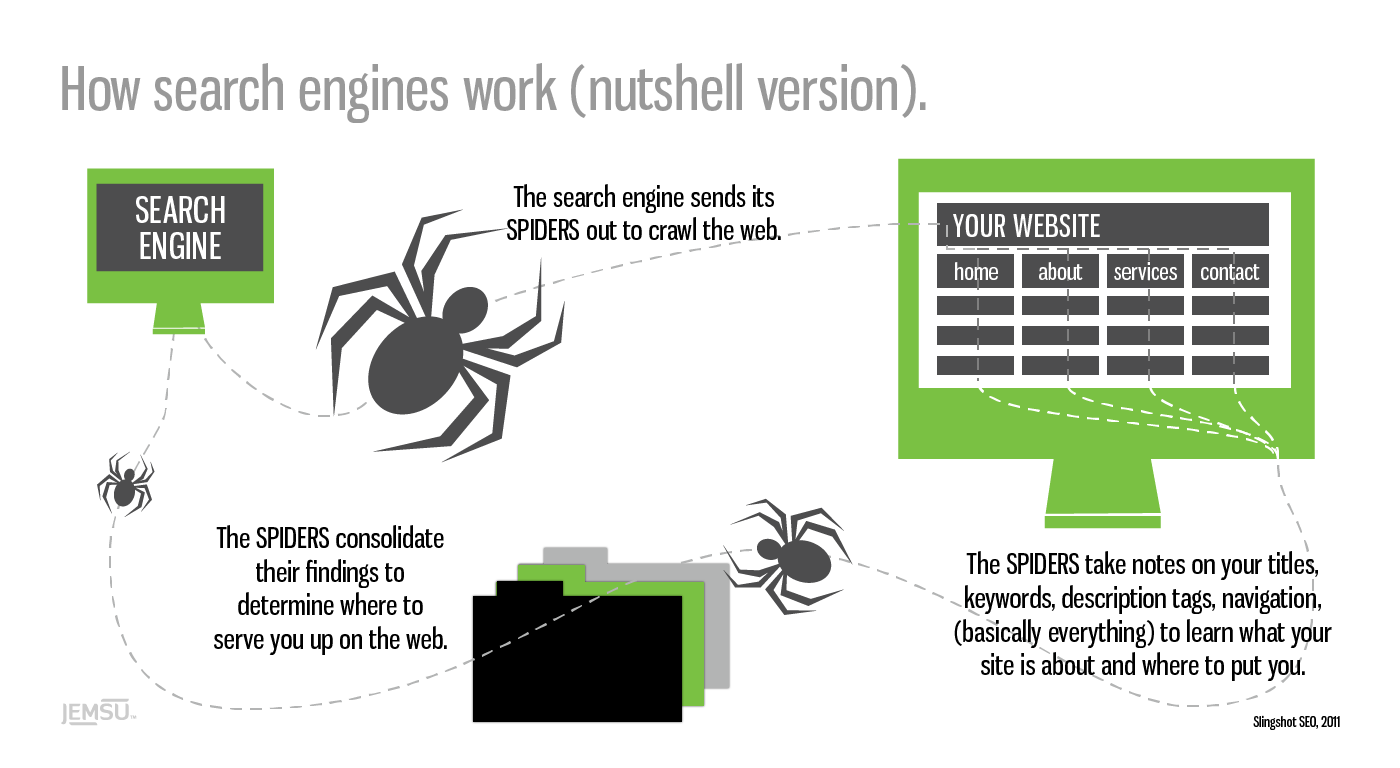
INFO 6205

Ranking System Algorithm Project Report



1. *RANKING SYSTEM:*

* Ranking system algorithm is the process of comparing given data set based on various factors and deciding the most expected outcome for certain event under given circumstances. Conclusion is made depending upon finding probabilities considering certain factors with past data and predicting future events. For an example, with the amount of information available on the web, finding what you need would be nearly impossible without some help sorting through it. Google ranking systems are designed to do just that: sort through hundreds of billions of webpages in our Search index to find the most relevant, useful results in a fraction of a second, and present them in a way that helps you find what you’re looking for.



1. *PROJECT AIM:*

* This project is about analyzing dataset of Cricket game which comprises of matches played between different teams, finding probabilities considering various factors and coming up with the likelihood of wining the match by any team against opponent based on past series data.
* Factors taken into account while calculating average probability are as follows:

1. Matches won/lost
2. Runs Per Over (RPO)

* In [Cricket](https://en.wikipedia.org/wiki/Cricket), Runs Per Over (RPO), is the [average](https://en.wikipedia.org/wiki/Arithmetic_mean) number of [runs](https://en.wikipedia.org/wiki/Run_(cricket)) a batting side scores per [over](https://en.wikipedia.org/wiki/Over_(cricket)). It includes all runs made by the batting side in the [innings](https://en.wikipedia.org/wiki/Innings) to that point of the game, both the runs scored by the batsmen and [extras](https://en.wikipedia.org/wiki/Extra_(cricket)) conceded by the bowling team.

1. Average

* In [cricket](https://en.wikipedia.org/wiki/Cricket), a player's bowling average is the number of [runs](https://en.wikipedia.org/wiki/Run_(cricket)) they have conceded per [wicket](https://en.wikipedia.org/wiki/Wicket#Dismissal_of_a_batsman) taken. The lower the bowling average is, the better the [bowler](https://en.wikipedia.org/wiki/Bowling_(cricket)) is performing. It is one of a number of statistics used to compare bowlers and  to judge the overall performance of a bowler.
* Given data makes the combination of discrete and continuous values, making us to explore various available probability distribution functions and methods, and choosing the best one to arrive at conclusion of deciding the wining team if match is played in future.
* For discrete values like matches won/lost, we have considered Binomial Distribution (Probability Mass Function), while T-distribution (Probability Density Function) comes to rescue when dealing with continuous values.

1. *DATA STRUCTURED USED:*

* Data structure used here is HashMap and that’s due to following reasons:
  + According to the dataset and required data to be stored for analysis and calculations, HashMap provides the best implementation where we can store key as combination of winning team and opponent team and values in the form of match results considering factors like win/loss, average and RPO.

Declaration: HashMap<String, Object>

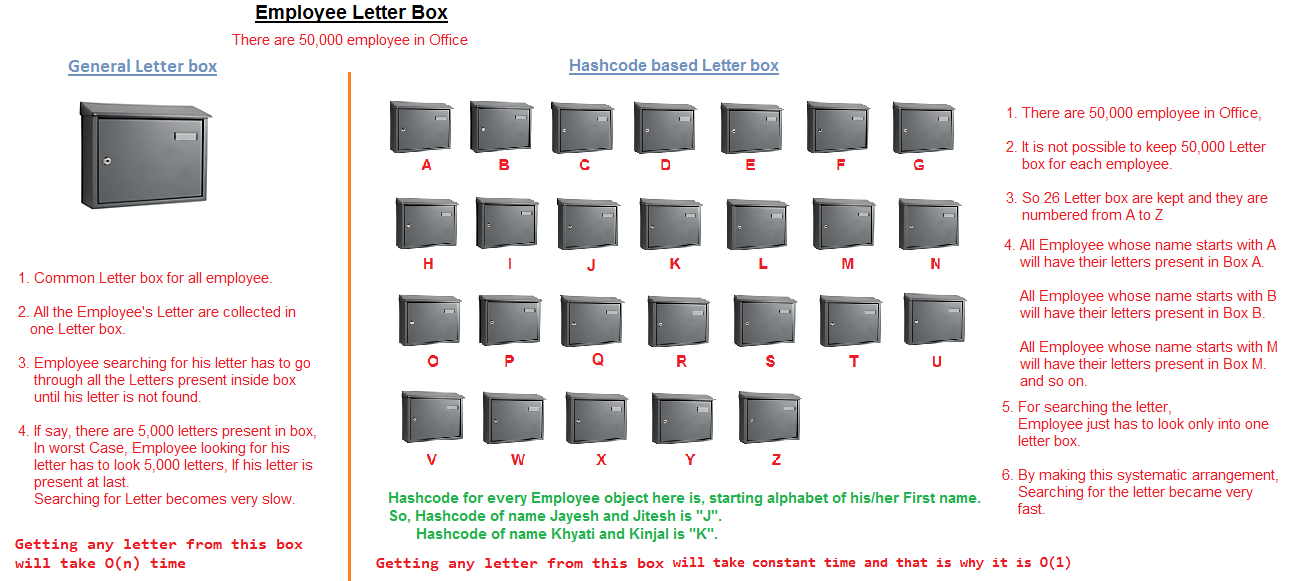
Where, Key = String = Concatenated String of winning team and opponent’s name

Value = Object = Statistical data of series between teams

Ex – If series has been played between India and Australia then Key will be India-Australia

and value will be statistical data of series between them

* + HashMap provides time complexity of O(1) for both put and get operations, making it efficient data structure when such operations happen frequently. Ultimately, it provides best efficient solution in terms of time.



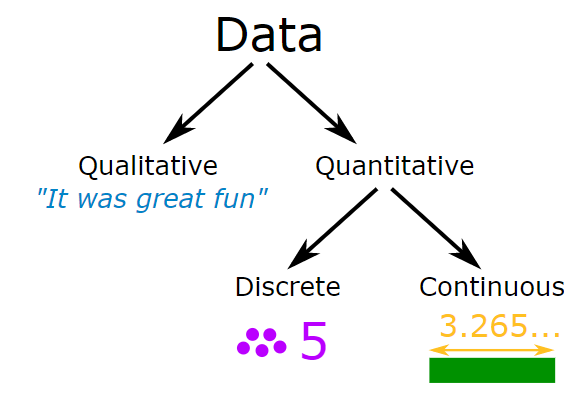
1. *DISCRETE VS CONTINUOUS DATASET*

* **Discrete Data:** Data that can only take certain values.

For example: the number of students in a class (you can't have half a student).

* **Continuous Data:** Data that can take any value (within a range).

Example - People's heights could be any value (within the range of human heights), not just certain fixed heights.



1. *T-DISTRIBUTION VS NORMAL DISTRIBUTION FOR CONTINUOUS VALUES*

* In concept of probability distribution, the sample mean, and sample standard deviation are estimates of the population mean and population standard deviation. When our observations are fewer then sample mean and sample standard deviation become less accurate estimates, which may cause in inaccurate probability and predicting wrong result. To avoid such issue and make our prediction efficient, T-Distribution plays vital role while dealing with small amount of data.

1. *BINOMIAL DISTRIBUTION:*

* A binomial distribution is a [statistical experiment](https://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:
* The experiment consists of n repeated trials
* Each trial results in only two possible outcomes – a Success (Win) and a Failure (Loss)
* The probability of ‘success’ p is the same for each outcome
* If these conditions are met, then X has a binomial distribution with parameters n and p, abbreviated B(n,p).

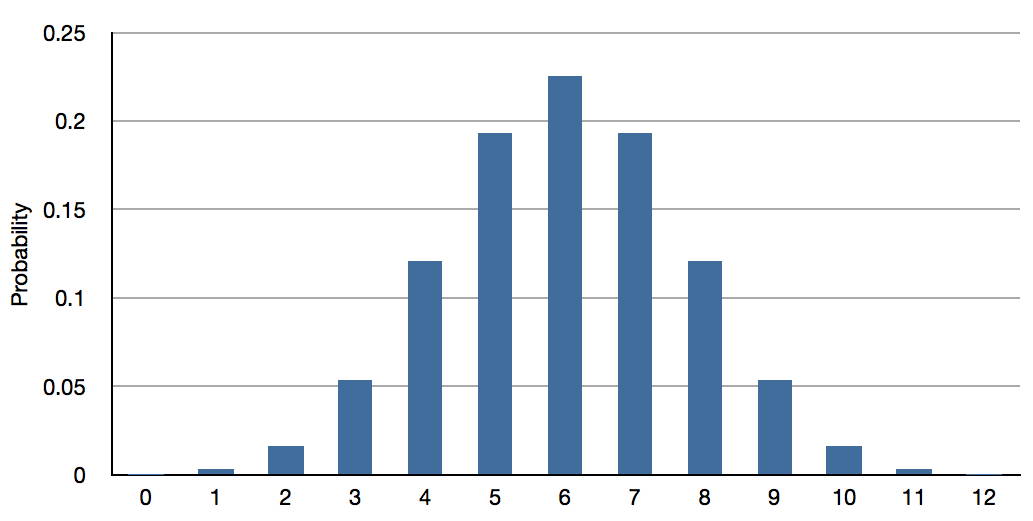


Where, n = number of total observations

k = number of successes

p = success probability for each trial

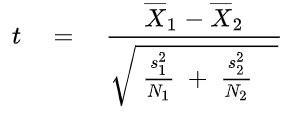
q = 1 – p



1. *T-DISTRIBUTION*

* The T-Distribution is used instead of the normal distribution when you have small samples.
* T-Distribution compensates inaccuracy that arises due to small number of observations.
* In case of normal distribution, the mean and standard deviation are based on only few observations, and so are not very accurate measures of the true mean and standard deviation.
* ***Use of Welch’s t-test:***

Welch's t-test performs better than Student's t-test whenever sample sizes and variances are unequal between groups and gives the same result when sample sizes and variances are equal. As per our dataset, Welch’s t-test yield more accurate results.



Where,

X¯1 = the sample mean

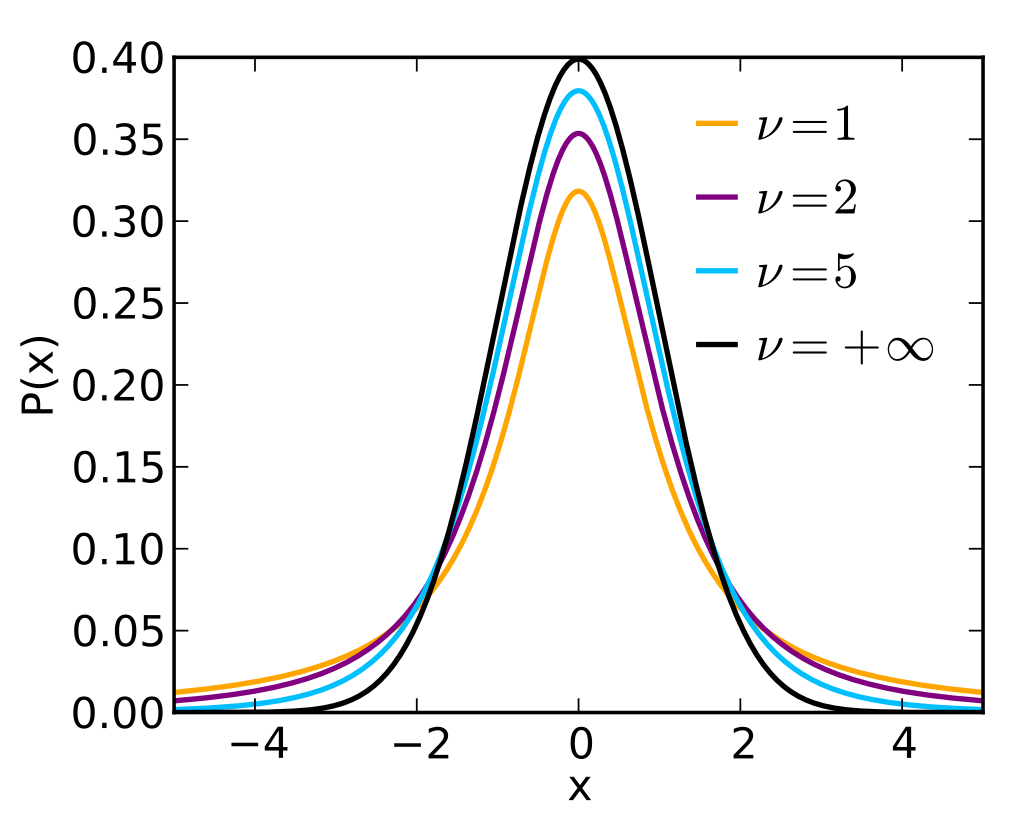
X¯2 = the population mean

S1 = standard deviation of the sample

S2 = standard deviation of the population

N1 = sample size

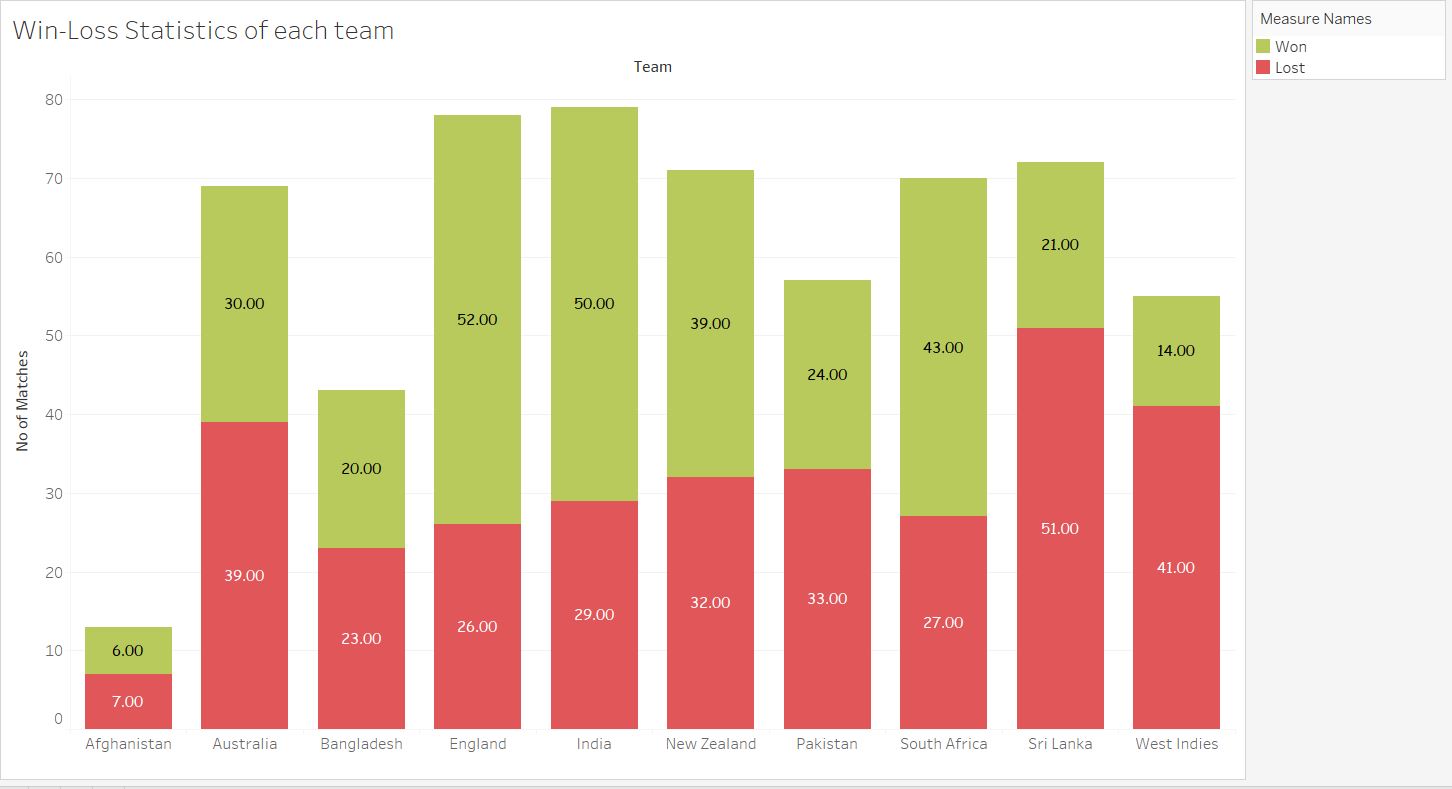
N2 = population size



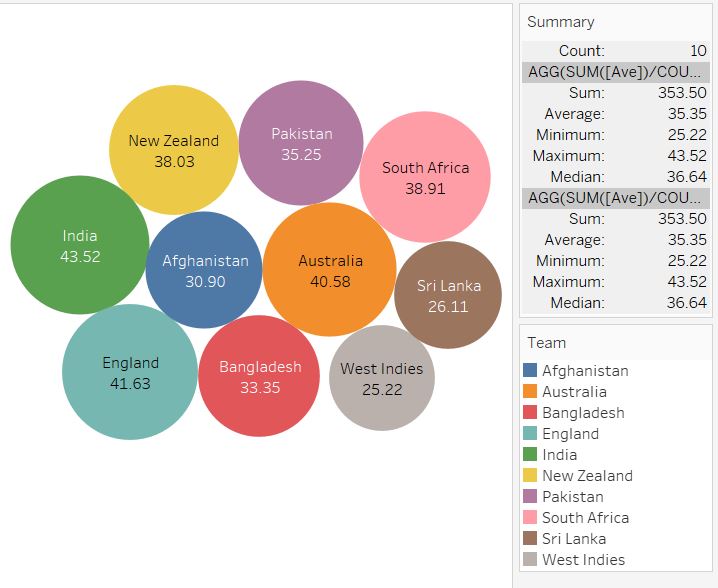
1. *DATASET INTRODUCTION*

* Dataset indicates the number of teams and statistics for teams’ matches with remaining teams
* Dataset provides below information
* Dataset has total 81 records for statistics of matched played between teams over last 3 years
* Data consider following parameters with its nature of data
* Each team played series of how many matches with remaining teams – Discrete data
* Matches won out of total matches - Discrete data
* Average during the series – Continuous data
* RPO during the series - Continuous data

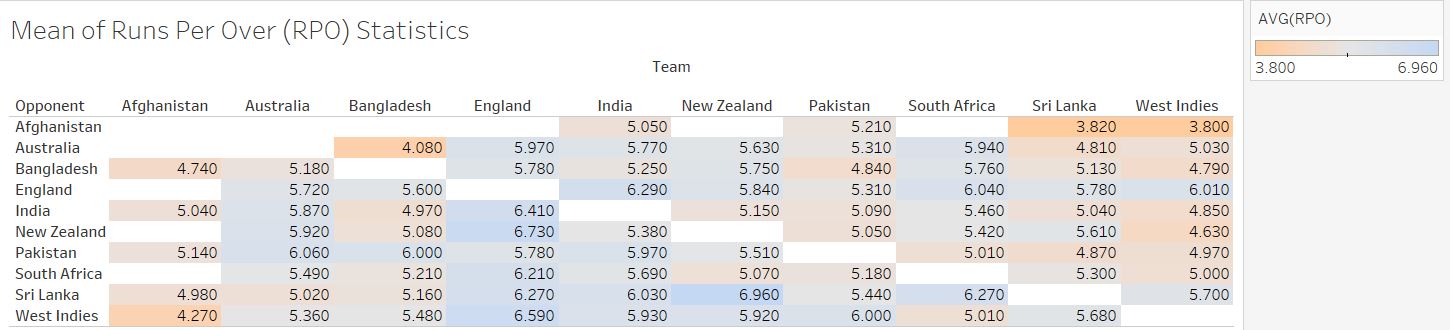
1. Win/Loss statistics for each team



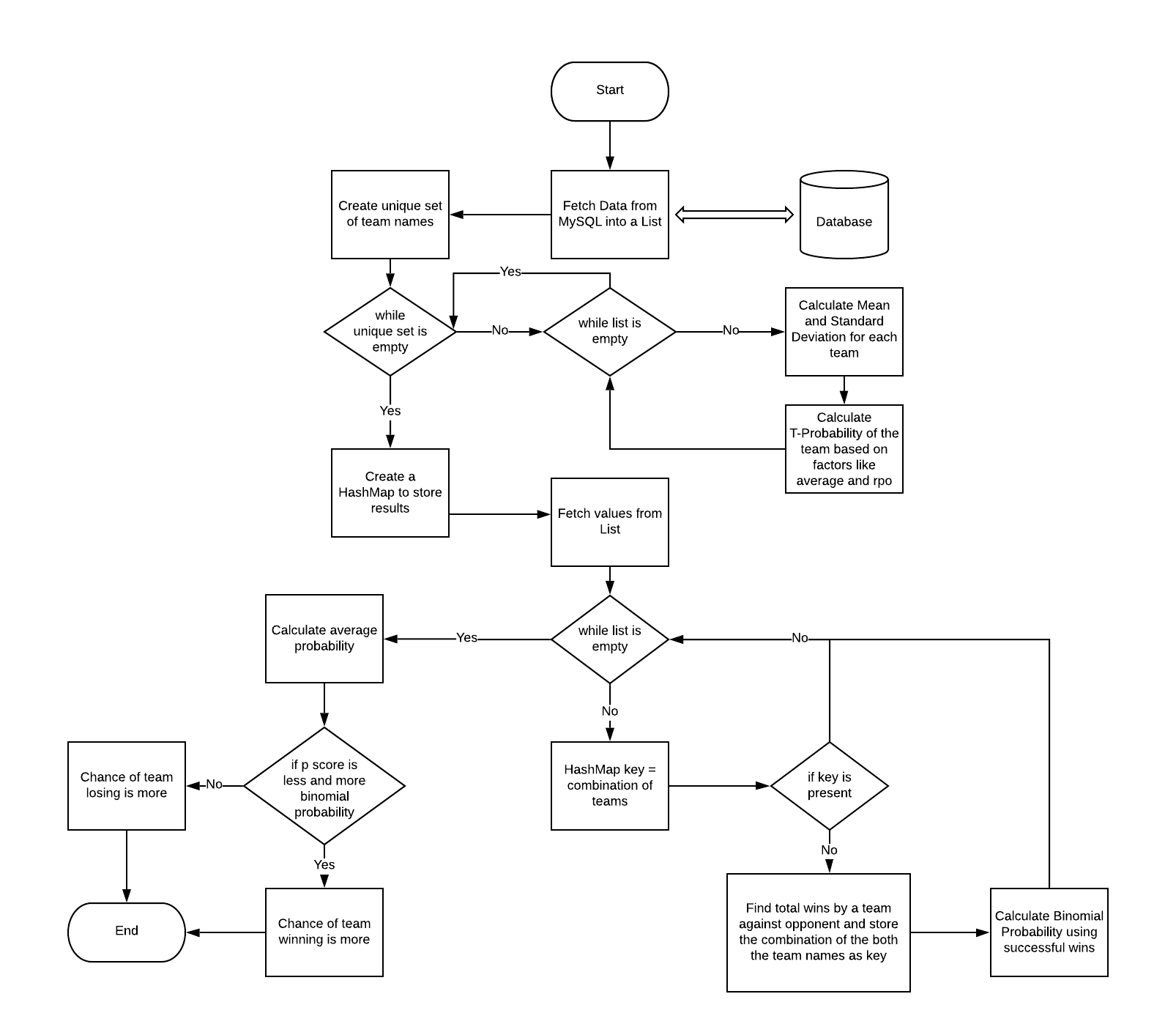
1. Mean of the *Average* of each team



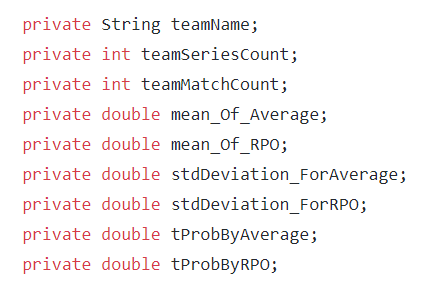
1. Mean of *RPO* of each team



1. *PROGRAM AND ALGORITHM FLOW:*



* Program execution starts by fetching data from MySQL database and populating List containing object of winning team, opponent, matches played in series, matches won/lost, Average and RPO.
* Now that we have data ready List, we are iterating List objects to find statistics.
* While iterating List objects, combination of winning team and opponent name will be stored in HashMap as key.
* Value of the HashMap contains the Java object of variables as below.
* Team name
* Team Series Count
* Team Match Count
* Mean of Average factor
* Mean of RPO factor
* Standard Deviation(SD) for Average
* Standard Deviation(SD) for RPO
* T-probability by Average
* T-Probability by RPO



* Now calculate the required parameters for finding p-score such as mean of *Average*, mean of *RPO,* SD for *Average,* SD for *RPO.*
* Find p-score by T-Distribution based on above calculated parameters and save in the object as value in the HashMap.
* Repeat the above step to calculate parameters and p-score for each combination of teams and prepare complete HashMap with all statistical data loaded.
* Calculate the probability from p-score for each team for *Average* and *RPO* factor

***The lower the value of the p-score, the higher the probability of success.***

* Calculate the Probability for each team by Binomial Probability function based on win/loss data
* Now that, we have probability for all factors calculated above, find overall mean probability for each team which is used to compare the performance and win prediction of any team when matches will be played in future.
* This overall probability is also used to rank the team based on factors considered.

1. *OVERALL PROBABILITY CALCULATION*

Formula:

Overall wining probability of any team P =

(probability calculated for *matches won* factor using binomial formula P1) + (probability calculated for *Average* factor using p-score from Welch’s t-test formula P2) + (probability calculated for *RPO* factor using p-score from Welch’s t-test formula P3)

* 1. *Calculate probability for matches won by any team (X) against other team (Y) using binomial probability formula:*

P1 = 

Example: For winning team Australia against Sri Lanka, probability can be calculated by putting data in above example where,

n = 5 (total number of matches)

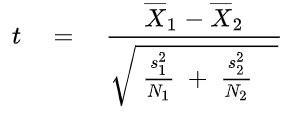
k = 4 (matches won)

p = 0.666 (probability of success)

q = 1 – 0.666 = 0.334 (probability of failure)

P1 = 0.32855904

* 1. *Calculate probability of Average factor for any team (X) against other team (Y) using T-Distribution Welch's t-test**formula:*



**X¯1**= the sample mean = Mean value of *Average* factor for team X, where X played matches against other teams

**X¯2** = the population mean = Mean value of *Average* factor for all records in dataset

**S1** = standard deviation of the sample = S.D calculated taking all *Average* values for team X against remaining teams

**S2** = standard deviation of the population = S.D calculated taking all *Average* values for all teams

N1 = sample size = total Series played by team X = number of teams as opposition

N2 = population size = total number of series including aa teams in dataset

***NOTE:*** *Here, output of above formula indicates the p-score, which is not actually the probability. As per the statistical science, the lower the value of the p-score, the higher the probability of success. Hence, if we compare p-score for two teams between whom match will be played, the probability of wining the match by team is more whose p-score is lower.*

Example – For winning team Australia,

X¯1 = mean of all Average values of Australia against other teams = 40.57

X¯2 = mean of all Average values of all teams = 35.3753

S1 = SD of Average of Australia against remaining team= 16.71099

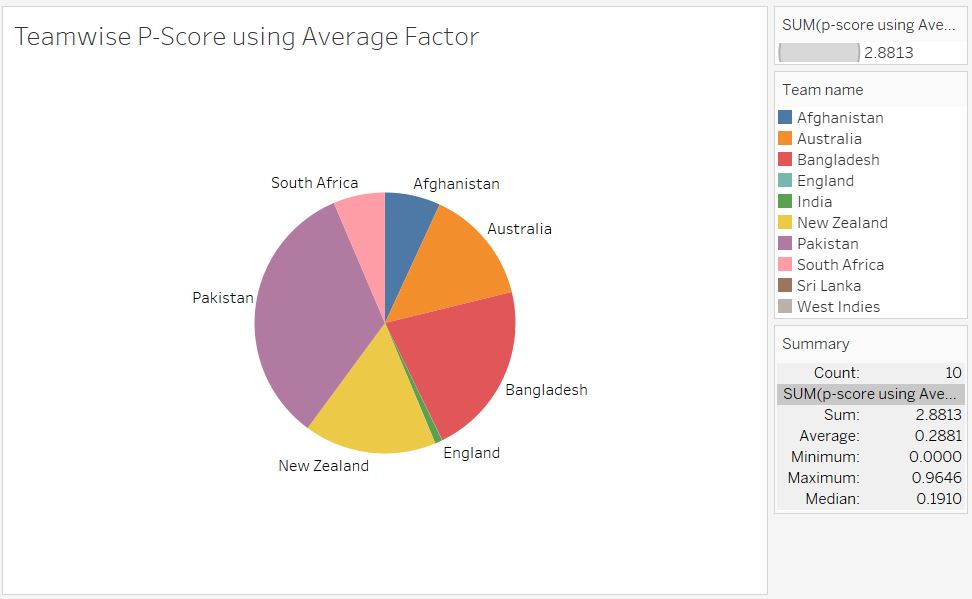
S2 = SD of Average of all teams= 8.7589

N1 = Number of series Australia played against other teams = 8

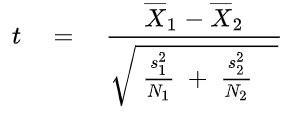
N2 = Total number of series for combination of teams in dataset = 81

p-score from Welch’s test formula = 0.4123

* P-score calculated for *Average* from above formula for each team can be demonstrated as below



* 1. *Calculate probability of RPO factor for any team (X) against other team (Y) using T-Distribution Welch's t-test**formula:*



**X¯1**= the sample mean = Mean value of *RPO* factor for team X, where X played matches against other teams

**X¯2** = the population mean = Mean value of *RPO* factor for all records in dataset

**S1** = standard deviation of the sample = S.D calculated taking all *RPO* values for team X against remaining teams

**S2** = standard deviation of the population = S.D calculated taking all *RPO* values for all teams

N1 = sample size = total Series played by team X = number of teams as opposition

N2 = population size = total number of series including all teams in dataset

***NOTE:*** *Here, output of above formula indicates the p-score, which is not actually the probability. As per the statistical science, the lower the value of the p-score, the higher the probability of success. Hence, if we compare p-score for two teams between whom match will be played, the probability of wining the match by team is more whose t-score is lower.*

X¯1 = mean of all *RPO* values of Australia against other teams = 5.5775

X¯2 = mean of all *RPO* values of all teams = 5.4377

S1 = SD of *RPO* of Australia against remaining team= 0.34995

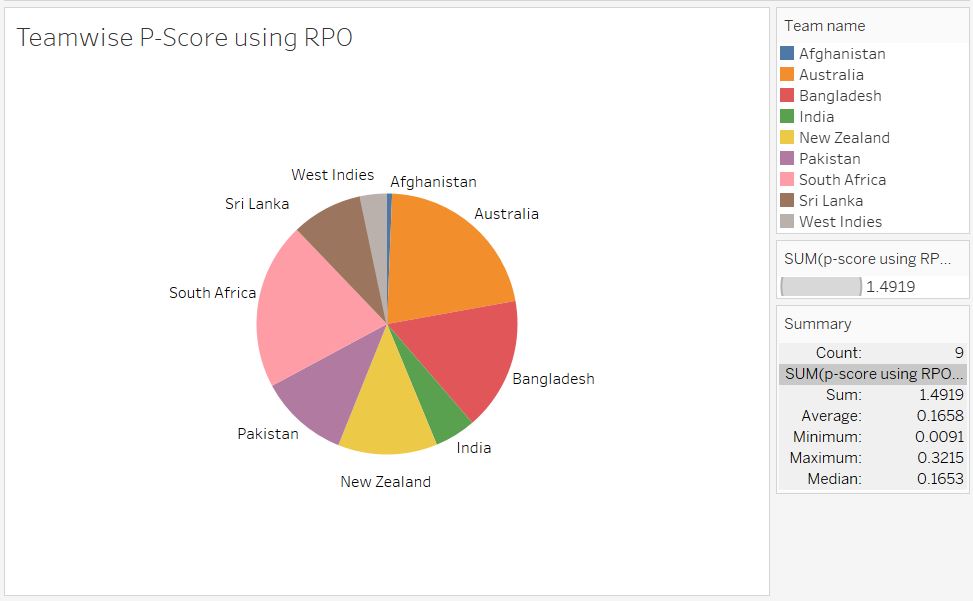
S2 = SD of *RPO* of all teams= 0.20564

N1 = Number of series Australia played against other teams = 8

N2 = Total number of series for combination of teams in dataset = 81

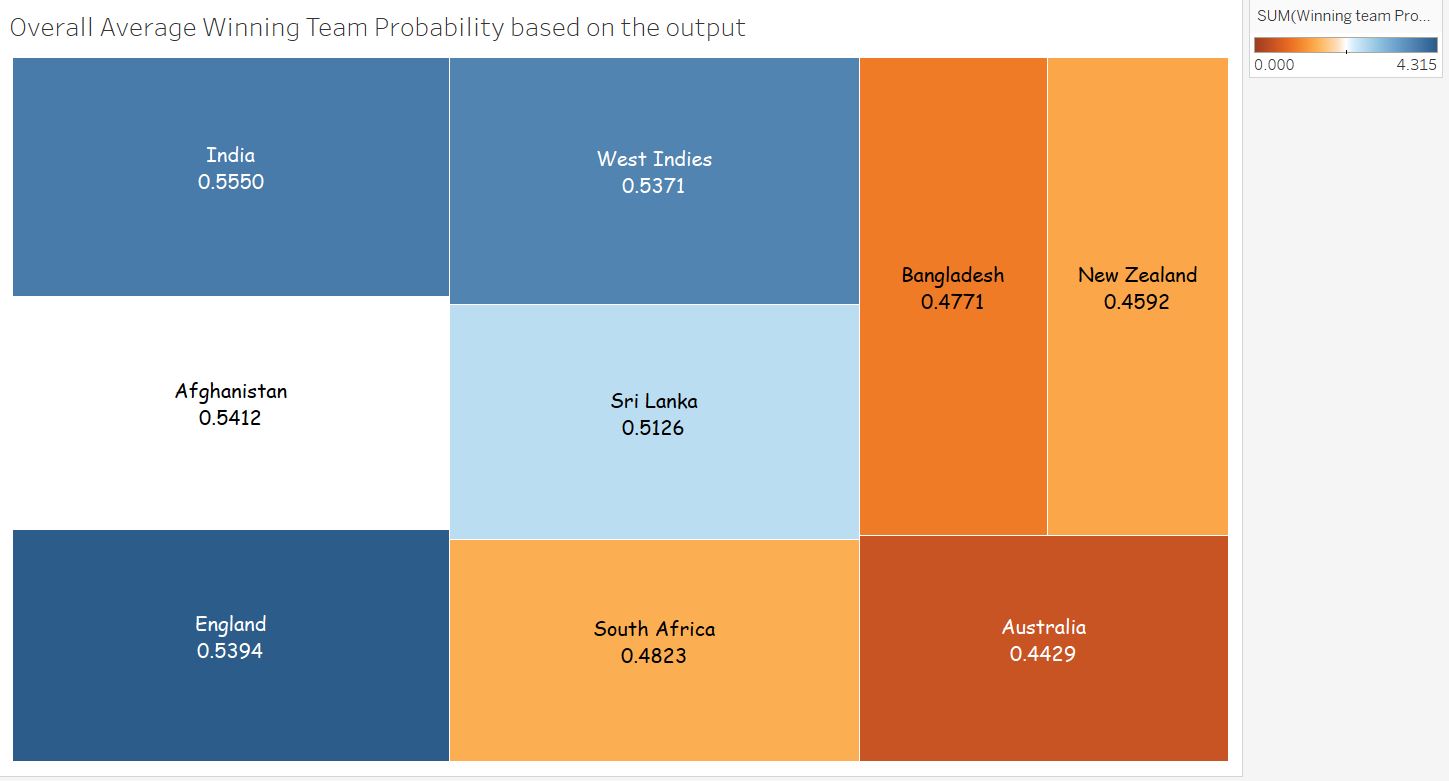
p-score from Welch’s test formula = 0.32152

* P-score calculated for *RPO* from above formula for each team can be demonstrated as below



1. *OUTPUT*

Considering Overall Average Winning Probability of team winning from Output HashMap



Let’s play the prediction game and input two teams:

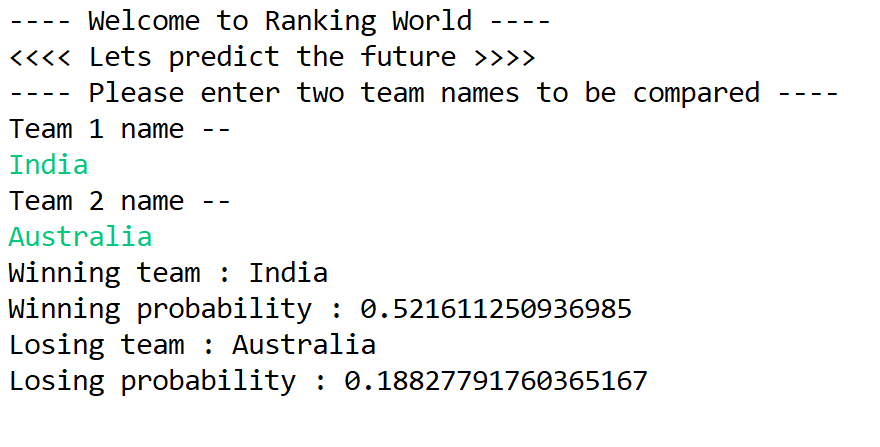
Team 1 name -

India

Team 2 name -

Australia

After considering all factors, the output shows that winning team is India showing fair chance based on win/loss, average of team, RPO of team.



1. *FUTURE PROJECT SCOPE*

The project can be taken on a large scale by considering following cases:

1. A more well-versed dataset including climatic parameters like dew, pitch/ground parameters like level of scales with which pitch made, place where match played (home ground/visiting ground) may make the estimation more accurate.
2. A minute probability of favoring luck could be add on following Monte Carlo Page Rank algorithm which uses a random number but always moves in correct direction.
3. *CONCLUSION*

* Algorithm has been implemented using HashMap data structure which is having time complexity of O(1) for both get and put method. Hence, data insertion and retrieval process become fast compared to other data structures, which reduces overall running time of algorithm.
* From the winning probability for each team calculated from Binomial Probability function, we can predict the winning chance of any team when match is played between them in future. It also indicates the standing of each team with comparison.
* P-score calculated from Welch’s t-test for continuous data such as *Average* and *RPO* help to find which team can perform better in term of considered factor when match will be played between them in future. More lower the p-score for a team, the higher the chances of winning the match.
* *REFERENCES:*
* <https://www.nature.com/news/scientific-method-statistical-errors-1.14700>
* <https://www.kaggle.com/akshayreddykotha/team-performance-before-cricket-world-cup-2019#recent_team_history.csv>
* <https://www.youtube.com/watch?v=7kqZ-G1DI9E&t=2s>
* <https://online-learning.harvard.edu/course/data-science-probability?delta=0>