain the query results as we can tell why this particular batsman or bowler is there in a particular list or not.
- ❏ It can help us to preprocess data as this dataset has a large number of columns so we can easily split the table into multiple dimension tables.
- ❏ Also we could tell by looking at the dataset that it is quite detailed, it had collected information about many parameters, so we hoped to get interesting patterns from this data.

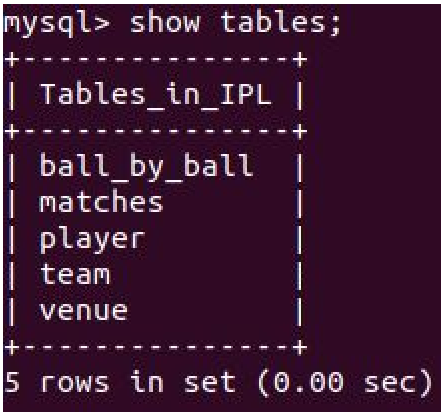
Data Warehouse schema



(The above picture shows the schema of our data warehouse)

- ❏ We have chosen a **snowflake schema** for our warehouse as 'matches' dimension table has been normalised and split into 'venue' dimension table.
- ❏ We have 5 tables(4 dimension tables and 1 fact table).
  - ❏ Team dimension table stores team names as there are fixed number of teams in the tournament.
  - ❏ Matches dimension table stores information about a match such as date, venue etc.

  - Venue dimension table stores information about the stadium such as name, city etc.
  - Player dimension table stores information about the player.
  - The fact table stores information about every ball. A ball is uniquely identified by (match_id,innings_no,over_no,ball_no) and contains measures such as runs scored, extras, out types etc.
- Since the transactions are on the ball, we have separated out other info such as info about players, teams etc in separate dimension tables as they contain fixed data, and also to reduce redundancy.
- All the data in dimension tables is fixed and does not change and only ball information is added in our fact table.



```
mysql> show tables;
+---------------+
| Tables_in_IPL |
+---------------+
| ball_by_ball  |
| matches       |
| player        |
| team          |
| venue         |
+---------------+
5 rows in set (0.00 sec)
```

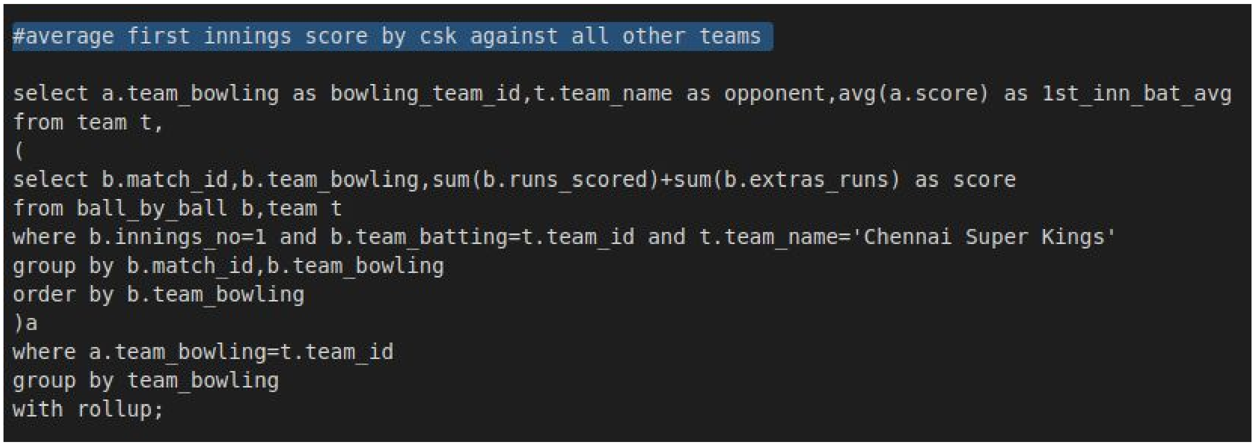
(Tables in our DB)

Database System used

- The database system we used is MySQL.
- One of the main reasons we used this is because we have already studied about it and used it in our Introduction to Databases course, so we were proficient in it already.
- Also, it’s open source and well documented, so we could learn to use new commands and apply OLAP queries in it very easily.
- MySQL also allows us to load data from csv files very easily in just one line of code. Hence it reduced our workload on that part.
- It is well known to handle even very large amounts of data very efficiently, hence we chose to use it.

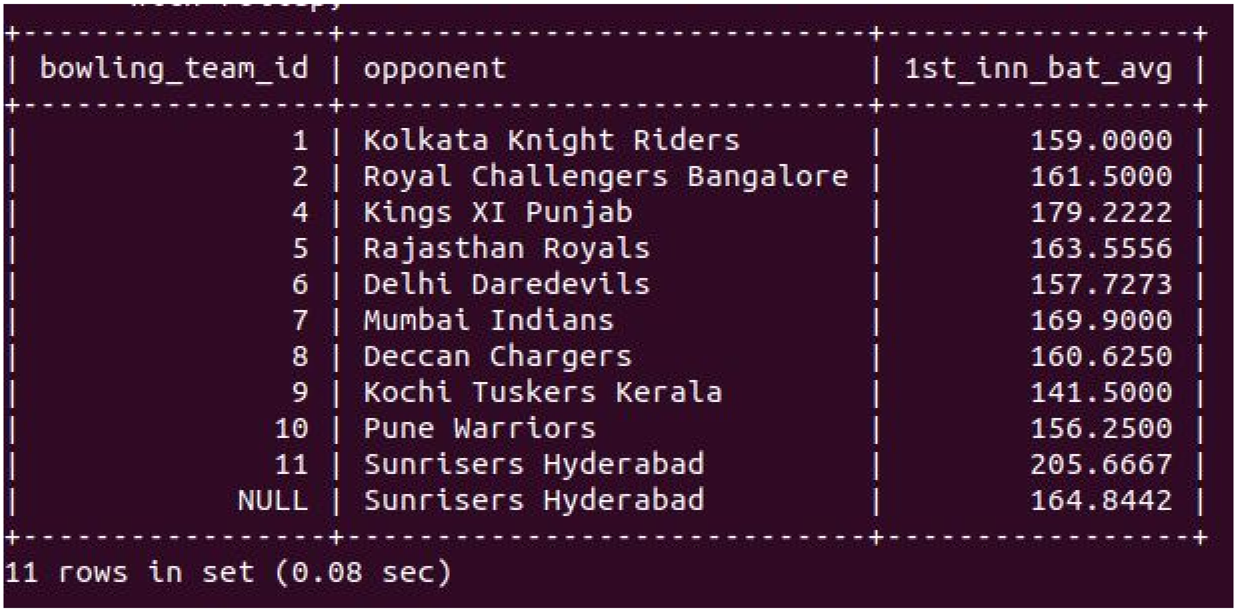
OLAP queries and results

Query 1: This query gives average first innings score by csk against all other teams. Using this we can analyse how this team has played against all other teams(batting first) over multiple seasons of IPL this dataset covers. As we can see that the batting averages are very good against all the teams and hence we can explain why CSK has very good record in IPL.



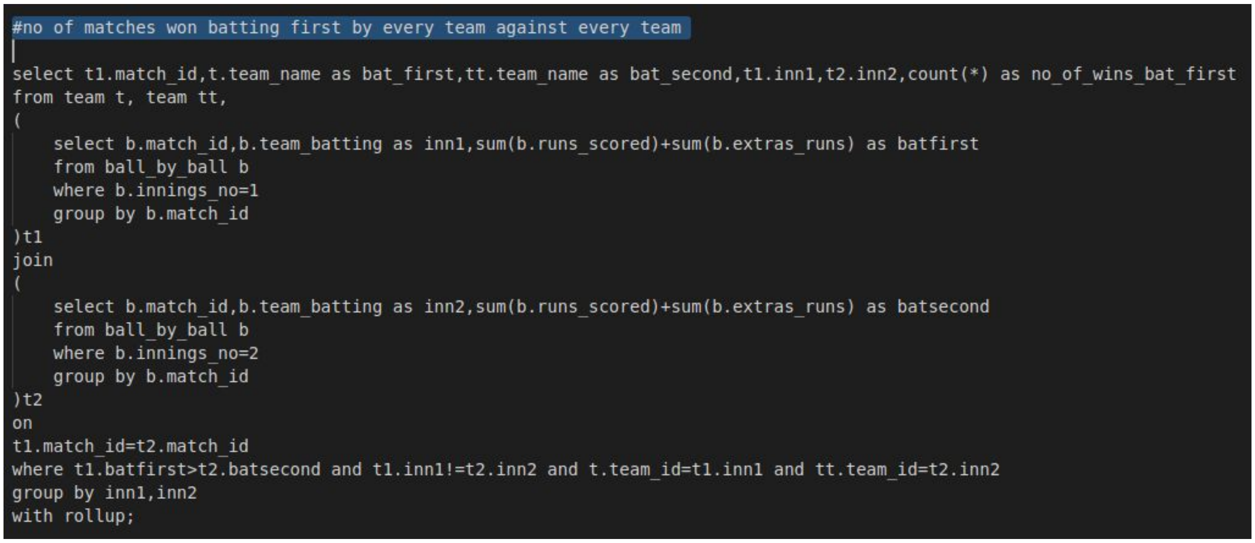
```
#average first innings score by csk against all other teams

select a.team_bowling as bowling_team_id,t.team_name as opponent,avg(a.score) as 1st_inn_bat_avg
from team t,
(
select b.match_id,b.team_bowling,sum(b.runs_scored)+sum(b.extras_runs) as score
from ball_by_ball b,team t
where b.innings_no=1 and b.team_batting=t.team_id and t.team_name='Chennai Super Kings'
group by b.match_id,b.team_bowling
order by b.team_bowling
)a
where a.team_bowling=t.team_id
group by team_bowling
with rollup;
```



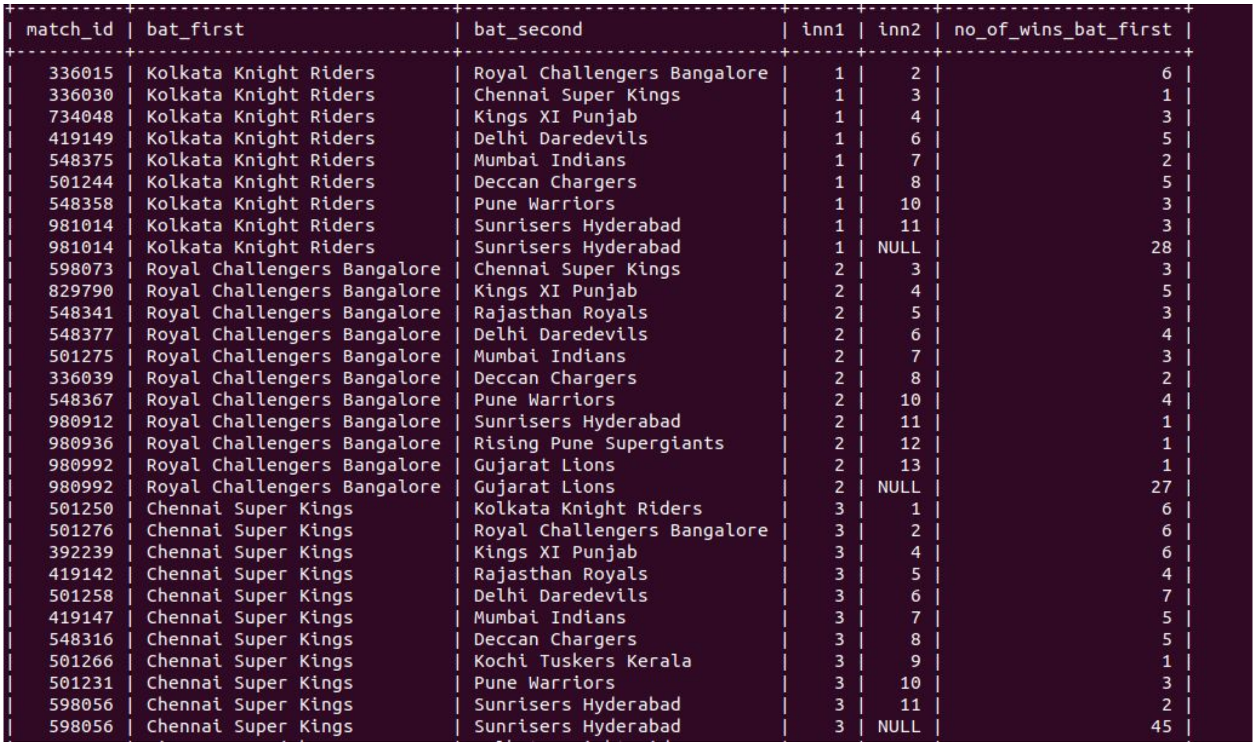
```
+-----------------+-----------------------------+-----------------+
| bowling_team_id | opponent                    | 1st_inn_bat_avg |
+-----------------+-----------------------------+-----------------+
|               1 | Kolkata Knight Riders       |        159.0000 |
|               2 | Royal Challengers Bangalore |        161.5000 |
|               4 | Kings XI Punjab             |        179.2222 |
|               5 | Rajasthan Royals            |        163.5556 |
|               6 | Delhi Daredevils            |        157.7273 |
|               7 | Mumbai Indians              |        169.9000 |
|               8 | Deccan Chargers             |        160.6250 |
|               9 | Kochi Tuskers Kerala        |        141.5000 |
|              10 | Pune Warriors               |        156.2500 |
|              11 | Sunrisers Hyderabad         |        205.6667 |
|            NULL | Sunrisers Hyderabad         |        164.8442 |
+-----------------+-----------------------------+-----------------+
11 rows in set (0.08 sec)
```

Query 2: This query gives stats about no. of matches won batting first by every team against every other team. This analysis gives us insights into how well the teams play when they bat first. Using this we can predict who will win when batting first in a game.



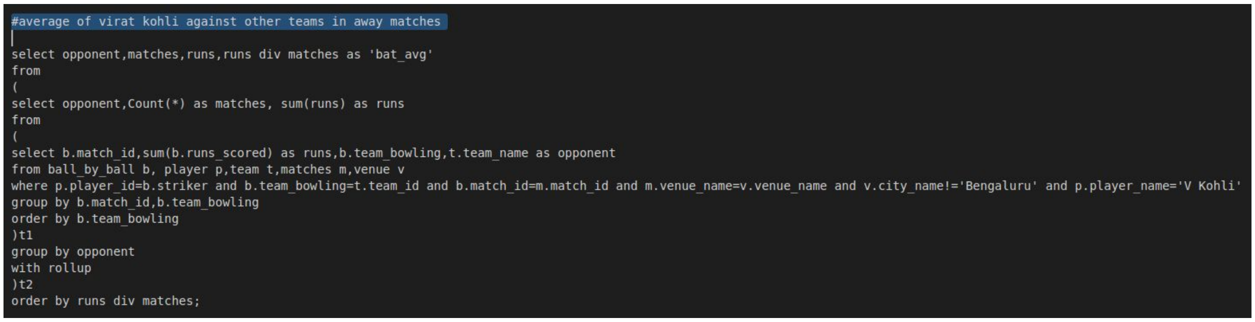
```
#no of matches won batting first by every team against every team

select t1.match_id,t.team_name as bat_first,tt.team_name as bat_second,t1.inn1,t2.inn2,count(*) as no_of_wins_bat_first
from team t, team tt,
(
    select b.match_id,b.team_batting as inn1,sum(b.runs_scored)+sum(b.extras_runs) as batfirst
    from ball_by_ball b
    where b.innings_no=1
    group by b.match_id
)t1
join
(
    select b.match_id,b.team_batting as inn2,sum(b.runs_scored)+sum(b.extras_runs) as batsecond
    from ball_by_ball b
    where b.innings_no=2
    group by b.match_id
)t2
on
t1.match_id=t2.match_id
where t1.batfirst>t2.batsecond and t1.inn1!=t2.inn2 and t.team_id=t1.inn1 and tt.team_id=t2.inn2
group by inn1,inn2
with rollup;
```



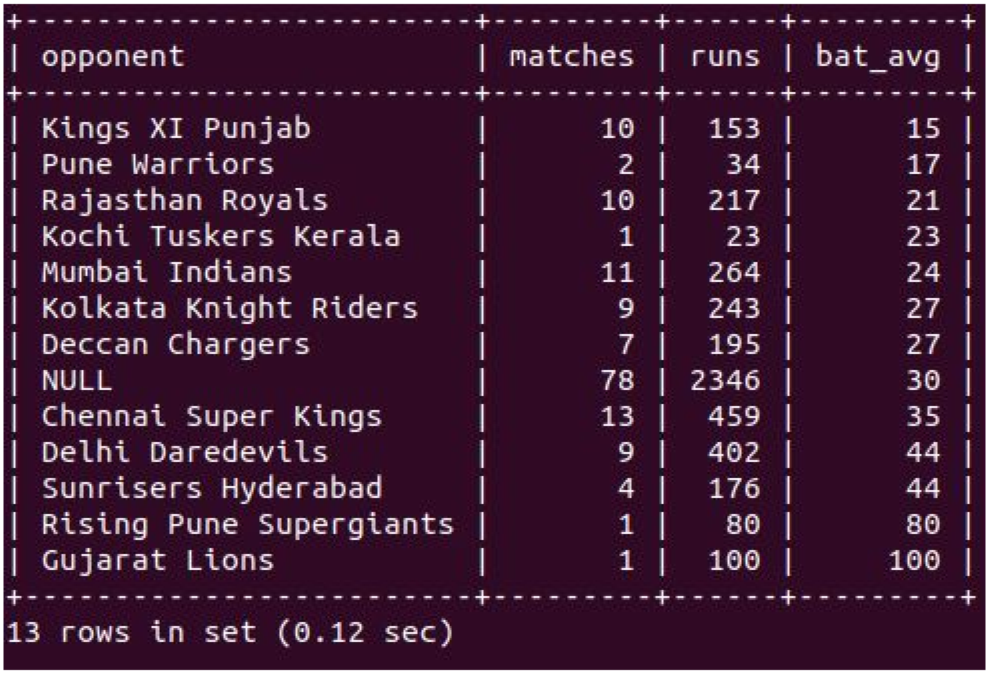
```
+----------+-----------------------------+-----------------------------+------+------+----------------------+
| match_id | bat_first                   | bat_second                  | inn1 | inn2 | no_of_wins_bat_first |
+----------+-----------------------------+-----------------------------+------+------+----------------------+
|   336015 | Kolkata Knight Riders       | Royal Challengers Bangalore |    1 |    2 |                    6 |
|   336030 | Kolkata Knight Riders       | Chennai Super Kings         |    1 |    3 |                    1 |
|   734048 | Kolkata Knight Riders       | Kings XI Punjab             |    1 |    4 |                    3 |
|   419149 | Kolkata Knight Riders       | Delhi Daredevils            |    1 |    6 |                    5 |
|   548375 | Kolkata Knight Riders       | Mumbai Indians              |    1 |    7 |                    2 |
|   501244 | Kolkata Knight Riders       | Deccan Chargers             |    1 |    8 |                    5 |
|   548358 | Kolkata Knight Riders       | Pune Warriors               |    1 |   10 |                    3 |
|   981014 | Kolkata Knight Riders       | Sunrisers Hyderabad         |    1 |   11 |                    3 |
|   981014 | Kolkata Knight Riders       | Sunrisers Hyderabad         |    1 | NULL |                   28 |
|   598073 | Royal Challengers Bangalore | Chennai Super Kings         |    2 |    3 |                    3 |
|   829790 | Royal Challengers Bangalore | Kings XI Punjab             |    2 |    4 |                    5 |
|   548341 | Royal Challengers Bangalore | Rajasthan Royals            |    2 |    5 |                    3 |
|   548377 | Royal Challengers Bangalore | Delhi Daredevils            |    2 |    6 |                    4 |
|   501275 | Royal Challengers Bangalore | Mumbai Indians              |    2 |    7 |                    3 |
|   336039 | Royal Challengers Bangalore | Deccan Chargers             |    2 |    8 |                    2 |
|   548367 | Royal Challengers Bangalore | Pune Warriors               |    2 |   10 |                    4 |
|   980912 | Royal Challengers Bangalore | Sunrisers Hyderabad         |    2 |   11 |                    1 |
|   980936 | Royal Challengers Bangalore | Rising Pune Supergiants     |    2 |   12 |                    1 |
|   980992 | Royal Challengers Bangalore | Gujarat Lions               |    2 |   13 |                    1 |
|   980992 | Royal Challengers Bangalore | Gujarat Lions               |    2 | NULL |                   27 |
|   501250 | Chennai Super Kings         | Kolkata Knight Riders       |    3 |    1 |                    6 |
|   501276 | Chennai Super Kings         | Royal Challengers Bangalore |    3 |    2 |                    6 |
|   392239 | Chennai Super Kings         | Kings XI Punjab             |    3 |    4 |                    6 |
|   419142 | Chennai Super Kings         | Rajasthan Royals            |    3 |    5 |                    4 |
|   501258 | Chennai Super Kings         | Delhi Daredevils            |    3 |    6 |                    7 |
|   419147 | Chennai Super Kings         | Mumbai Indians              |    3 |    7 |                    5 |
|   548316 | Chennai Super Kings         | Deccan Chargers             |    3 |    8 |                    5 |
|   501266 | Chennai Super Kings         | Kochi Tuskers Kerala        |    3 |    9 |                    1 |
|   501231 | Chennai Super Kings         | Pune Warriors               |    3 |   10 |                    3 |
|   598056 | Chennai Super Kings         | Sunrisers Hyderabad         |    3 |   11 |                    2 |
|   598056 | Chennai Super Kings         | Sunrisers Hyderabad         |    3 | NULL |                   45 |
```

Query 3: This query gives batting average of virat kohli against other teams in away matches(i.e. Not played in Bengaluru which is the home ground of the team). By looking at this we can judge how well Virat plays on other grounds against all other teams. Since he is one of the main players in his team and his batting average is pretty ordinary in away matches, we can explain RCB's poor performance in away matches.



```
#average of virat kohli against other teams in away matches

select opponent,matches,runs,runs div matches as 'bat_avg'
from
(
select opponent,Count(*) as matches, sum(runs) as runs
from
(
select b.match_id,sum(b.runs_scored) as runs,b.team_bowling,t.team_name as opponent
from ball_by_ball b, player p,team t,matches m,venue v
where p.player_id=b.striker and b.team_bowling=t.team_id and b.match_id=m.match_id and m.venue_name=v.venue_name and v.city_name!='Bengaluru' and p.player_name='V Kohli'
group by b.match_id,b.team_bowling
order by b.team_bowling
)t1
group by opponent
with rollup
)t2
order by runs div matches;
```



```
+-------------------------+---------+------+---------+
| opponent                | matches | runs | bat_avg |
+-------------------------+---------+------+---------+
| Kings XI Punjab         |      10 |  153 |      15 |
| Pune Warriors           |       2 |   34 |      17 |
| Rajasthan Royals        |      10 |  217 |      21 |
| Kochi Tuskers Kerala    |       1 |   23 |      23 |
| Mumbai Indians          |      11 |  264 |      24 |
| Kolkata Knight Riders   |       9 |  243 |      27 |
| Deccan Chargers         |       7 |  195 |      27 |
| NULL                    |      78 | 2346 |      30 |
| Chennai Super Kings     |      13 |  459 |      35 |
| Delhi Daredevils        |       9 |  402 |      44 |
| Sunrisers Hyderabad     |       4 |  176 |      44 |
| Rising Pune Supergiants |       1 |   80 |      80 |
| Gujarat Lions           |       1 |  100 |     100 |
+-------------------------+---------+------+---------+
13 rows in set (0.12 sec)
```