CAPSTONE PROJECT

BUSINESS REPORT

ABSTRACT

The dataset contains information about the matches team India has played. The BCCI wants to make predictions regarding the winning prospects to make team India win.

Yashveer Kothari. A
Post Graduate Programme in Data Science & Business Analytics

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PROBLEM STATEMENT

BCCI has hired an external analytics consulting firm for data analytics. The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win. Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team. Once a model is developed then you have to extract actionable insights and recommendation. Also, below are the details of the next 10 matches, India is going to play. You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction. You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy. Hence for all the below 5 matches you have to suggest unique strategies to make India win. The suggestions should be in-line with the variables that have been mentioned in the given data set. Do consider the feasibility of the suggestions very carefully as well.

1. Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.
2. T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.
3. ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

DATA DICTIONARY

VARIABLES	DESCRIPTION
Game_number	Unique ID for each match
Result	Final result of the match
Avg_team_Age	Average age of the playing 11 players for that match
Match_light_type	type of match: Day, night or day & night
Match_format	Format of the match: T20, ODI or test
Bowlers_in_team	how many full-time bowlers has been player in the team
Wicket_keeper_in_team	how many full-time wicket keepers has been player in the team
All_rounder_in_team	how many full-time all-rounders has been player in the team
First_selection	First inning of team: batting or bowling
Opponent	Opponent team in the match
Season	What is the season of the city, where match has been played
Audience_number	Total number of audiences in the stadium
Offshore	Match played within country or outside of the country
Max_run_scored_1over	Maximum run scored in 1 over by team
Max_wicket_taken_1over	Maximum wicket taken in 1 over by team
Extra_bowls_bowled	Total number of extras bowled by team
Min_run_given_1over	Minimum run given by the bowler in one over
Min_run_scored_1over	Minimum run scored in 1 over by team
Max_run_given_1over	Maximum run given by the bowler in one over
extra_bowls_opponent	Total number of extras bowled by opponent
player_highest_run	Highest score in the match by one player
Players_scored_zero	Number of players out on zero run
player_highest_wicket	Highest wickets taken by single player in match

TABLE 1: TOP 5 DATASET SAMPLE

	Game_number	Result	Avg_team_Age	Match_light_type	Match_format	Bowlers_in_team	Wicket_keeper_in_team	All_rounder_in_team	First_selection
0	Game_1	Loss	18.0	Day	ODI	3.0	1	3.0	Bowling
1	Game_2	Win	24.0	Day	T20	3.0	1	4.0	Batting
2	Game_3	Loss	24.0	Day and Night	T20	3.0	1	2.0	Bowling
3	Game_4	Win	24.0	NaN	ODI	2.0	1	2.0	Bowling
4	Game_5	Loss	24.0	Night	ODI	1.0	1	3.0	Bowling

Opponent	Season	Audience_number	Offshore	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled	Min_run_given_1over
Srilanka	Summer	9940.0	No	13.0	3	0.0	2
Zimbabwe	Summer	8400.0	No	12.0	1	0.0	0
Zimbabwe	NaN	13146.0	Yes	14.0	4	0.0	0
Kenya	Summer	7357.0	No	15.0	4	0.0	2
Srilanka	Summer	13328.0	No	12.0	4	0.0	0

Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
3.0	6.0	0	54.0	3	1
3.0	6.0	0	69.0	2	1
3.0	6.0	0	69.0	3	1
3.0	6.0	0	73.0	3	1
3.0	6.0	0	80.0	3	1

TABLE 2: LAST 5 DATASET SAMPLE

	Game_number	Result	Avg_team_Age	Match_light_type	Match_format	Bowlers_in_team	Wicket_keeper_in_team	All_rounder_in_team	First_selection
2925	Game_2926	Win	30.0	Day	T20	3.0	1	4.0	Batting
2926	Game_2927	Win	30.0	Day	ODI	4.0	1	3.0	Bowling
2927	Game_2928	Win	30.0	Day and Night	ODI	4.0	1	3.0	Bowling
2928	Game_2929	Win	30.0	Day	ODI	4.0	1	3.0	Batting
2929	Game_2930	Win	30.0	Day	ODI	4.0	1	3.0	Batting

Opponent	Season	Audience_number	Offshore	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled	Min_run_given_1over
South Africa	Summer	33950.0	No	15.0	3	8.0	0
Kenya	Summer	19663.0	No	14.0	4	8.0	2
Pakistan	Rainy	39823.0	Yes	14.0	4	10.0	2
Kenya	Rainy	14007.0	No	14.0	2	20.0	2
Kenya	Rainy	20839.0	No	12.0	4	4.0	5

Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
3.0	6.0	3	50.0	3	2
3.0	6.0	2	52.0	2	1
4.0	10.0	2	80.0	3	2
3.0	6.0	3	98.0	3	1
3.0	6.0	3	62.0	1	1

TABLE 3: SHAPE OF THE DATASET

```
The number of rows and columns in the dataset is (2930, 23) respectively
```

TABLE 4: DATASET INFORMATION

```
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Game_number             2930 non-null   object
 1   Result                  2930 non-null   object
 2   Avg_team_Age            2833 non-null   float64
 3   Match_light_type        2878 non-null   object
 4   Match_format            2860 non-null   object
 5   Bowlers_in_team         2848 non-null   float64
 6   Wicket_keeper_in_team   2930 non-null   int64
 7   All_rounder_in_team     2890 non-null   float64
 8   First_selection         2871 non-null   object
 9   Opponent                2894 non-null   object
 10  Season                  2868 non-null   object
 11  Audience_number         2849 non-null   float64
 12  Offshore                2866 non-null   object
 13  Max_run_scored_1over    2902 non-null   float64
 14  Max_wicket_taken_1over  2930 non-null   int64
 15  Extra_bowls_bowled      2901 non-null   float64
 16  Min_run_given_1over     2930 non-null   int64
 17  Min_run_scored_1over    2903 non-null   float64
 18  Max_run_given_1over     2896 non-null   float64
 19  extra_bowls_opponent    2930 non-null   int64
 20  player_highest_run      2902 non-null   float64
 21  Players_scored_zero     2930 non-null   object
 22  player_highest_wicket   2930 non-null   object
dtypes: float64(9), int64(4), object(10)
```

1. There are 2930 rows and 13 columns in the dataset.
2. There are 10 variables with object data type.
3. There are 4 datatypes with integer data type.
4. There are 9 datatypes with float data type.

TABLE 5: DATASET DESCRIPTION

	count	mean	std	min	25%	50%	75%	max
Avg_team_Age	2833.0	29.242852	2.264230	12.0	30.0	30.0	30.00	70.0
Bowlers_in_team	2848.0	2.913624	1.023907	1.0	2.0	3.0	4.00	5.0
Wicket_keeper_in_team	2930.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0
All_rounder_in_team	2890.0	2.722491	1.092699	1.0	2.0	3.0	4.00	4.0
Audience_number	2849.0	46267.960688	48599.581459	7063.0	20363.0	34349.0	57876.00	1399930.0
Max_run_scored_1over	2902.0	15.199862	3.661010	11.0	12.0	14.0	18.00	25.0
Max_wicket_taken_1over	2930.0	2.713993	1.080623	1.0	2.0	3.0	4.00	4.0
Extra_bowls_bowled	2901.0	11.252671	7.780829	0.0	6.0	10.0	15.00	40.0
Min_run_given_1over	2930.0	1.952560	1.678332	0.0	0.0	2.0	3.00	6.0
Min_run_scored_1over	2903.0	2.762659	0.705759	1.0	2.0	3.0	3.00	4.0
Max_run_given_1over	2896.0	8.669199	5.003525	6.0	6.0	6.0	9.25	40.0
extra_bowls_opponent	2930.0	4.229693	3.626108	0.0	2.0	3.0	7.00	18.0
player_highest_run	2902.0	65.889387	20.331614	30.0	48.0	66.0	84.00	100.0

TABLE 6: DATASET DUPLICATES

```
The dataset contains 0 duplicate entries
```

- There are no duplicates in the dataset.

TABLE 7: MISSING DATA

There are 789 missing values in the dataset

Game_number	0
Result	0
Avg_team_Age	97
Match_light_type	52
Match_format	70
Bowlers_in_team	82
Wicket_keeper_in_team	0
All_rounder_in_team	40
First_selection	59
Opponent	36
Season	62
Audience_number	81
Offshore	64
Max_run_scored_1over	28
Max_wicket_taken_1over	0
Extra_bowls_bowled	29
Min_run_given_1over	0
Min_run_scored_1over	27
Max_run_given_1over	34
extra_bowls_opponent	0
player_highest_run	28
Players_scored_zero	0
player_highest_wicket	0

- We replace the missing value in the dataset with median and mode for Numerical variables and categorical variables respectively. (Refer Table 10)

TABLE 8: UNIQUE COUNT

```
unique count of Game_number
['Game_1' 'Game_2' 'Game_3' ... 'Game_2928' 'Game_2929' 'Game_2930']

unique count of Result
['Loss' 'Win']

unique count of Match_light_type
['Day' 'Day and Night' nan 'Night']

unique count of Match_format
['ODI' 'T20' 'Test' '20-20' nan]

unique count of First_selection
['Bowling' 'Batting' 'Bat' nan]

unique count of Opponent
['Srilanka' 'Zimbabwe' 'Kenya' 'Australia' 'England' 'South Africa'
 'Pakistan' 'West Indies' 'Bangladesh' nan]

unique count of Season
['Summer' nan 'Winter' 'Rainy']

unique count of Offshore
['No' 'Yes' nan]

unique count of Players_scored_zero
[3 2 1 4 'Three']

unique count of player_highest_wicket
[1 2 3 4 'Three' 5]
```

- **Modifying few unique data names:**

1. 20-20: ‘T20’
2. Bat: ‘Batting’
3. Three: ‘3’ in both 'Players_scored_zero' and 'player_highest_wicket' columns

- **Modified names in the dataset:**

```
['ODI' 'T20' 'Test' nan]
['Bowling' 'Batting' nan]
[3 2 1 4]
[1 2 3 4 5]
```

- **Modifying few data types:**

1. 'Players_scored_zero’ and 'player_highest_wicket' to integer type (‘int64’)

TABLE 9: DATA INFORMATION AFTER MODIFICATION

```
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Game_number             2930 non-null   object
 1   Result                  2930 non-null   object
 2   Avg_team_Age            2833 non-null   float64
 3   Match_light_type        2878 non-null   object
 4   Match_format            2860 non-null   object
 5   Bowlers_in_team         2848 non-null   float64
 6   Wicket_keeper_in_team   2930 non-null   int64
 7   All_rounder_in_team     2890 non-null   float64
 8   First_selection         2871 non-null   object
 9   Opponent                2894 non-null   object
 10  Season                  2868 non-null   object
 11  Audience_number         2849 non-null   float64
 12  Offshore                2866 non-null   object
 13  Max_run_scored_1over    2902 non-null   float64
 14  Max_wicket_taken_1over  2930 non-null   int64
 15  Extra_bowls_bowled      2901 non-null   float64
 16  Min_run_given_1over     2930 non-null   int64
 17  Min_run_scored_1over    2903 non-null   float64
 18  Max_run_given_1over     2896 non-null   float64
 19  extra_bowls_opponent    2930 non-null   int64
 20  player_highest_run      2902 non-null   float64
 21  Players_scored_zero     2930 non-null   int64
 22  player_highest_wicket   2930 non-null   int64
dtypes: float64(9), int64(6), object(8)
```

TABLE 10: REPLACING THE MISSING VALUES

```
Result                    0
Avg_team_Age              0
Match_light_type          0
Match_format              0
Bowlers_in_team           0
All_rounder_in_team       0
First_selection           0
Opponent                  0
Season                    0
Audience_number           0
Offshore                  0
Max_run_scored_1over      0
Max_wicket_taken_1over    0
Extra_bowls_bowled        0
Min_run_given_1over       0
Min_run_scored_1over      0
Max_run_given_1over       0
extra_bowls_opponent      0
player_highest_run        0
Players_scored_zero       0
player_highest_wicket     0
dtype: int64
```

- The missing data in the dataset has been replaced as per the datatype using the median and mode. Hence There are no missing values in the dataset.

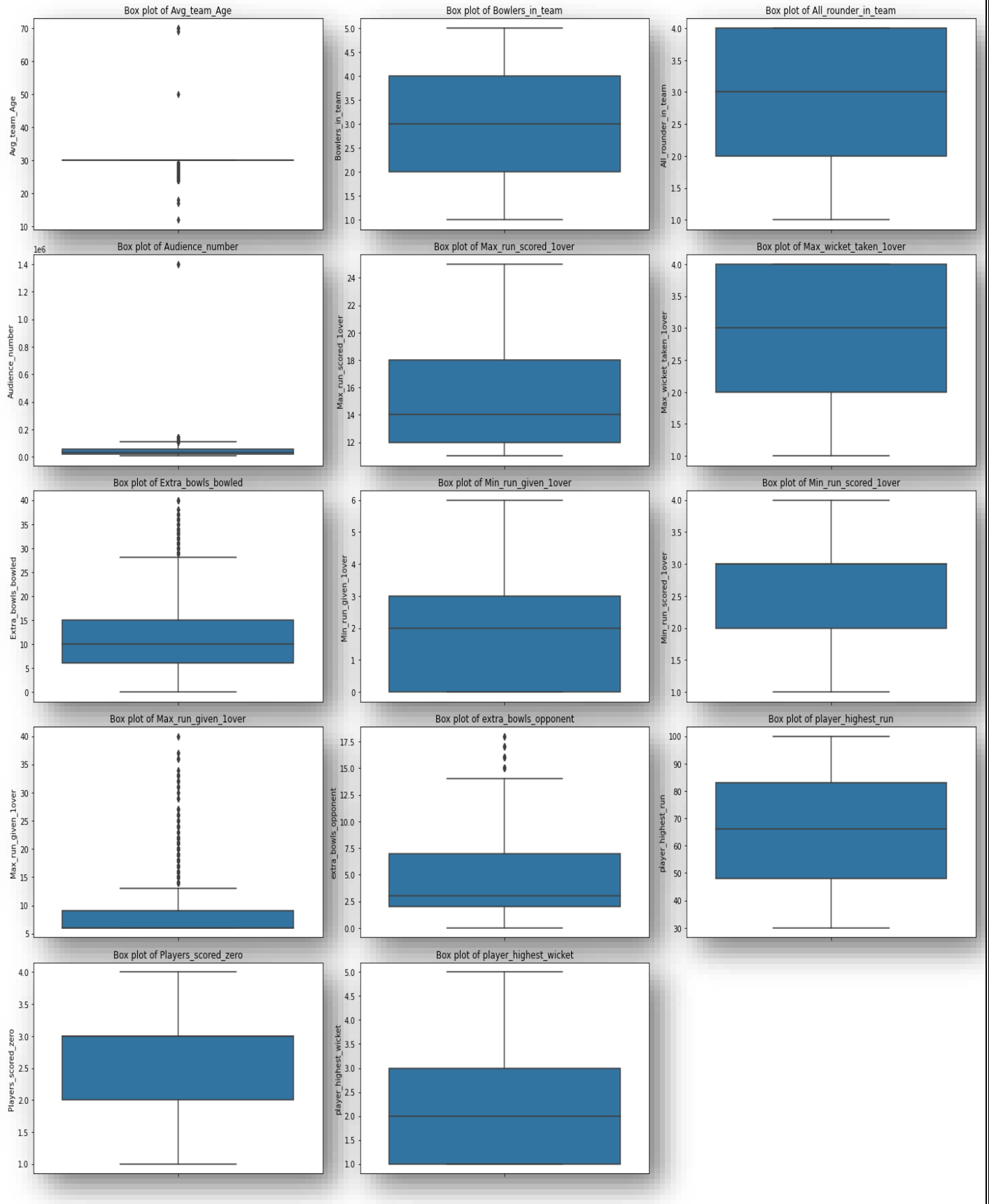
Numerical data

```
Avg_team_Age              0
Bowlers_in_team           0
All_rounder_in_team       0
Audience_number           0
Max_run_scored_1over      0
Max_wicket_taken_1over    0
Extra_bowls_bowled        0
Min_run_given_1over       0
Min_run_scored_1over      0
Max_run_given_1over       0
extra_bowls_opponent      0
player_highest_run        0
Players_scored_zero       0
player_highest_wicket     0
```

Categorical data

```
Match_light_type    0
Match_format        0
First_selection     0
Opponent            0
Season              0
Offshore            0
Result              0
```

FIGURE 1: BOXPLOT



1. Variable ‘avg_team_age’ is close to normal distribution as mean and median are similar though it has outliers.
2. Variable ‘Bowlers_in_team’ is normally distributed as mean and median are almost same, no outliers are present.
3. Variable ‘All_rounder_in_team’ is slightly left skewed as there is difference in mean and median and also 75 percentile and the maximum value are equal, no outliers are present.
4. Variable ‘audience_number’ is right skewed as mean is affected due to outliers present.
5. Variable ‘Max_run_scored_1over’ is slightly right skewed, no outliers are present.
6. Variable ‘Max_wicket_taken_1over’ is slightly left skewed as mean is less than median and also 75 percentile and the maximum value are same, no outliers are present.
7. Variable ‘Extra_bowls_bowled’ is right skewed as mean is higher than median, outliers are present.
8. Variable ‘Min_run_given_1over’ is close to normal and minimum value and 25 percentiles are equal, no outliers are present.
9. Variable ‘Min_run_scored_1over’ is slightly left skewed, 50 and 75 percentiles are equal, no outliers are present.
10. Variable ‘Max_run_given_1over’ is right skewed, minimum value,25,50 percentile has same values, outliers are present.
11. Variable ‘extra_bowls_opponent’ is right skewed, outliers are present.
12. Variable ‘player_highest_run’ is normally distributed, no outliers are present.
13. Variable ‘Players_scored_zero’ is slightly left skewed, 50 and 75 percentiles have the same value, no outliers are present.
14. Variable ‘player_highest_wicket’ is normally distributed, minimum value and 25 percentiles are equal.

TABLE 11: SKEWNESS

Avg_team_Age	5.068403
Bowlers_in_team	-0.296492
All_rounder_in_team	-0.335012
Audience_number	15.782867
Max_run_scored_1over	0.838907
Max_wicket_taken_1over	-0.305597
Extra_bowls_bowled	1.132432
Min_run_given_1over	0.433859
Min_run_scored_1over	-0.568821
Max_run_given_1over	2.692147
extra_bowls_opponent	0.916295
player_highest_run	-0.031472
Players_scored_zero	-0.505491
player_highest_wicket	1.026090

FIGURE 2: MATCH LIGHT GRAPH

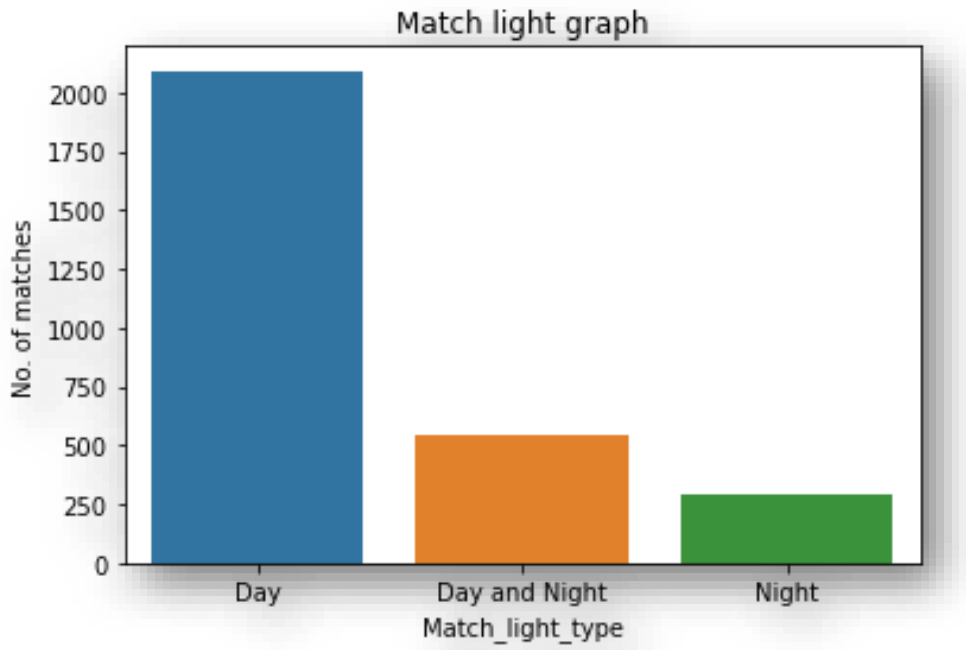


FIGURE 3: MATCH FORMAT GRAPH

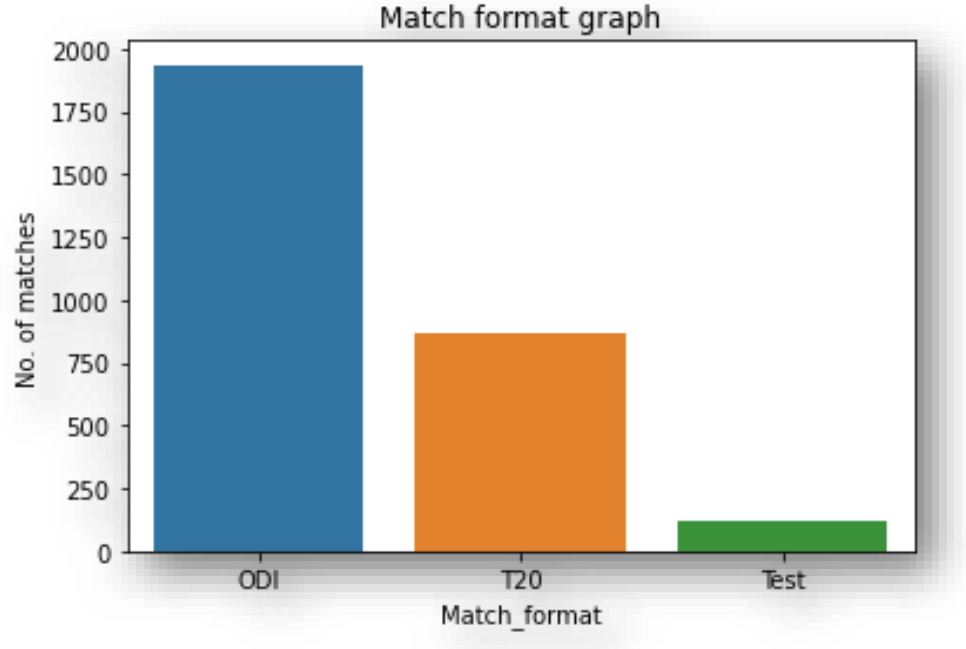


FIGURE 4: FIRST SELECTION GRAPH

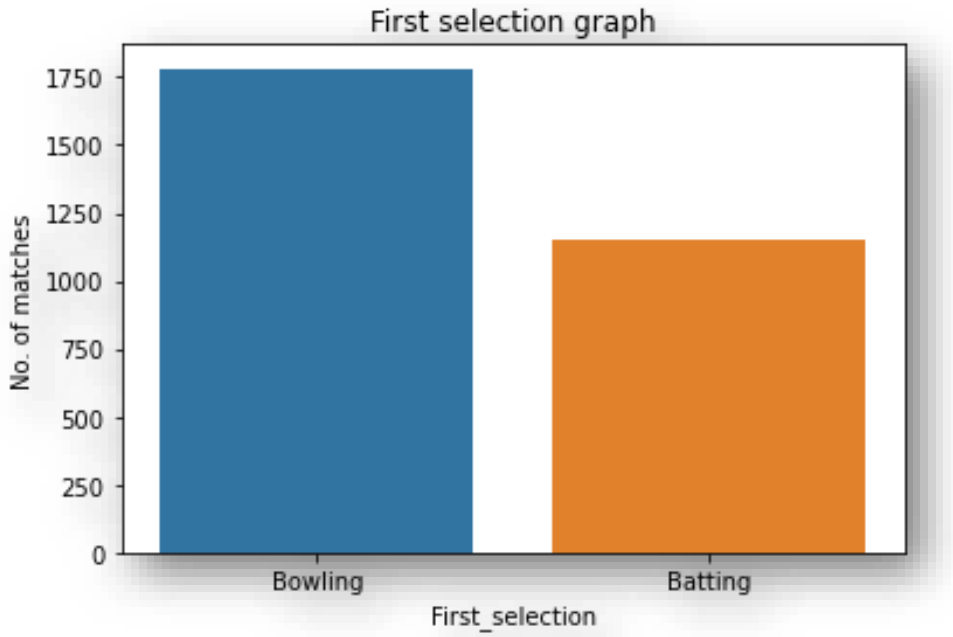


FIGURE 5: OPPONENT GRAPH

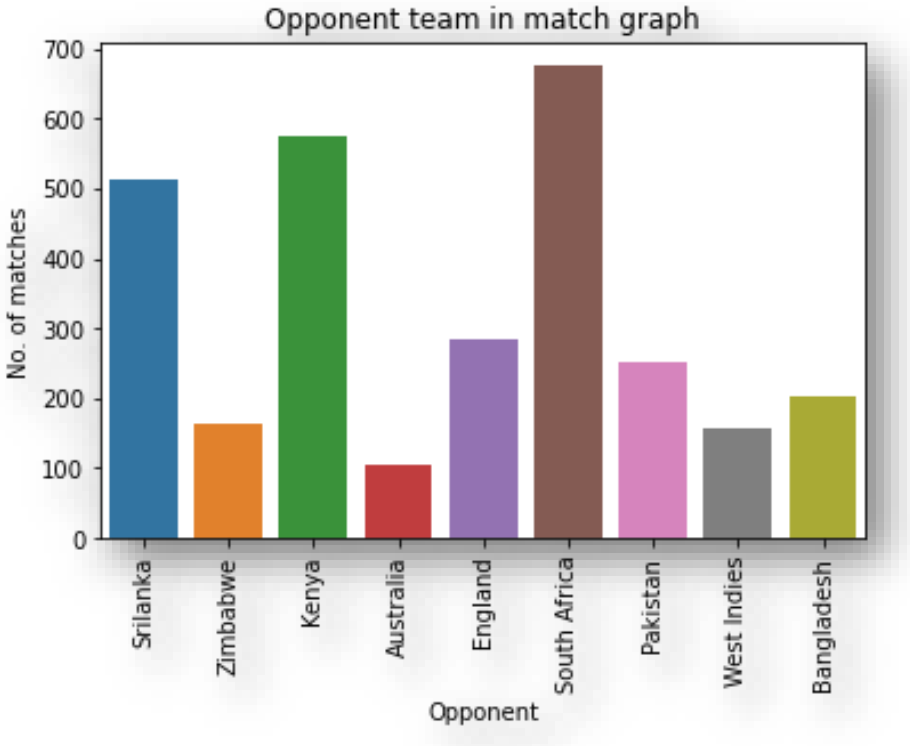


FIGURE 6: SEASON OF THE CITY GRAPH

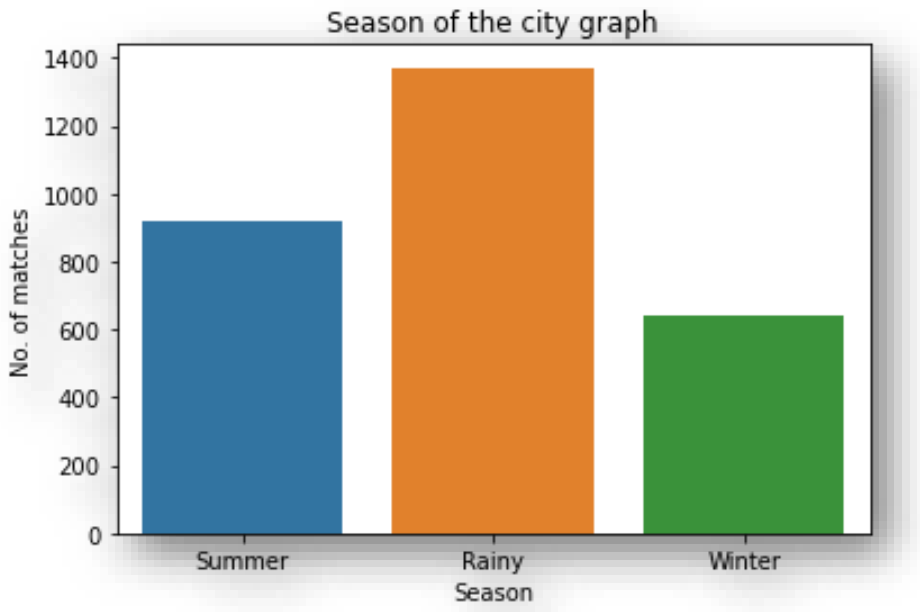


FIGURE 7: OFFSHORE GRAPH

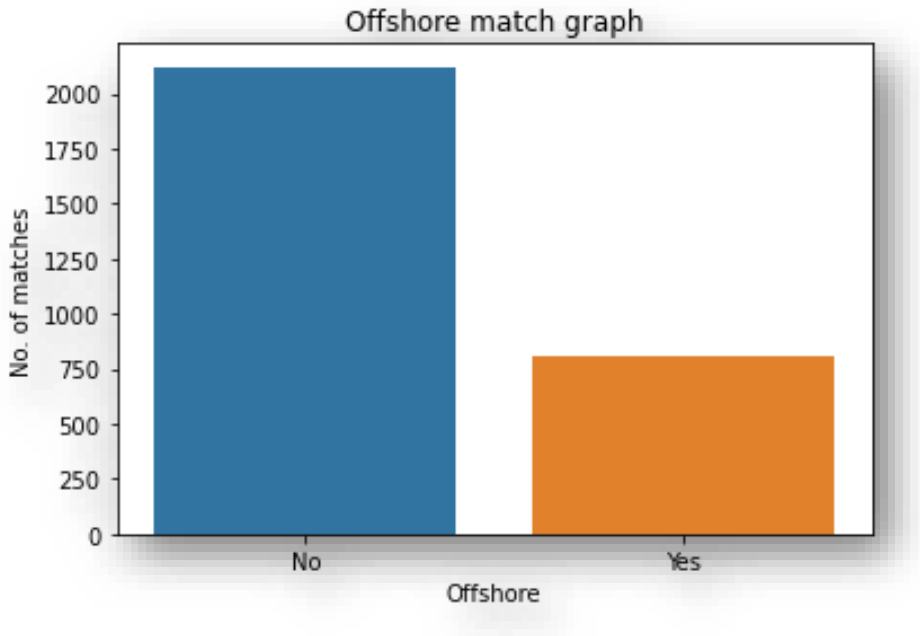
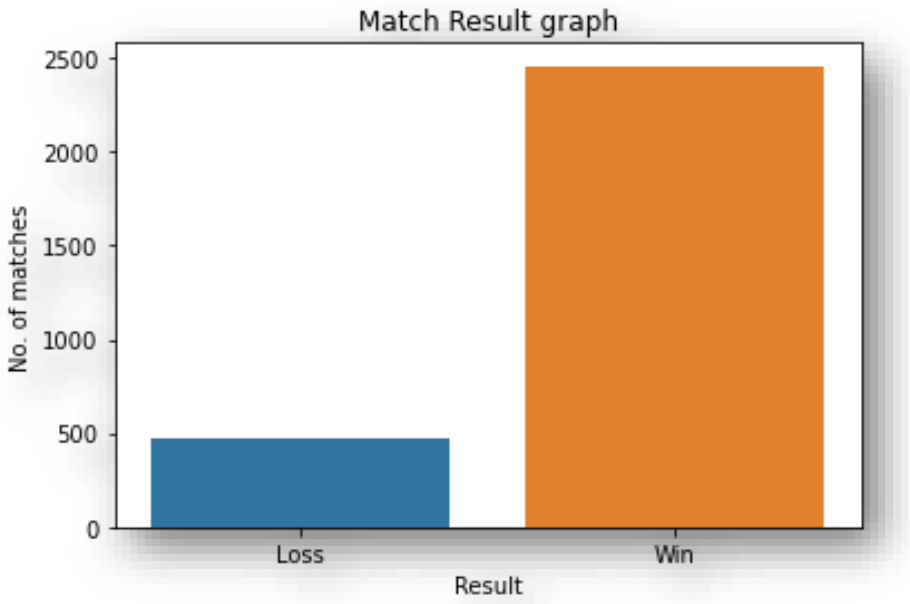
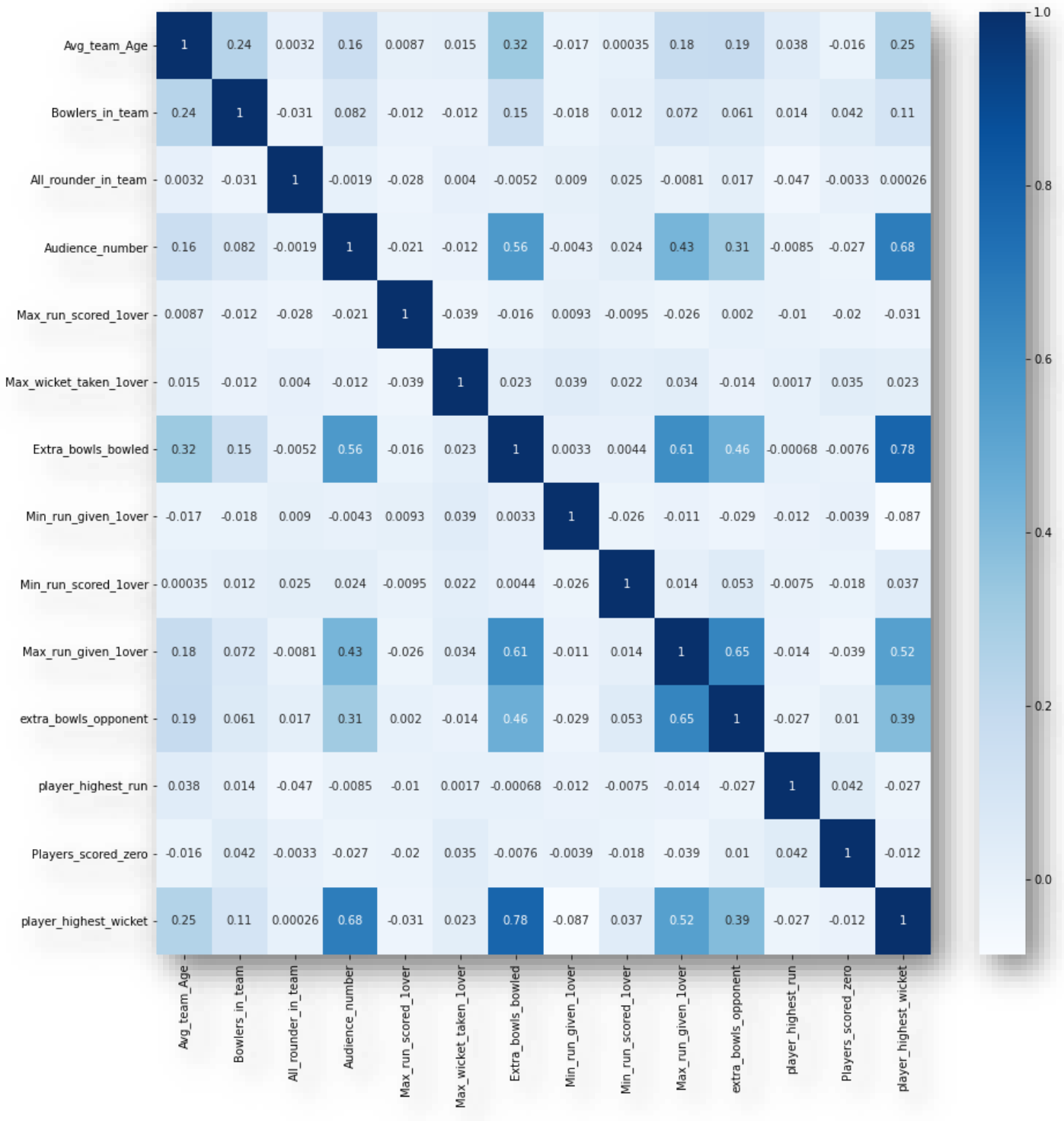


FIGURE 8: MATCH RESULT GRAPH



1. **Match_light_type**: The number of matches played during the day is the highest and the lowest matches played during the night.
2. **Match_format:** The number of ODI matches played is the highest and the lowest played is for the test format.
3. **First selection:** The highest number of the matches played with bowling first and lowest with the batting as the first selection.
4. **Opponent:** Highest number of matches are played against south Africa and lowest matches played against Australia.
5. **Season**: highest number of matches are played during rainy season and lowest matches are played during winter season.
6. **Offshore:** Less matches are played offshore, most of the matches are played on the home ground.
7. **Result**: Among all the matches most of the matches are won, less matches are lost.

FIGURE 9: HEATMAP/CORRELATION PLOT



1. There are is multi collinearity as per the above heatmap among 10 variables.
2. Variables ‘extra bowls bowled’ and ‘Players_highest_wicket’ has the highest correlation 0.78.
3. Variables ‘Audience number’ and ‘Players_highest_wicket’ has the strong correlation 0.68.

4. Variables ‘Maximum runs given in one over’ and ‘extra bowls’ opponents have the correlation of 0.65.
5. Variables ‘Maximum runs given in one over’ and ‘extra bowls bowled’ has the correlation of 0.61.

FIGURE 10: PAIRPLOT

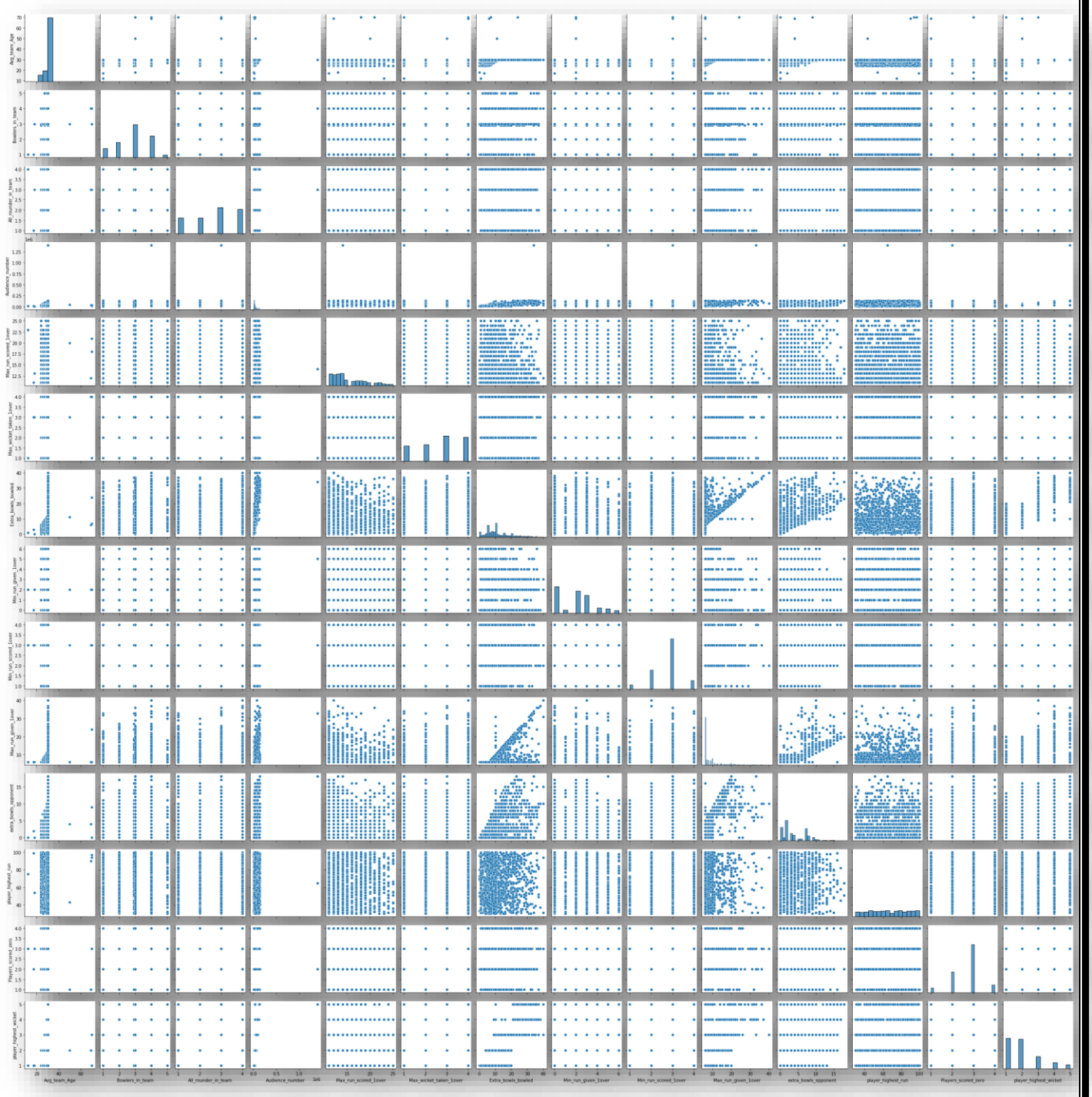
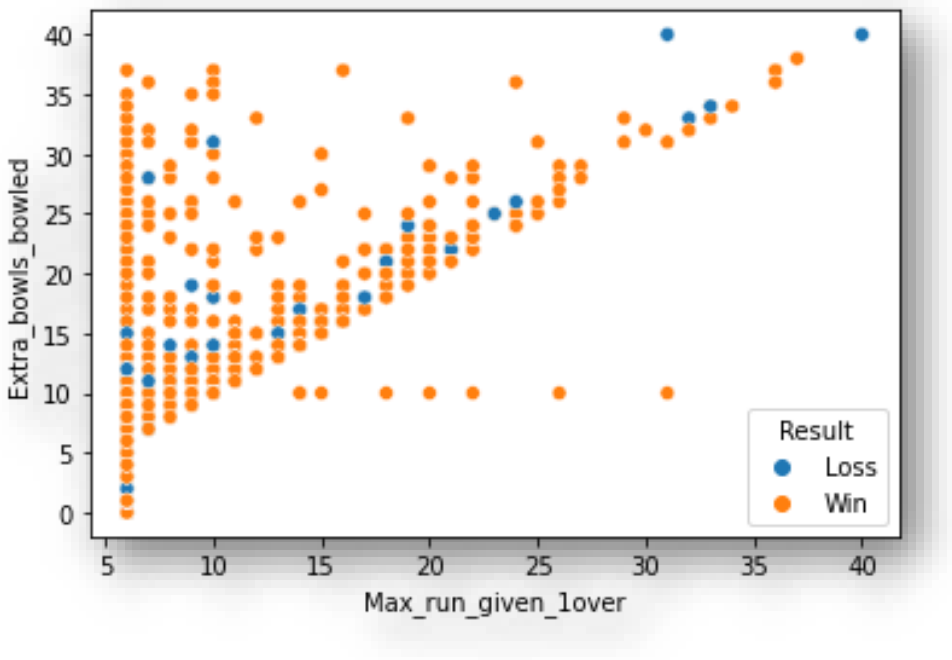


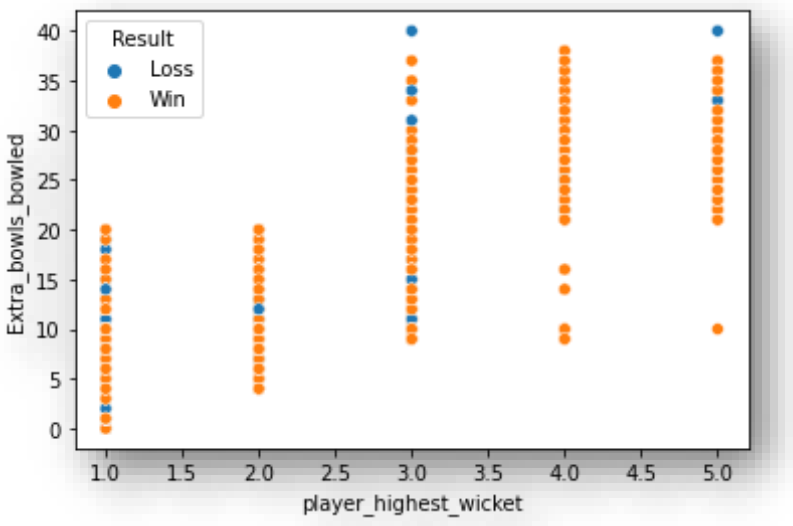
FIGURE 11: SCATTER PLOT

- **MAX_RUN_GIVEN_1OVER & EXTRA_BOWLS_BOWLED**



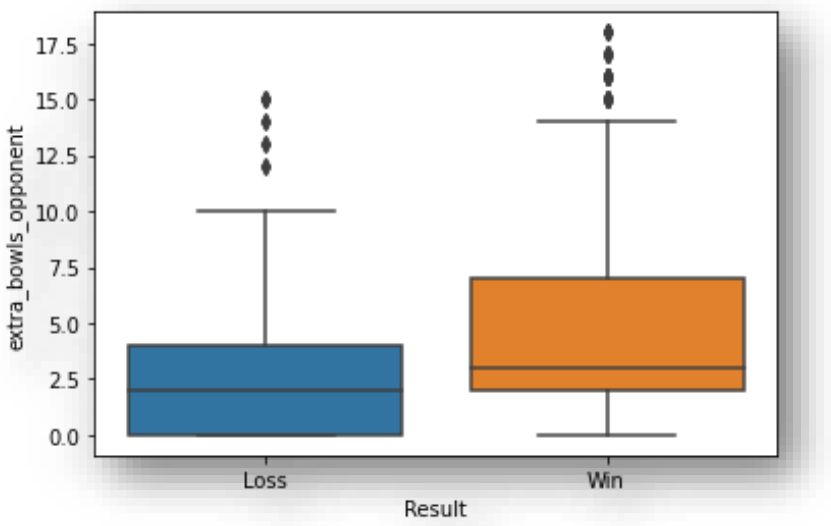
- It shows the linear relationship between maximum run given in one over and extra bowls bowled. Highest value of both the variables results in loss.

- **PLAYER_HIGHEST_WICKET & EXTRA_BOWLS_BOWLED**



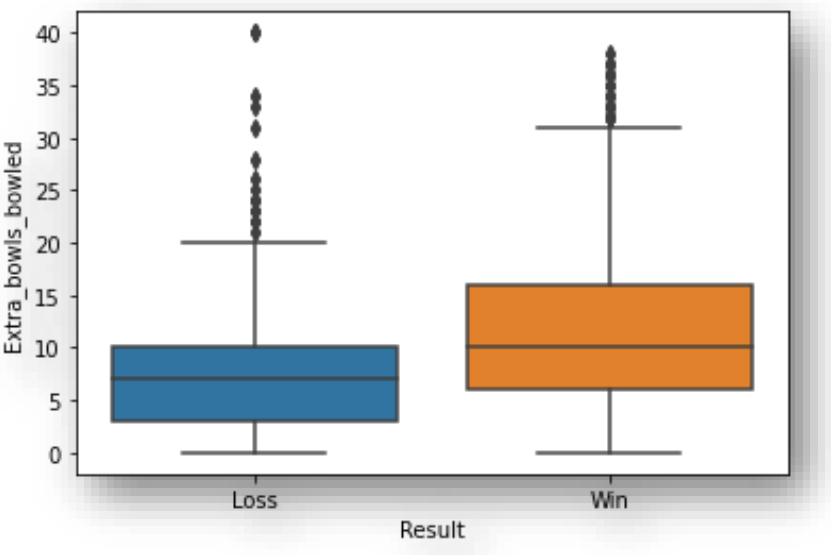
- The highest results also end up in a loss.

FIGURE 12: BOXPLOT – RESULT AND EXTRA BOWLS OPPONENT



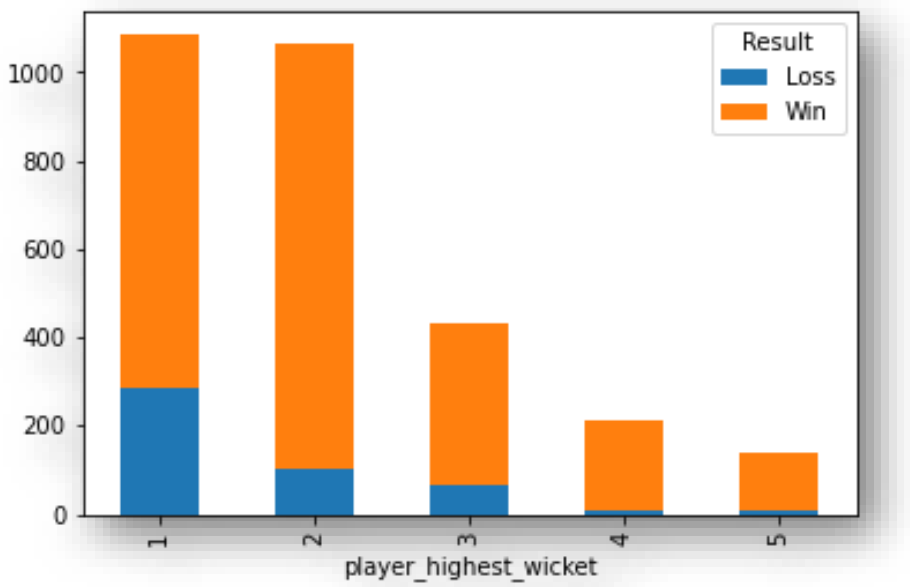
- The median of extra bowls opponents is higher for win, If the extra bowls opponents is higher than 16 then India will win the match as per the data set.
- If the extra bowls opponents are higher than 10 then chances of winning the match is more.

FIGURE 13: BOXPLOT – RESULT AND EXTRA BOWLS BOWLED



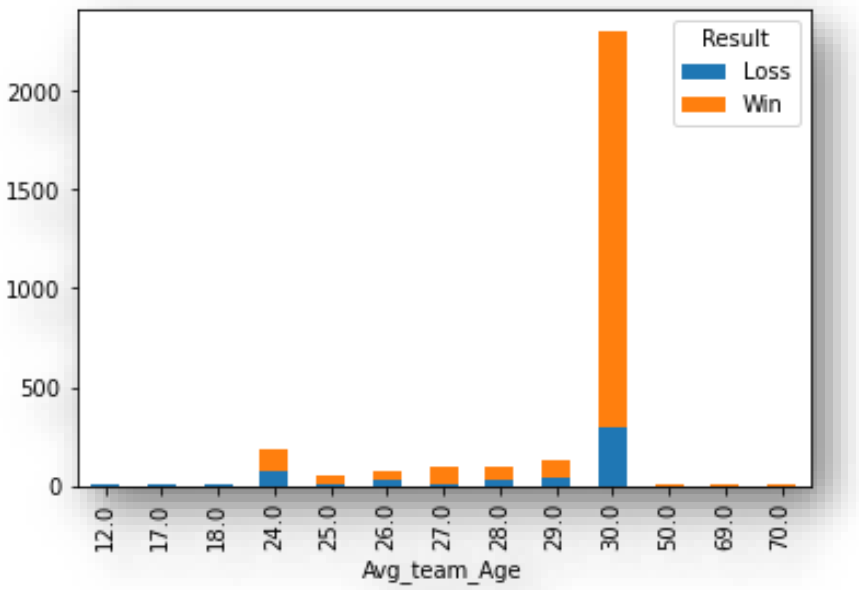
- As per data when you bowl 40 extra bowls India is definitely will lose the match.

FIGURE 14: BARPLOT – RESULT AND PLAYER HIGHEST WICKET



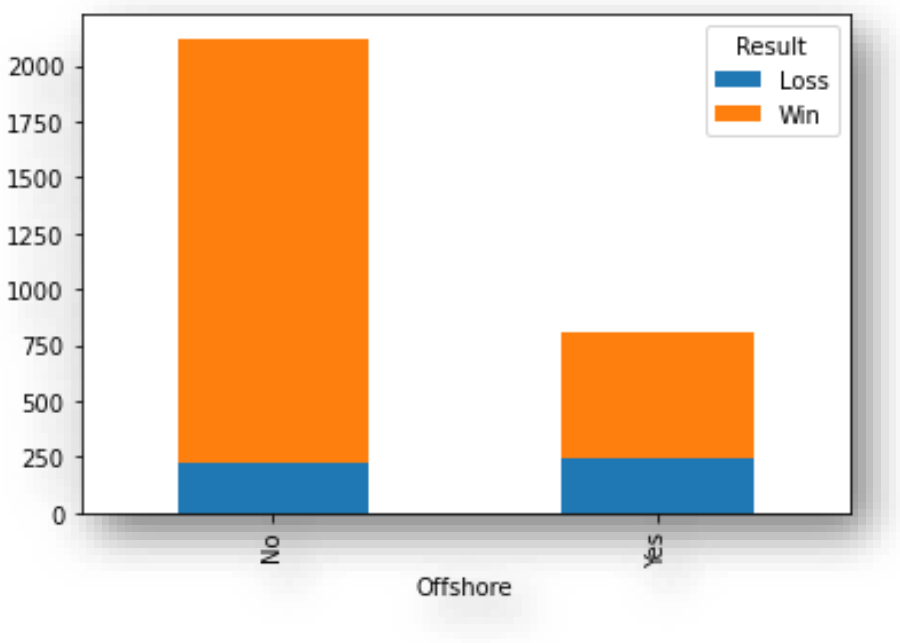
- The chances of winning the match are high when the 2 wickets are taken by a single player.

FIGURE 15: BARPLOT – RESULT AND AVG TEAM AGE



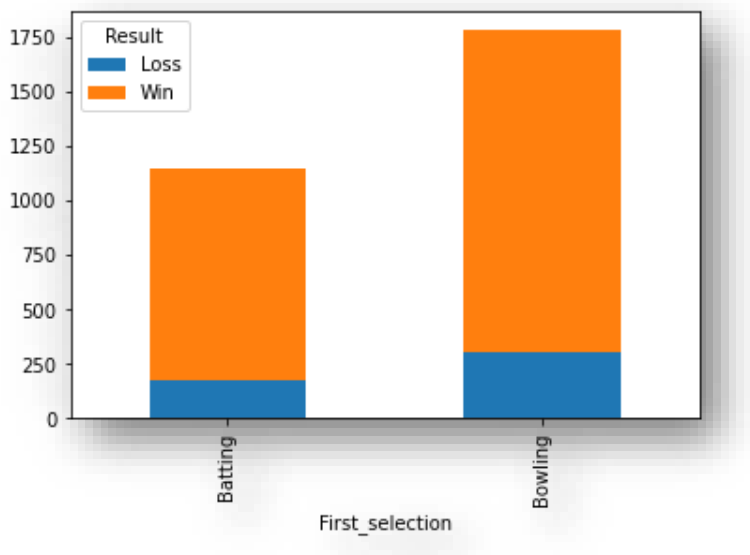
- India has the highest win with average team age at 30.

FIGURE 16: BARPLOT – RESULT & OFFSHORE



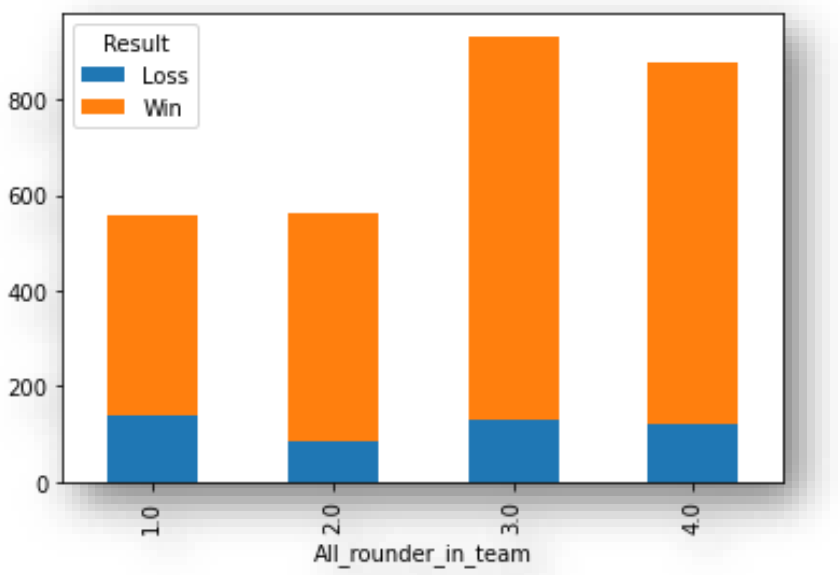
- The performance of the Indian team is good when played on the home ground.

FIGURE 17: BARPLOT – RESULT & FIRST SELECTION



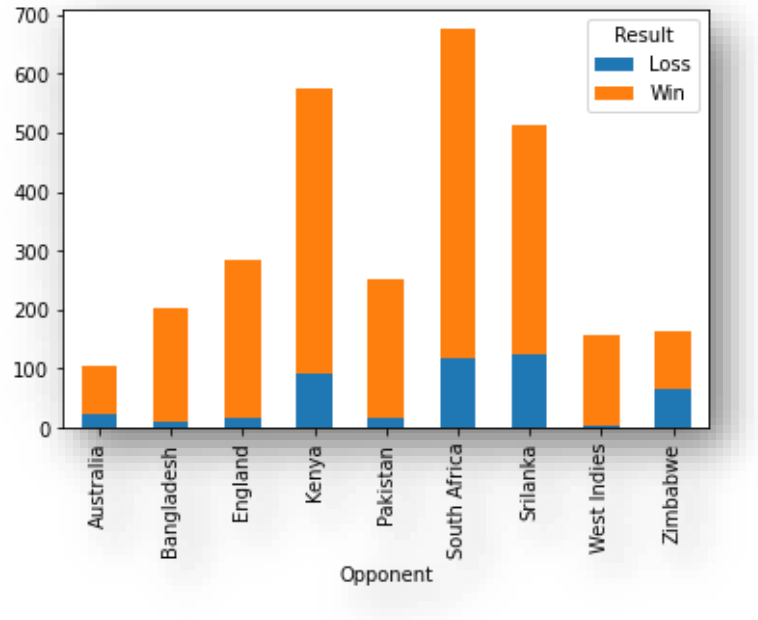
- Most of the matches are won when selected to bowl first.

FIGURE 18: BARPLOT – RESULT & ALL ROUNDER IN TEAM



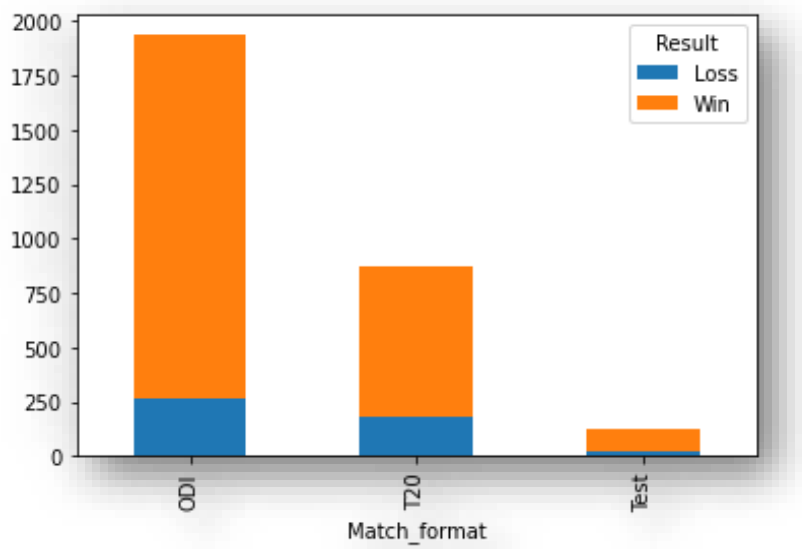
- Winning percentage is higher when matches are played with 3 to 4 all-rounders in the team.

FIGURE 19: BARPLOT – RESULT & OPPONENT IN TEAM



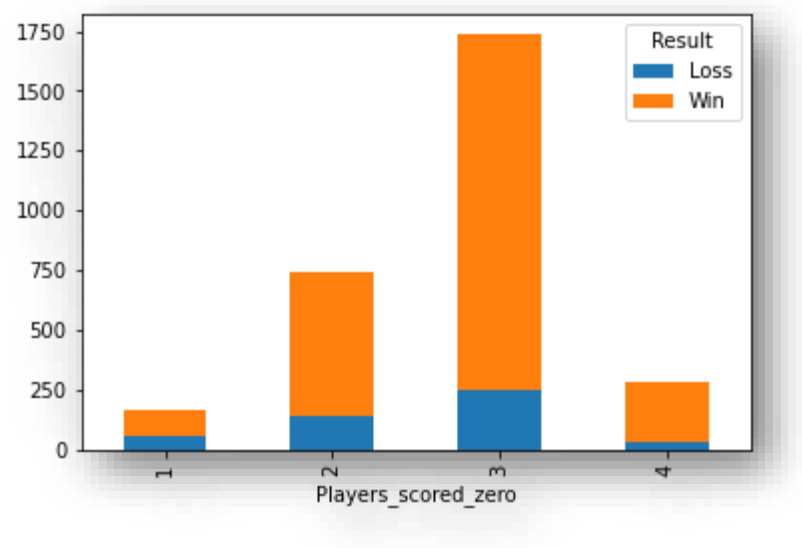
- India team is performing well against West Indies, Bangladesh, England and Pakistan India team wining rate is less against South Africa, Sri Lanka, Zimbabwe and Australia.

FIGURE 20: BARPLOT – RESULT & MATCH FORMAT IN TEAM



- Indian team has a good performance in ODI format as compared to both the formats Winning rate is higher in ODI format and lesser in T20 and test.

FIGURE 21: BARPLOT – RESULT & PLAYER SCORED ZERO

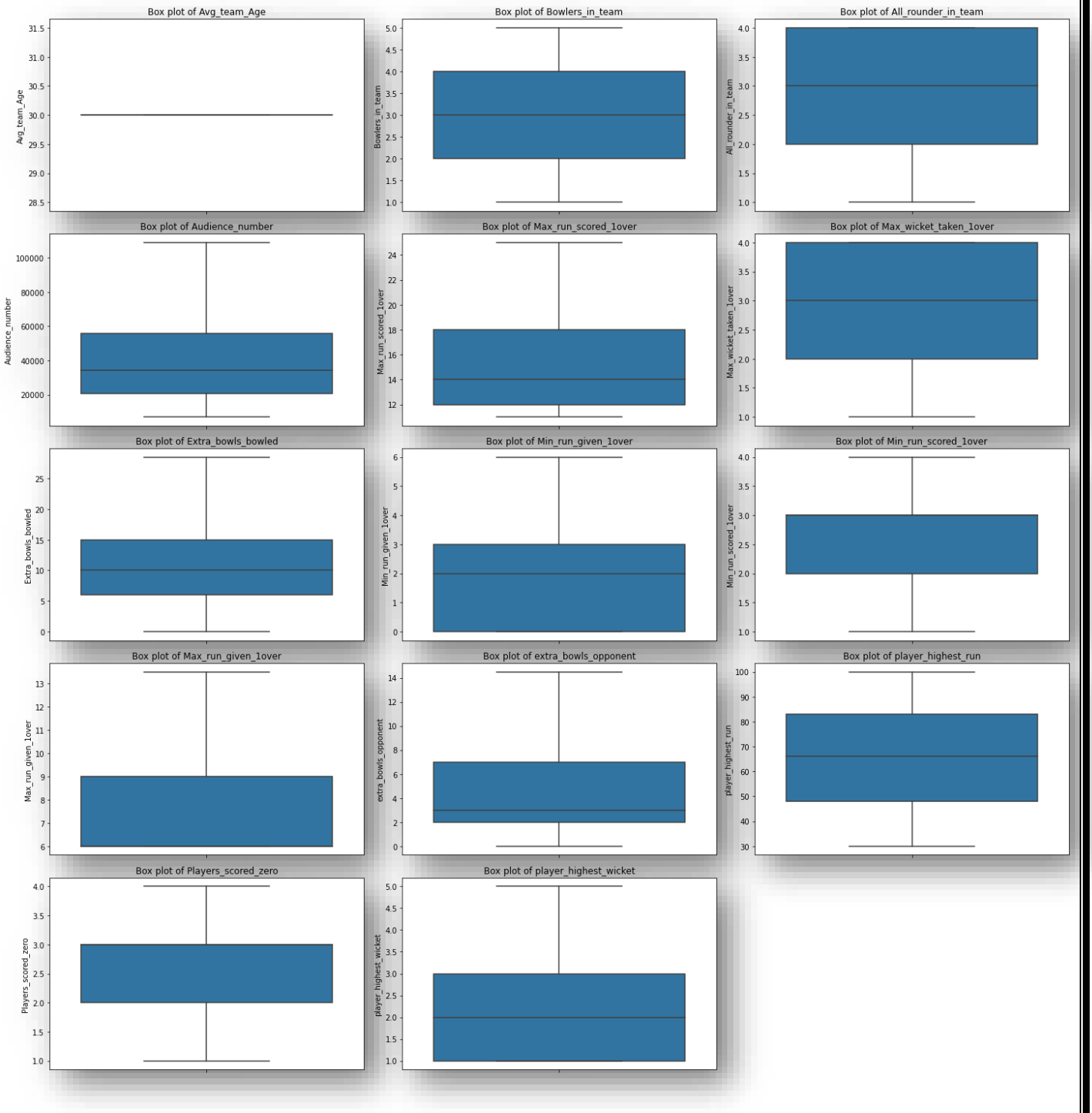


- Indian team has the highest wining rate when there are 3 players scored zero and lowest when there is one player scored zero.

BUSINESS INSIGHTS

1. Variables maximum runs given in one over' and extra bowls bowled have a good relation which results in win or loss of the match. More number of runs given in one over and extra bowls bowled chances of losing is high
2. Variable extra bowl's opponent can add value to the prediction. More extra bowls opponent more is chance of the winning
3. Variable offshore also predict the win. If the number of matches played on home ground are more chances of winning is high
4. Variable first selection also impact on the result. If the first selection is bowling the chances of winning is high
5. Variable all-rounder in team has impact on the result. Having 3 to 4 all-rounders in the team may result into win

FIGURE 22: OUTLIER TREATMENT



1. Outliers have effect on mean. It increases the mean. If it is a distance-based calculations, then the modelling may be affected. So, it is necessary to treat outliers.
2. Box plot is used to find the outliers.
3. Variables avg_team_age, audience number, extra bowls bowled, maximum run given in one over, extra bowls opponent have outliers. We use Inter Quartile Range method to treat the outliers.

TABLE 12: LABEL ENCODING

	Avg_team_Age	Bowlers_in_team	All_rounder_in_team	Audience_number	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled
0	30.0	3.0	3.0	9940.0	13.0	3.0	0.0
1	30.0	3.0	4.0	8400.0	12.0	1.0	0.0
2	30.0	3.0	2.0	13146.0	14.0	4.0	0.0
3	30.0	2.0	2.0	7357.0	15.0	4.0	0.0
4	30.0	1.0	3.0	13328.0	12.0	4.0	0.0

Min_run_given_1over	Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
2.0	3.0	6.0	0.0	54.0	3.0	1.0
0.0	3.0	6.0	0.0	69.0	2.0	1.0
0.0	3.0	6.0	0.0	69.0	3.0	1.0
2.0	3.0	6.0	0.0	73.0	3.0	1.0
0.0	3.0	6.0	0.0	80.0	3.0	1.0

Match_light_type	Match_format	First_selection	Opponent	Season	Offshore	Result
Day	ODI	Bowling	Srilanka	Summer	No	0
Day	T20	Batting	Zimbabwe	Summer	No	1
Day and Night	T20	Bowling	Zimbabwe	Rainy	Yes	0
Day	ODI	Bowling	Kenya	Summer	No	1
Night	ODI	Bowling	Srilanka	Summer	No	0

TABLE 13: ONE HOT ENCODING

	Avg_team_Age	Bowlers_in_team	All_rounder_in_team	Audience_number	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled
0	30.0	3.0	3.0	9940.0	13.0	3.0	0.0
1	30.0	3.0	4.0	8400.0	12.0	1.0	0.0
2	30.0	3.0	2.0	13146.0	14.0	4.0	0.0
3	30.0	2.0	2.0	7357.0	15.0	4.0	0.0
4	30.0	1.0	3.0	13328.0	12.0	4.0	0.0

Min_run_given_1over	Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
2.0	3.0	6.0	0.0	54.0	3.0	1.0
0.0	3.0	6.0	0.0	69.0	2.0	1.0
0.0	3.0	6.0	0.0	69.0	3.0	1.0
2.0	3.0	6.0	0.0	73.0	3.0	1.0
0.0	3.0	6.0	0.0	80.0	3.0	1.0

Result	Match_light_type_Day and Night	Match_light_type_Night	Match_format_T20	Match_format_Test	First_selection_Bowling	Opponent_Bangladesh	Opponent_England
0	0	0	0	0	1	0	0
1	0	0	1	0	0	0	0
0	1	0	1	0	1	0	0
1	0	0	0	0	1	0	0
0	0	1	0	0	1	0	0

Opponent_Kenya	Opponent_Pakistan	Opponent_South Africa	Opponent_Srilanka	Opponent_West Indies	Opponent_Zimbabwe	Season_Summer	Season_Winter	Offshore_Yes
0	0	0	1	0	0	1	0	0
0	0	0	0	0	1	1	0	0
0	0	0	0	0	1	0	0	1
1	0	0	0	0	0	1	0	0
0	0	0	1	0	0	1	0	0

TABLE 14: SHAPE OF THE DATASET AFTER ONE HOT ENCODING

- The shape of the dataset after one hot encoding is: (2930, 31) rows and columns respectively.

TABLE 15: LOGISTIC REGRESSION

1. We split the result with all independent variables
2. We then split the data into 70 and 30. 70 for training and 30 for testing using train test split
3. For logistics regression one hot encoding is done
4. Logistic regression model internally uses linear equation to find the intercept and coefficient and then it is converted to the classes using activation function. It uses sigmoid curve to calculate the probability depending on the defined threshold. Any value greater than threshold will be considered as 1 and the value less than threshold will be considered as 0. Threshold value is usually 0.5 and it can be adjusted accordingly. It uses log of odds to convert into the probability. Log of odds is the linear equation having intercept and coefficient.

Logit Regression Results

Dep. Variable:	Result	No. Observations:	2930
Model:	Logit	Df Residuals:	2900
Method:	MLE	Df Model:	29
Date:	Tue, 17 Jan 2023	Pseudo R-squ.:	0.2501
Time:	15:32:34	Log-Likelihood:	-971.21
converged:	True	LL-Null:	-1295.2
Covariance Type:	nonrobust	LLR p-value:	7.077e-118

	coef	std err	z	P>|z|	[0.025	0.975]
Avg_team_Age	-0.1074	0.048	-2.234	0.025	-0.202	-0.013
Bowlers_in_team	-0.0322	0.058	-0.555	0.579	-0.146	0.081
All_rounder_in_team	0.3381	0.054	6.253	0.000	0.232	0.444
Audience_number	2.67e-06	6.77e-06	0.394	0.693	-1.06e-05	1.59e-05
Max_run_scored_1over	0.0241	0.016	1.467	0.142	-0.008	0.056
Max_wicket_taken_1over	0.1737	0.054	3.231	0.001	0.068	0.279
Extra_bowls_bowled	0.0567	0.016	3.560	0.000	0.025	0.088
Min_run_given_1over	0.2330	0.060	3.854	0.000	0.114	0.351
Min_run_scored_1over	0.2621	0.083	3.174	0.002	0.100	0.424
Max_run_given_1over	-0.1418	0.042	-3.416	0.001	-0.223	-0.060
extra_bowls_opponent	0.1444	0.028	5.111	0.000	0.089	0.200
player_highest_run	-0.0007	0.003	-0.254	0.800	-0.006	0.005
Players_scored_zero	0.5206	0.080	6.474	0.000	0.363	0.678
player_highest_wicket	-0.1148	0.185	-0.621	0.535	-0.477	0.247
Match_light_type_Day and Night	-0.6757	0.137	-4.927	0.000	-0.945	-0.407
Match_light_type_Night	0.7951	0.244	3.255	0.001	0.316	1.274
Match_format_T20	0.5067	0.405	1.250	0.211	-0.288	1.301
Match_format_Test	1.3228	1.326	0.997	0.319	-1.277	3.922
First_selection_Bowling	-0.2514	0.122	-2.059	0.040	-0.491	-0.012
Opponent_Bangladesh	2.3718	1.378	1.722	0.085	-0.328	5.072
Opponent_England	2.7566	1.364	2.020	0.043	0.082	5.431
Opponent_Kenya	2.1283	1.338	1.590	0.112	-0.495	4.751
Opponent_Pakistan	2.6804	1.365	1.963	0.050	0.005	5.356
Opponent_South Africa	1.8050	1.364	1.324	0.186	-0.867	4.478
Opponent_Srilanka	1.3367	1.337	1.000	0.317	-1.283	3.956
Opponent_West Indies	3.6085	1.481	2.436	0.015	0.705	6.512
Opponent_Zimbabwe	0.9704	1.367	0.710	0.478	-1.709	3.650
Season_Summer	-0.9185	0.129	-7.129	0.000	-1.171	-0.666
Season_Winter	0.2839	0.172	1.653	0.098	-0.053	0.621
Offshore_Yes	-1.6736	0.124	-13.543	0.000	-1.916	-1.431

- Significant variables are whose p value is less than alpha which is 0.05

1. Avg_team_Age
2. All_rounder_in_team
3. Max_wicket_taken_1over
4. Extra_bowls_bowled

5. Min_run_given_1over
6. Min_run_scored_1over
7. Max_run_given_1over
8. Extra_bowls_opponent
9. Players_scored_zero
10. Match_light_type_Day and Night
11. Match_light_type_Night
12. Season_Summer
13. Offshore_yes.

LOGISTIC REGRESSION

- **EFFORTS TO IMPROVE THE MODEL PERFORMANCE:**

1. Taking threshold as 0.5:

a. If the value is greater than 0.5 it is will be 1 and less than 0.5 it will 0. To increase the precision value, we need to decrease the false positive cases we can do by shifting the threshold point to the right

b. In this case precision value should be high, so we reduce the false positive rate by shifting the threshold value from 0.5 to 0.6.

2. Taking threshold as 0.6:

a. If the value is greater than 0.6 it is will be 1 and less than 0.6 it will 0

b. We can further increase the precision value we shifting the threshold from 0.6 to 0.7.

3. Taking threshold as 0.7

a. We can consider threshold point as 0.7 where we get the maximum precision value which is 0.91 and lower false positive cases.

b. We can use this threshold point for predictions.

TABLE 16: LOGISTIC REGRESSION – CONFUSION MATRIX – TEST DATA

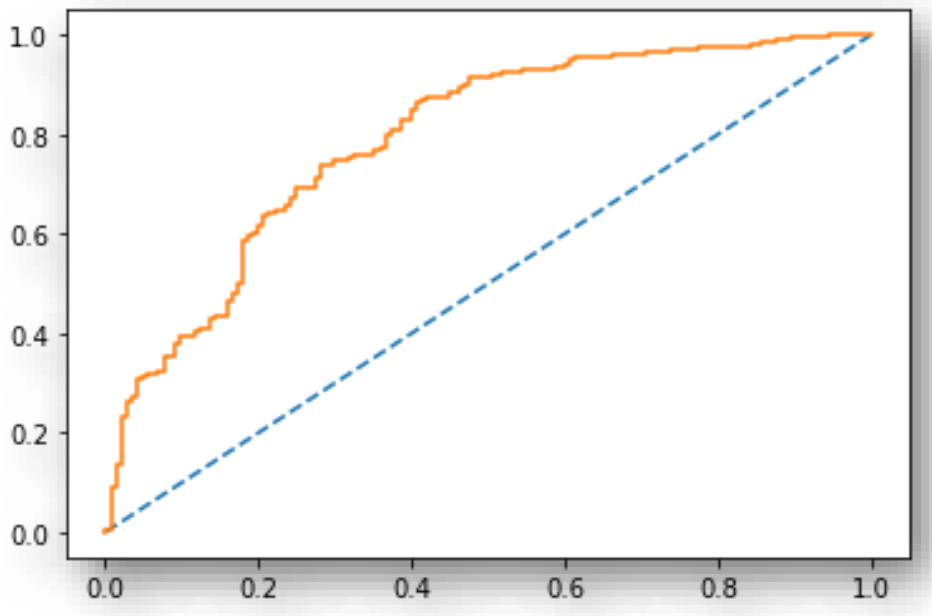
```
[[ 82  63]
 [ 77 657]]
```

TABLE 17: LOGISTIC REGRESSION – CLASSIFICATION REPORT – TEST DATA

	precision	recall	f1-score	support
0	0.52	0.57	0.54	145
1	0.91	0.90	0.90	734
accuracy			0.84	879
macro avg	0.71	0.73	0.72	879
weighted avg	0.85	0.84	0.84	879

FIGURE 23: LOGISTIC REGRESSION – AUC ROC CURVE – TEST DATA

AUC: 0.791



RANDOM FOREST

- **EFFORTS TO IMPROVE THE MODEL PERFORMANCE:**

1. Random forest model uses ensemble technique.
2. Random forest using grid search. Grid search cross validation is a technique to find the best combination of parameters.
3. Getting the best parameters using grid search (Refer Jupyter Notebook).

TABLE 18: RANDOM FOREST– CONFUSION MATRIX – TEST DATA

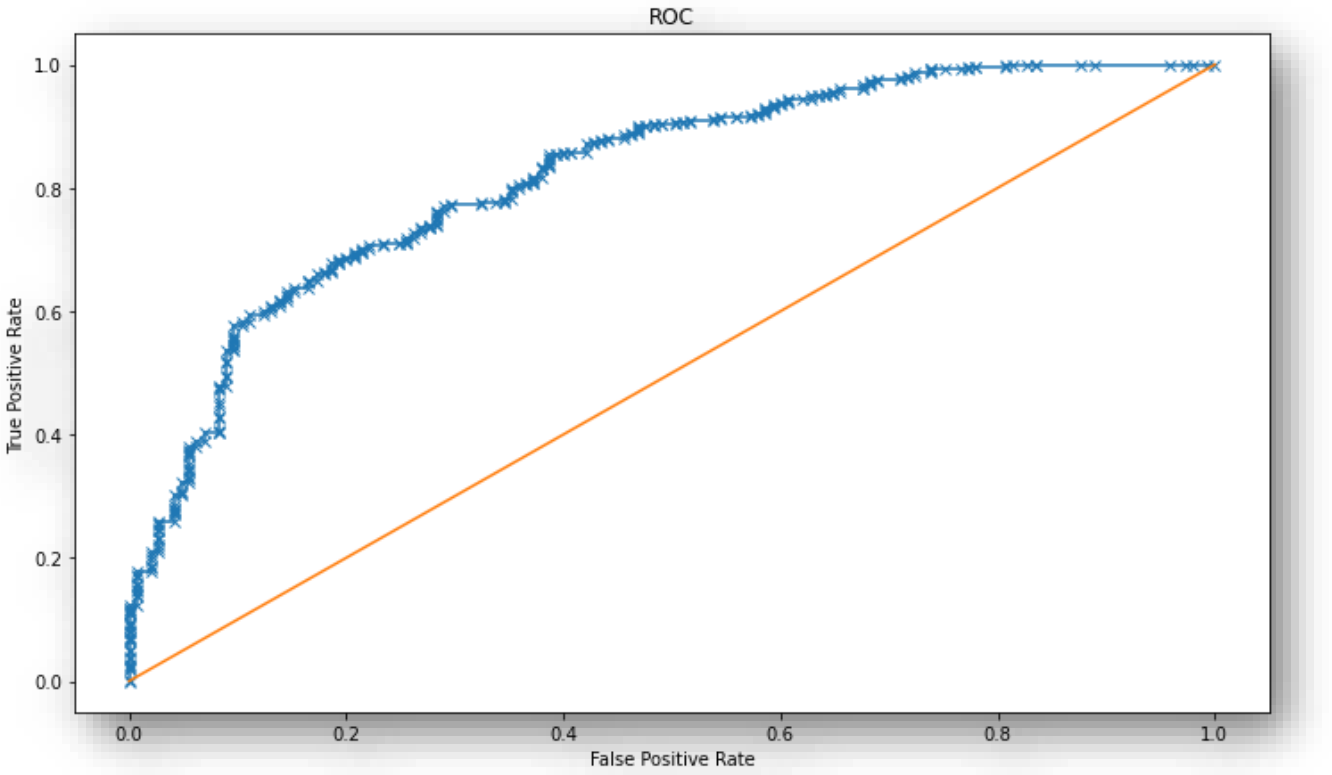
```
array([[ 24, 121],
       [  0, 734]], dtype=int64)
```

TABLE 19: RANDOM FOREST– CLASSIFICATION REPORT – TEST DATA

	precision	recall	f1-score	support
0	1.00	0.17	0.28	145
1	0.86	1.00	0.92	734
accuracy			0.86	879
macro avg	0.93	0.58	0.60	879
weighted avg	0.88	0.86	0.82	879

FIGURE 24: RANDOM FOREST - AUC ROC CURVE – TEST DATA

```
Area under Curve is 0.8216386357230104
```



DECISION TREE

- **EFFORTS TO IMPROVE THE MODEL PERFORMANCE:**

1. Gini index is used to take out the best/important features

TABLE 20: IMPORTANT FEATURES

	imp
player_highest_run	0.135164
Audience_number	0.133344
Extra_bowls_bowled	0.075105
player_highest_wicket	0.069106
Players_scored_zero	0.067666
Offshore	0.062639
Season	0.060780
Max_run_scored_1over	0.048033
Bowlers_in_team	0.047098
Min_run_scored_1over	0.046439
All_rounder_in_team	0.043080
Opponent	0.041045
Max_wicket_taken_1over	0.040904
Min_run_given_1over	0.039963
extra_bowls_opponent	0.037076
Max_run_given_1over	0.020022
Match_light_type	0.018768
Match_format	0.008770
First_selection	0.004997
Avg_team_Age	0.000000

TABLE 21: DESCISION TREE – CONFUSION MATRIX – TEST DATA

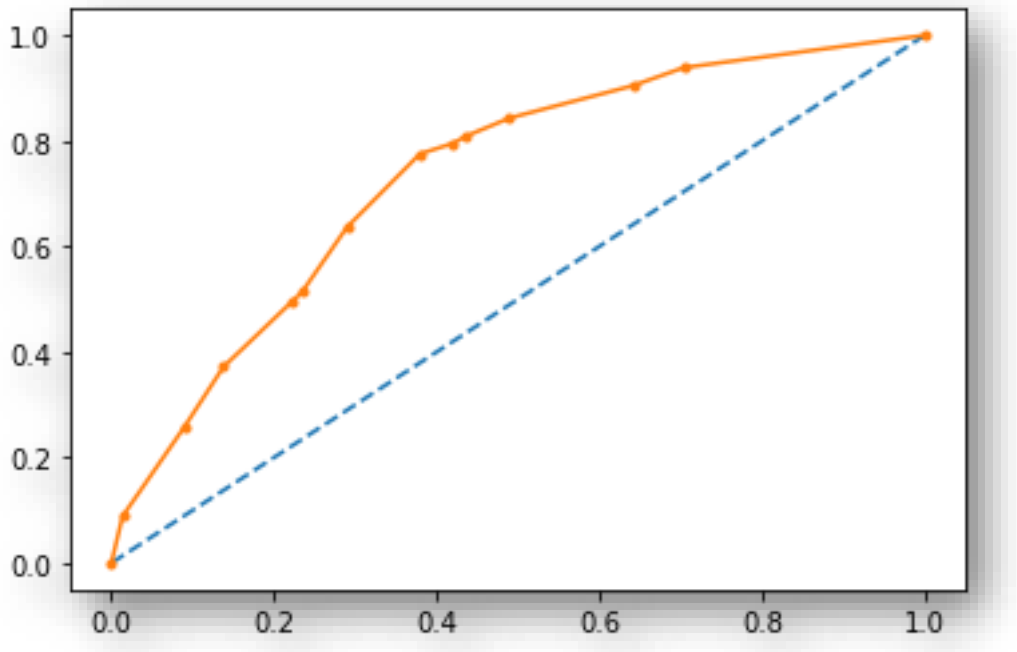
```
array([[ 43, 102],
       [ 45, 689]],
```

TABLE 22: DESCISION TREE – CLASSIFICATION REPORT– TEST DATA

	precision	recall	f1-score	support
0	0.49	0.30	0.37	145
1	0.87	0.94	0.90	734
accuracy			0.83	879
macro avg	0.68	0.62	0.64	879
weighted avg	0.81	0.83	0.82	879

FIGURE 25: DESCISION TREE – AUC ROC CURVE – TEST DATA

AUC: 0.733



ARTIFICIAL NEURAL NETWORK

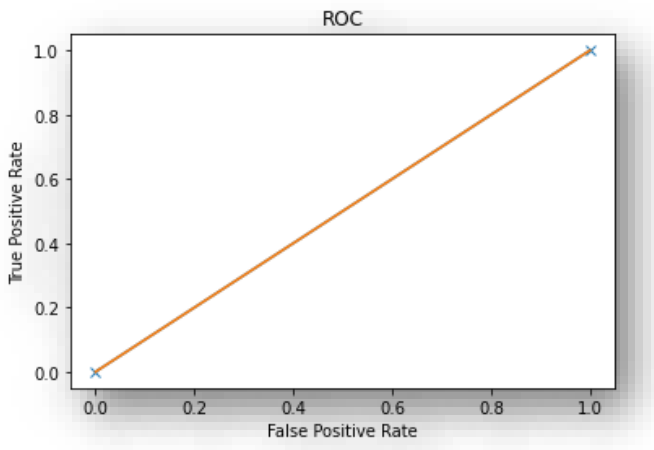
TABLE 23: ANN – CONFUSION MATRIX – TEST DATA

```
array([[  0, 145],
       [  0, 734]],
```

TABLE 24: ANN – CLASSIFICATION REPORT – TEST DATA

	precision	recall	f1-score	support
0	0.00	0.00	0.00	145
1	0.84	1.00	0.91	734
accuracy			0.84	879
macro avg	0.42	0.50	0.46	879
weighted avg	0.70	0.84	0.76	879

FIGURE 26: ANN – AUC ROC CURVE– TEST DATA




```
Area under Curve is 0.5
```

TABLE 25: MODEL COMPARISION

	logistic Train(sm)	logistic Test(sm)	CART train	CART test	\
Accuracy	0.87	0.84	0.85	0.83	
AUC	0.85	0.80	0.82	0.72	
Precision	0.93	0.91	0.88	0.87	
Recall	0.92	0.90	0.94	0.94	
F1 score	0.92	0.90	0.91	0.90	

	RANDOM FOREST train	RANDOM FOREST test	NEURAL NETWORK train	\
Accuracy	0.86	0.86	0.84	
AUC	0.87	0.79	0.50	
Precision	0.87	0.86	0.84	
Recall	0.99	0.99	1.00	
F1 score	0.92	0.92	0.91	

	NEURAL NETWORK test
Accuracy	0.84
AUC	0.50
Precision	0.84
Recall	1.00
F1 score	0.91

MODEL INTERPRETATION

- **LOGISTIC REGRESSION:**

1. Accuracy is 0.87 and 0.84 for train and test set.
2. As false positive cases are not acceptable. Hence precision is important. It has precision 0.93 and 0.91 for train and test set. It is right fit model.
3. Area under curve for train and test is 0.85 and 0.80.

- **CART MODEL:**

1. Accuracy is 0.85 and 0.83 for train and test set.
2. As false positive cases are not acceptable. Hence precision is important. It has precision 0.88 and 0.87 for train and test set. It is right fit model.
3. Area under curve for train and test is 0.82 and 0.72.

- **RANDOM FOREST:**

1. Accuracy is 0.86 and 0.86 for train and test set.
2. As false positive cases are not acceptable. Hence precision is important. It has precision 0.87 and 0.86 for train and test set. It is right fit model.

3. Area under curve for train and test is 0.87 and 0.79.

- **ARTIFICIAL NEURAL NETWORK:**

1. Accuracy is 0.84 and 0.84 for train and test set.
2. As false positive is not acceptable. Hence precision is important. It has precision 0.84 and 0.84 for train and test set. It is right fit model.
3. Area under curve for train and test is 0.5 and 0.5.

- **From above all the models we select logistic regression using stats model. It has highest accuracy and precision value. Logistic regression using stats model gives us approach to make the strategy by making changes to the values of the variables with respect to the coefficient.**

BUSINESS RECOMMENDATION

1. **Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.**

STRATEGIES RECOMMENDED:

a. **Average team age:** Team average age should not be above 34. above age 34 we may lose the match.
b. **All-rounder's in team**: There should be at least 3 all-rounders' in team.
c. **First selection**: The first selection should be bowling.
d. **Bowlers in team:** There should be at least one bowler in the team.
e. **Players scored zero:** There should be no player scored zero.

2. **T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.**

1ST STRATEGY RECOMMENDED:

a. **Average team age:** Team average age should be 31. If it is greater 31 then we may lose the match.
b. **All-rounder's in team:** There should be at least 3 all-rounders' in team, less than 3 all-rounders may lose the match.
c. **First selection:** The first selection should be bowling.
d. **Bowlers in team:** There should be 3 bowlers in team.
e. **Extra bowls opponents:** It should be greater than 14. if it is 14 we may lose the match 31.
f. **Maximum runs given in one over**: It should be less than 9 runs. If it 9 or greater then we may lose the match.
g. **Player scored zero:** There should be at least 2 players scored zero.

2nd STRATEGY RECOMMENDED:

a. **Average team age:** Team average age should be 32. If it is greater than 32 we may lose the match.
b. **All-rounder's in team:** There should be at least one all-rounder's in team. Playing with no all-rounders can lose the match.
c. **First selection:** The first selection should be batting.
d. **Bowlers in team:** There should be 2 bowlers in team.
e. **Extra bowls opponents:** It should be greater than 14. If it is 14 we may lose the match.
f. **Maximum runs given in one over**: It should not be greater than 4. If it is greater than 4 then we may lose the match.
g. **Player scored zero:** There should be at least 3 players scored zero.

3. ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

1st STRATEGY RECOMMENDED:

a. Average team age: Team average age should be less than 36. if it is 36 we may lose the match
b. All-rounder's in team: There should be at least 2 all-rounders' in team.
c. First selection: The first selection should be bowling.
d. Bowlers in team: There should be 2 bowlers in team.
e. Maximum runs given in one over: It should be less than 13. If it is 13 or greater than that we may lose the match.
f. Extra bowls opponents: It should be greater than 4. If it is 5 we may lose the match.
g. Player scored zero: We should be having at least two players scored zero. With one player scored zero we may lose the match.

2nd STRATEGY RECOMMENDED:

a. Average team age: Team average age should be 34. If it is greater than 34 we may lose the match.
b. All-rounder's in team: There should be at least 3 all-rounders' in team.
c. First selection: The first selection should be batting.
d. Bowlers in team: There should be 3 bowlers in team.
e. Maximum runs given in one over: It should be 23. more than 23 runs will result into lost.
f. Extra bowls opponents: It should be greater than 19. If it is 19 we may lose the match.
g. Player scored zero: We should be having at least one player scored zero. With no players scored zero we may lose the match.

>>THE END<<