DATA ANALYSIS AND VISUALIZATION

Final Project

Final Project

https://www.kaggle.com/shivanshsinghal107/cricbuzz-players-data

TANAY SONI

Context:

Cricket is a bat-and-ball game played between two teams of eleven players each on a cricket field, at the center of which is a rectangular 20-meter (22-yard) pitch with a target at each end called the wicket (a set of three wooden stumps upon which two bails sit). Each phase of play is called an innings, during which one team bats, attempting to score as many runs as possible, whilst their opponents bowl and field, attempting to minimize the number of runs scored. When each innings ends, the teams usually swap roles for the next innings (i.e., the team that previously batted will bowl/field, and vice versa). The teams each bat for one or two innings, depending on the type of match. The winning team is the one that scores the most runs, including any extras gained (except when the result is not a win/loss result).
Source: https://www.kaggle.com/shivanshsinghal107/cricbuzz-players-data.

Content:

The dataset contains two files: bat.csv and bowl.csv
Each file consists of both batting and bowling data of players from 12 different teams in IPL, ODI, T20, and Test Cricket till 16 Jan 2021.

The batting data (bat.csv) consists of the following columns:

Name	Description
Player	Player Name.
Team	Team for which the player plays.
Type	The match type it is.
M	Number of matches played.
Inn	Number of Innings.
NO	Not Out.
Runs	Total Runs scored.
HS	Highest Score.
Avg	Batting Average.
BF	Balls Faced.
SR	Strike Rate.
4s	Total number of 4 runs scored.
6s	Total number of 6 runs scored.
50	Total number of 50 runs scored in a single inning. (Half century)
100	Total number of 100 runs scored in a single inning. (Full century)
200	Total number of 200 runs scored in a single inning. (Double century)

In cricket, a player's **batting average (Avg)** is the total number of runs they have scored divided by the number of times they have been out, usually given to two decimal places. Since the number of runs a player scores and how often they get out are primarily measures of their own playing ability, and largely independent of their teammates, batting average is a good metric for an individual player's skill as a batter (although the practice of drawing comparisons between players on this basis is not without criticism). The number is also simple to interpret intuitively. If all the batter's innings were completed (i.e. they were out every innings), this is the average number of runs they score per innings. If they did

not complete all their innings (i.e. some innings they finished not out), this number is an estimate of the unknown average number of runs they score per innings.

Batting **strike rate** (SR) is defined for a batsman as the average number of runs scored per 100 balls faced. The higher the **strike rate**, the more effective a batsman is at scoring quickly.
Here in my dataset the type columns refers to the different formats of the game, like in Test Cricket match there are indefinite overs (6 balls in an over), in One-Day International (ODI) format the overs are limited to 50 overs (300 balls in an inning), in T-20 (Twenty-20) as the name suggests the overs are limited to 20 overs (120 balls in an inning), in Indian-Premier League (IPL) this is a Indian cricket series with limited over to 20 overs (120 balls in an inning).

Here, in the above data I have 16 variable and I have performed various exploratory data analysis techniques. Here I have taken Runs as my dependent variable and rest of the 15 variables are independent variable.

The Bowling data (bowl.csv) consists of the following columns:

Name	**Description**
Player	Player Name.
Team	Team for which the player plays.
Type	The match type it is.
M	Number of matches played.
Inn	Number of Innings.
B	Number of Balls thrown.
Runs	Total Runs scored.
Wkts	Total number of wickets.
BBI	Best bowling in Innings.
BBM	Best bowling in Match.
Econ	Economy Rate.
Avg	Bowling Average.
SR	Strike Rate.
5W	Total time of 5 wickets taken in a single inning.
10W	Total time of 10 wickets taken in a single inning.

Bowling Strike Rate is a measurement of a bowler's average number of balls bowled for every wicket taken. A lower strike rate is preferable – it means that the bowler can get more batsmen out with fewer balls. The statistic is considered to be more important in longer games than shorter 1-day matches. Bowling strike rate is a complimentary statistic to the more popular Batting strike rate.

In cricket, a player's **economy rate** is the average number of runs they have conceded per over bowled. In most circumstances, the lower the economy rate is, the better the bowler is performing. It is one of a number of statistics used to compare bowlers, commonly used alongside bowling average and strike rate to judge the overall performance of a bowler.

Here in the above dataset I have 15 variable out of which I removed two columns which I though of no relevance (i.e., BBI and BBM), here my dependent variable is Wkts (wickets) and the rest 14 columns are my independent variable.

Data has been extracted from cricbuzz using python scripting and web scraping.

Acknowledgements:

Source: Cricbuzz
Data is available on the website in the team's column, wherein the teams are specified, and for each team, each player has an individual page having batting and bowling data (in form of tables).

Inspiration:

Research purpose: To build some scale, on which each player can be measured combining his batting and bowling performances in various types of matches.

Graphs:

For plotting different charts of the above datasets I had to use the tableau's join function and I joined the above two datasets on the name of players.

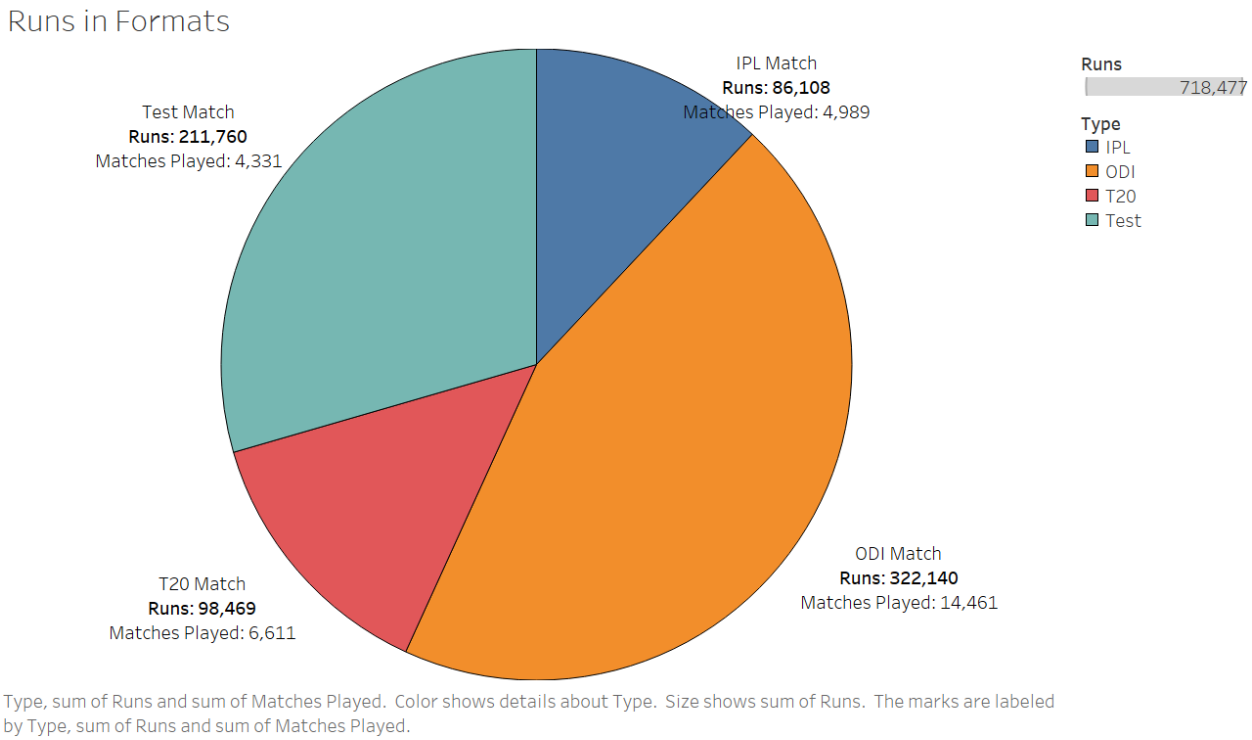
The graphs which I tried to plot are on the next page.

Details:

For details on the game of cricket you can go on the following link.

Summary of game Cricket.

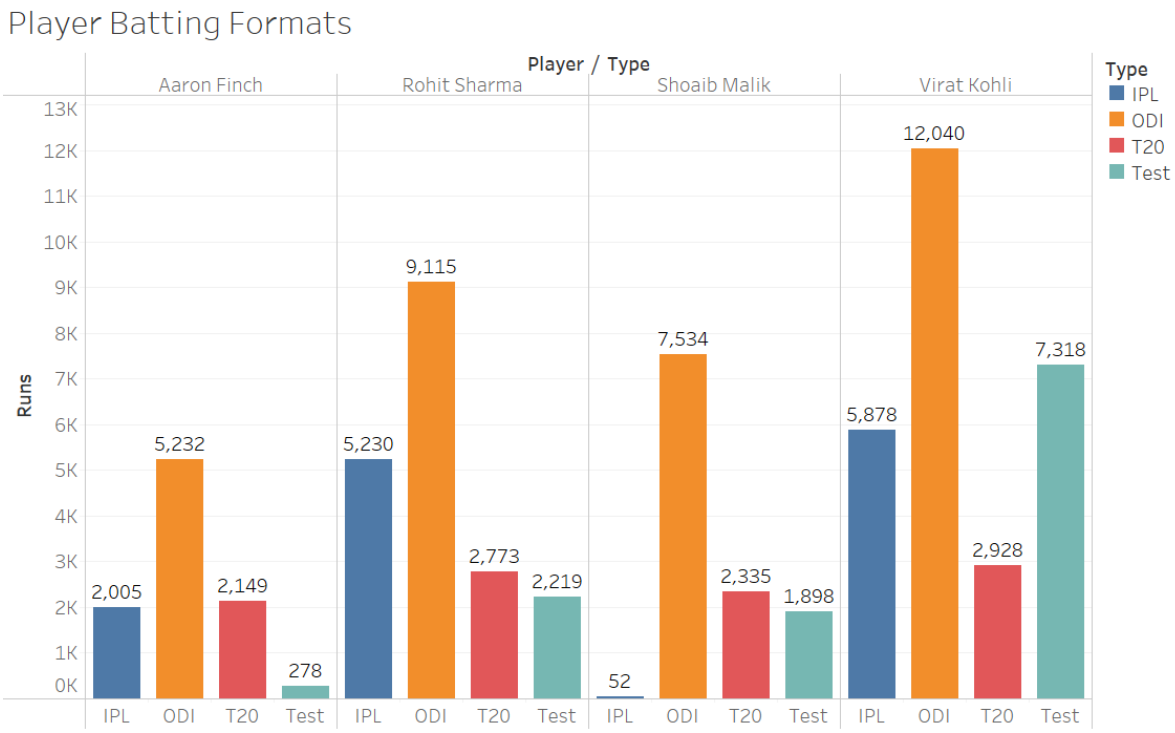
1. **Pie-Chart Comparision between the number of matchs played in each format and the runs scored in each format.**



In the above pie-chart, the angle of each section is determined by the number of runs scored in that format. Here, it very evident that maximum runs are scored in the One-Day International (ODI) format, then there is the Test format. Here we can also see the number of matches played in that format, which is also a very good factor affecting the above pie-chart as it is very basic that the more a player plays the more, he/she is going to score. Same goes here, the more matches played the more runs are scored in that format.

Here for the test match anomaly, it is due to the no restrictions in overs (1 over = 6 balls), but in other formats there is over restrictions (i.e., in ODI one inning comprises of 50 overs (300 balls), in T-20 one inning comprises of 20 overs (120 balls) and same in IPL. That is the share of runs is more in test match with respect to the total match played.

2. Player wise comparison between the runs scored in different formats of cricket.

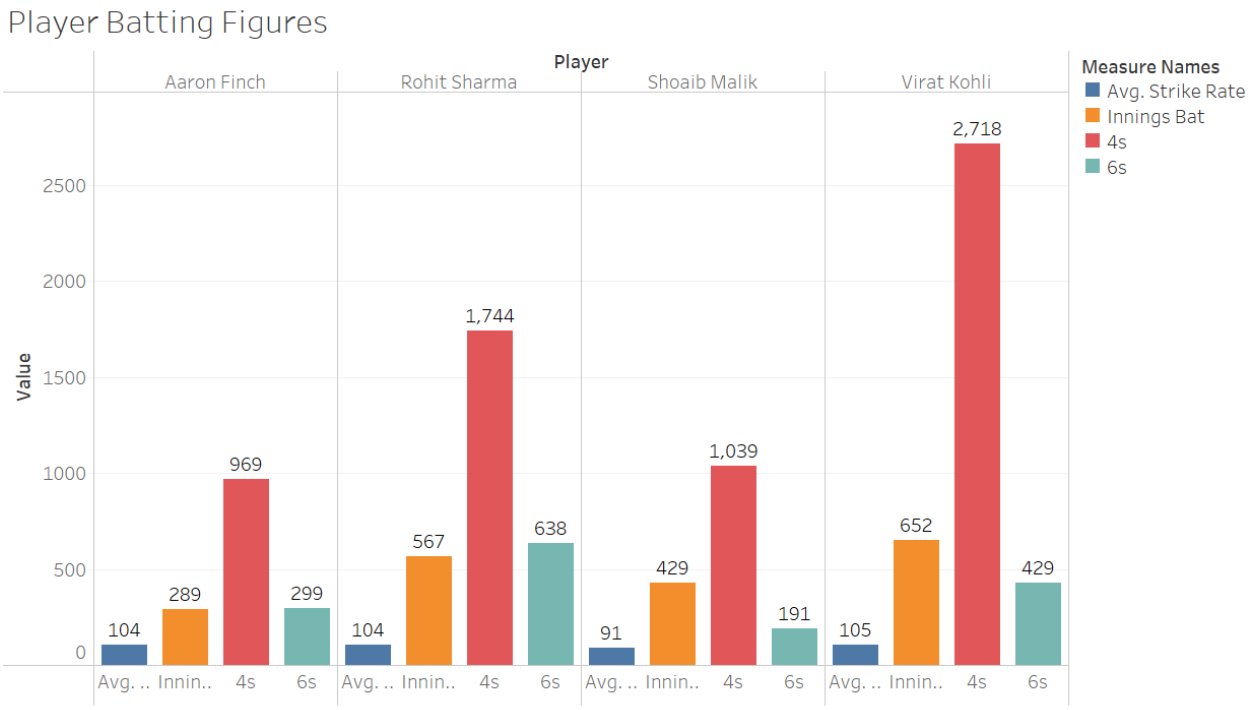


In this side-by-side bar chart I tried to show the different formats of cricket and the runs scored by the player in that format. This is an interactive chart which is very interactive when seen in the dashboard named as Batting. In that dashboard one can select the players to compare their runs scored in different formats of cricket.

Here just for reference I compared four players (Aaron Finch, Rohit Sharma, Shoaib Malik and Virat Kohli). For the low run in IPL format of Shoaib Malik can be because of the ban of Pakistani players after the Mumbai Terror attacks (26/11) in Indian Premier League (IPL) as he only played seven matches in the IPL and in those, he only scored 52 runs.

The chart tells us the story that maximum runs are either scored in Test or ODI format as they are the only format where one can face more balls when compared to other formats.

3. Player wise comparison of their batting performance.



Avg. Strike Rate, Innings Bat, 4s and 6s for each Player. Color shows details about Avg. Strike Rate, Innings Bat, 4s and 6s. The marks are labeled by Avg. Strike Rate, Innings Bat, 4s and 6s. The data is filtered on Type, which keeps IPL, ODI, T20 and Test. The view is filtered on Player, which keeps Aaron Finch, Rohit Sharma, Shoaib Malik and Virat Kohli.

This side-by-side interactive bar chart for the presentation purpose I took only four player and their batting statistics.

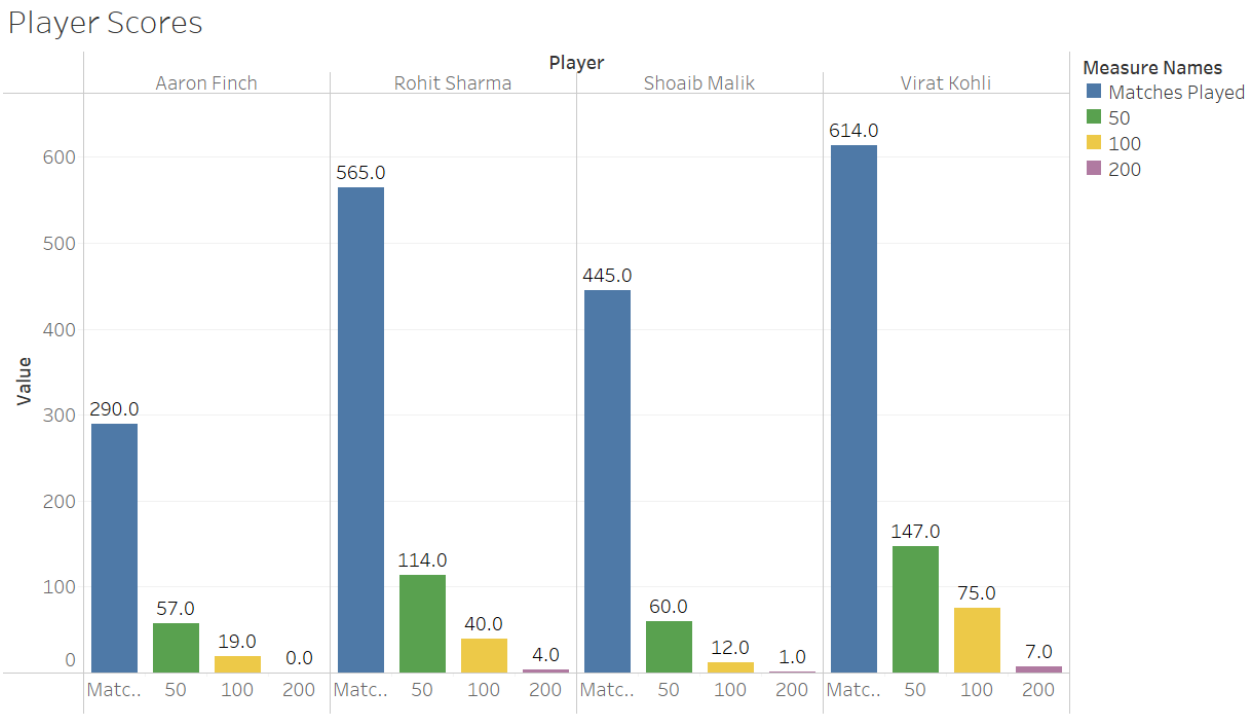
The stike rate is defined for a batsman as the average number of runs scored per 100 balls faced. The higher the **strike rate**, the more effective a batsman is at scoring quickly. Here, I took the average of the strike rate of the player's performance in all the formats of cricket. In this way I got a good insight of the overall performance of the players.

The innings are the times the player has played a ball in the match. There are two innings in a cricket game (ODI, T-20 and IPL) and 4 innings in a Test game. That's why it was a important variable in defining a players performance.

4s and 6s means the this much of runs scored in a single ball. Over here in the chart we can clearly see that Virat Kohli loves to deal runs in 4s when compared to 6s as discussed in **(Homework 3).**

This chart can be best viewed in the dashboard name as Batting.

4. Player wise comparison of their batting statistics.



Matches Played, 50, 100 and 200 for each Player. Color shows details about Matches Played, 50, 100 and 200. The marks are labeled by Matches Played, 50, 100 and 200. The data is filtered on Type, which keeps IPL, ODI, T20 and Test. The view is filtered on Player, which keeps Aaron Finch, Rohit Sharma, Shoaib Malik and Virat Kohli.

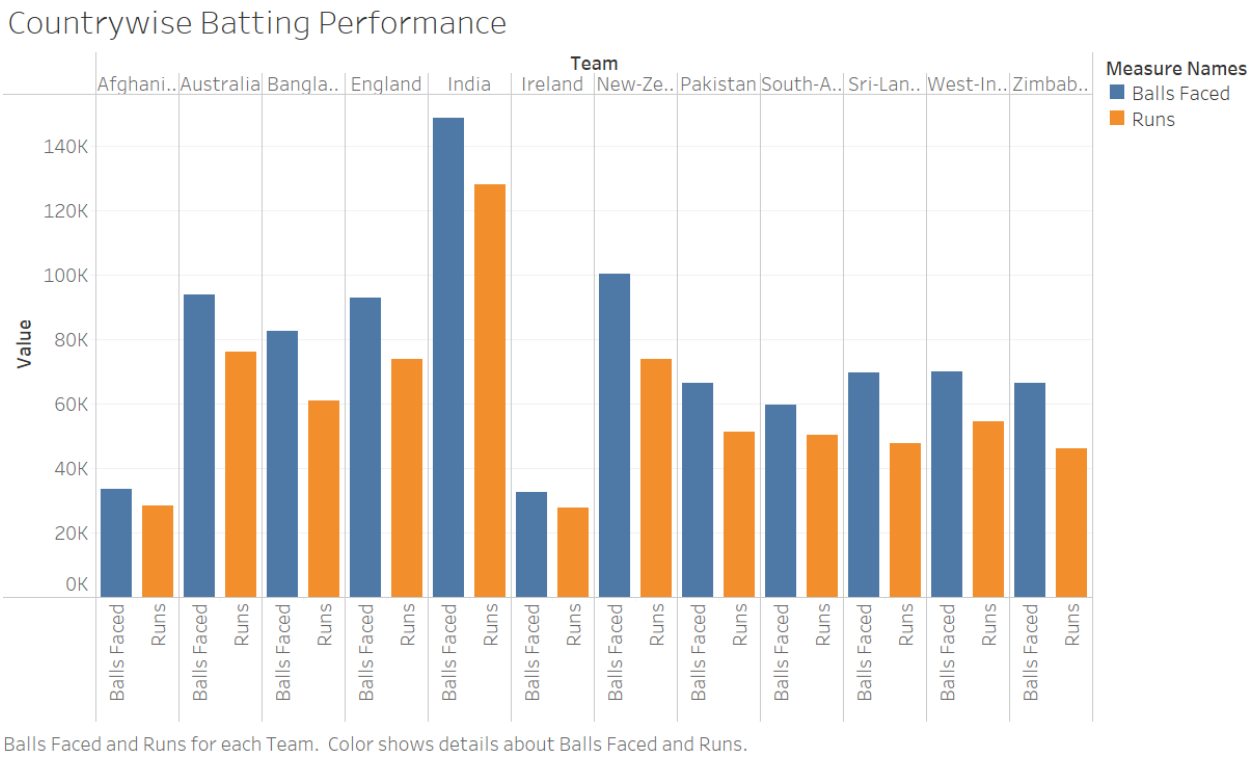
This side-by-side chart is also an interactive chart which can be best viewed in the dashboard named as batting.

For presentation purpose I choose four player and showed their batting statistics.

Here, Matches Played means the total number of matches played by a player in different formats of the game. 50 means the total of 50 runs scored by a player in a single inning without getting out in all the formats of the game. 100 means the total of 100 runs scored by a player in a single inning without getting out in all the four formats of the game. While 200 means the total of 200 runs scored in a single inning in any format of the game.

Here we can clearly see that why Virat Kohli is the world's best batsman. He played a total of 614 matches out of which he got 147 fifties, 75 centuries and 7 double centuries, which is in itself a record.

5. Country-wise batting performance.

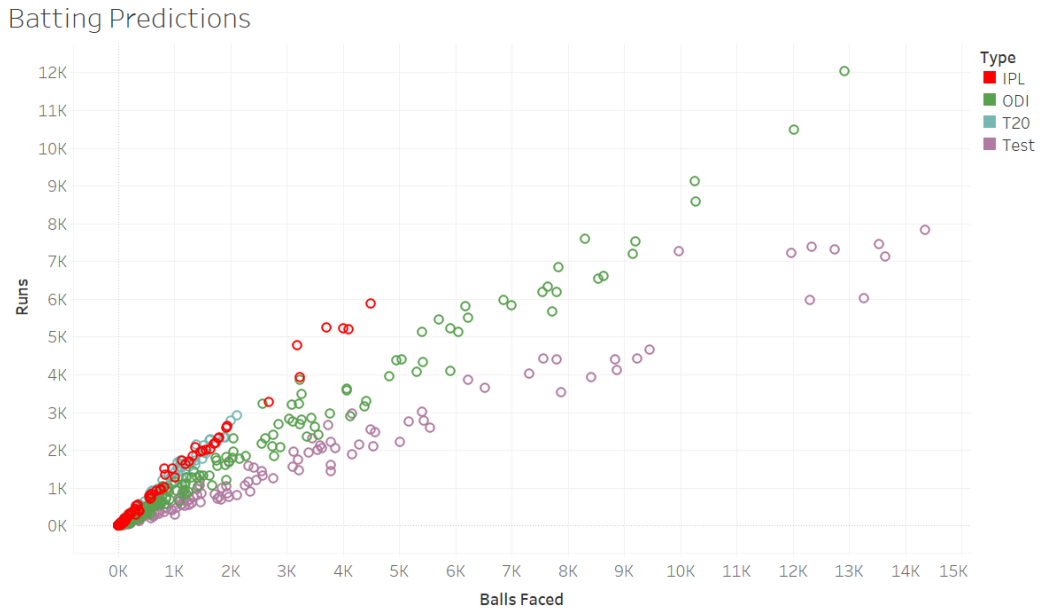


This is a side-by-side bar chart which shows the data of 7 different countries of the number of balls faced and the number of runs they scored in the balls faced.

The close the two bar charts to each other the faster they score runs.

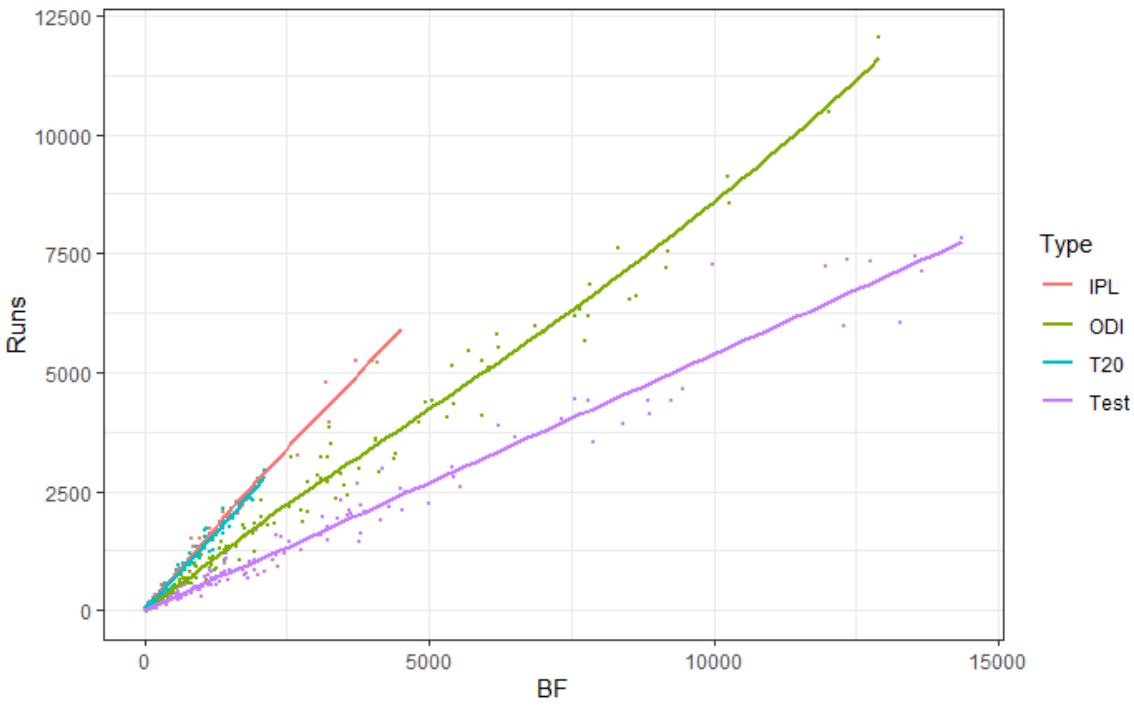
Here the values of Afghanistan, and Ireland is low because they do not play much in the international formats.

6. Batting Predictions on different types of formats.



This is a scatter plot between the two continuous variables ball faced and the runs scored from the different formats of the game of cricket.

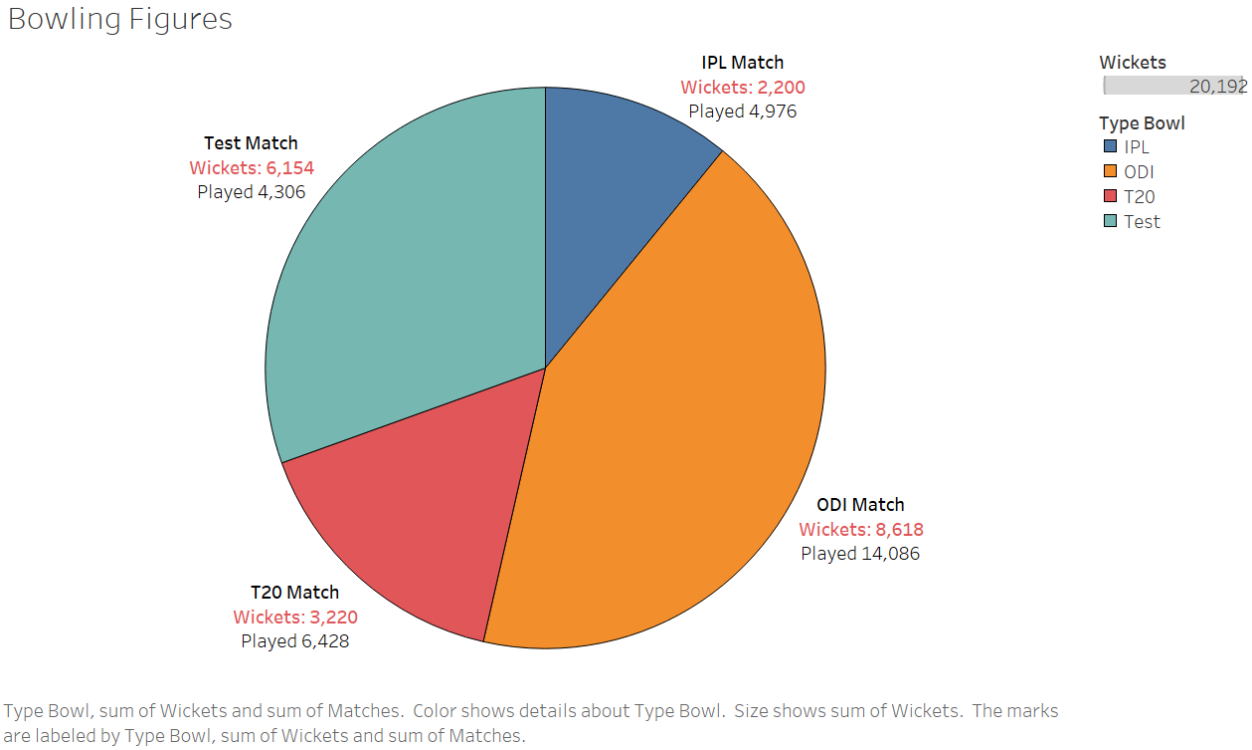
Here I can clearly see that there is some linear relation between the balls faced and the runs scored that's why I used the R language, to get a linear regression chart.



In the above linear chart I can clearly see that the rate of scoring runs is less in Test, then ODI, then T-20 and then IPL. In IPL players don't play for their country, instead they play for different clubs like in soccer. So that can be the reason about the high scoring rate in the IPL as players don't have much pressure of playing for their country. While in Test there are infinite balls so, players don't take risks and try to hit the bad deliveries only.

Furthermore, from the linear regression chart it is very evident that the more balls played in any format the more runs are scored. The only difference is the rate at which the runs are scored.

7. **Pie-Chart comparison between the number of matches played and the wickets taken in different formats of the game of cricket.**

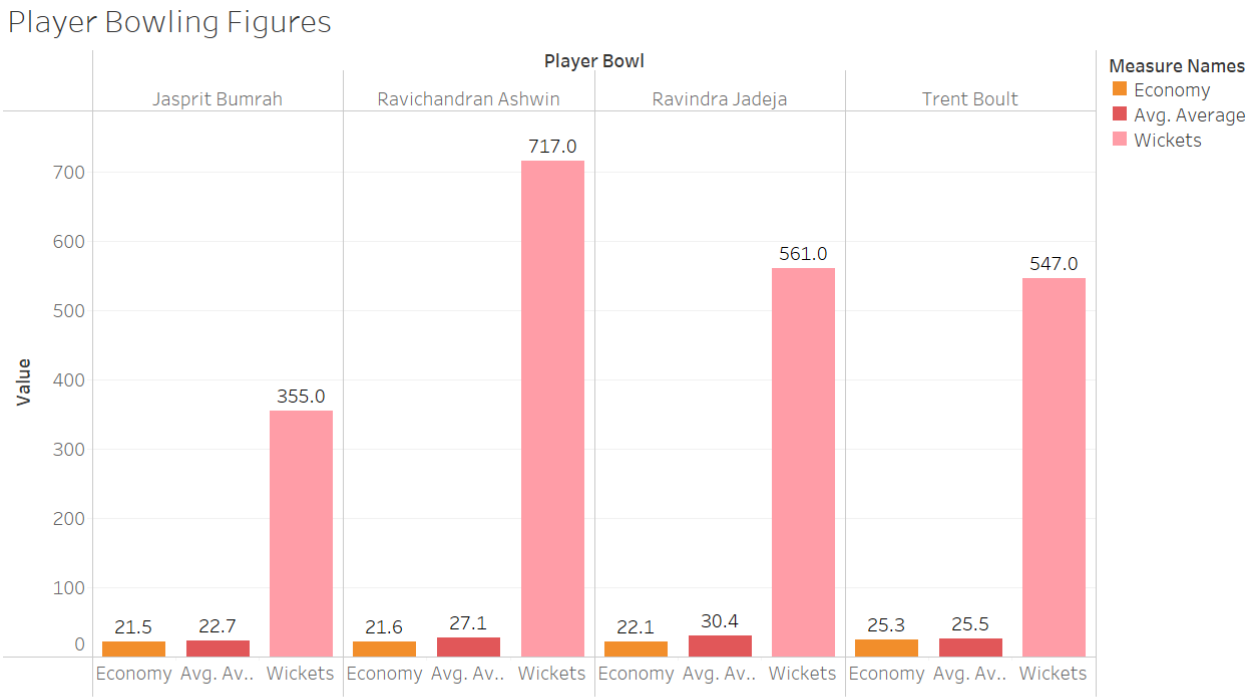


Type Bowl, sum of Wickets and sum of Matches. Color shows details about Type Bowl. Size shows sum of Wickets. The marks are labeled by Type Bowl, sum of Wickets and sum of Matches.

This pie-chart clearly shows the amount of wickets (getting one player out) taken in the different formats with respect to the matches played.

Here the angles are decided by the amount of wickets in that format. There are more wickets in ODI because the number of matches are also more. This means that matches also plays an important role in deciding the number of wickets.

8. Player-wise comparison of their bowling statistics.



Economy, Avg. Average and Wickets for each Player Bowl. Color shows details about Economy, Avg. Average and Wickets. The marks are labeled by Economy, Avg. Average and Wickets. The data is filtered on Type Bowl, which keeps IPL, ODI, T20 and Test. The view is filtered on Player Bowl, which keeps Jasprit Bumrah, Ravichandran Ashwin, Ravindra Jadeja and Trent Boult.

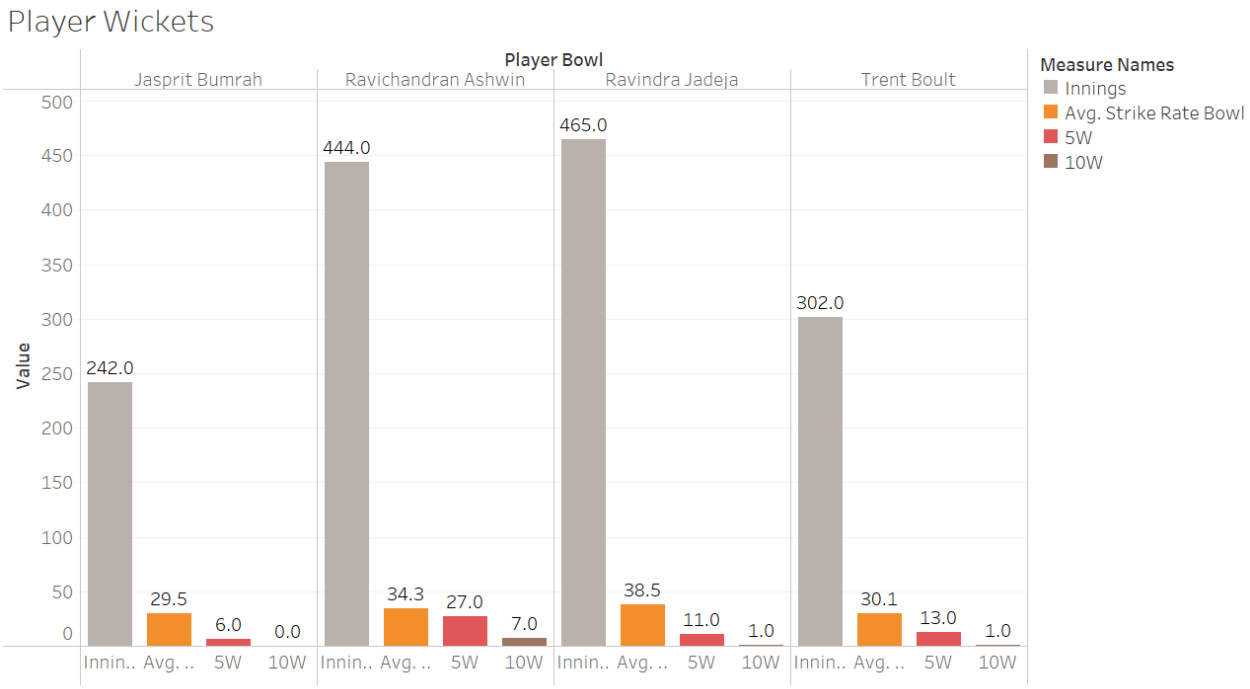
This is a side-by-side interactive bar chart, which can be best seen in the bowling dashboard. For presentation purpose, I have taken the stats of four players Jasprit Bumrah, Ravichandran Ashwin, Ravindra Jadeja and Trent Boult. They are chosen randomly.

In cricket, a player's **economy rate** is the average number of runs they have conceded per over bowled. In most circumstances, the lower the economy rate is, the better the bowler is performing. It is one of a number of statistics used to compare bowlers, commonly used alongside bowling average and strike rate to judge the overall performance of a bowler.

Furthermore, in cricket, a player's **bowling average (Avg.Average)** is the number of runs they have conceded per wicket taken. The lower the bowling average is, the better the bowler is performing. It is one of a number of statistics used to compare bowlers, commonly used alongside the economy rate and the strike rate to judge the overall performance of a bowler.

Here in the above graph the values are from all the four formats of the game of cricket to use the filter and get most out of the graph please go to the bowling dashboard. Where one can enter any number of player to get the most out of the comparison, with respect to the different formats.

9. Player-wise comparison of their bowling performance.



Innings, Avg. Strike Rate Bowl, 5W and 10W for each Player Bowl. Color shows details about Innings, Avg. Strike Rate Bowl, 5W and 10W. The marks are labeled by Innings, Avg. Strike Rate Bowl, 5W and 10W. The data is filtered on Type Bowl, Action (Player (bowling cleaned.csv)), Action (Player (bowling cleaned.csv),Type (bowling cleaned.csv)) and Action (Player Bowl). The Type Bowl filter keeps IPL, ODI, T20 and Test. The Action (Player (bowling cleaned.csv)) filter keeps 205 members. The Action (Player (bowling cleaned.csv),Type (bowling cleaned.csv)) filter keeps 820 members. The Action (Player Bowl) filter keeps 205 members. The view is filtered on Player Bowl, which keeps Jasprit Bumrah, Ravichandran Ashwin, Ravindra Jadeja and Trent Boult.

This is also an interactive chart which can be best seen in the dashboard named bowling.

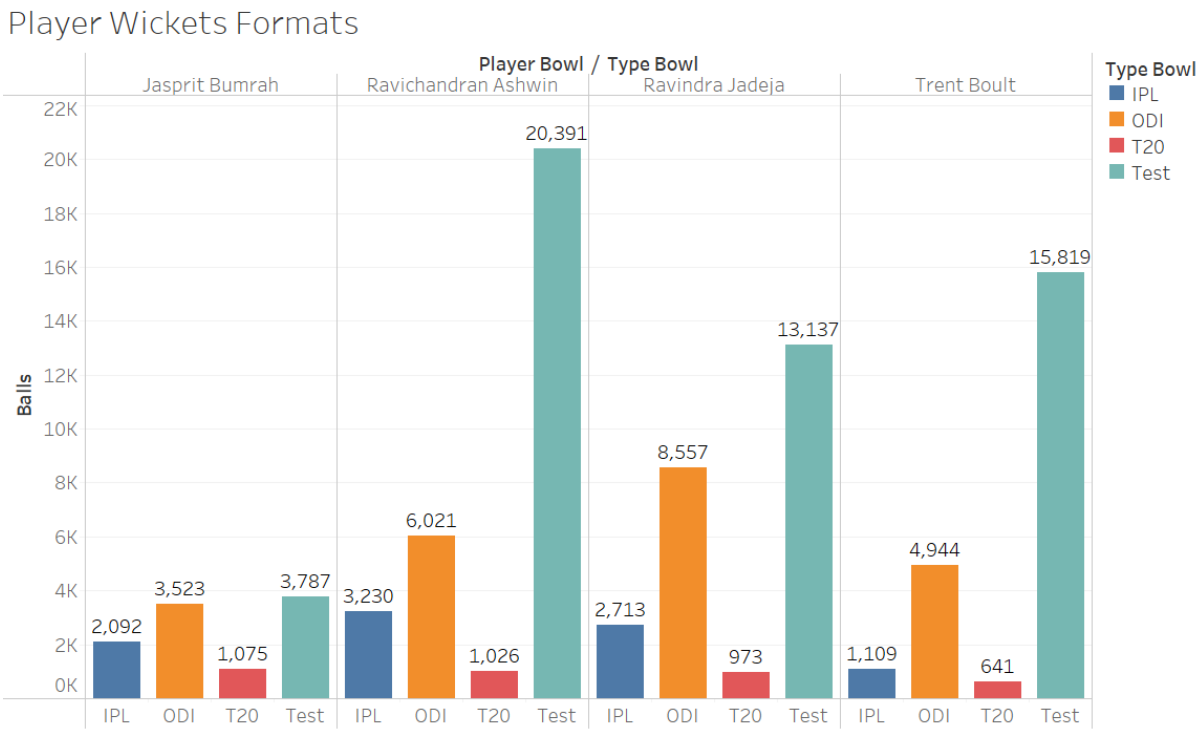
Bowling strike rate is defined for a bowler as the average number of balls bowled per wicket taken. The lower the strike rate, the more effective a bowler is at taking wickets quickly.

Here in the graph, I tried to plot the number of innings the player played, with average strike rate, times he/she has taken 5 wickets in a single innings and the times he/she has 10 wickets in a single innings.

For presentation purpose I have taken four players and compared their performances.

This graph is interlinked with other graphs which get change by selecting the type of format and the player.

10. Player-wise comparison of balls thrown in different formats of the game.

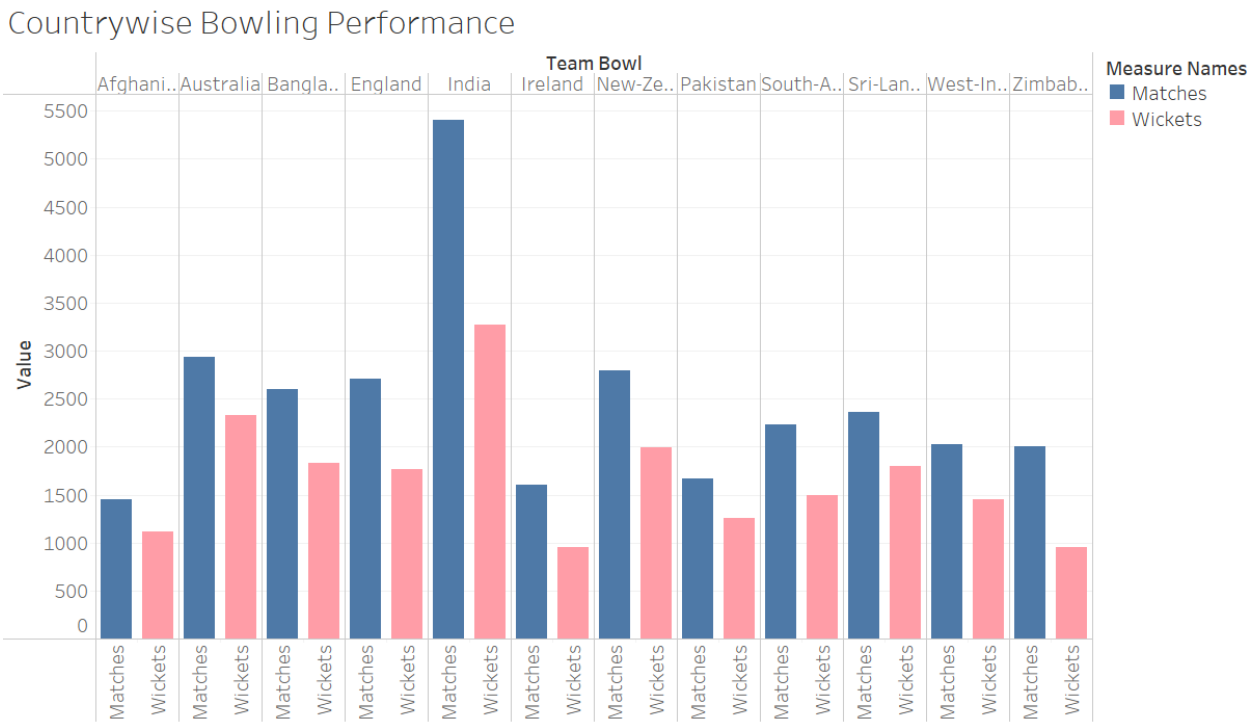


This is also an interactive chart which can be best seen in the bowling dashboard.

Here I tried to plot the total number of balls a player thrown in different formats of the game. And from this chart it is very evident that in test format there are a huge number of balls thrown by every player when compared to the other formats because in test format of cricket there is no limit of overs, means till the opposition teams all players don't get out the other team has to throw the overs, or the opposition team don't decide to declare their inning. The test format cricket match is of five days.

That is why I thought it is a good point to show in the dashboard.

11. Country-wise bowling performance.

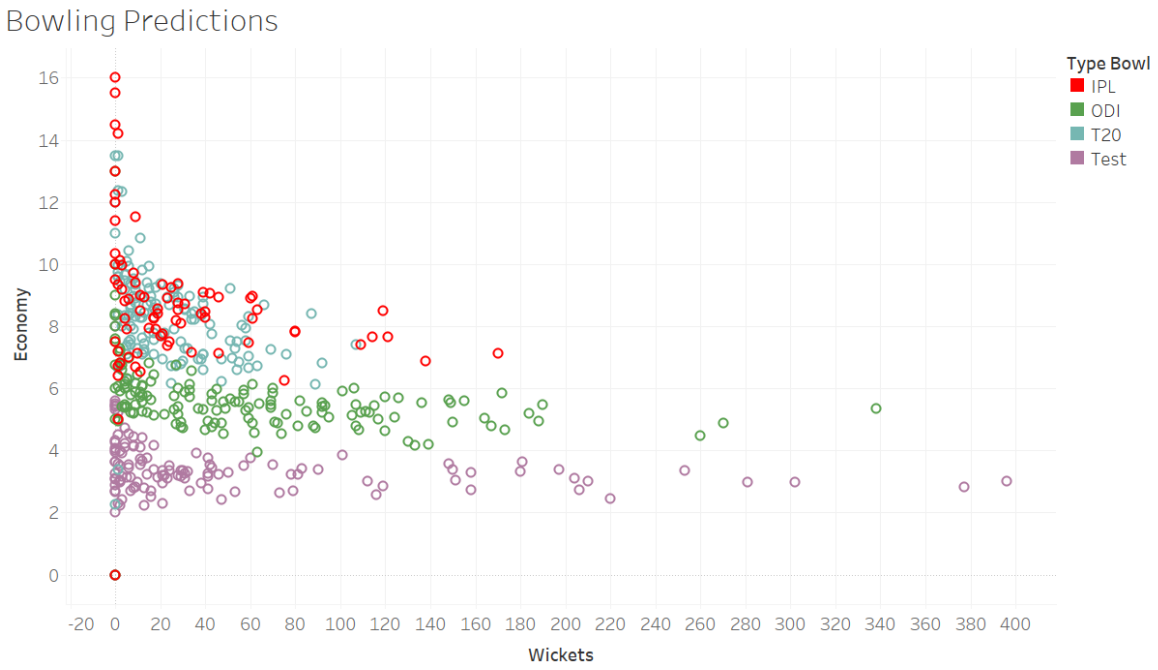


This is a side-by-side bar chart of countries and their bowling performances.

Here, in this graph I can conclude that teams like Afghanistan and Ireland has not played much if the matches on the international platforms. And team India's bowling performance is not that great when compared to the other countries as the wickets to match ratio is almost equals to half.

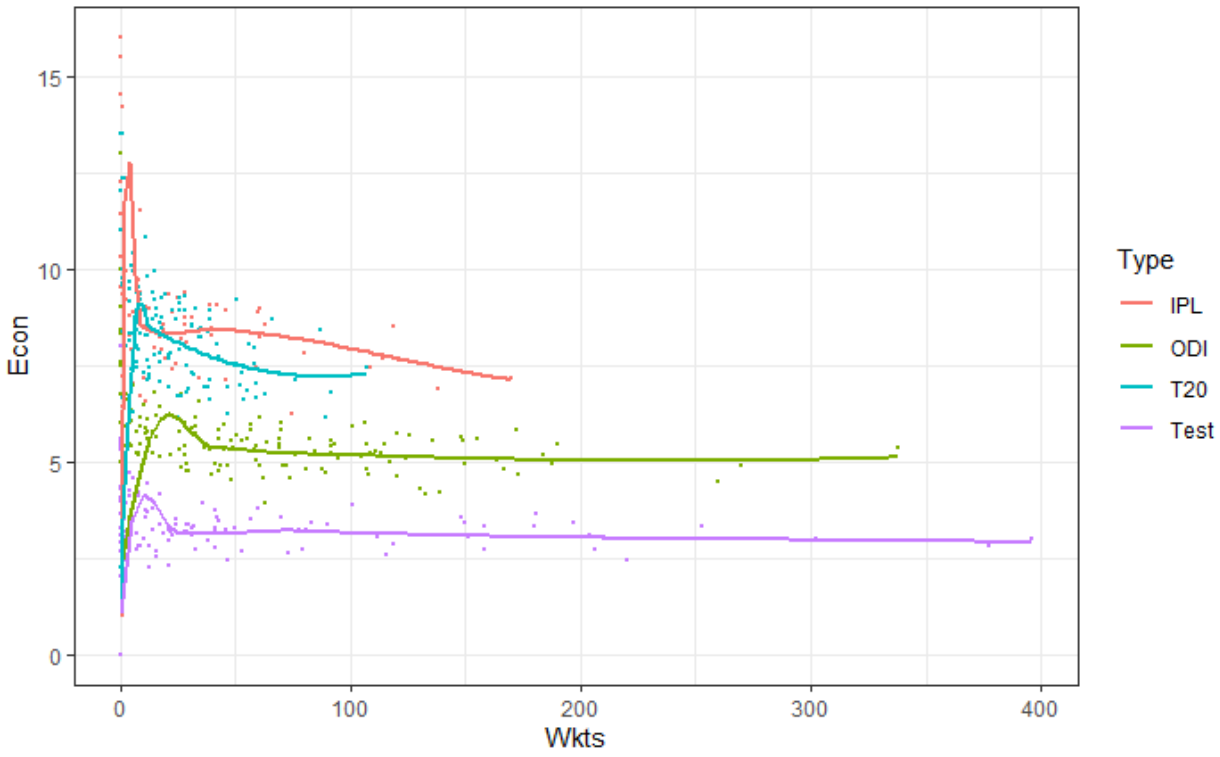
This plot is a good estimator of bowling performances of different countries.

12. Predictions using the wickets and the economy.



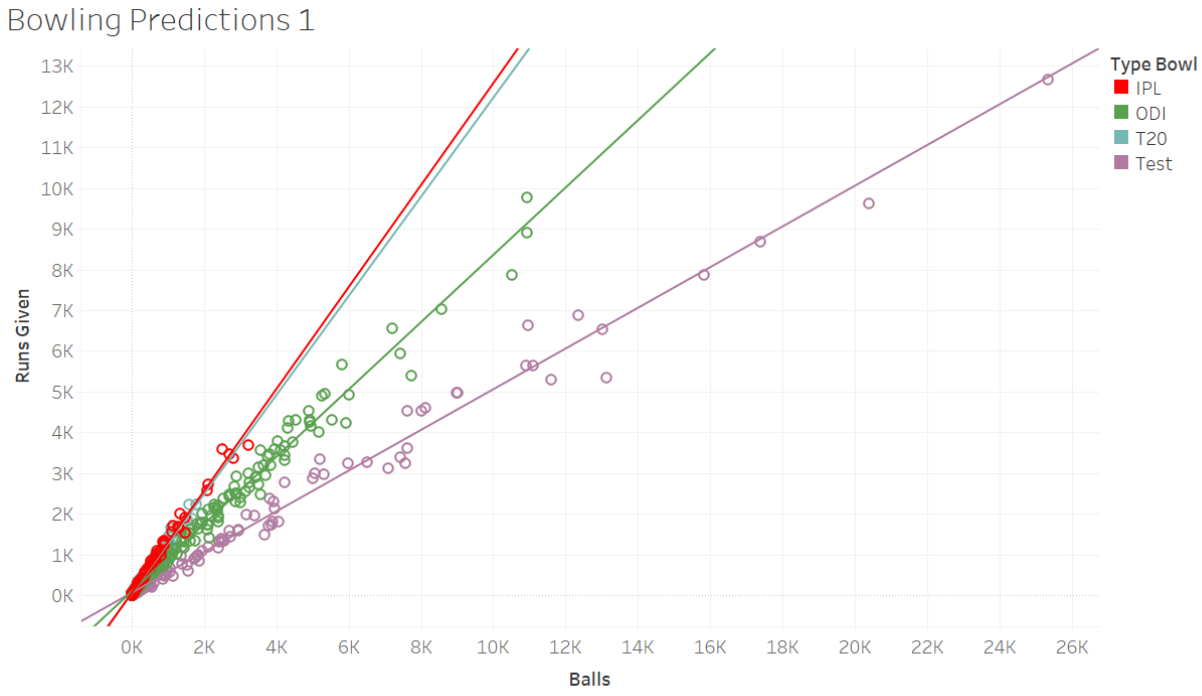
Wickets vs. Economy. Color shows details about Type Bowl. The data is filtered on Player Bowl, which keeps 205 of 205 members.

This is a scatter plot between the two continuous variables i.e., wickets and the economy. And here I can see a logistic regression going on.



Here in the logistic regression which I performed on R Studio and I can clearly see a pattern going on like in the case of ODI the economy at max reaches a value of 6-7 and then almost a constant at 5. This can be because of the number of wickets rise the economy starts to settle. Bowling economy is a very good factor in deciding the bowling abilities of any player.

13. Bowling predictions using balls thrown and runs given.

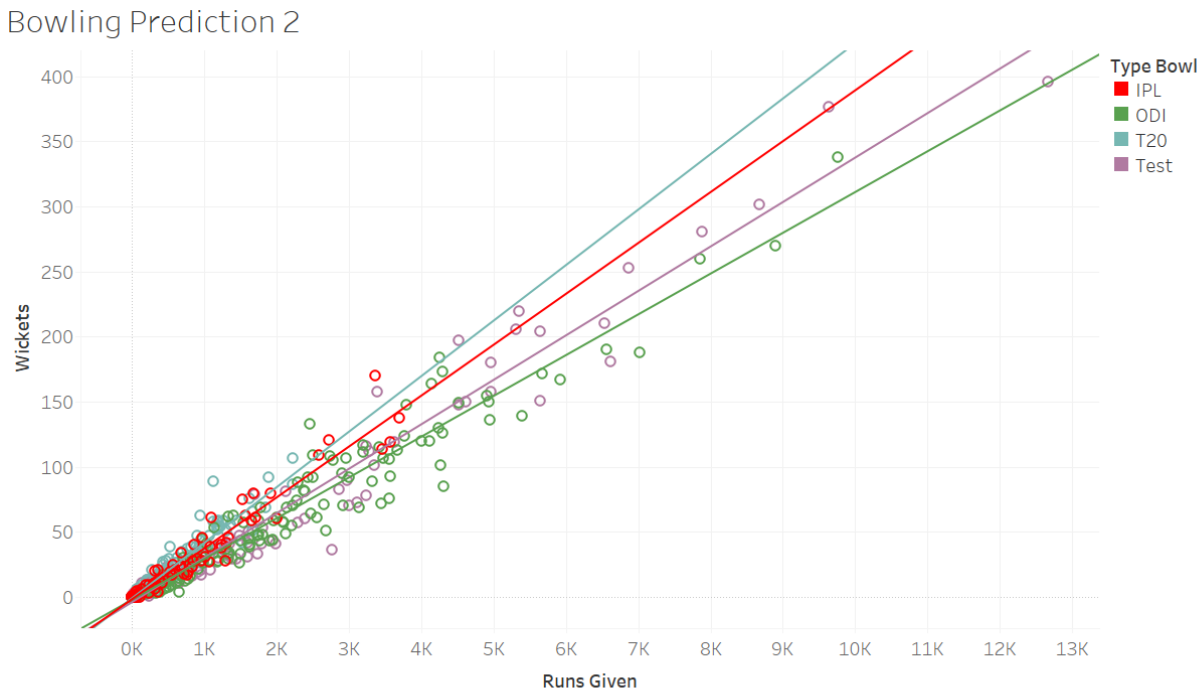


Balls vs. Runs Given. Color shows details about Type Bowl. The view is filtered on Type Bowl, which keeps IPL, ODI, T20 and Test.

This is a scatter plot between the two continuous and very important variables i.e., Runs given vs Balls. Here I used a tableau integrated function “trend lines” to show the trend going on in my dataset. I plotted a correlation matrix at first in R Studio and came with this graph.

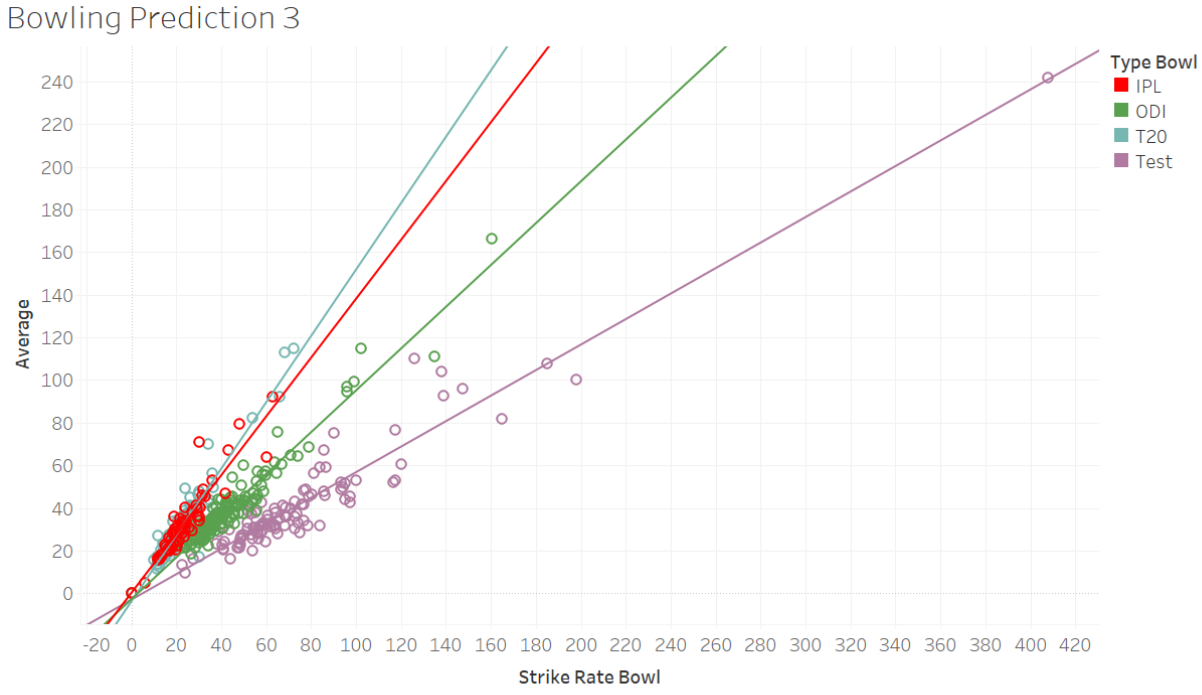
As from the graph the numbers of balls increase the number of runs also increases but vary with respect the format of the game. Like in test batsman don’t usually takes risks and try to play safer that’s why the rate of change in runs with respect to balls is low in test format of cricket. While in T-20 and IPL there is less balls to play and the more the runs one team score the probability of their winnings increases. So therefore batsman takes risks and try to hit 6s and 4s more as compared to 1s,2s,3s. That is the main reason behind the steep slope of the trend line of IPL and T-20 formats of cricket.

14. Wicket Predictions using Runs Given vs Wickets



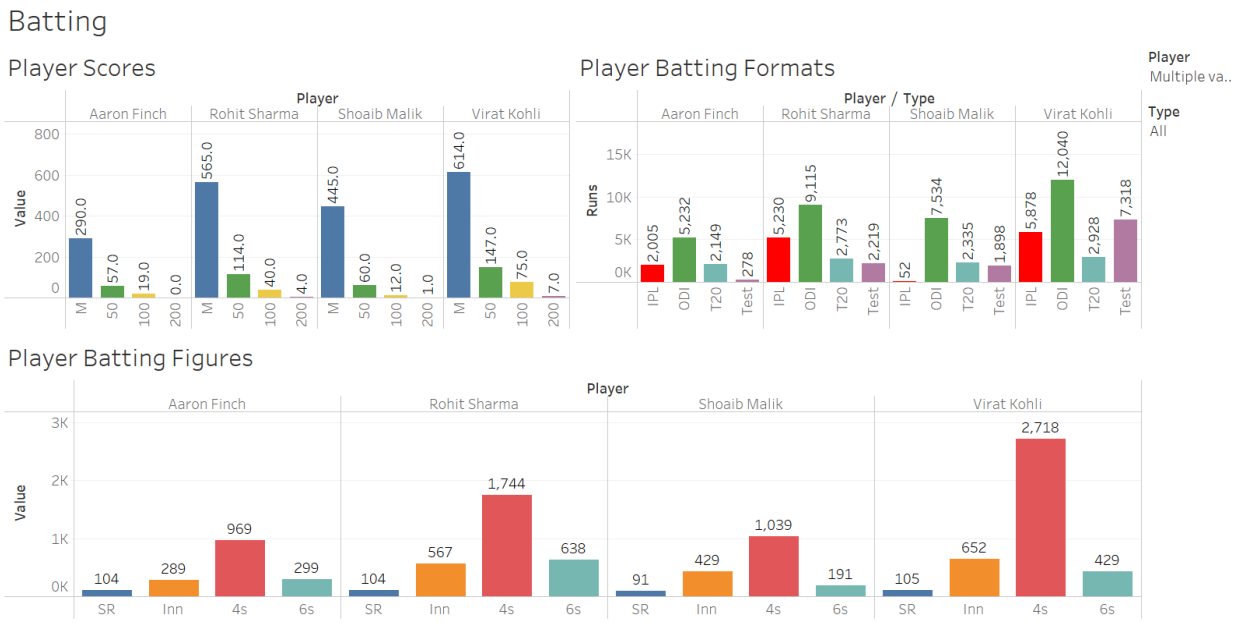
Runs Given vs. Wickets. Color shows details about Type Bowl. The view is filtered on Type Bowl, which keeps IPL, ODI, T20 and Test.

15. Predictions using Bowling Strike Rate and Average.



Strike Rate Bowl vs. Average . Color shows details about Type Bowl. The view is filtered on Type Bowl, which keeps IPL, ODI, T20 and Test.

16. Batting Dashboard by Players and by Type.

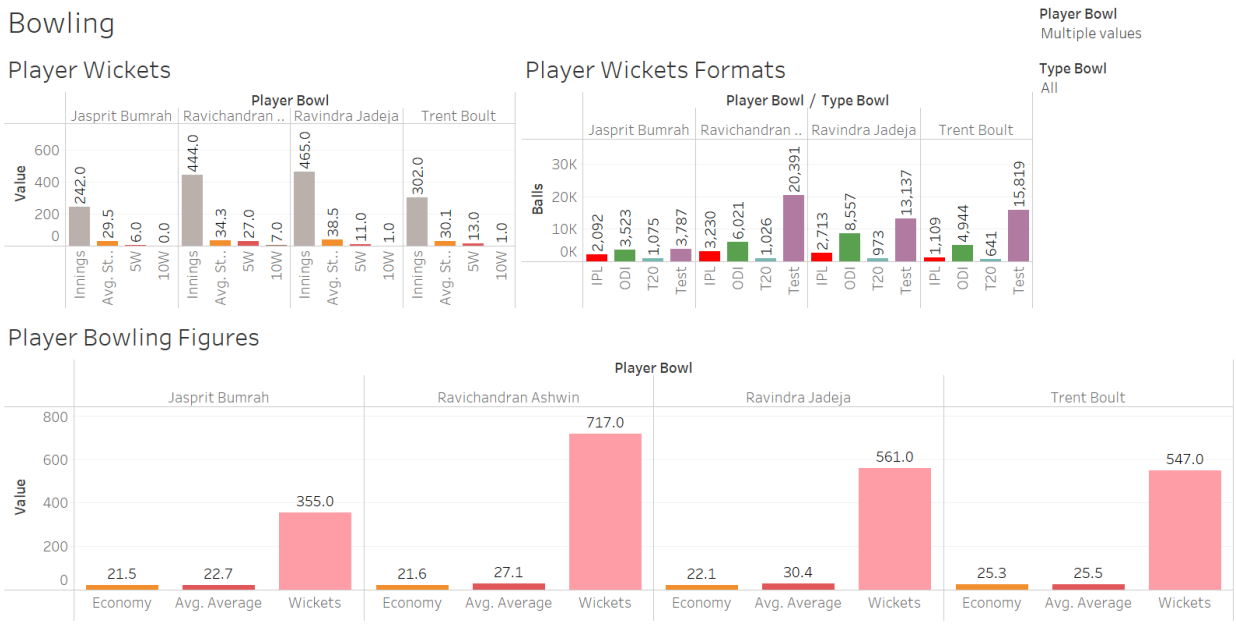


In the batting dashboard it gets very easy to compare different players according to the format. Here, one can get many useful insights like the number of matches played by that player with how much runs, 5s, 100s, 200s scored by that player in how much balls. His Strike Rate, number of runs delt in 4s and 6s etc.

For representation purposes I used only four players in different formats of cricket. And one can easily change the player according to the user.

Over here in the dashboard, one can easily find out the best player on comparing it with all other variables.

17. Bowling Dashboard by Players and by Type.

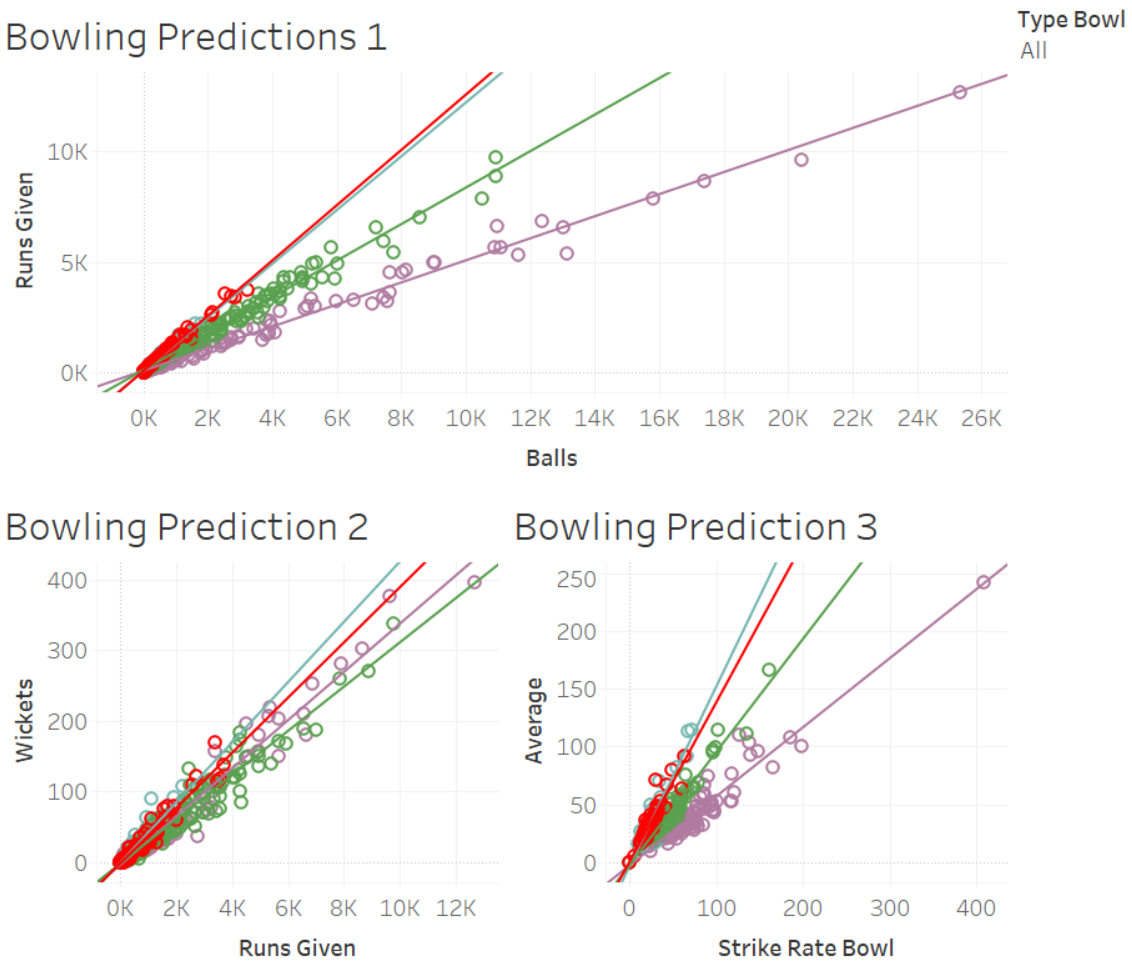


The above dashboard creates an user friendly platform to compare different bowler according to their bowling stats.

As one can get many useful insights from the above dashboard which can help them comparing.

Here, we have the wickets, economy, average, number of balls thrown by the player in different formats, number of innings played, and many more.

18. Bowling Predictions Dashboard by Type.



This is a prediction dashboard with respect to the type of game format. Here one can get some good insight about a particular format's bowling performances of all the dataset.

THE END