Problem Statement

As the IPL wraps up this year, we remember the great times we had with our friends and families watching the matches. One of the most fun activities we do during a match is to predict what the score might be at the end of the innings. Generally, we utilize manual intuition to come up with a solution. In this problem, we provide you with detailed data and give you a chance to come up with something much better.

Your task is to construct a function which can predict the final score of the **first** innings of an IPL match given detailed information of what has happened in the innings till a randomly chosen point between 8th and 12th over.

Data Format

The data format for both training and consumption by the final function is a csv with the following columns.

match_id	Match ID for help distinguish matches
batting_team	Batting Team Name
bowling_team	Bowling Team Name
over	Current Over (starting from 1)
ball	Current Ball
batsman	Batsman Name
non_striker	Non Striker Name
bowler	Bowler Name
wide_runs	runs due to wide ball
bye_runs	runs due to byes
legbye_runs	runs due to leg byes
noball_runs	runs due to no ball
penalty_runs	Other types of penalty runs
batsman_runs	Runs by batsman
extra_runs	Runs due to extras
total_runs	Total runs on ball
player_dismissed	If a batsman is dismissed, their name
dismissal_kind	Type of dismissal
fielder	If a fielder is involved in dismissal, their name

For training of your model, we shall provide you data for completed **first** innings for 300 IPL matches. We shall also provide you partial data for 100 IPL matches to test your predictions. The data format to return prediction is a csv with 2 columns, match_id and prediction.

Submission Format

You shall submit the following files

- A word document describing your approach to the problem in a single page at the maximum.
- A python file containing a function which consumes file paths for the data, prediction & model and saves your predictions to prediction file. A sample function is provided below.
- Optionally a pickled file containing all information you need from your trained model.
- A prediction file generated for the 100 matches we provided for test

Put the files in a zip and upload the zip file. Do note that you need not put the files in a folder when zipping.

Evaluation Criteria

We shall measure submissions on the criteria:

$\Sigma_{\{all\ matches\}}(PredictedScore - ActualScore)^2$

Primarily we shall use the prediction file provided for the 100 matches for test to form a cutoff. We shall use extra matches from our data not provided to you for final scoring.

Sample function

This simple sample function provides you with the format of function we expect in your python file.

```
import pandas as pd
def predictInnings(inningFile,predictionFile,modelFile=''):
    innings = pd.read_csv(inningFile)
    if (modelFile!=''):
        model = pd.read_pickle(modelFile)
    innings["balls_done"] = 6 * innings["over"] + innings["ball"] - 6
    predicted_runs = innings.groupby("match_id").apply(lambda
x:np.round((120.0 *
x["total_runs"].sum())/x["balls_done"].max())).reset_index(name="predi
ction")
    predicted_runs.to_csv(predictionFile)
```