PES University, Bangalore
UE15CS322 - Data Analytics

Session: Aug – Dec 2017
Week 1 – Assignment 1

Date of Submission: August 18, 2017 **Max Marks:** 20

NOTE: *In your assignment books, write the question, R commands required to get the desired output and the obtained output. For questions that require you to plot graphs, you will have to also print the obtained output (graphs/charts) and attach it in your assignment book. The solutions to the assignment must be* ***hand-written****. You will also need to upload your code (.R file) on your cloud storage and share the link of the same on the Google forms that will be sent out to you (****must be neatly commented****).*

TOPIC: Computation of summary statistics using R

Indian Premier League (IPL) *is the perhaps the biggest professional Twenty20 cricket league across the globe. It was launched in 2008 and is held annually in India. Its 10th season recently concluded on May 21, 2017 marking a decade of success for the flashy T20 competition beloved by India. It follows the double round robin format followed by the playoffs. The shorter-format league has about 60 matches every season (56 round robin, 4 playoffs) and has evolved into India's most popular sporting event, filling stadiums and attracting TV audiences well beyond that enjoyed by Test and one-day competitions in the cricket-mad subcontinent. Ever since its inception in 2008, it has been known for its magnificence, pyrotechnics, huge signing bonuses, glamour, richness and controversies.*

Excerpts from 'Indian Premier League – the tournament has seen it all', Zee News

The dataset **deliveries.csv** is the ball-by-ball data of all IPL cricket matches of the first nine seasons. The data was obtained from Cricsheet in YAML format and was made available in the CSV format on Kaggle. Perform the following explanatory data analysis tasks:

Question 1 (4 points)

Chinnaswamy Stadium, Bangalore witnessed the finals of season 9 between Royal Challengers Bangalore and Sunrisers Hyderabad on May 29, 2016 (match_id 577). Sunrisers Hyderabad won the toss and elected to bat. What was the target set by the team? What was the run rate (number of runs per over) required by Royal Challengers to ensure that they at least drew the match with their opponent?

Plot a histogram indicating the number of runs scored by Royal Challengers Bangalore in each over and state what type of distribution (modality) it follows. (**Reference**)

Question 2 (4 points)

For each match, compute the total number of runs scored by each team in their respective innings. Print the result in the following format:

match_id	inning	batting_team	total_runs
1	1	Kolkata Knight Riders	222

..	..	..	..

Report your result for matches with match ID 7, 27, 67, 171 and 414.

Also, report the team that scored the least number of runs and the team that scored the maximum number of runs in any match throughout the tournament along with their scores. (You may have to use **results.csv** as well for this question – what is expected is the max/min score for a normal/tied match with no Duckworth-Lewis method (D/L) applied – do not consider matches with no results.)

Question 3 (4 points)

Compute the total runs scored, mean, median, mode, 1st quartile, 3rd quartile, IQR and standard deviation of runs scored by AB de Villiers, MS Dhoni, RA Jadeja, SK Raina and V Kohli throughout the tournament. Season 11 of IPL will see a complete shuffle of all teams. Hypothetically, if you were the owner of a team in IPL 11, which three players of the five would you choose to bid-on for your team? Give relevant reasons. (You don't have to consider intricate cricket rules such as 'leg bye runs are not added to the score of the player', etc.)

Question 4 (4 points)

Based on all the performances given by V Kohli and MS Dhoni, draw the boxplot of the total number of runs scored by them in each match. Comment about their outstanding and underwhelming performances, if any. Also, calculate the 90th percentile of their scores.

If you were to choose a captain for your team based on consistency, whom would you choose among the two and why?

Question 5 (2 point)

'Suresh Raina is a key player of Gujarat Lions and his performance in every match is directly reflected in the total score of the team' – Comment.

Question 6 (2 point)

We studied that for a normal density function, if the variance is high, the spread is more and if the variance is low, the bell curve has a sharper peak. We also studied that if the kurtosis is low (platykurtic), it indicates heavier tails and if the kurtosis is high (leptokurtic), it indicates an increasing concentration of data from the tails towards the center/ peak. Does this mean if kurtosis is high, the variance is low? Briefly explain your answer.