
# 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


| QUESTIONS / PROCESS | CONTENT | PAGE # |
|---|---|---|
| **INTRODUCTION** | PROBLEM STATEMENT | 3 |
| | DATA DICTIONARY | 3 |
| | NEED TO SOLVE THE PROBLEM | 4 |
| **DATA CLEANING & PRE-PROCESSING** | VARIABLE MODIFICATION | 9 |
| | MISSING VALUE TREATMENT | 11 |
| | OUTLIER TREATMENT | 12 |
| **EXPLORATORY DATA ANALYSIS** | UNIVARIATE AND BI-VARIATE ANALYSIS | 15 |
| | MULTIVARIATE ANALYSIS | 24 |
| **MODEL BUILDING** | ENCODING THE DATASET | 25 |
| | SPLITTING THE DATASET | 25 |
| | FEATURE SELECTION | 25 |
| **MODEL VALIDATION** | ALL MODEL COMPARISION AND INSIGHTS | 26 |
| **INTERPRETATION** | BUSSINESS RECOMMENDATION | 27 |


## LIST OF FIGURES


| FIGURE # | FIGURE NAME | PAGE # |
|---|---|---|
| **1** | BOXPLOT | 12 |
| **2** | MATCH LIGHT | 14 |
| **3** | MATCH FORMAT | 14 |
| **4** | FIRST SELECTION | 14 |
| **5** | OPPONENT | 15 |
| **6** | SEASON | 15 |
| **7** | OFFSHORE | 16 |
| **8** | SCATTER PLOT | 17 |
| **9** | BOXPLOT – RESULT AND EXTRA BOWLS OPPONENT | 18 |
| **10** | BOXPLOT – RESULT AND EXTRA BOWLS BOWLED | 18 |
| **11** | BARPLOT – RESULT AND PLAYER HIGHEST WICKET | 19 |
| **12** | BARPLOT – RESULT AND AVG TEAM AGE | 19 |
| **13** | BARPLOT – RESULT & OFFSHORE | 20 |
| **14** | BARPLOT – RESULT & FIRST SELECTION | 20 |
| **15** | BARPLOT – RESULT & ALL ROUNDER IN TEAM | 21 |
| **16** | BARPLOT – RESULT & OPPONENT IN TEAM | 21 |
| **17** | BARPLOT – RESULT & MATCH FORMAT IN TEAM | 22 |
| **18** | BARPLOT – RESULT & PLAYER SCORED ZERO | 22 |
| **19** | HEATMAP | 23 |

| LIST OF TABLES | | |
|---|---|---|
| **TABLE #** | **TABLE NAME** | **PAGE #** |
| **1** | TOP 5 DATA SAMPLE | 4 |
| **2** | LAST 5 DATA SAMPLE | 5 |
| **3** | SHAPE OF THE DATASET | 6 |
| **4** | DATASET INFORMATION | 6 |
| **5** | DATASET DESCRIPTION | 7 |
| **6** | DATASET DUPLICATES | 7 |
| **7** | MISSING DATA | 8 |
| **8** | UNIQUE COUNT | 9 |
| **9** | DATA INFORMATION AFTER MODIFICATION | 10 |
| **10** | REPLACING THE MISSING VALUES | 11 |
| **11** | SKEWNESS | 13 |
| **12** | MATCH LIGHT | 14 |
| **13** | MATCH FORMAT | 14 |
| **14** | FIRST SELECTION | 14 |
| **15** | OPPONENT | 15 |
| **16** | SEASON | 15 |
| **17** | OFFSHORE | 16 |
| **18** | PLAYERS SCORED ZERO | 16 |
| **19** | PLAYER HIGHEST WICKET | 16 |
| **20** | ALL MODEL COMPARISION | 25 |

# 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.

# 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 |

# NEED TO SOLVE THE PROBLEM

1. To increase the chances of winning
2. Help in selecting the right team for the right match
3. It also helps the management to know which conditions are favourable for the team, so that the chances of winning are high.

# 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’ (The original name can be understood as a numerical datatype by the system automatically, this helps in understanding the variable precisely and easily)
2. Bat: ‘Batting’ (Since the major data values are “batting”, to get accurate results we are converting the same.
3. Three: ‘3’ in both 'Players_scored_zero' and 'player_highest_wicket' columns (Since the major data values are numerical “3”, to get accurate results we are converting the same.)

- **Modified names in the dataset:**

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

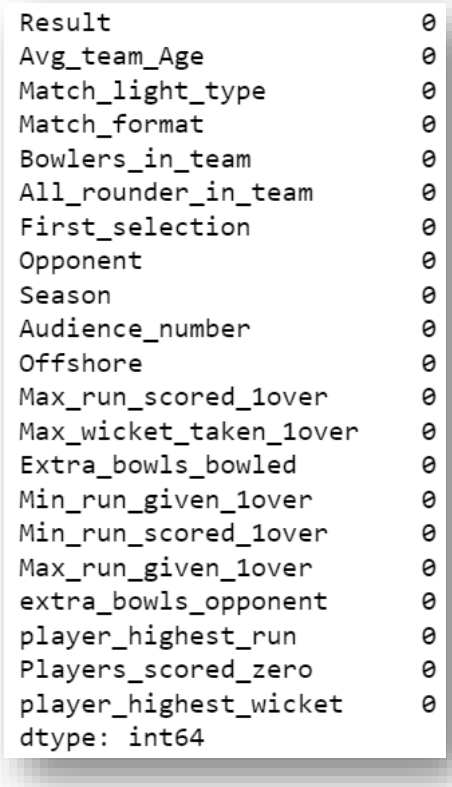
- **Modifying few data types:**

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)
```

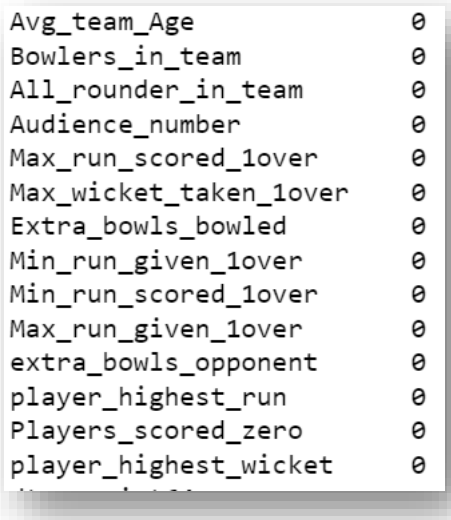
# TABLE 10: REPLACING THE MISSING VALUES



```
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
```

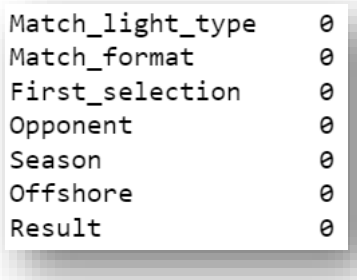
- The missing data in the dataset has been replaced as per the datatype using the median and mode. Hence There are no missing values in the dataset.

Numerical data



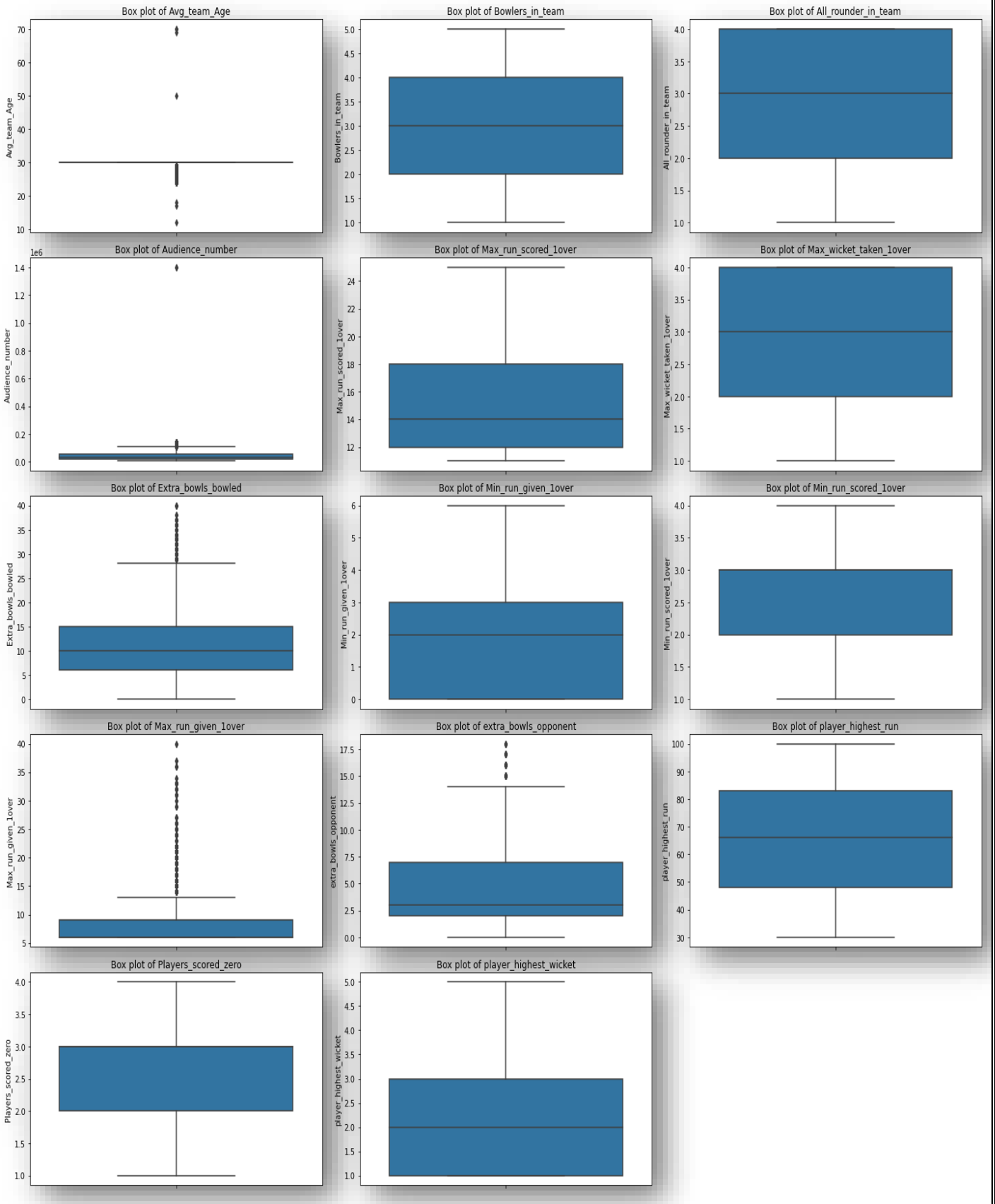
```
Avg_team_Age              0
Bowlers_in_team           0
All_rounder_in_team       0
Audience_number           0
Max_run_scored_1over      0
Max_wicket_taken_1over    0
Extra_bowls_bowled        0
Min_run_given_1over       0
Min_run_scored_1over      0
Max_run_given_1over       0
extra_bowls_opponent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

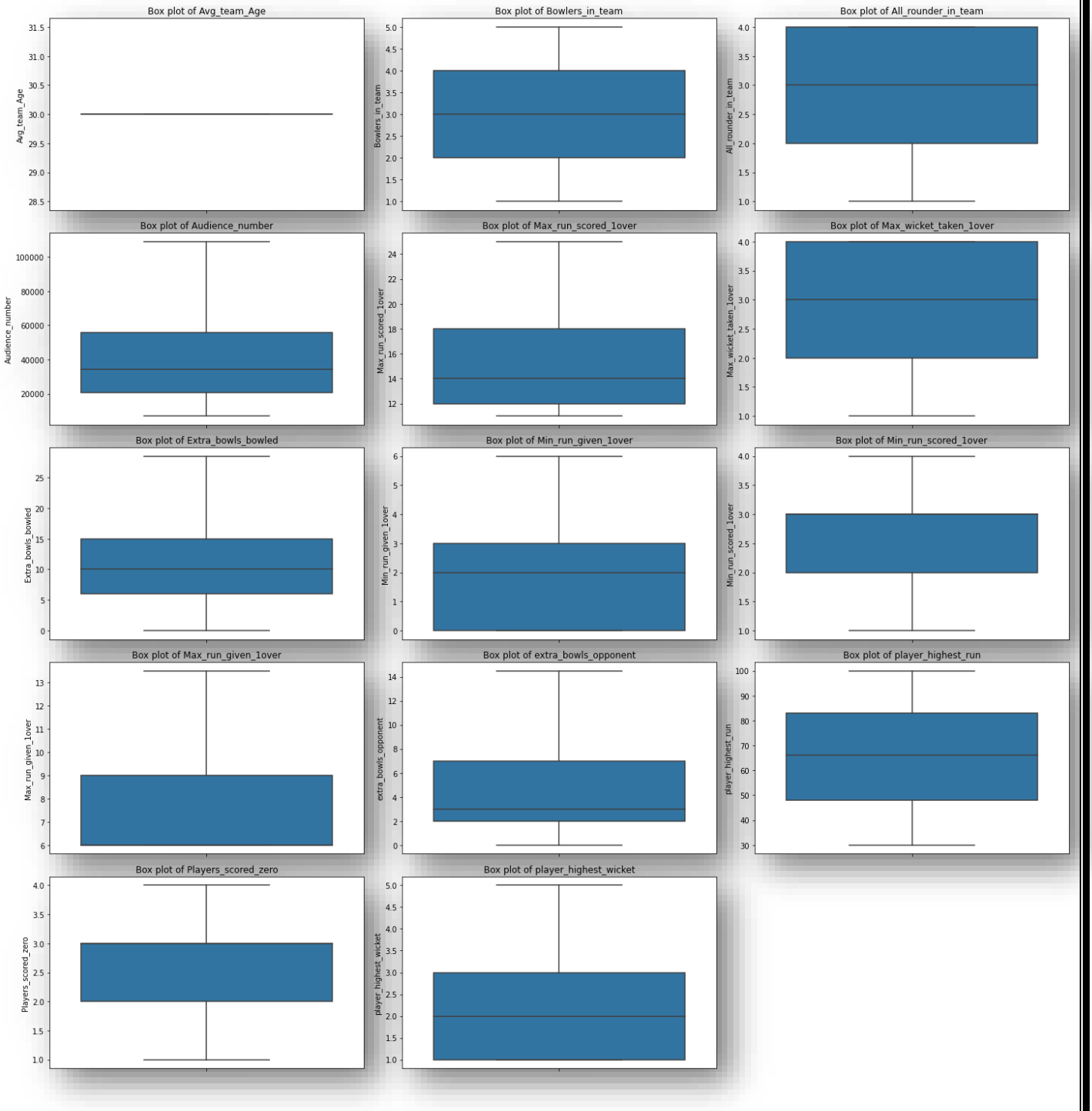


1. Variable ‘avg_team_age’ is close to normal distribution as mean and median are similar though it has outliers.
2. Variable ‘Bowlers_in_team’ is normally distributed as mean and median are almost same, no outliers are present.
3. Variable ‘All_rounder_in_team’ is slightly left skewed as there is difference 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 |

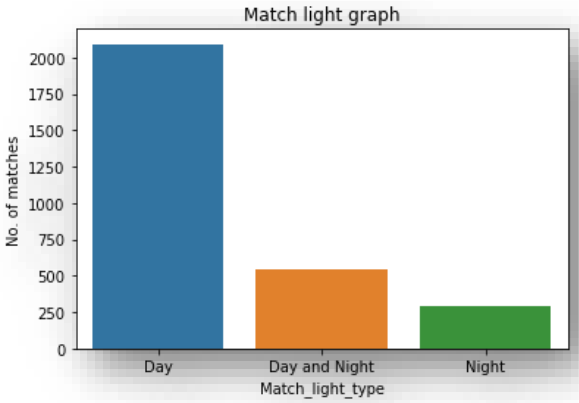
# FIGURE 22: OUTLIER TREATMENT




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.

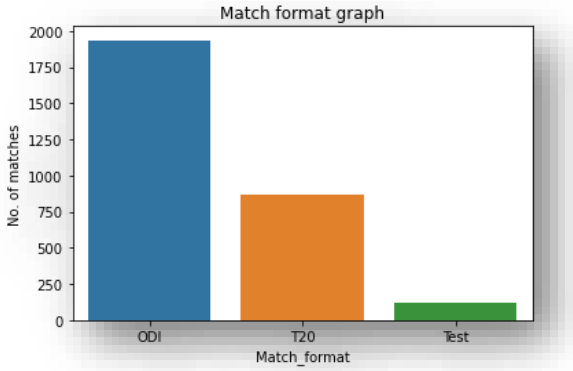
# TABLE 12 & FIGURE 2: MATCH LIGHT

| Result | Loss | Win |
|---|---|---|
| Match_light_type | | |
| Day | 314 | 1779 |
| Day and Night | 135 | 406 |
| Night | 24 | 272 |



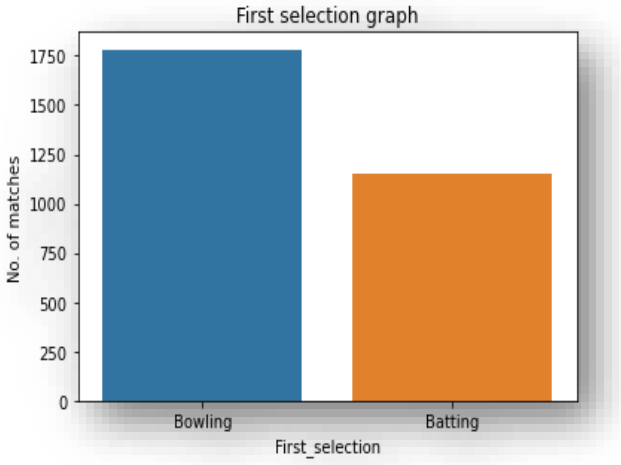

# TABLE 13 & FIGURE 3: MATCH FORMAT

| Result | Loss | Win |
|---|---|---|
| Match_format | | |
| ODI | 269 | 1666 |
| T20 | 180 | 690 |
| Test | 24 | 101 |



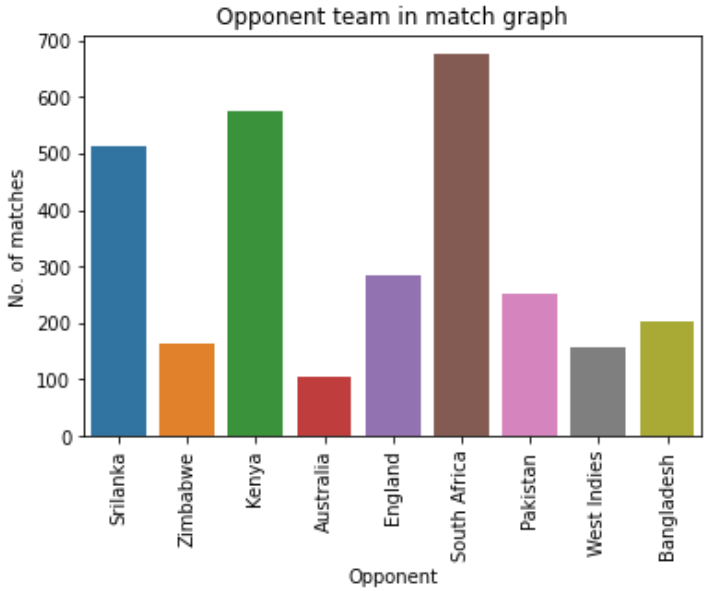

# TABLE 14 & FIGURE 4: FIRST SELECTION

| Result | Loss | Win |
|---|---|---|
| First_selection | | |
| Batting | 172 | 977 |
| Bowling | 301 | 1480 |



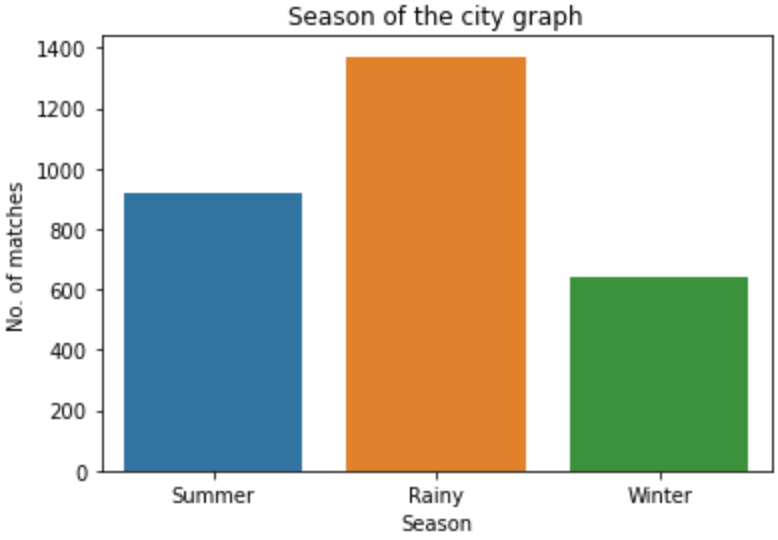
# TABLE 15 & FIGURE 5: OPPONENT

| Result<br>Opponent | Loss | Win |
|---|---|---|
| Australia | 24 | 80 |
| Bangladesh | 10 | 194 |
| England | 18 | 265 |
| Kenya | 93 | 483 |
| Pakistan | 17 | 236 |
| South Africa | 117 | 559 |
| Srilanka | 124 | 389 |
| West Indies | 4 | 154 |
| Zimbabwe | 66 | 97 |



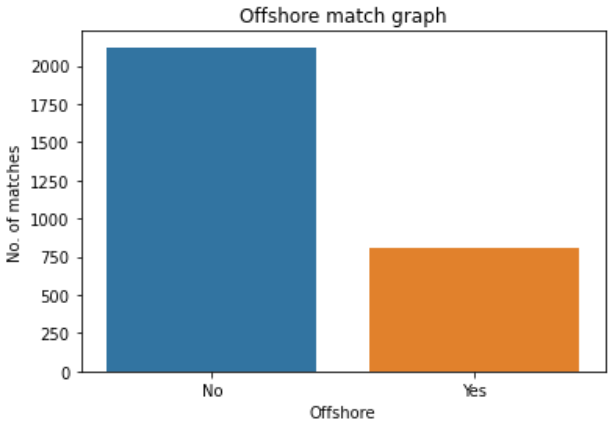

# TABLE 16 & FIGURE 6: SEASON

| Result<br>Season | Loss | Win |
|---|---|---|
| Rainy | 170 | 1201 |
| Summer | 238 | 680 |
| Winter | 65 | 576 |



# TABLE 17 & FIGURE 7: OFFSHORE

| Result | Loss | Win |
|---|---|---|
| Offshore | | |
| No | 227 | 1894 |
| Yes | 246 | 563 |




# TABLE 18: PLAYERS SCORED ZERO

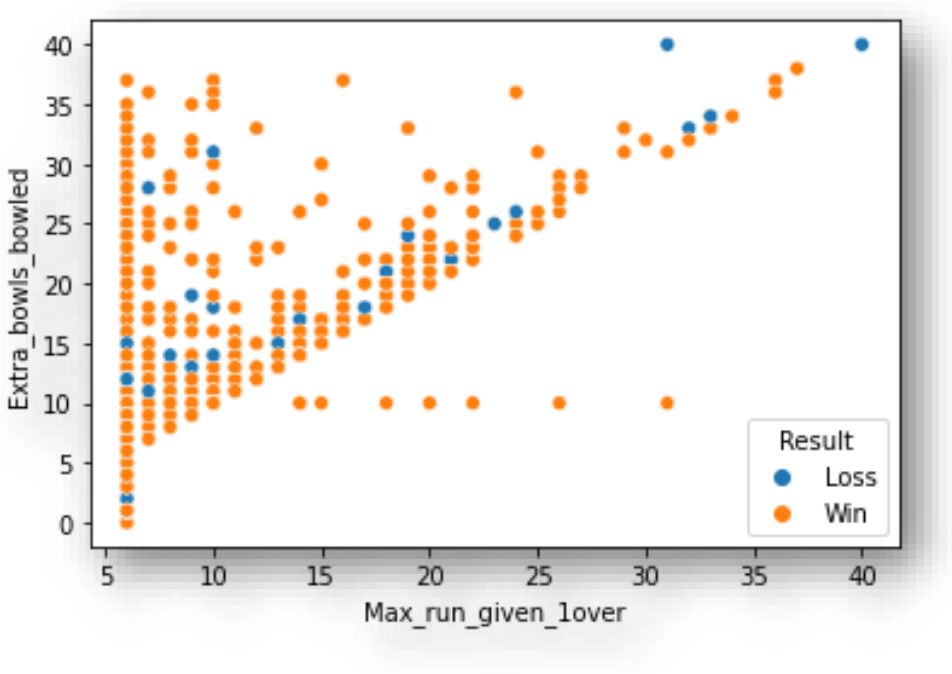
| Result | Loss | Win |
|---|---|---|
| Players_scored_zero | | |
| 1 | 56 | 110 |
| 2 | 141 | 603 |
| 3 | 250 | 1485 |
| 4 | 26 | 259 |

# TABLE 19: PLAYER HIGHEST WICKET

| Result | Loss | Win |
|---|---|---|
| player_highest_wicket | | |
| 1 | 286 | 798 |
| 2 | 104 | 959 |
| 3 | 63 | 371 |
| 4 | 10 | 201 |
| 5 | 10 | 128 |

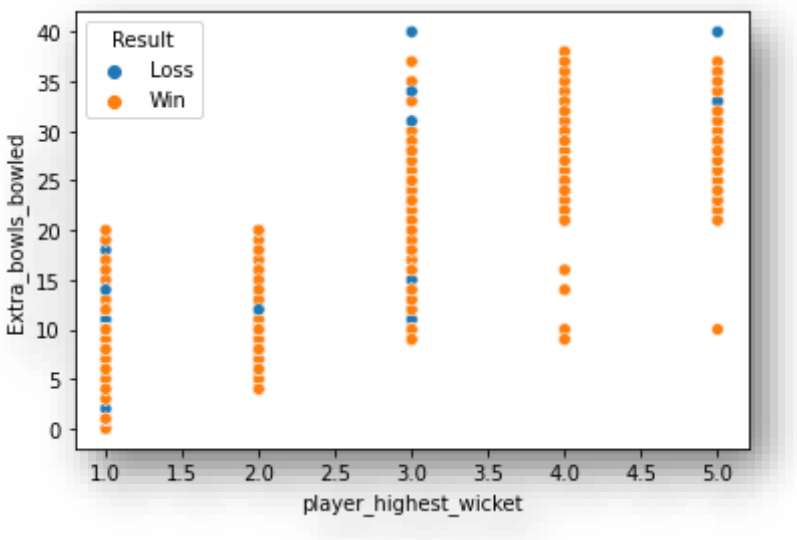
# FIGURE 8: SCATTER PLOT

- **MAX_RUN_GIVEN_1OVER & EXTRA_BOWLS_BOWLED**



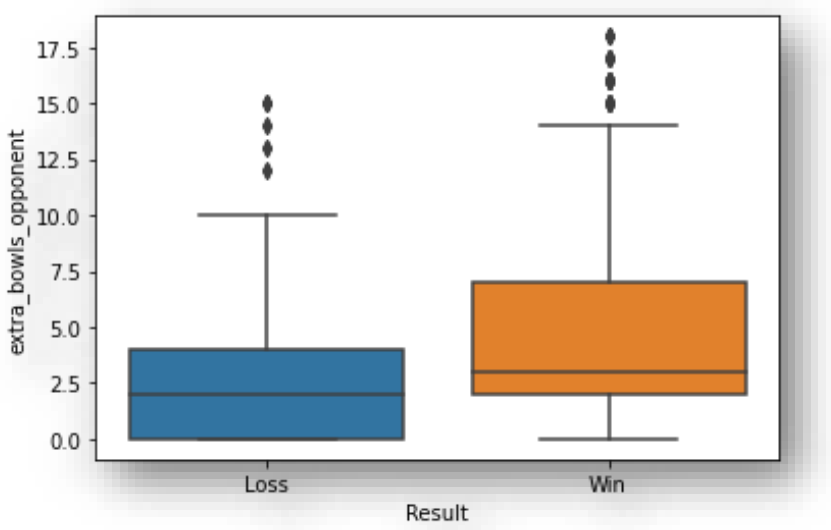

- It shows the linear relationship between maximum run given in one over and extra bowls bowled. Highest value of both the variables results in loss.

- **PLAYER_HIGHEST_WICKET & EXTRA_BOWLS_BOWLED**



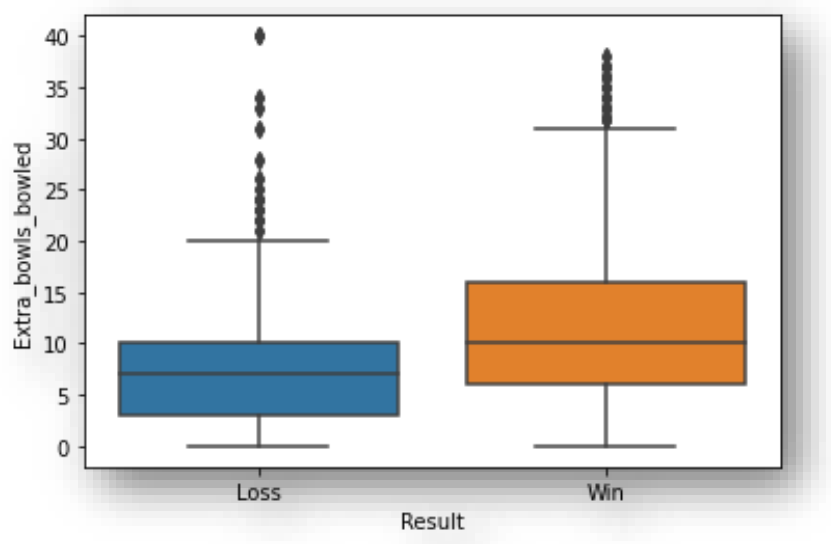

- The highest results also end up in a loss.

# FIGURE 9: BOXPLOT – RESULT AND EXTRA BOWLS OPPONENT



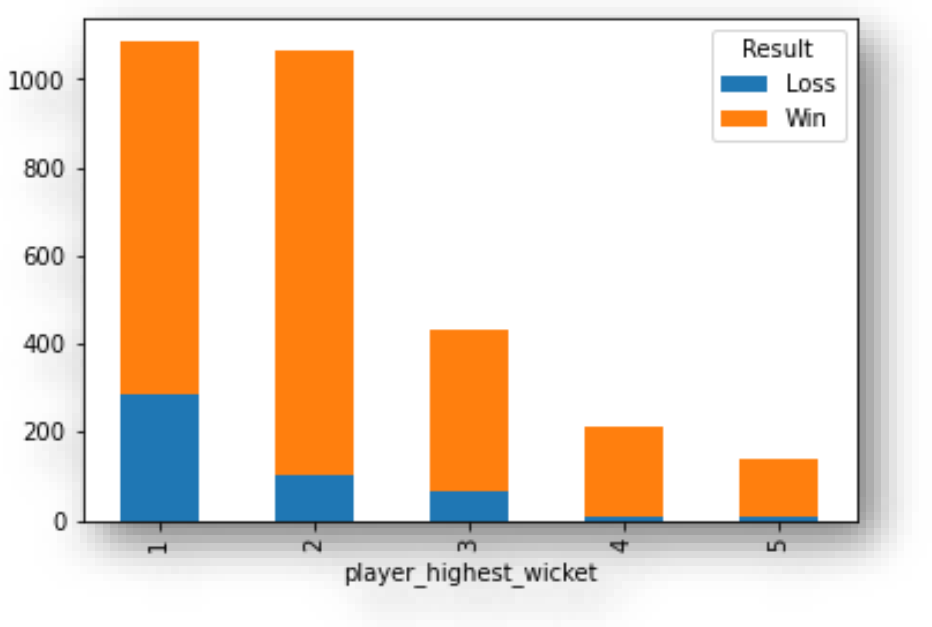

- The median of extra bowls opponents is higher for win, If the extra bowls opponents is higher than 16 then India will win the match as per the data set.
- If the extra bowls opponents are higher than 10 then chances of winning the match is more.

# FIGURE 10: BOXPLOT – RESULT AND EXTRA BOWLS BOWLED



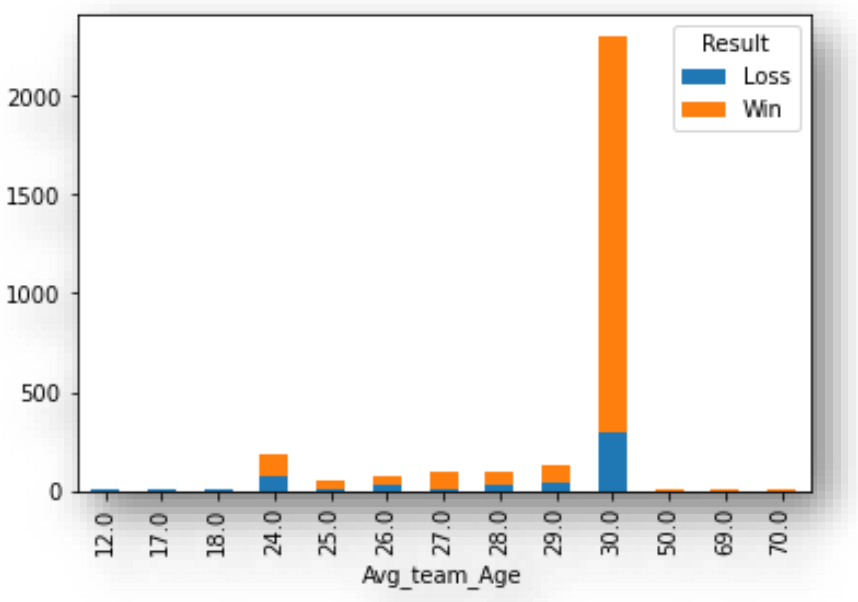

- As per data when you bowl 40 extra bowls India is definitely will lose the match.

# FIGURE 11: BARPLOT – RESULT AND PLAYER HIGHEST WICKET



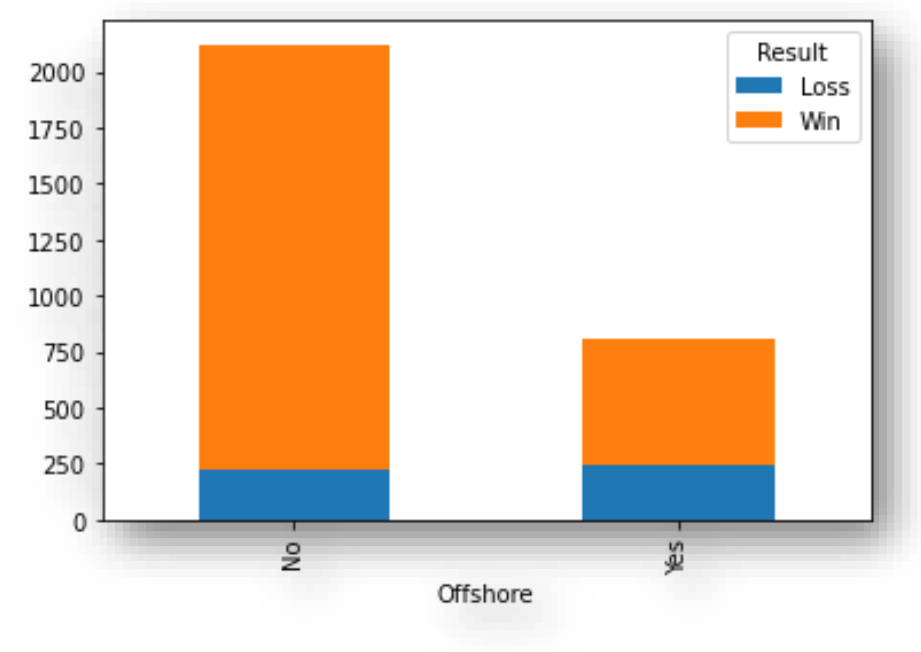

- The chances of winning the match are high when the 2 wickets are taken by a single player

# FIGURE 12: BARPLOT – RESULT AND AVG TEAM AGE



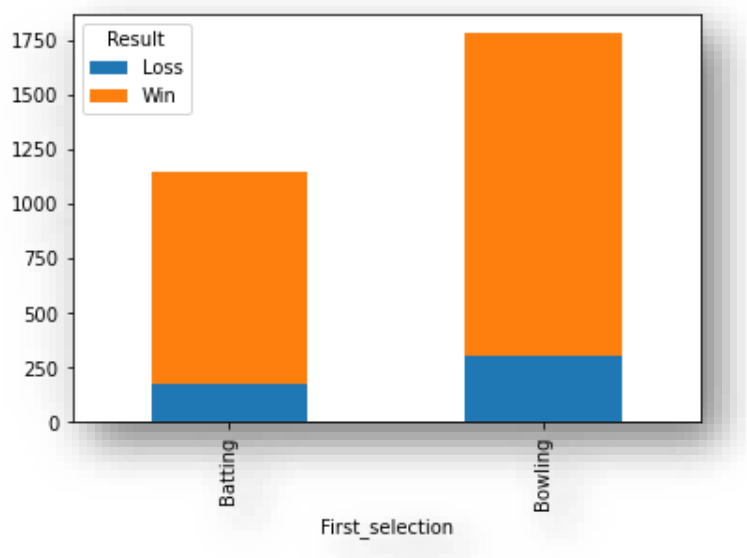

- India has the highest win with average team age at 30.

# FIGURE 13: BARPLOT – RESULT & OFFSHORE



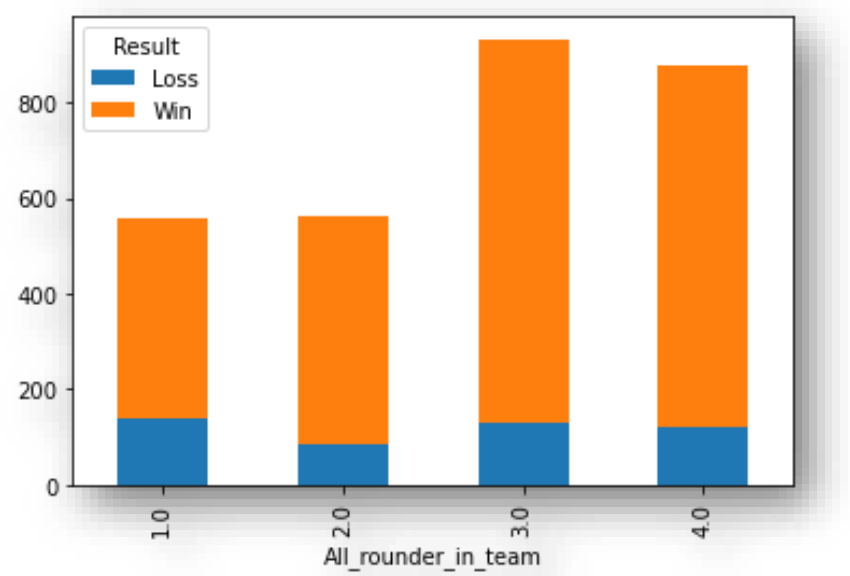

- The performance of the Indian team is good when played on the home ground.

# FIGURE 14: BARPLOT – RESULT & FIRST SELECTION



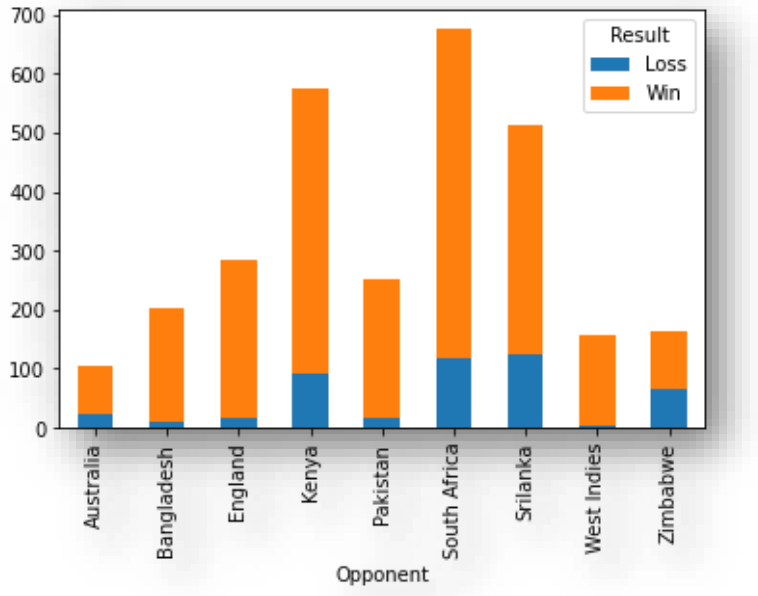

- Most of the matches are won when selected to bowl first.

# FIGURE 15: BARPLOT – RESULT & ALL ROUNDER IN TEAM



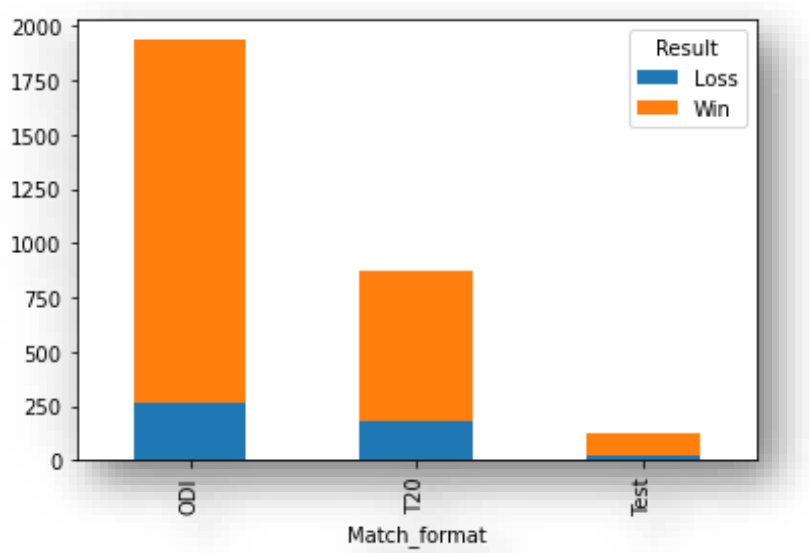

- Winning percentage is higher when matches are played with 3 to 4 all-rounders in the team.

# FIGURE 16: BARPLOT – RESULT & OPPONENT IN TEAM



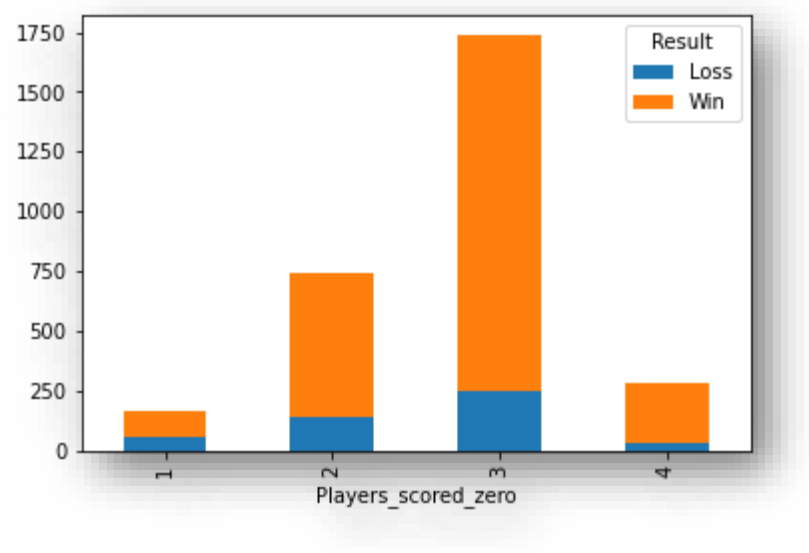

- India team is performing well against West Indies, Bangladesh, England and Pakistan India team wining rate is less against South Africa, Sri Lanka, Zimbabwe and Australia.

# FIGURE 17: BARPLOT – RESULT & MATCH FORMAT IN TEAM



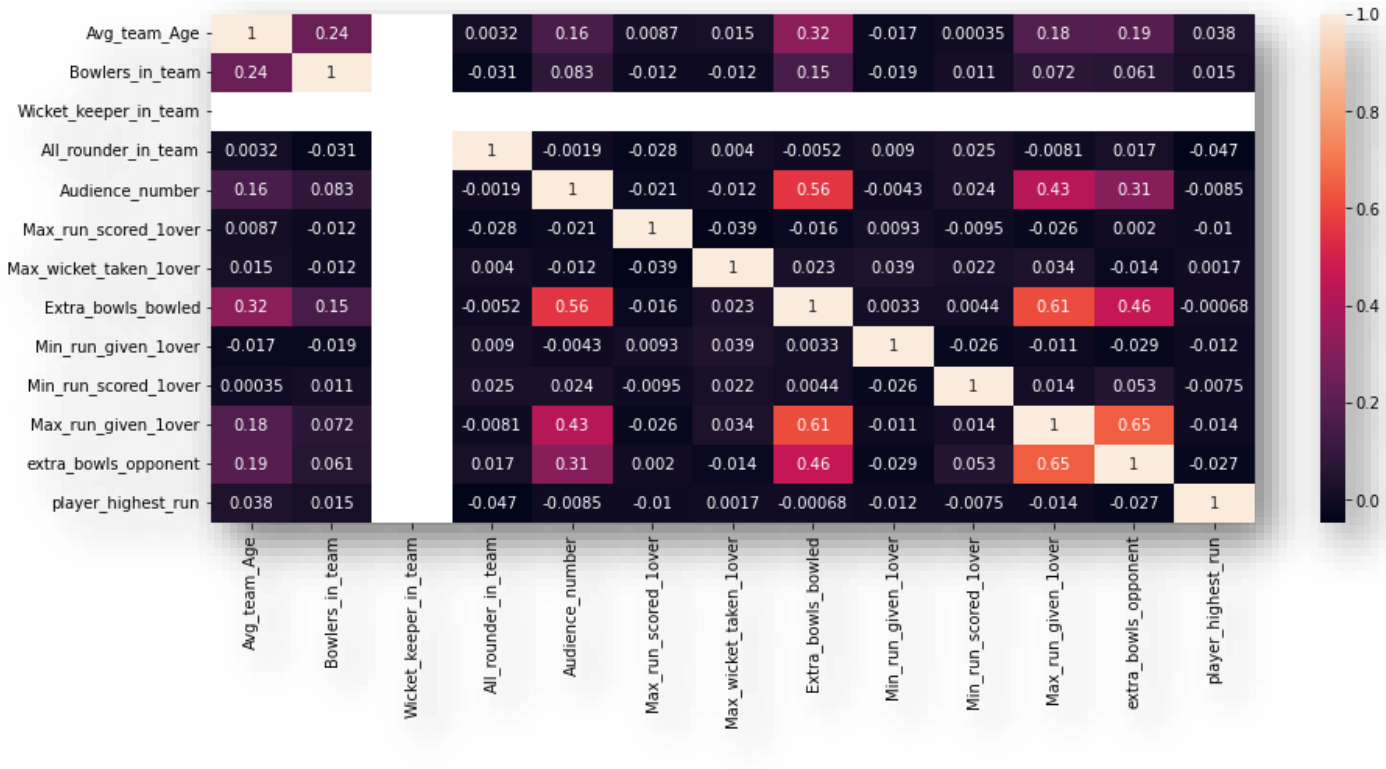

- Indian team has a good performance in ODI format as compared to both the formats Winning rate is higher in ODI format and lesser in T20 and test.

# FIGURE 18: 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.

# FIGURE 19: HEATMAP




# INSIGHTS ON EDA

- Around 72% of matches are played in India and only 28% are played out off India.
- Most of the matches are played in Rainy Season.
- Majority of the matches are played against South Africa (676).
- Around 71% ODI Matches are played in Day light.
- Team have win around 83% of the matches.
- 60% of the time team gets chance to bat first.
- Most of the matches are played against south Africa in which 531 times India secure to win.
- 1781-time team have opted to bowl first out of which 1480 have win the match.
- On average, 19% of the time when playing outside the country, and 65% when playing within the country, team manage to win.
- Extra_bowls_opponent has help team to win the matches.
- Inexperienced team (young player) has the higher chances to lose the match.
- We can see data is imbalance we have to do smote to resolve this issue. (EFFECT OF EDA ON THE PROCESS)
- During the bowling contest, the team won 51% Matches while Batting 33%.
- Team has won 57% of ODI matches and 24% of T20 matches.

- 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
- Variable extra bowl's opponent can add value to the prediction. More extra bowls opponent more is chance of the winning
- Variable offshore also predict the win. If the number of matches played on home ground are more chances of winning is high
- Variable first selection also impact on the result. If the first selection is bowling the chances of winning is high
- Variable all-rounder in team has impact on the result. Having 3 to 4 all-rounders in the team may result into win.

# TABLE 20: ENCODING THE DATASET

- As many machine learning models cannot work with string values we will encode the categorical variables and convert their datatypes to integer type.
- For variable like 'Result', 'Offshore' and 'First_selection' I have used simple categorical conversion technique This will convert the values into 0 and 1. As there is no level or order in the subcategory any encoding will give the same result.
- For remaining variable we have used dummies encoding technique.
- We can see the value count of this variable after encoding as below.

# SPLITTING THE DATASET

- Our target variable is 'Result. As we can see that there is a data imbalance in the variable.
- So, here I will be using the oversampling technique (i.e., SMOTE) and check whether it improves our model performance or not.
- We have stored all the predictor in x and target variable in y.

# FEATURE SELECTION

- Chi-square test is used to determine the relationship between the predictor and target variable.
- In Feature selection, we aim to select the features which are highly dependent on the target variable.
- Higher the chi-square value indicate that the feature is more dependent on the target variable and can be select for model training.

- Chi-square score for Game number is null. So, we eliminate non-significant variable Game number.
- After second iteration we find Wicket keeper as non-significant variable as per chi-square test and same we can in the heat map. So, both the variable has been eliminated to train our model with remaining predictor.

# MODEL SELECTION

- For this classification the following models have been used:
    i. Logistic Regression
    ii. Linear Discriminant Analysis (LDA)
    iii. KNN
    iv. Decision Tree (CART)
    v. Random Forest
    vi. Naïve Bayes
    vii. ANN.

# TABLE 20: ALL MODEL COMPARISION

| | LR-TRAIN | LR-TEST | LR(Tune)-TRAIN | LR(Tune)-TEST | LDA-Train | LDA-Test | LDA(Tune)-Train | LDA(Tune)-Test | KNN-Train | KNN-Test | NB-Train | NB-Test |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 84 | 84 | 88 | 87 | 87 | 87 | 87 | 87 | 90 | 87 | 76 | 78 |
| **F1 Score** | 48 | 47 | 70 | 70 | 70 | 71 | 70 | 70 | 77 | 71 | 62 | 66 |
| **Recall** | 51 | 51 | 66 | 66 | 66 | 67 | 66 | 66 | 72 | 68 | 65 | 70 |
| **Precision** | 78 | 92 | 83 | 79 | 80 | 79 | 80 | 78 | 88 | 79 | 61 | 65 |
| **AUC** | 73 | 73 | 83 | 82 | 82 | 83 | 82 | 83 | 95 | 83 | 76 | 78 |

| | CART-TR | CART-TEST | RF-TRAIN | RF-TEST | ANN-rain | ANN-Test | Bagging-Train | Bagging-Test | Ada-train | Ada-Test | GB-Train | GB-Test | Smote-rf | Smote-rf-test |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 86 | 85 | 87 | 86 | 84 | 84 | 88 | 87 | 88 | 87 | 93 | 90 | 91 | 83 |
| **F1 Score** | 60 | 57 | 63 | 59 | 46 | 46 | 70 | 70 | 71 | 70 | 85 | 76 | 91 | 70 |
| **Recall** | 58 | 56 | 60 | 58 | 50 | 50 | 66 | 66 | 67 | 66 | 80 | 70 | 91 | 71 |
| **Precision** | 80 | 81 | 91 | 91 | 42 | 42 | 83 | 79 | 82 | 82 | 94 | 89 | 91 | 70 |
| **AUC** | 83 | 82 | 92 | 89 | 76 | 75 | 90 | 87 | 87 | 85 | 95 | 89 | 96 | 83 |

1. In this classification problem the most important measurement matrix we see is Recall, precision, accuracy, and F1-Score.

2. In this case, precision is the total predicted win and loss. Recall is total Actually win and loss.
3. F1- score is the harmonic mean of precision and recall.
4. In this case our most important matrix is Recall because we must predict winning for the Indian team and must reduce the false positive rate.
5. **Comparing all models, going with 'Gradient Boosting Model' for this Case study.**
6. This model is just like the ADA Boosting model. Gradient Boosting works by sequentially adding the misidentified predictors and under-fitted predictions to the ensemble, ensuring the errors identified previously are corrected.
7. This method tries to fit the new predictor to the residual errors made by the previous one.
8. Gradient Boost Model have less False '+ve' and False '-ve' for both win and loss Classes. Compare to other model it has Higher Precision, Recall and Accuracy for both Train and Test.

# BUSINESS RECOMMENDATION

1. Try to collect more some more predictor, like total score, bowling style etc. for better Model.
2. Try to add more than 3 all-rounders in the team that will improve the team performance.
3. **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.

4. **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 is 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.

5. **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.

6. **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 <<