**DATA ANALYSIS PROBLEM: IPL**

**Data Set Given :** <https://docs.google.com/spreadsheets/d/1ctWoqqDg2sA77qclZlHGSK-TvcuO-UYJYOEw4TQDSGk/edit#gid=672337210>

**Ask**

Create a personal copy of the data to analyze the data to answer the following questions.

The choice of tool for analysis is not a restriction

* Balls per Wicket, Balls per Six, and Runs per Wicket
* The ratio of total balls to total dots for each batsman
* Percentage of boundary runs to total runs for each batsman
* Batsman responsible for the most wins
* Best Death Over Bowler
* Most Fifties by Batsmen

**About the Dataset given:**

The data set given is the IPL Dataset which has records of IPL matches – Ball by ball with certain features like match\_id, innings, batting\_team, bowling\_team, striker, non\_striker etc..,

It has **179078** observations and **21** features.

**Approach:**

The approach for the problem statement, is quite simple, since the data set is clean, as there is no any need for Data cleaning, and pre-processing.

Data Manipulation tasks plays significant role in the approach to the problem statement. It includes:

* Filter
* Group by
* Sorting/order by
* Concat/Join
* Etc..,

The approach to the problem is, to prepare data for different requirements which has been asked as questions and that is done through creating sub sets for the given data.

I created subsets for data from the root data/ given data, and it includes:

**IPL DATA SET**

**Batsmen Data Bowlers Data Fielders Data Match Data**

**Death & Powerplay Data Batsmen runs with results**

**Tools Used:**

* Python
* Pandas
* Numpy
* Seaborn/Matplotlib

I also include the notebook I worked and the sub sets of the data created, with this report.

**Notebook: [link 1: Analysis, link 2: Data Manipulation]**

<https://colab.research.google.com/drive/1nWe2bVS38TtPnfV40ftUTHkMR8COpOhg?usp=sharing>

<https://colab.research.google.com/drive/1useMJUSTD8Zs1dowj01cAvvCvYn0TiM6?usp=sharing>

**Sub sets:**

<https://drive.google.com/drive/folders/1CL7eJT1CDvjOz0j_zB_9a-6Z6VIs_MTD?usp=share_link>

**Information on the Subsets:**

Subsets have been created through Data manipulation and Feature Extraction, it includes,

* combing two features
* applying functions on features
* mapping values to features
* calculations on the features etc..,

**Batsmen Data set:**

It consists of features like:

'Batsmen', 'Runs', 'Balls Faced', 'Strike Rate', 'Matches Played',

'Sixes', 'Fours', 'Triples', 'Doubles', 'Singles', 'Dismissals',

'Caught', 'Bowled', 'RunOut', 'LBW', '50s', '100s', '150s', '200s',

'Bp6', 'Bp4', 'Best', 'Wins', 'Wins Percentage', 'Dots', 'Dots Rate',

'Boundary\_Runs', 'BR Percent'.

**Bowlers Data set:**

It consists of features like:

'Bowlers', 'Wickets', 'Balls Bowled', 'Overs Bowled', 'Strike Rate',

'Extras', 'Runs Conceded', 'Economy', 'Dots', 'Best'

**Match Data set:**

It consists of features like:

'match\_id', 'First Innings', 'Second Innings', 'First Innings Score',

'Second Innings Score', 'Winner', 'Batting wins', 'Bowling wins'

**Fielder Data set:**

It consists of features like:

'Fielders', 'Catches', 'Run Outs'

**Death & Power Play Overs Data set:**

It consists of features like:

'Bowlers', 'Runs Conceded', 'Wickets', 'Economy', 'Balls', 'Overs', 'Dots'

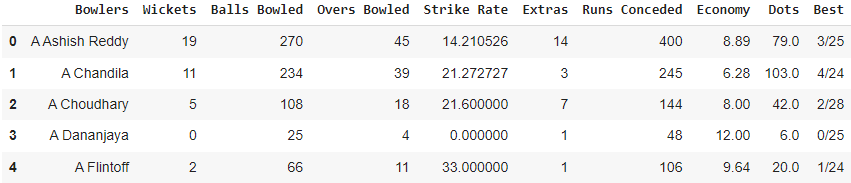
**Question 1:**

Balls per Wicket, Balls per Six, and Runs per Wicket

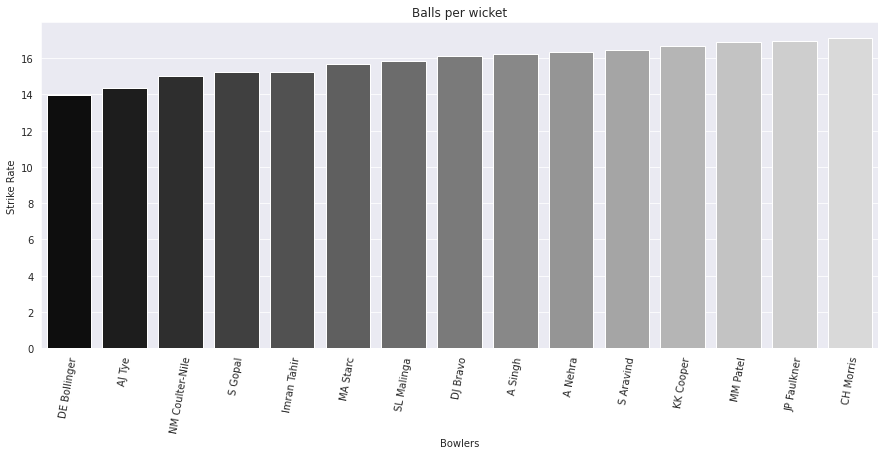
**Balls per Wicket:**

It is the Bowling Strike Rate of Bowlers, that shows the number of balls been bowled by a bowler after every wicket he takes.

**Data: Bowlers Data:**

****

**Strike Rate = Balls Bowled / Wickets**

****

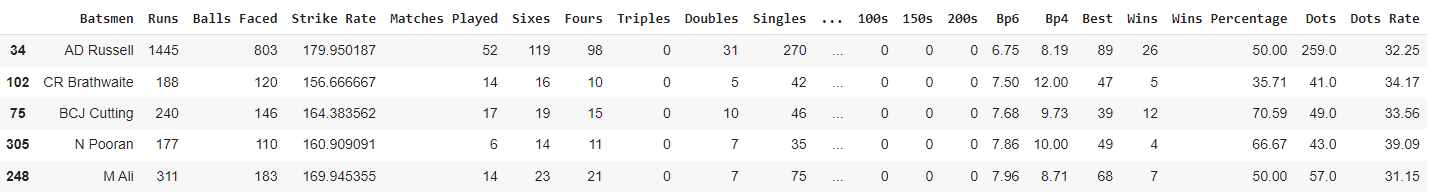
DE Bollinger has good Bowling Strike Rate

The working of creating this data set has been given through a notebook link above. Please refer to it.

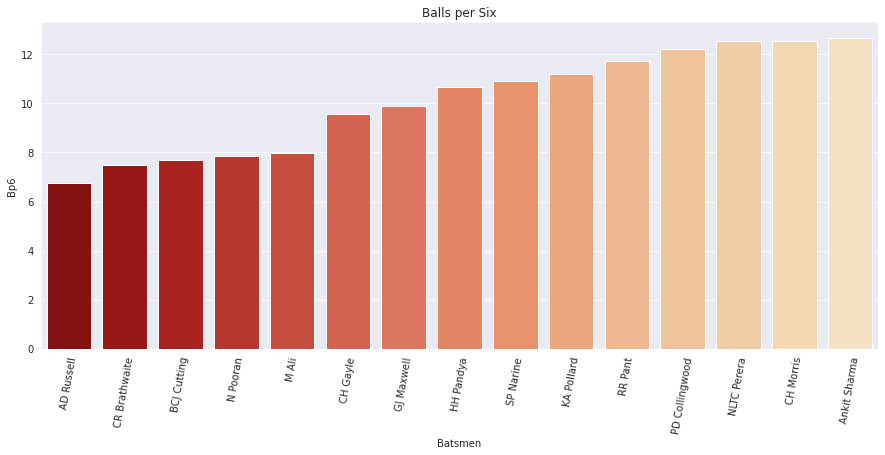
**Balls per Six:**

It is the average number of balls a batsman takes to hit sixes.

**Data: Batsmen Data:**



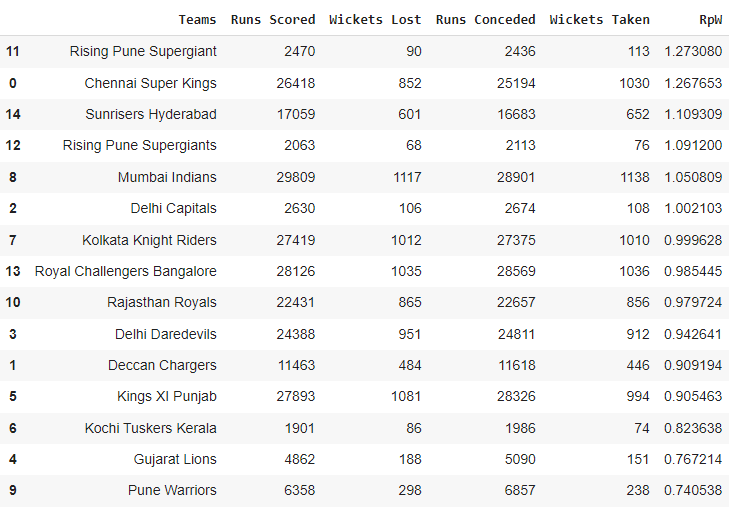
**Bp6 = Balls Faced / Sixes**

****

These are the top 15 players who are highly skilled in hitting sixes who are having Balls per six rate within 15.0

* It shows that AD Russell a destructive Player who has the chances of hitting a six in every 7(6.75) balls he faces.

**Runs per Wicket:**

****

**Runs per Wicket = [Runs Scored/ Wickets Lost] = [Runs Conceded/ Wickets taken]**

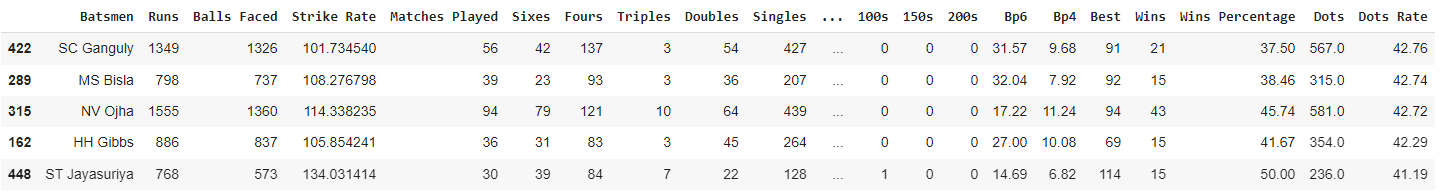
It is a Ranking metric, which is used to rank team, based on the runs and wickets they take, lose and score, conceded.

Chennai Super Kings should be ranked 1st in Run Quotient table [RPW is greater than 1], since they scored more runs than they conceded and they played more number of matches.

**Question 2:**

The ratio of total balls to total dots for each batsman

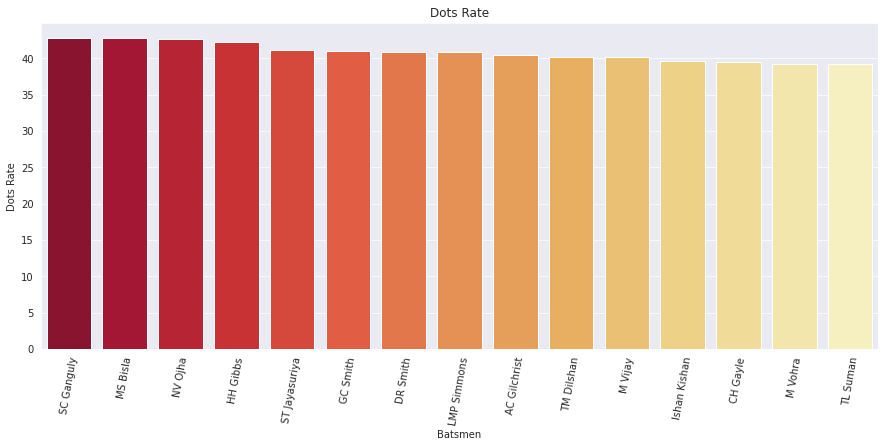
**Data: Batsmen Data:**

****

The working of creating this data set has been given through a notebook link above. Please refer to it.

The aim of this question is to see who are all the experienced players who have faced at least 500 Balls has higher Dots rate(Total Dot Balls in Total Balls faced)

* Balls faced is an important factor in this, since the dot rate for inexperienced players who have faced minimum number of balls still can have higher Dots Rate.



We could see players like

* SC Ganguly
* MS Bisla
* NV Oijha
* M Vijay etc..

And also some big hitters like

* DR Smith
* Ishan Kishan
* CH Gayle

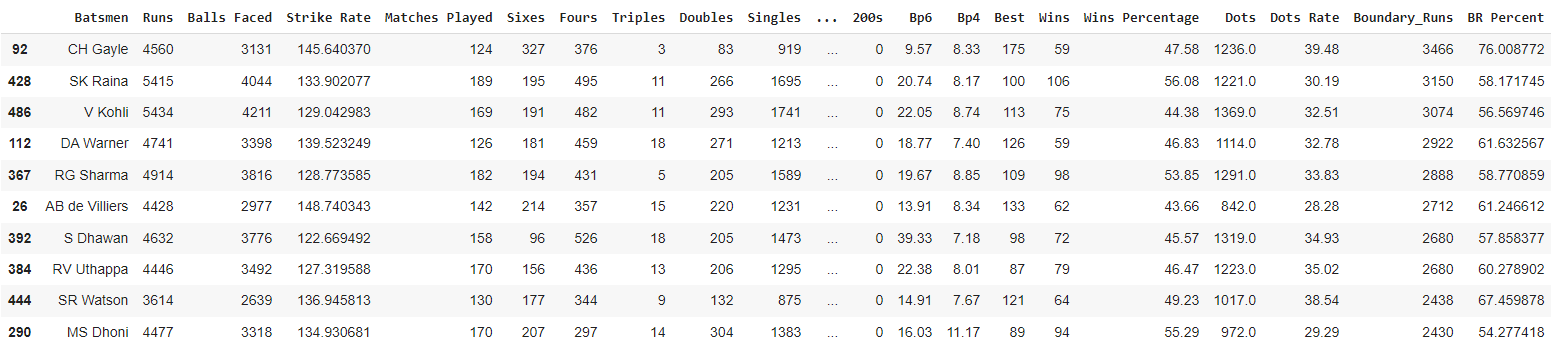
also comes under the list, who made 35-40 % of Balls they faced as Dot Balls.

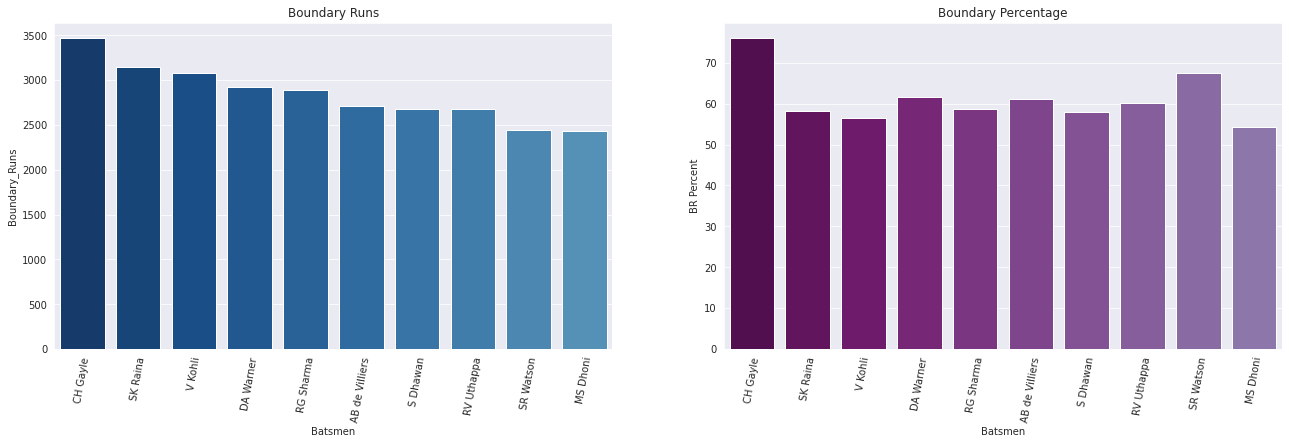
But on the other hand, we cannot make any inferences on the performances of these players based on their Dot Balls Rate, still there are players in this list who can go for a big chase and capable of scoring high scores with shorter balls.

**Question 3:**

Percentage of boundary runs to total runs for each batsman

**Data: Batsmen Data:**

****

****

In this above figure we could see the top **10** players who scored higher runs through boundaries. Such a way, we have

* **CH Gayle** on the top who scored **3466** runs only through boundaries

following him

* **SK Raina** on the top 2 who scored **3150** runs only through boundaries.

following him we also have players like

* V Kohli
* DA Warner
* RG Sharma
* AB de Villiars
* S Dhawan
* RV Uthappa
* SR Watson
* MS Dhoni

Looking on the Boundary Percentage:

We have again

* **CH Gayle** on the top, who has **76%** of his runs scored only through boundaries.

following him, we have

* **SR Watson** on the top 2, who has **67%** of his runs scored only through boundaries
* DA Warner
* AD de Villiars
* RV Uthappa etc..,

are the players we have following the top 2

**Question 4:**

Batsman responsible for the most wins

This question can be seen in many angles

* for example: Wins can be dependent on various factors not only on a good batting of Batsmen of a team. And those various factors include:
  + Performance of bowlers
  + Pitch
  + Toss
  + Weather and so on...

There are also certain situation/matches where a certain batsman scores high but the team will not win the game and also, a certain batsman won't perform well on that day, or they are only required to score few runs (if the first inning score is very low), the team will win on the other hand..

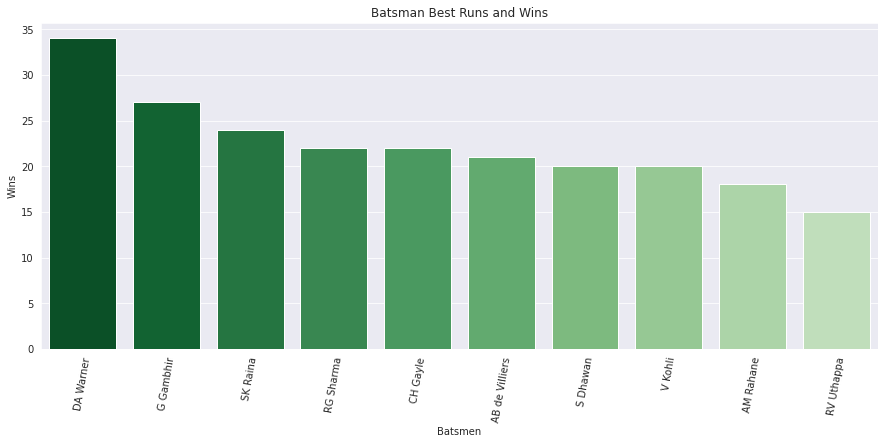
So, we can solve this through mapping best runs of a batsmen scored in a match to the result of the particular match.

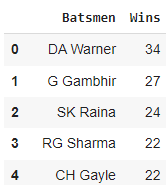
* We can frame condition like:

if Batsmen run == max(ie., scores 50+ and 100+) and Result == 'Win'

But again we cannot look this as pattern for win since, winning a match is not only dependent on a good batting of a batsman.

**Data: Batsmen runs with results:**

****

**** ****

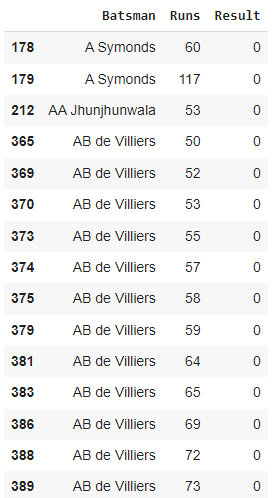
In the above figure we could see the player names like

* DA Warner
* G Gambhir
* SK Raina etc..,

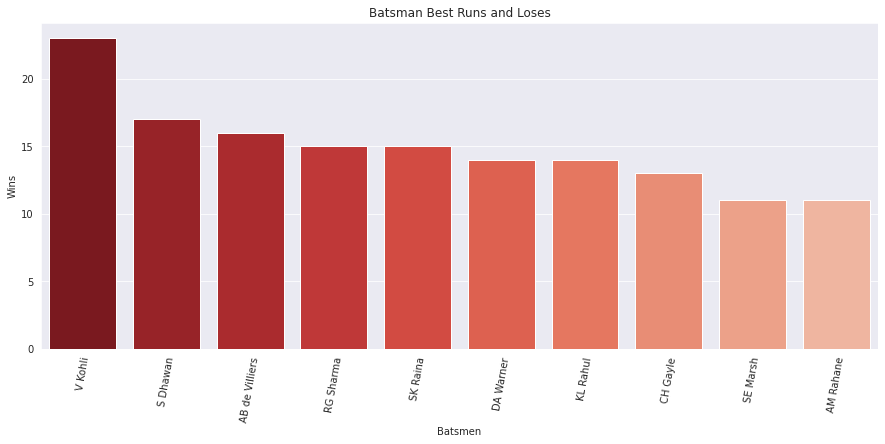
Where, they scored 50+ (more than 50 runs) and the number of wins they had by scoring so. Likewise

* **DA Warner** has **34** wins, when he scores more than 50 runs

To prove the irony which I mentioned before, we can also see players who scored more than 50 runs has the match result as 'Lose'.



See the consistency of AB de Villiars, who scores great in subsequent matches, but the results are disappointing.



This could be seen as the number of times a player been unlucky!

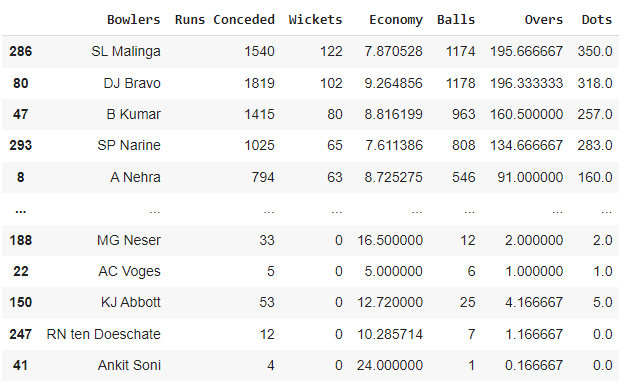
* V Kohli had good scores in **23** matches but RCB doesn't made it, in the end!

#### **Therefore winning is not only dependent on a Batsman Performance.**

**Question 5:**

Best Death Over Bowler

**Data: Death Over Data set**

****

Undoubtedly, the best death over bowler is **SL Malinga**

He conceded **1540** runs which is less than other bowlers in his class in the death over.

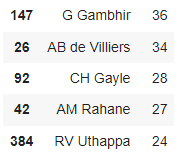
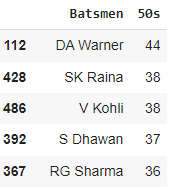
He had taken **122** wickets in death overs.

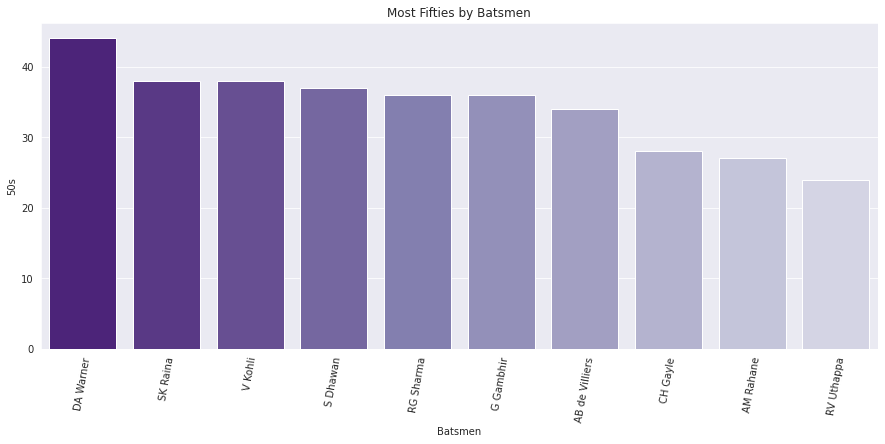
His Death over economy is **7.87**, which is again a comparatively a decent figure

**Question 6:**

Most fifties by Batsmen

**Data: Batsmen Data:**

****

****

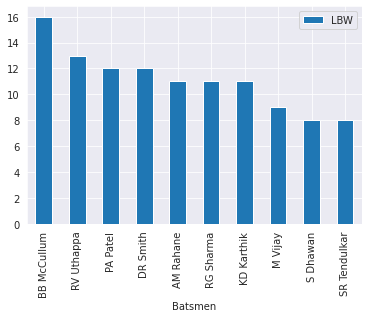
Here are the top 10 batsmen who has maximum 50s in IPL league matches.

* **DA Warner** is on the top, who made **44** half-centuries
* **SK Raina** is on the top 2 who made **38** half-centuries

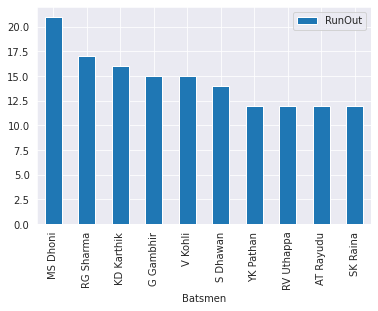
**Apart from the Questions:**

**Findings:**

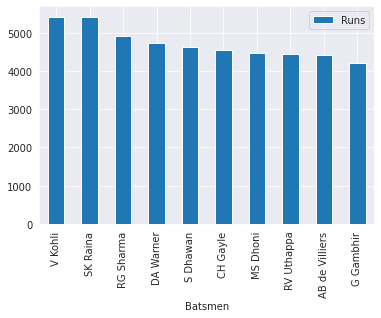
BB McCullum is the batsmen who have been dismissed mostly through LBW dismissals.

****

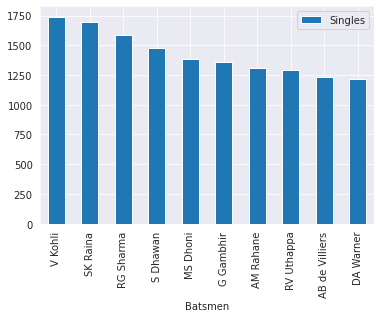
MS Dhoni is the batsmen who have been dismissed mostly through run outs.

****

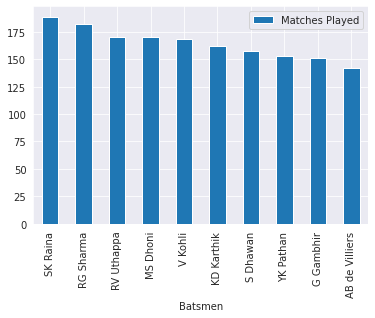
V Kohli is the Highest run scorer in IPL

****

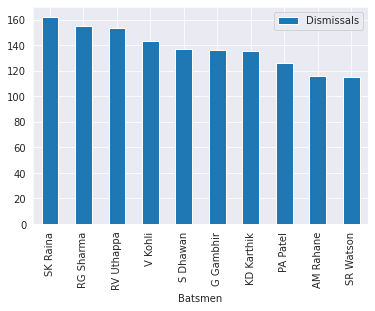
V Kohli is the batmen who made more singles running between the wickets.

****

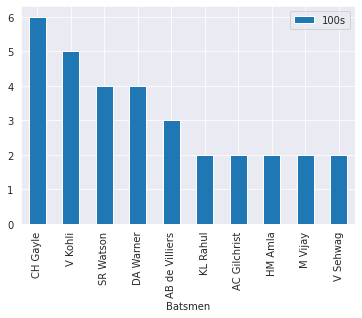
SK Raina is the batsmen who played a greater number of matches

****

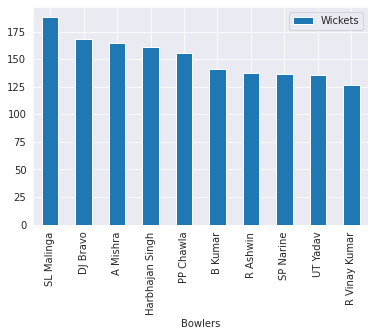
SK Raina is the Batsman who have been dismissed more.

****

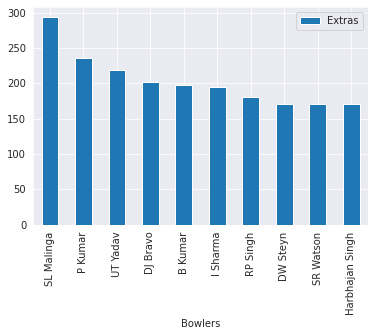
CH Gayle is the Batsman who have made more Centuries

****

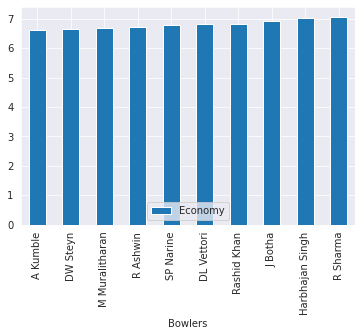
SL Malinga is the leading Wicket taker in IPL

****

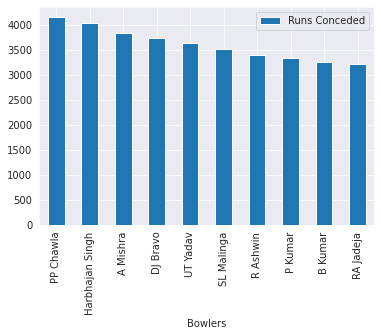
SL Mallinga is the bowler who bowled more Extras

****

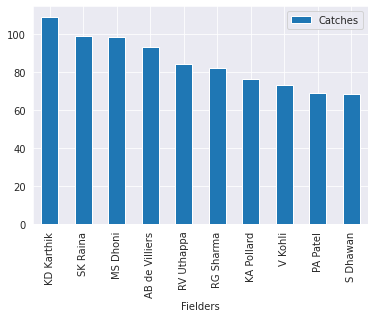
A Kumble is the bowler who has lowest Economy

****

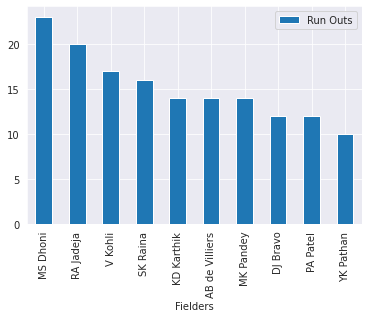
PP Chawla is the bowler who have conceded more runs in IPL as a Bowler

****

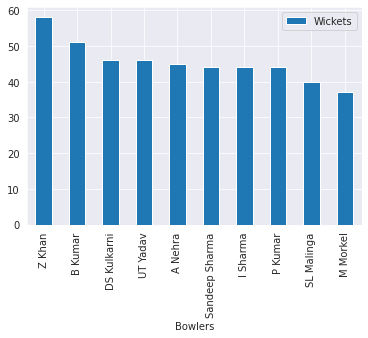
KD Karthik is the player who caught a greater number of catches

****

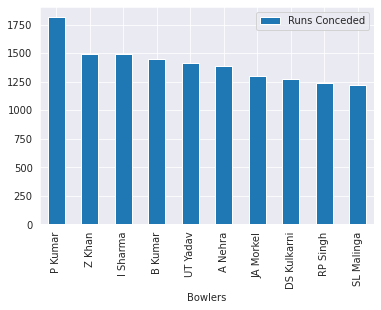
MS Dhoni is the Player who made more run out dismissals

****

Z Khan is the highest wicket taker in Power Play

****

P Kumar is the bowler who conceded more runs in Power Play

****

**End of the report**

**Conclusion:**

This IPL data set is a vast and a huge data set which has enormous of information, that can be extracted based on the problem and the requirements.

So far, we only Manipulated the data and extracted information from it, and visualized through tables and bar graphs, beyond this there is a greater scope for Analytics in prediction and modelling and performing Statistical Analysis with the data.

Example:

* 1. We can predict the chances of dismissal/ the chances of a batsman to score century in the future matches with the past data like this. It involves a practical application of probability and statistics.
  2. Statistics on head-to-head like how many times a particular batsman has been dismissed by a particular bowler.
  3. We can also create interactive dashboards for this data using BI tools like Power BI or Tableau.

Thanking You.

Rajagopal S

MSc. Data Analytics

[rajag2001416@gmail.com](mailto:rajag2001416@gmail.com)