**Approach to the problem:**

Created a linear regression model that predicts the score of the inning based on the following factors:-

* Team on the batting side
* Team on the bowling side
* Total runs scored
* Total wickets taken
* Total number of balls played (till that moment of the match)

The first step was to check for any missing values in the data and check the description of all the variables. Irrelevant variables like batsman name, bowler name, fielder, the non-striker, and player dismissed were removed. A team with two names (Rising Pune Supergiant and Rising Pune Supergiants) was converted to 1 team. A new variable called wickets was created using dismissal kind by replacing Nan values with 0 and other values with 1. A cumulative sum was used to calculate wickets to point. The dismissal kind variable was dropped then. The total number of balls played was calculated using overs and balls till point. The total score was calculated to point using a cumulative sum, and the rest of all scores were dropped, including overs and balls. Batting and bowling teams were one hot encoded to convert them into numeric variables.

A new target variable score was created by calculating the total score of the full innings and was assigned to each row. Next, a linear regression model was used to fit the data. The model predicted the inning score given the total number of balls played to the moment, total wickets, total runs scored, and teams on the batting and bowling sides. These were calculated by preprocessing the data in the given format. Data below five overs of the match was dropped since it does not make sense to predict the inning score using the initial five overs. The scores were predicted after each ball of the match (from sixth over). The average of the scores for each ball of a match id was calculated to predict the innings score and rounded off to the next integer.