Data Analysis and Visual

Tanay Soni

Homework – 3

Virat Kohli
(Indian Cricket Team Captain)

https://www.kaggle.com/shadabhussain/virat-kohli-odi-dataset

This dataset contains the ODI (one day international) cricket statistics of Virat Kohli's batting performance from 2008 to 2017. The dataset has the following variables.

Name	Description
Runs	The number of runs scored by Virat in a particular cricket match.
Mins	Total time taken by Virat to score the runs. Essentially the time for which Virat batted in that match.
BF	Total number of balls faced by Virat to score the runs in that match.
X4s	Number of fours hit by Virat in that match
X6s	Number of sixes hit by Virat in that match
SR	The strike rate of Virat in that match.
Pos	The batting position of Virat in that match.
Dismissal	The mode of dismissal in that match.
Inns	Denotes whether team batted first or batted second
Opposition	The opposition country against which Virat played.
Ground	The ground where the match was played.
Start Date	The date when the match was played.

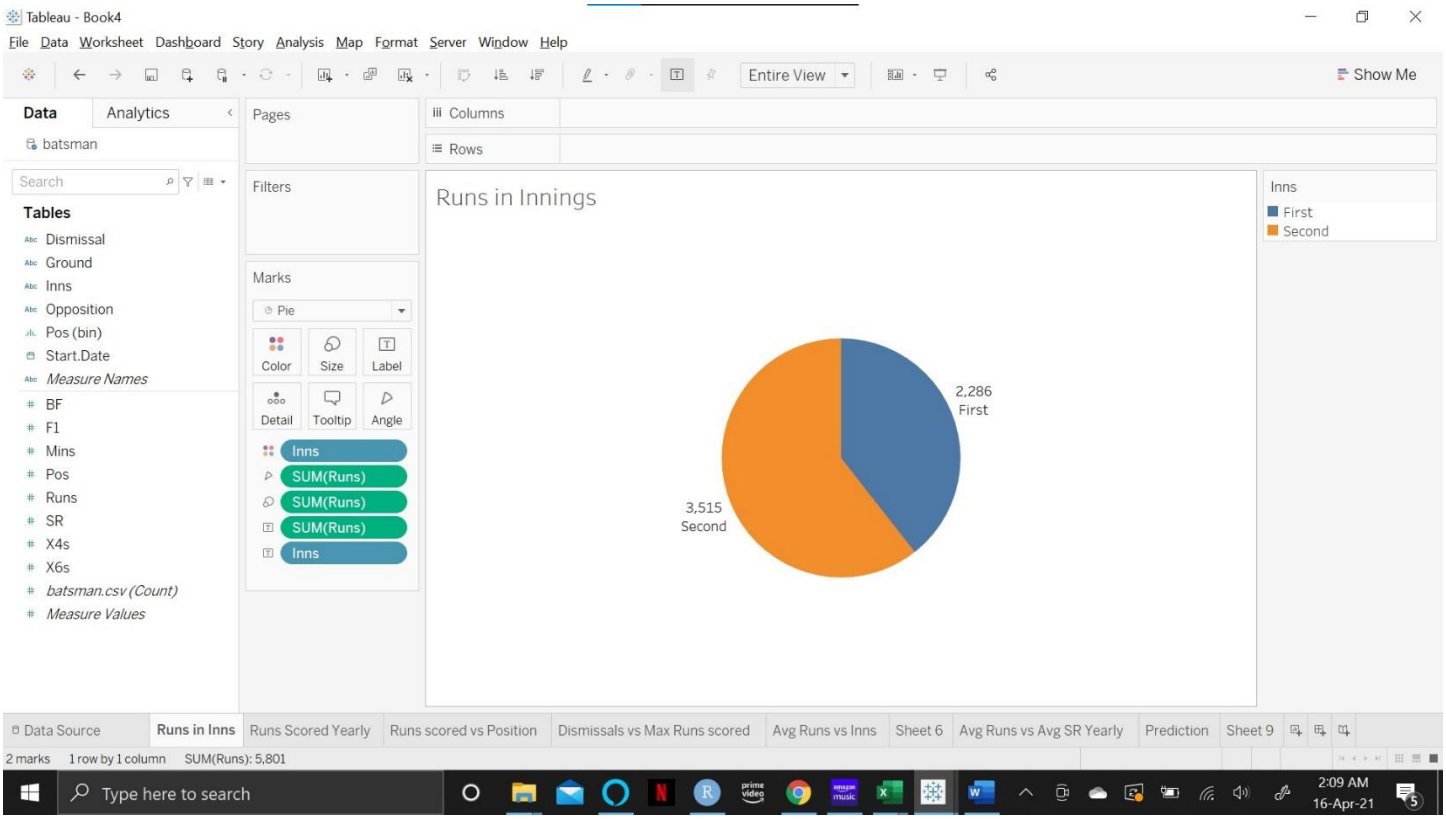
Virat Kohli's performance in the world is top class. In the International Cricket Council, he has been ranked number one batsman in the world and a very successful captain of Indian Cricket history. He has several records in cricket history, he is one of the lead runs scorers in a very short interval of time.

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 I have 12 variables, and, in this dataset, I have performed various exploratory data analysis, like I have cleared my dataset for null values, made changes in the type of columns, like for Mins column it was a character column which I converted it into numerical column because it was one of a important variable for my dataset evaluation. For the use of the dataset, I had to convert some columns into factors to get the most of it.

I'm attaching the graphs which I plotted from the dataset on the next page.

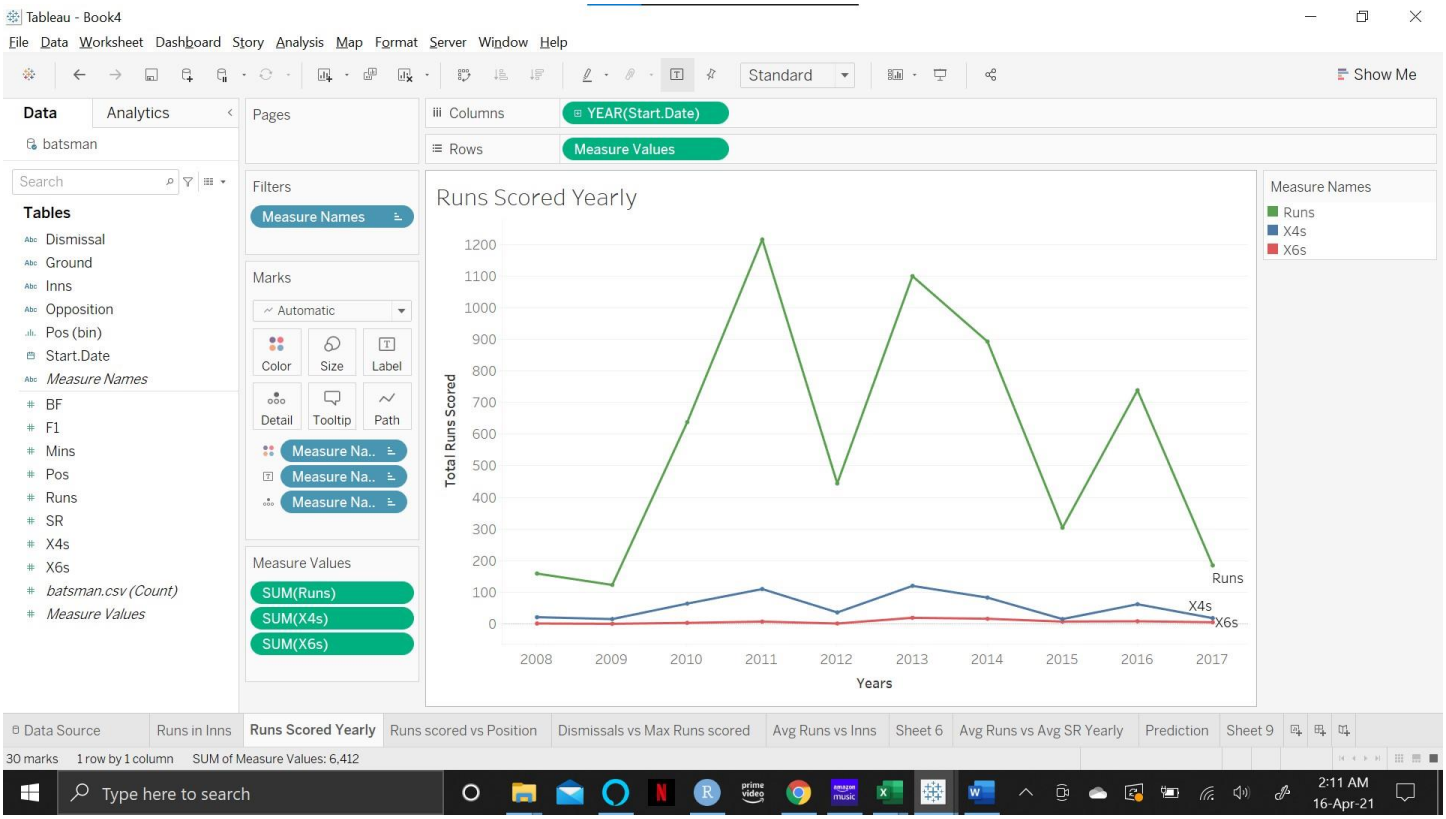
1. Pie Chart Comparission of the runs scored per innings.



In the game of cricket there are two teams who plays against each other, there is toss of coin which decides which team will play(bat/bowl) first, and the captain who wins the toss decides what his team is going to do first, either they bowl first or bat first. The game of cricket is of 50 overs and one over is of 6 balls. So here, in the above graph ***First*** stands for first innings which means Indian team first batted a made a score of certain runs and in the second innings they must defend the total. And ***Second*** means team India was given a certain total which they must make to win the match.

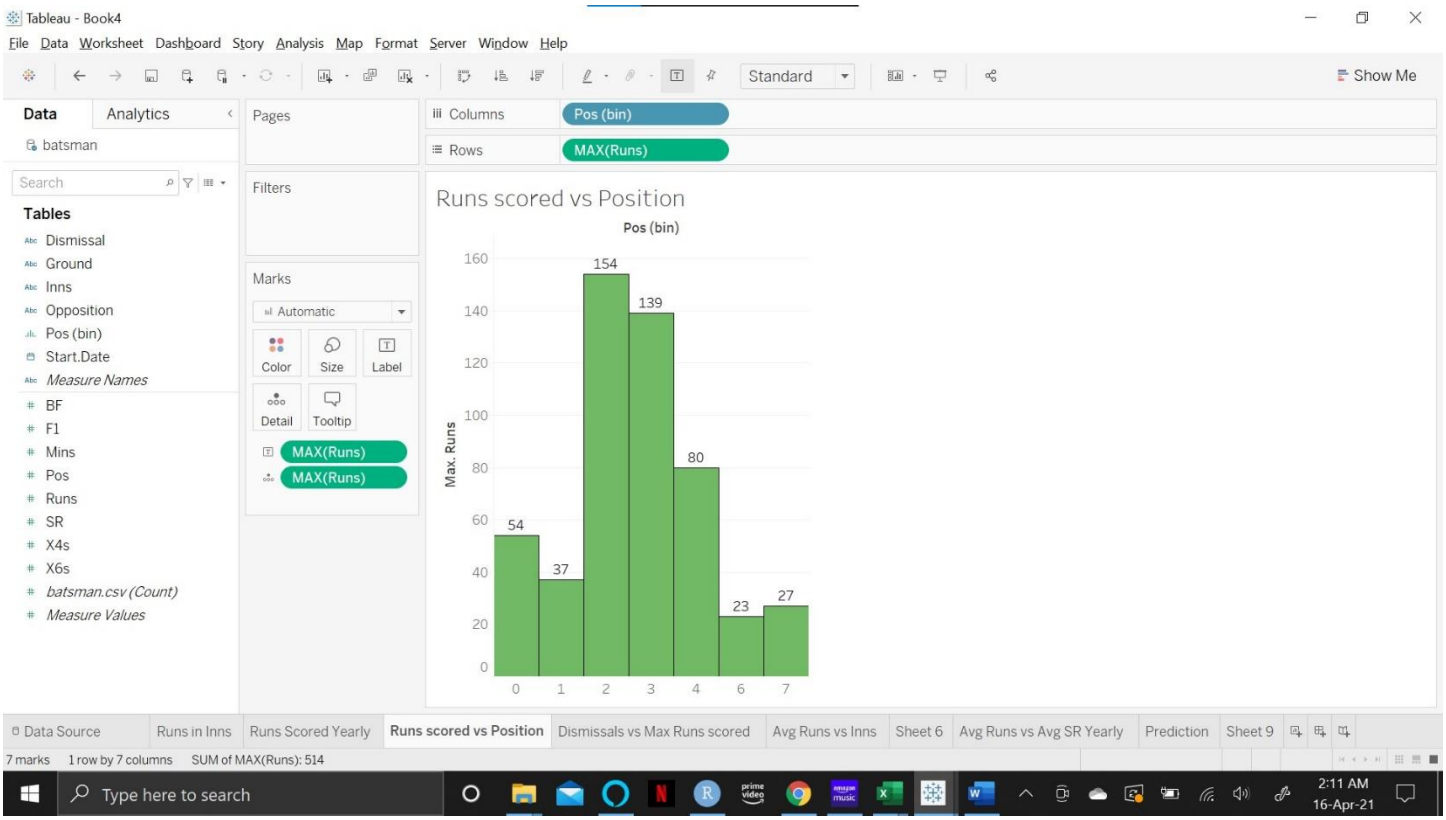
From the above pie-chart, one can easily conclude that Virat loves to chase, as he has scored almost a thousand more runs than the first innings.

2. Runs scored yearly



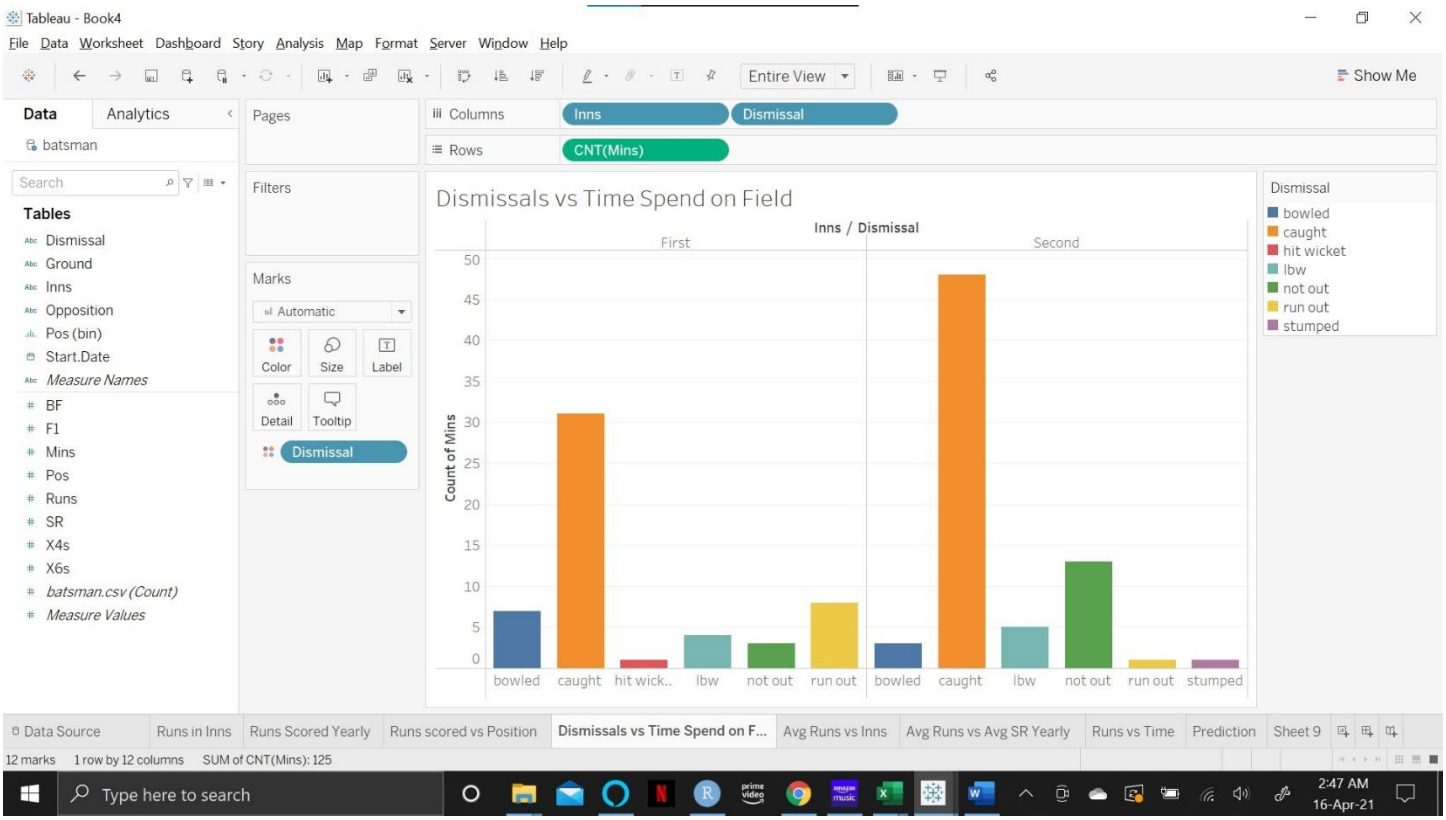
The chart tells us the story, that Virat loves to play in singles, doubles or triples as in cricket one can score runs in five categories, single run (one run from one end of the cricket pitch to the other), double run (two runs from one end of the cricket pitch to the other), triple run (three runs from one end of the cricket pitch to the other), fours(when the bowl touches the ground and crosses the boundary of the pitch) and sixes (when the bowl directly crosses the boundary). In sixes there is a huge risk of getting out if someone's catches the bowl before letting it crossing the line. So, in the graph the red line shows the sum of runs he scored in sixes which are very less when compared to the green line which shows the total of runs he scored using all the five categories. He usually likes to deals runs in fours, and running between the wickets.

3. Maximum runs scored at different positions.



In the game of cricket one can enter the pitch only single time in 11 different positions. Here zero stands for opening batsman, means he faced the first ball of that inning. Here in the chart one can easily conclude that on positions 2 and 3 he plays very well and scored a lot of run when compared to the other positions. He loves to play on position 2 where he scored 154 runs in a single inning.

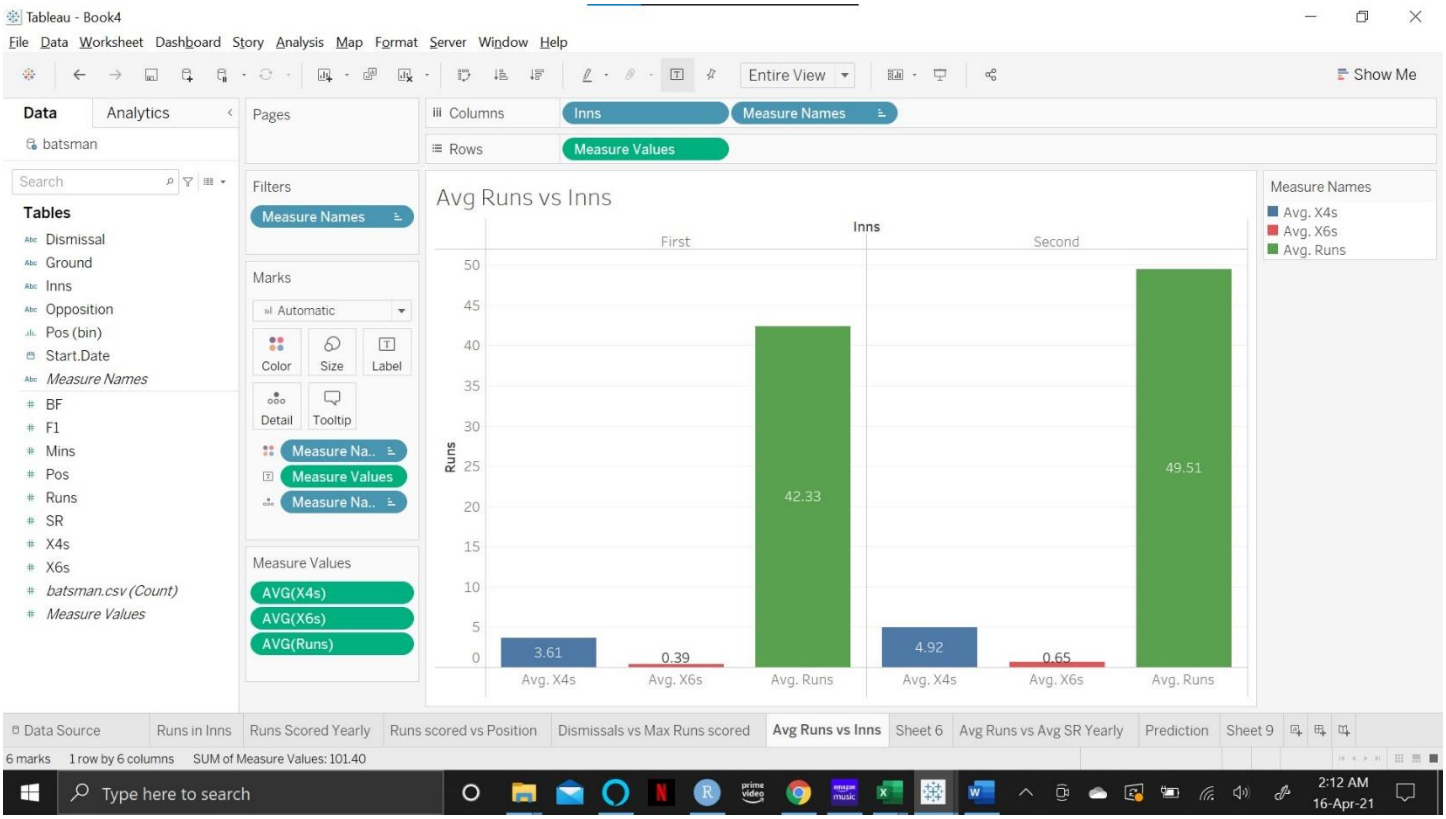
4. Dismissals vs Time spend on Field.



There are several ways of getting out in the game of cricket which means you will not play in that inning again if you get out. For your reference I am attaching the link which will direct you to different ways a batsman can get out in the game of cricket (https://en.wikipedia.org/wiki/Dismissal_(cricket)).

Here in the chart, it is very evident that the most common way to take the wicket of him is by getting him catch out. Hit wicket is a very rare way of getting a batsman out. In the second inning, he usually stays not out and completes the match for his team. In the first inning he was never stumped by any wicket keeper (according to the data).

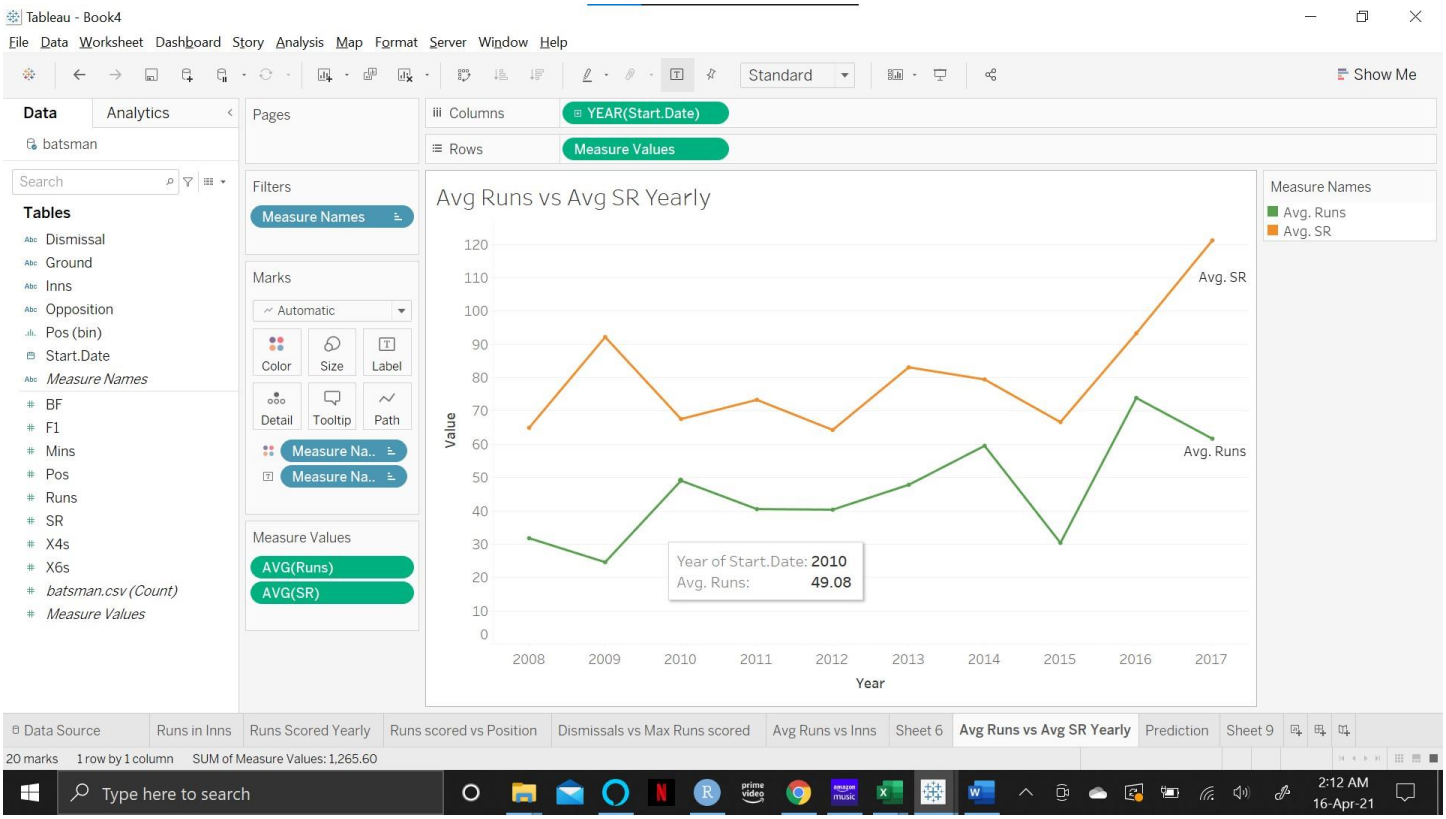
5. Average runs scored vs Innings.



Virat has a record of having the highest average runs scored by any Indian skipper in any innings. As in the chart, it is very evident that he loves to score the runs in singles, doubles or triples when compared to sizes and fours.

He has a staggering average of 42 runs in first innings and almost 50 (half-century) runs in second innings. He usually makes more runs through boundaries in the second innings of the match when compared to the first innings.

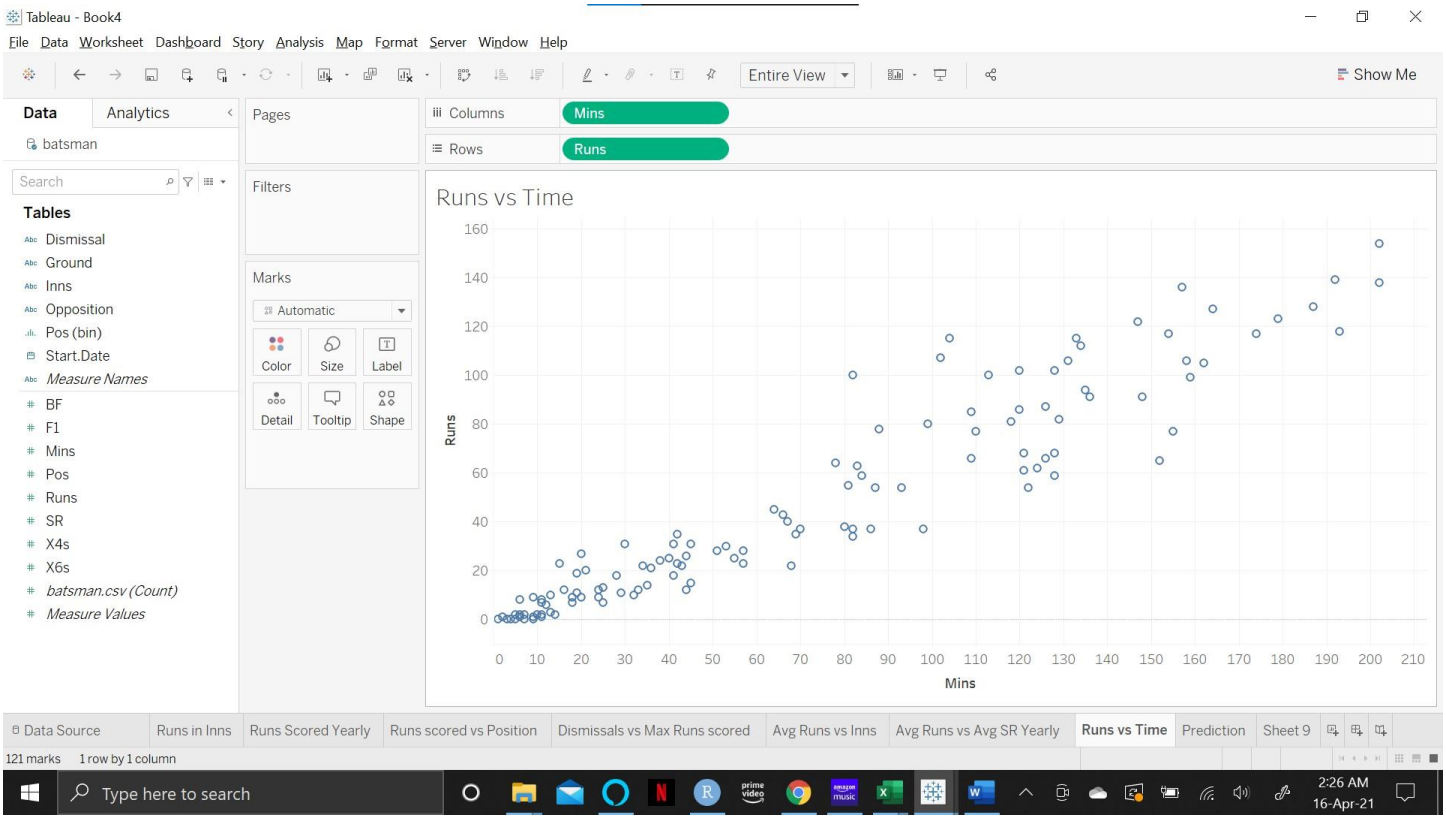
6. Avg Strike Rate (SR) vs Avg Runs Scored Yearly.



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.

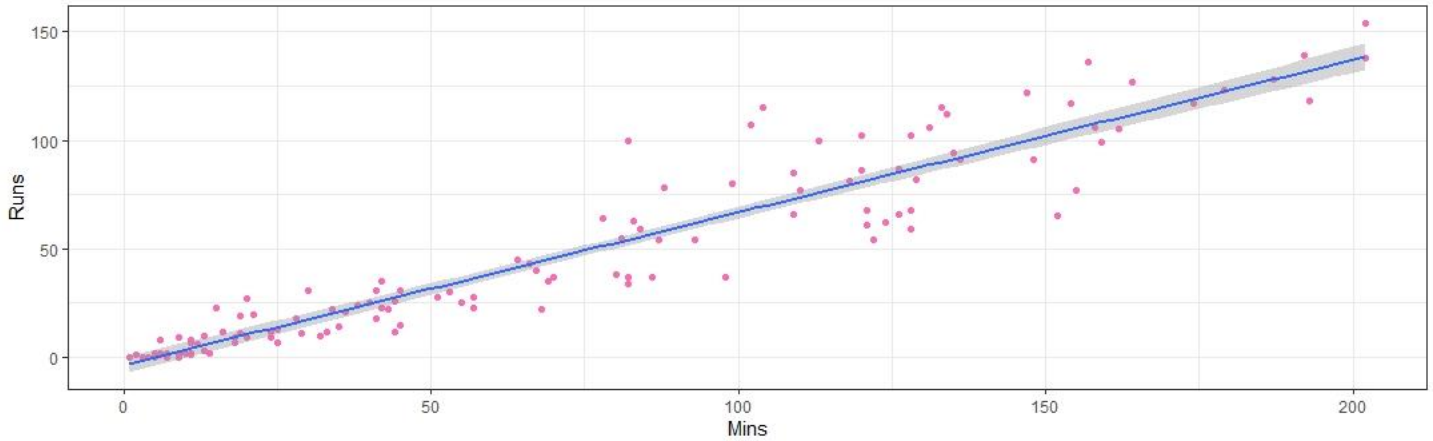
Here in the chart, from the year 2015 to 2017 we can see a sudden rise in average strike rate and score less it can be justified as he might have less time score a certain amount of runs which made him to go for more boundaries and sixes. He has an average of an approx. 50 runs throughout his career.

7. Runs Scored vs Time spend on the field.



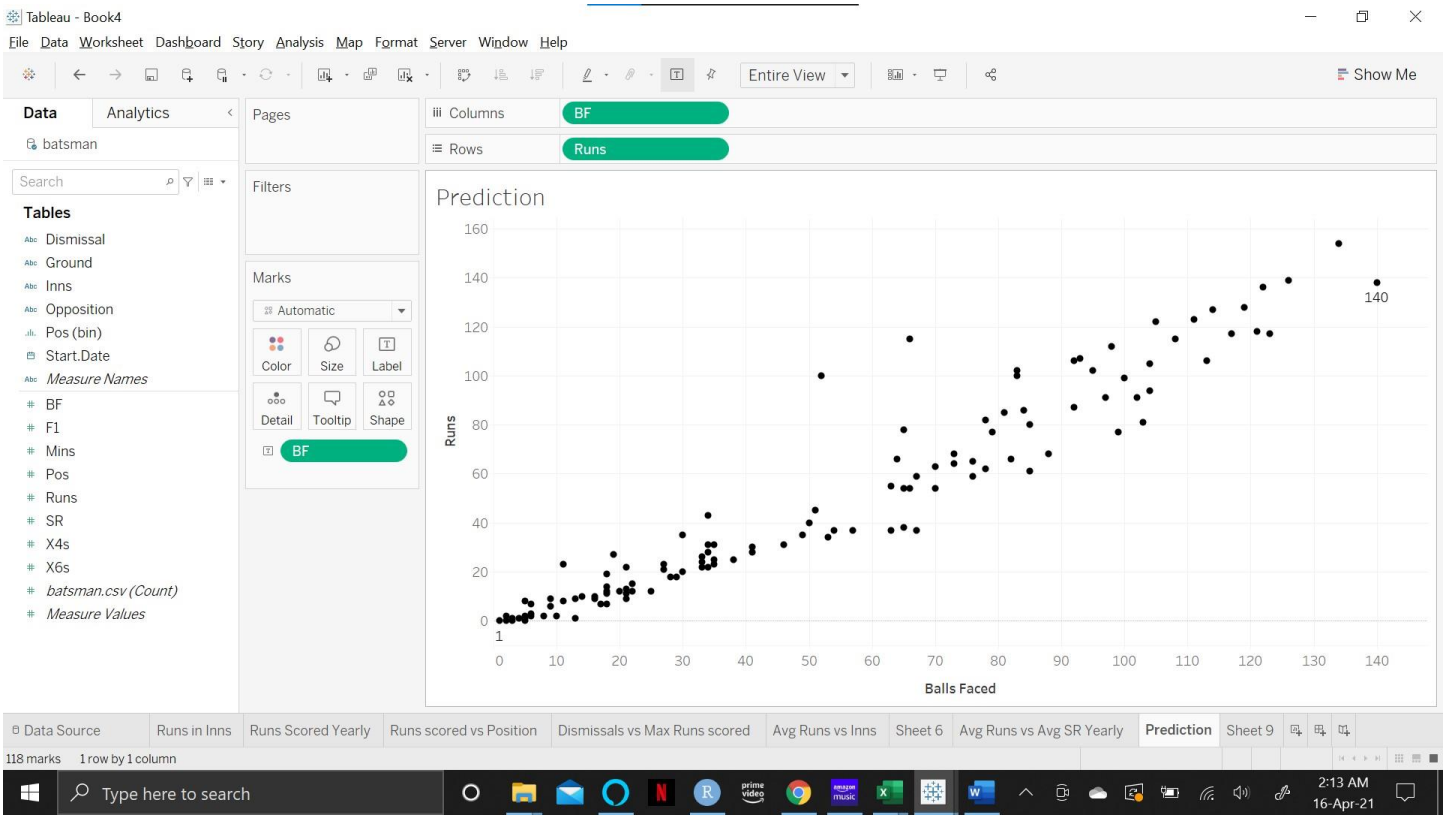
Here in the chart, I can see a positive linear relationship between the time spend by him on the field and the runs scored, which I will be supporting by my exploratory data analysis file in R.

He scored a max of 154 runs in his career after spending a time of almost 202 mins. Means he might have started the innings, or the earlier batsman might have got out early. Runs is my dependent variable and Mins are independent variable.

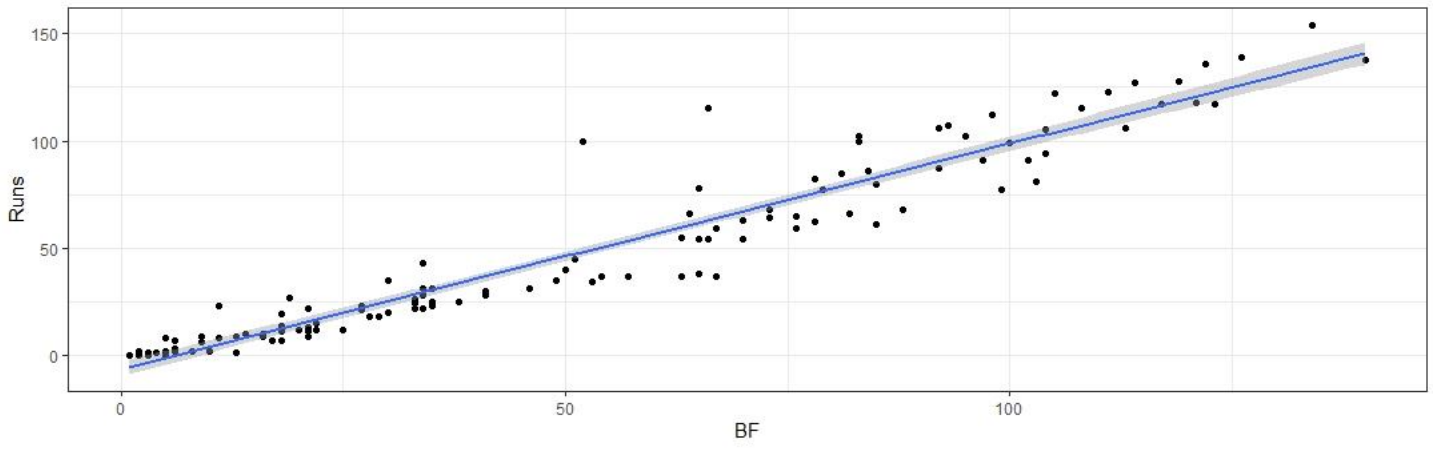


From the above chart, it is clear that the more time he spends on the field the more he score.

8. Balls faced vs Runs scored.



Here again we can clearly see a linear relationship between the variables. Here BF stands for number of balls he faced during the inning. The more balls he faces the more runs he score for his team.



Almost a 45° line, and a positive trend, which means the numbers of balls faced is almost equivalent to numbers of runs he scores.

With this file I am attaching my R source file with the dataset used for the above observations.