SAMPLE BUSINESS REPORT

Project Report

Contents

1 Exploratory Data Analysis

1.1 Introduction of the business problem

The major objective of this data set 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.

2: Structure of Data

2.1 Data Description

The data belongs to the team configuration and various performance aspects shown by the team in the past matched against different opposing teams. Certain weak attributes of the opposing teams are also captured.

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 keeper has been player in the team
All_rounder_in_team	how many full time all rounder 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 audience 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 player out on zero run
player_highest_wicket	Highest wickets taken by single player in match

2.2 Visual inspection of data (rows, columns, descriptive details)

```
#   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Game_number             2930 non-null   object
 1   Result                  2930 non-null   object
```

```
 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
The number of rows (observations) is 2930
The number of columns (variables) is 23
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
All_rounder_in_team	2890.00	NaN	NaN	NaN	2.72	1.09	1.00	2.00	3.00	4.00	4.00
Audience_number	2849.00	NaN	NaN	NaN	46267.96	48599.58	7063.00	20363.00	34349.00	57876.00	1399930.00
Avg_team_Age	2833.00	NaN	NaN	NaN	29.24	2.26	12.00	30.00	30.00	30.00	70.00
Bowlers_in_team	2848.00	NaN	NaN	NaN	2.91	1.02	1.00	2.00	3.00	4.00	5.00
Extra_bowls_bowled	2901.00	NaN	NaN	NaN	11.25	7.78	0.00	6.00	10.00	15.00	40.00
extra_bowls_opponent	2930.00	NaN	NaN	NaN	4.23	3.63	0.00	2.00	3.00	7.00	18.00
First_selection	2871	3	Bowling	1722	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Match_format	2860	4	ODI	1865	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Match_light_type	2878	3	Day	2041	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Max_run_given_1over	2896.00	NaN	NaN	NaN	8.67	5.00	6.00	6.00	6.00	9.25	40.00
Max_run_scored_1over	2902.00	NaN	NaN	NaN	15.20	3.66	11.00	12.00	14.00	18.00	25.00
Max_wicket_taken_1over	2930.00	NaN	NaN	NaN	2.71	1.08	1.00	2.00	3.00	4.00	4.00
Min_run_given_1over	2930.00	NaN	NaN	NaN	1.95	1.68	0.00	0.00	2.00	3.00	6.00
Min_run_scored_1over	2903.00	NaN	NaN	NaN	2.76	0.71	1.00	2.00	3.00	3.00	4.00
Offshore	2866	2	No	2057	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Opponent	2894	9	South Africa	640	NaN	NaN	NaN	NaN	NaN	NaN	NaN
player_highest_run	2902.00	NaN	NaN	NaN	65.89	20.33	30.00	48.00	66.00	84.00	100.00
player_highest_wicket	2930.00	6.00	1.00	1084.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Players_scored_zero	2930.00	5.00	3.00	1730.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Result	2930	2	Win	2457	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Season	2868	3	Rainy	1309	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2.3 Understanding of attributes (variable info, renaming if required)

	Column	Count	Dtype	Remarks
1	Game_number	2930	object	Redundant Column and can be removed
2	Result	2930	object	Categorical and Target Variable
3	Avg_team_Age	2833	float64	Numeric
4	Match_light_type	2878	object	Categorical
5	Match_format	2860	object	Categorical
6	Bowlers_in_team	2848	float64	Numeric
7	Wicket_keeper_in_team	2930	int64	Redundant Column as the values of the observation is same over the data set . The variable seems to be a common denominator and can be removed
8	All_rounder_in_team	2890	float64	Numeric
9	First_selection	2871	object	Categorical
10	Opponent	2894	object	Categorical
11	Season	2868	object	Categorical
12	Audience_number	2849	float64	Numeric
13	Offshore	2866	object	Categorical
14	Max_run_scored_1over	2902	float64	Numeric
15	Max_wicket_taken_1over	2930	int64	Numeric
16	Extra_bowls_bowled	2901	float64	Numeric
17	Min_run_given_1over	2930	int64	Numeric
18	Min_run_scored_1over	2903	float64	Numeric
19	Max_run_given_1over	2896	float64	Numeric
20	extra_bowls_opponent	2930	int64	Numeric
21	player_highest_run	2902	float64	Numeric
22	Players_scored_zero	2930	object	Numeric
23	player_highest_wicket	2930	object	Numeric

Dropped Column Game_number and Wicket_keeper_in_team

df.drop(['Game_number','Wicket_keeper_in_team'],axis=1,inplace=True)

The name of the columns seems to be fine with no special characters or spaces between them . They are quite long and can be shortened but this was avoided to reduce any mapping requirements later for presentation to BCCI

Unique values of various Categories

```
Result :  2
Loss     473
Win     2457
Name: Result, dtype: int64


Match_light_type :  3
Night              296
Day and Night      541
Day               2041
Name: Match_light_type, dtype: int64


Match_format :  4
20-20        6
```

```
T20       864
ODI      1865
Name: Match_format, dtype: int64


First_selection :  3
Bat          11
Batting    1138
Bowling    1722
Name: First_selection, dtype: int64


Opponent :  9
Australia       104
West Indies     158
Zimbabwe        163
Bangladesh      204
Pakistan        253
England         283
Srilanka        513
Kenya           576
South Africa    640
Name: Opponent, dtype: int64


Season :  3
Winter     641
Summer     918
Rainy     1309
Name: Season, dtype: int64


Offshore :  2
Yes     809
No     2057
Name: Offshore, dtype: int64


Players_scored_zero :  5
Three        5
1          166
4          285
2          744
3         1730
Name: Players_scored_zero, dtype: int64


player_highest_wicket :  6
Three        7
5          138
4          211
3          427
2         1063
1         1084
```

The highlighted data seems to be recorded incorrectly and required replacement and this was done to ensure the right categories are picked up by the model

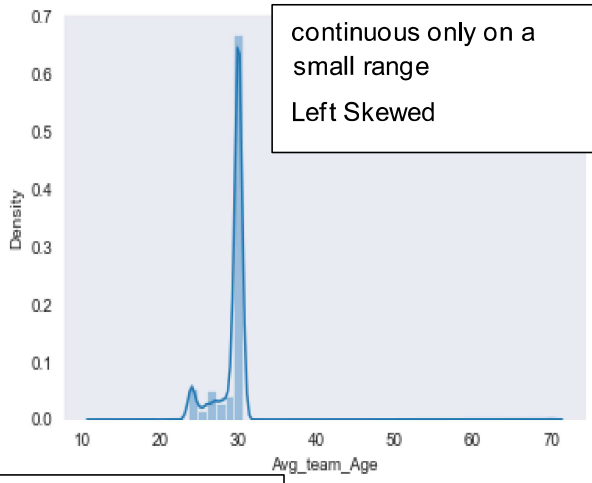
Post fixing of the data

```
Result :  2
Loss     473
Win     2457
Name: Result, dtype: int64


Match_light_type :  2
Day and Night     837
Day              2041
Name: Match_light_type, dtype: int64


Match_format :  3
Test     125
T20      870
ODI     1865
Name: Match_format, dtype: int64


First_selection :  2
Batting    1149
Bowling    1722
Name: First_selection, dtype: int64


Opponent :  9
Australia       104
West Indies     158
Zimbabwe        163
Bangladesh      204
Pakistan        253
England         283
Srilanka        513
Kenya           576
South Africa    640
Name: Opponent, dtype: int64


Season :  3
Winter     641
Summer     918
Rainy     1309
Name: Season, dtype: int64


Offshore :  2
Yes     809
No     2057
Name: Offshore, dtype: int64
```

The last three were actually numeric columns but perceived as Object because of incorrect data capture .. Fixing the inconsistencies fixed the type of the variable as well.

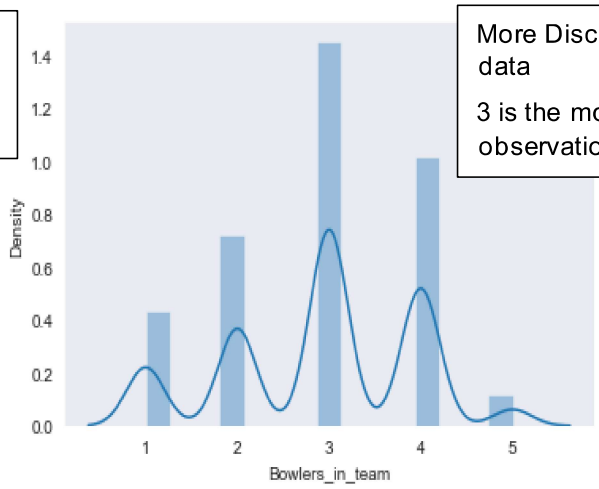
3. Predictive Power of Data

3.1 Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)



continuous only on a small range

Left Skewed

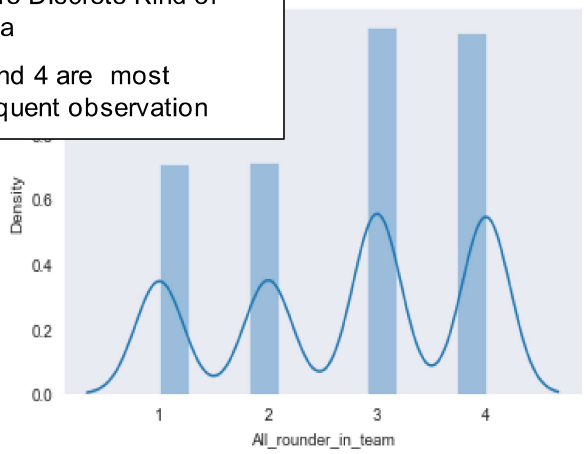


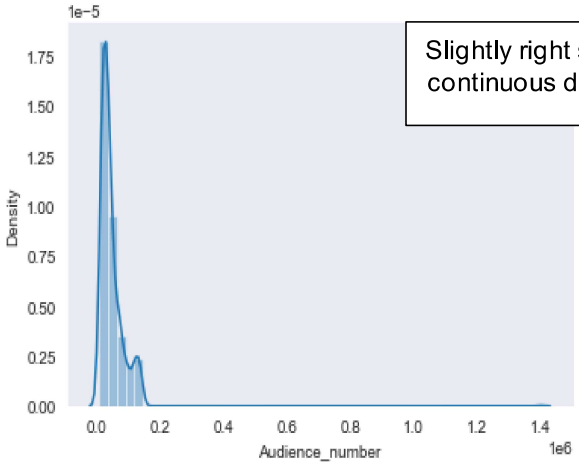
More Discrete Kind of data

3 is the most frequent observation

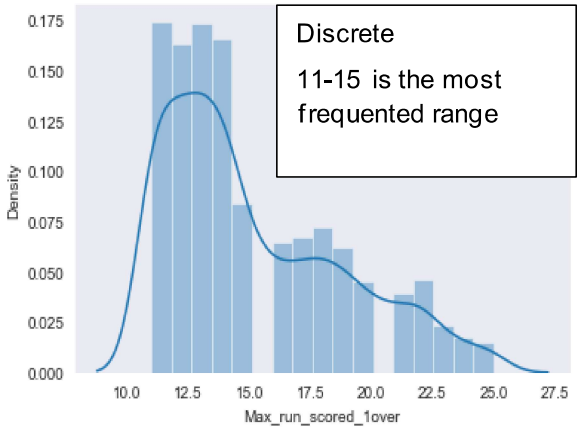
More Discrete Kind of data

3 and 4 are most frequent observation



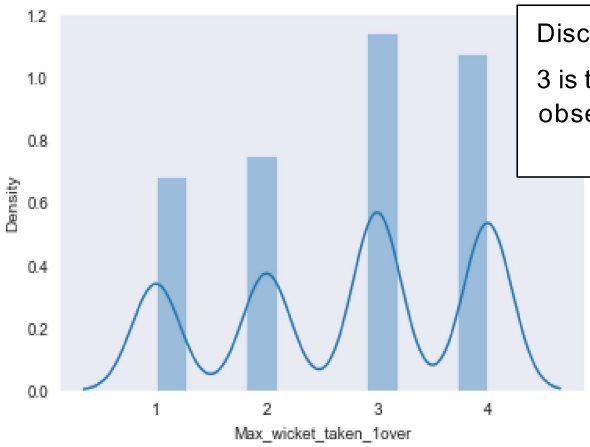


Slightly right skewed and continuous data



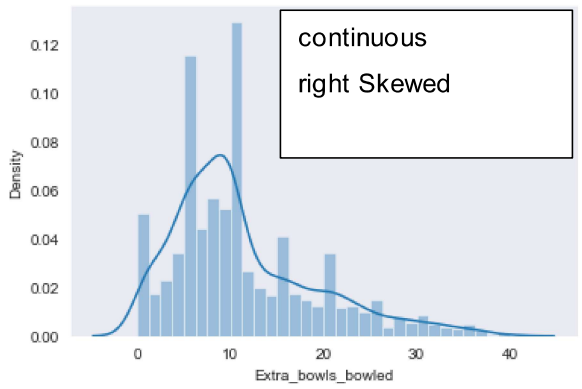
Discrete

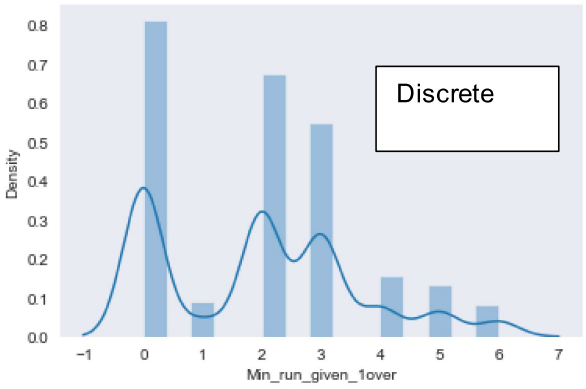
11-15 is the most frequented range

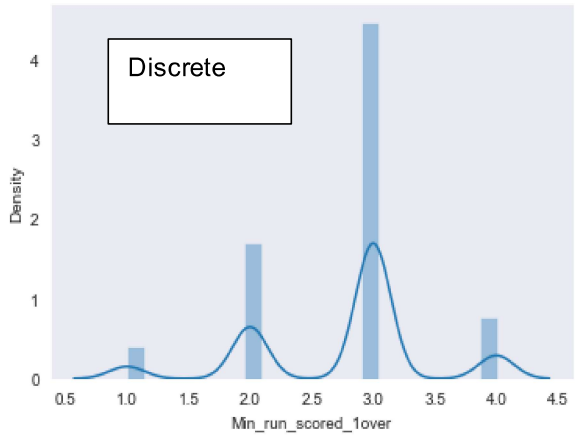


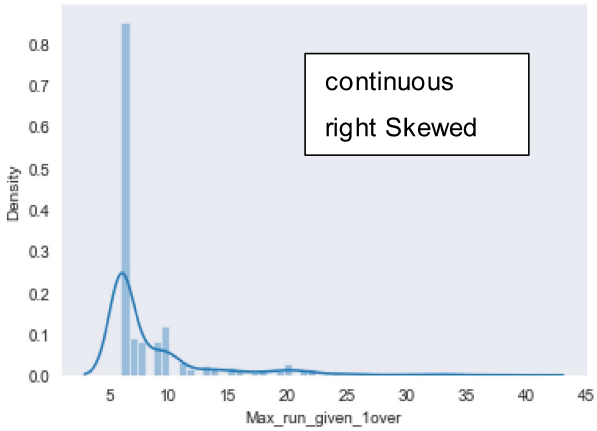
Discrete

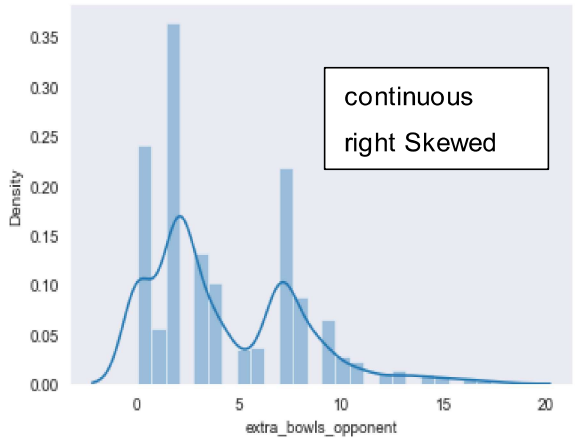
3 is the most frequent observation

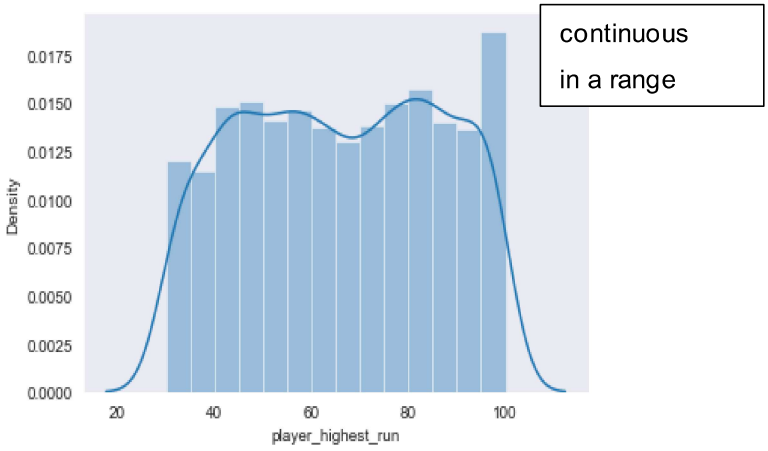


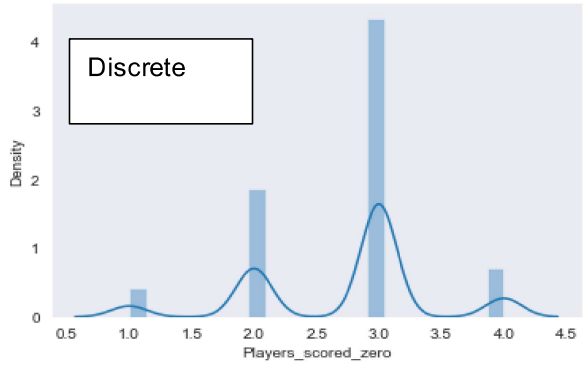


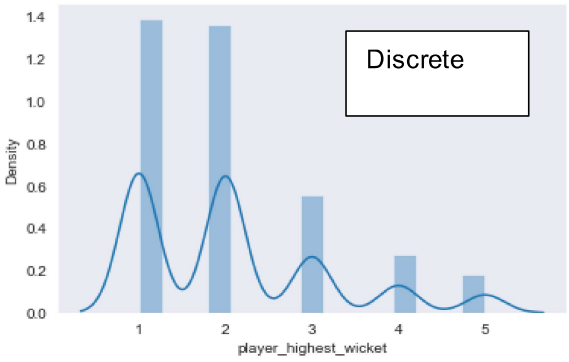






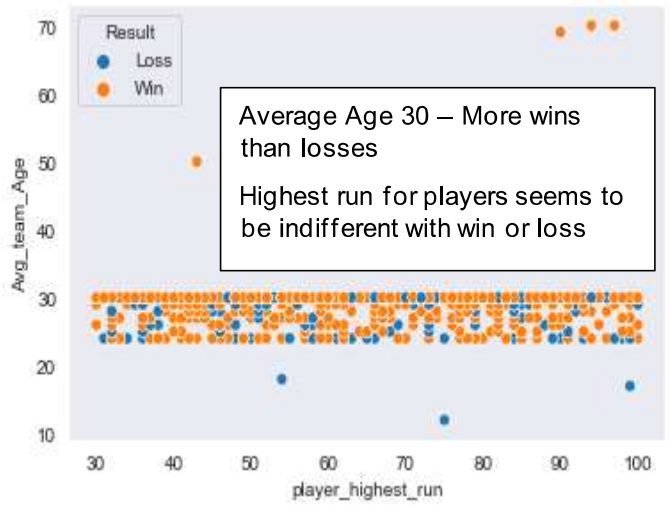


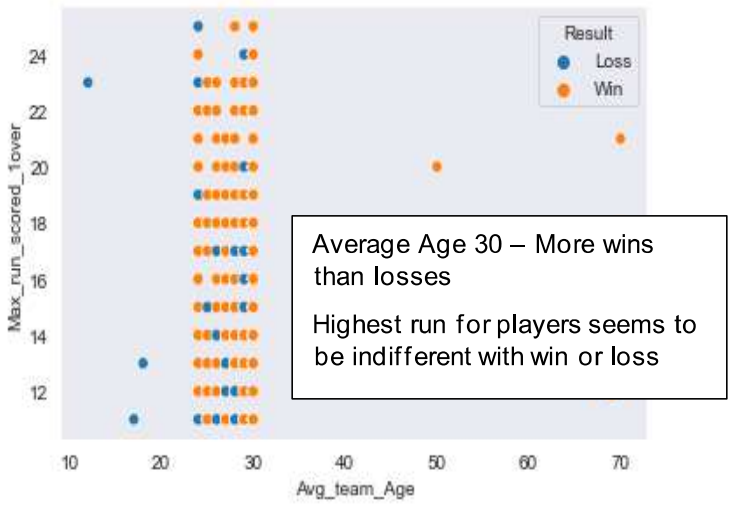


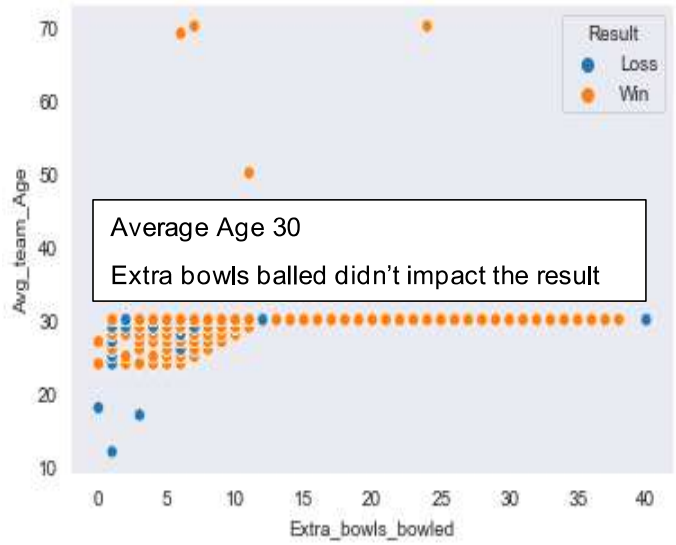


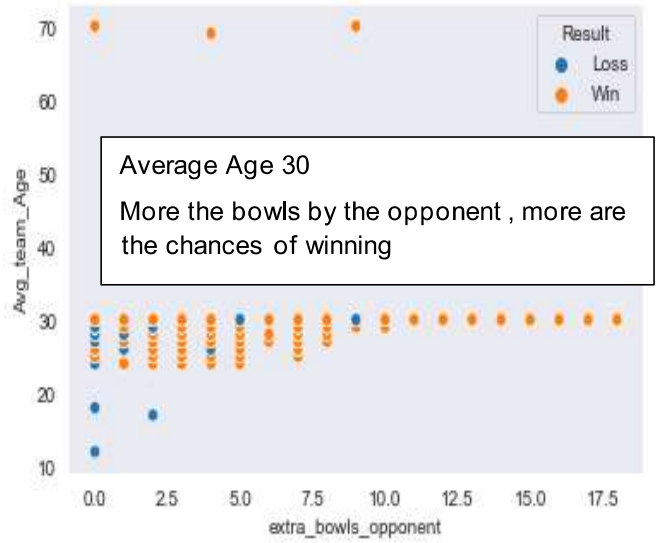
Most of the numerical data is discrete since the nature of the domain is such . Like the players will only make runs in a integer and there can only be 6 bowls in an over which limits the range in which the data values are available . So even if the data seems continuous but is limited to a range.

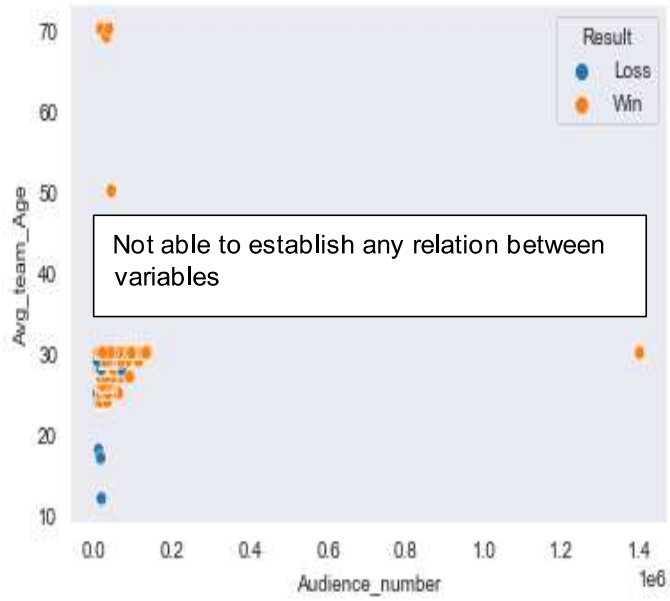
3.2 Bivariate analysis (relationship between different variables , correlations)

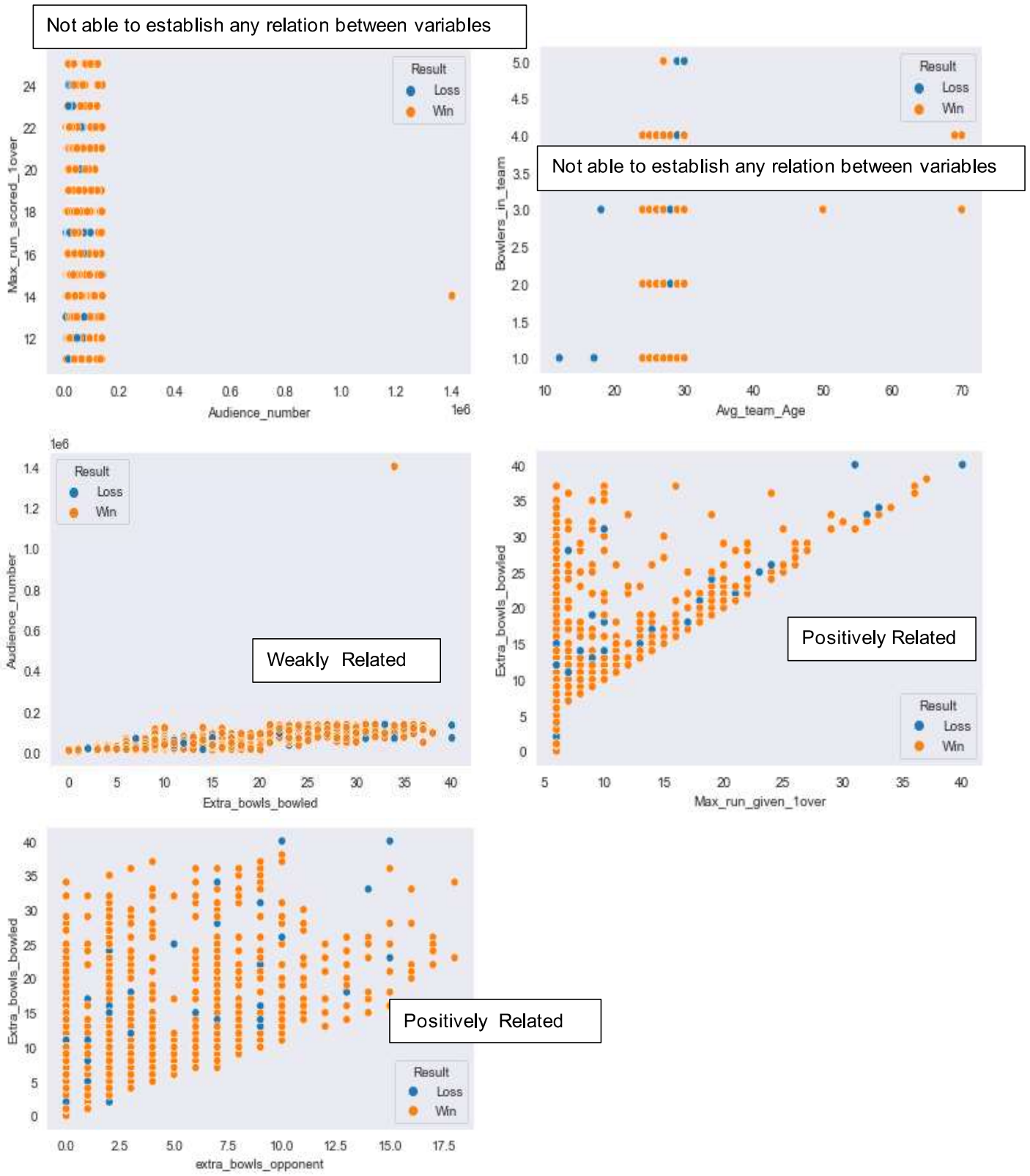






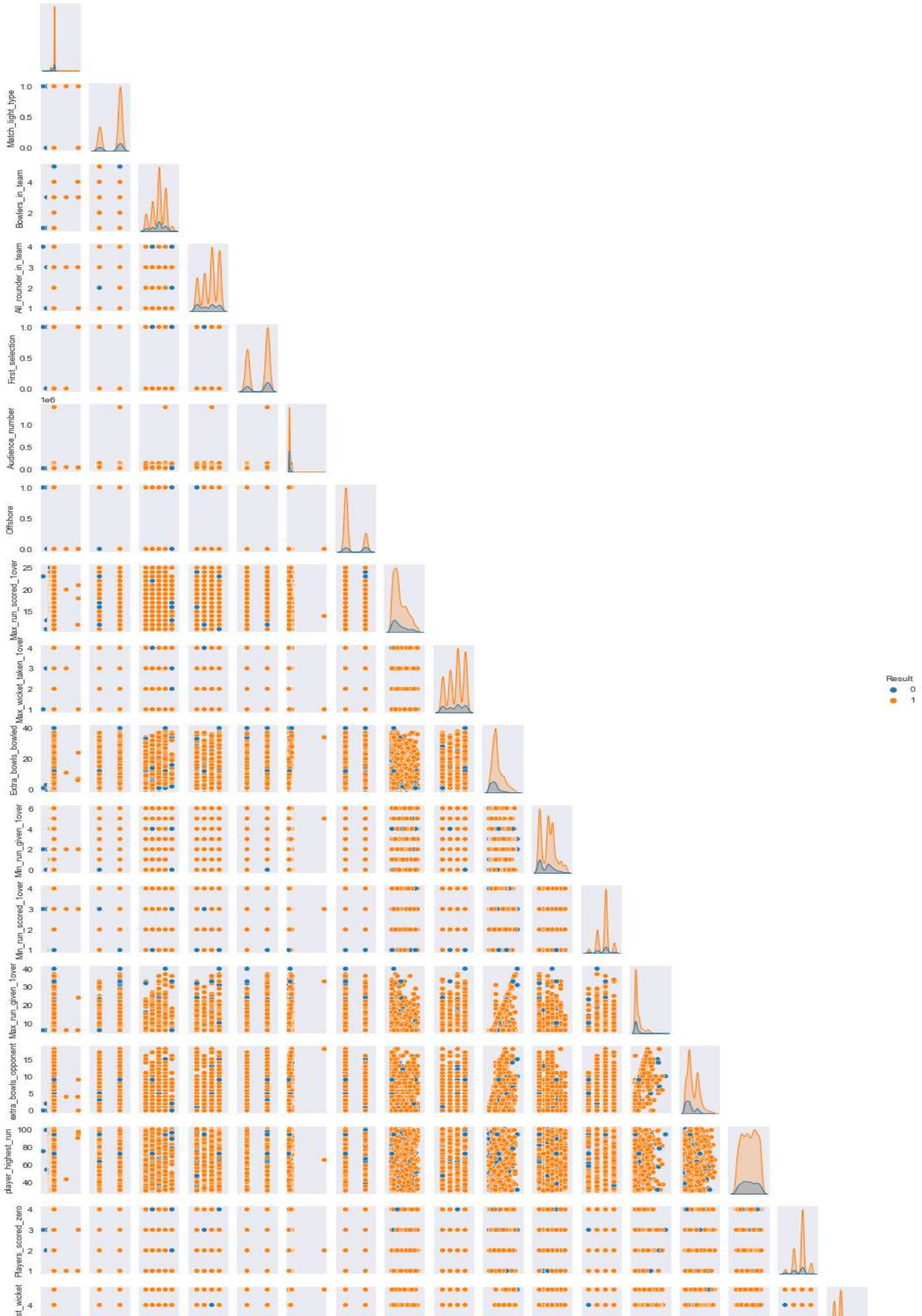


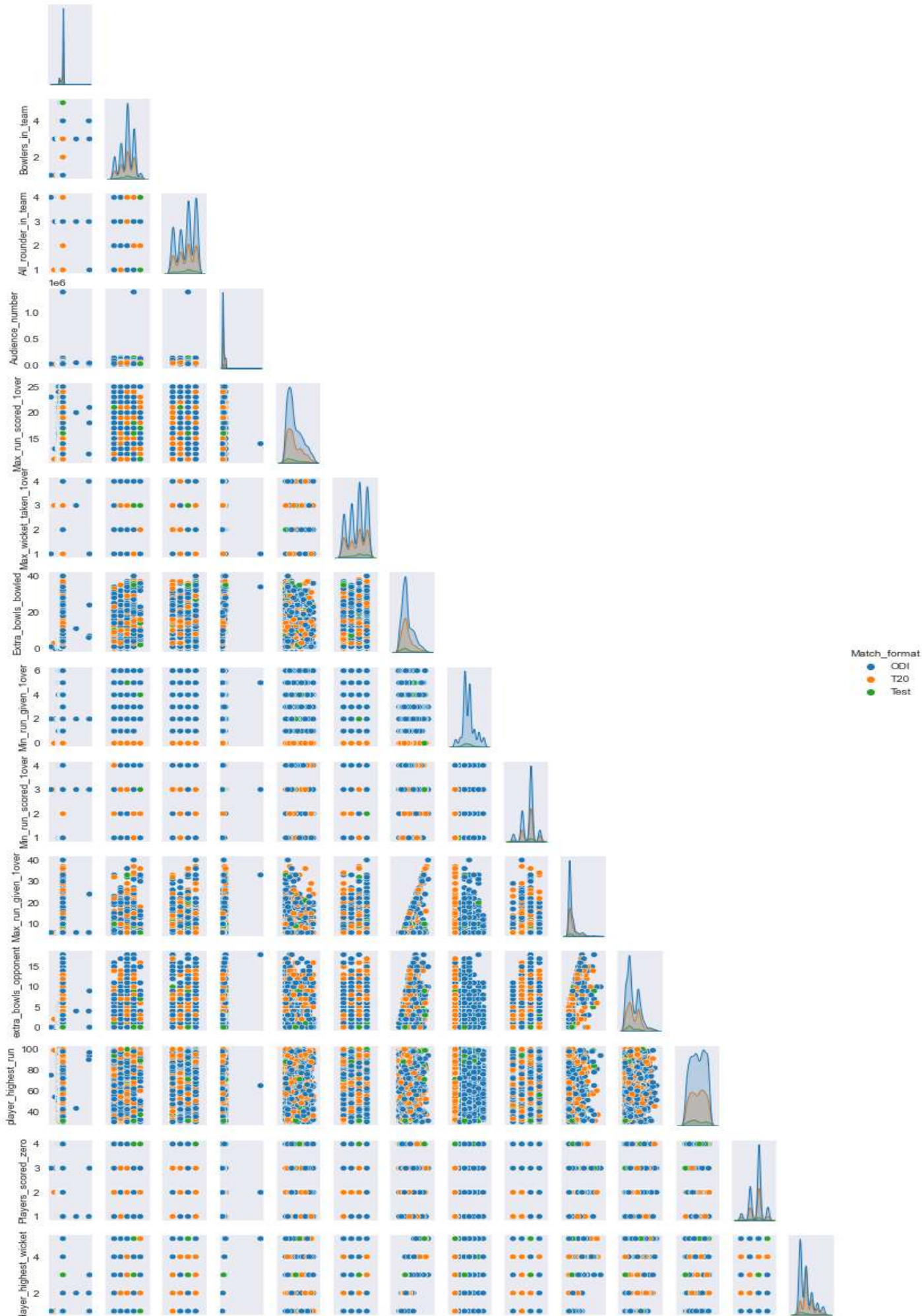




Most of the variables don't seem to be related closely to each other which means there is low multi-collinearity in the data and each feature would have its importance in building the right model . because of this we have not dropped any columns and would want to build the model to see the variable importance.

The pair plot also seems to suggest the same thing . But due to the huge number of columns pair plot was not providing very clear insight and hence resorted to bi variate plots with every combination possible.


Match_light_type
Bowlers_in_team
All_rounder_in_team
First_selection
Audience_number
1e6
Offshore
Max_run_scored_1over
Max_wicket_taken_1over
Extra_bowls_bowled
Min_run_given_1over
Min_run_scored_1over
Max_run_given_1over
extra_bowls_opponent
player_highest_run
Players_scored_zero
Result
0
1


Bowlers_in_team
All_rounder_in_team
1e6
Audience_number
Max_run_scored_1over
Max_wicket_taken_1over
Extra_bowls_bowled
Mn_run_given_1over
Mn_run_scored_1over
Max_run_given_1over
extra_bowls_opponent
player_highest_run
Players_scored_zero
player_highest_wicket
Match_format
ODI
T20
Test

4. Quality of Data

4.1 Missing Value treatment (if applicable)

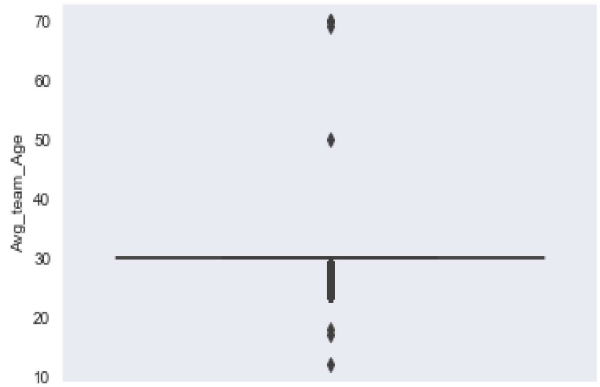
```
Avg_team_Age                97
Bowlers_in_team             82
Audience_number             81
Match_format                70
Offshore                    64
Season                      62
First_selection             59
Match_light_type            52
All_rounder_in_team         40
Opponent         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

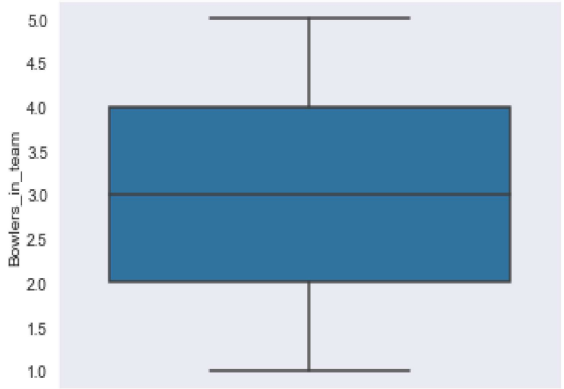
The missing values have been treated with most frequent values than median for numeric data including categorical data . The main reason of choosing mode or most frequent entry was it was making more sense considering the sports domain to which the problem belongs . More so as we have been in the various plots as well the numeric data has discrete pattern due to which we treated them as categorical data.

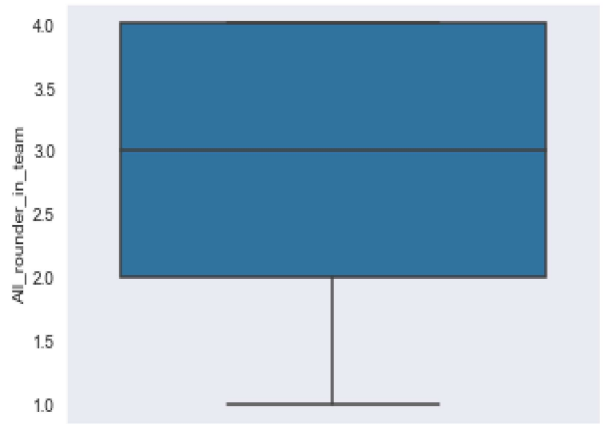
```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='most_frequent',missing_values=np.nan)
```

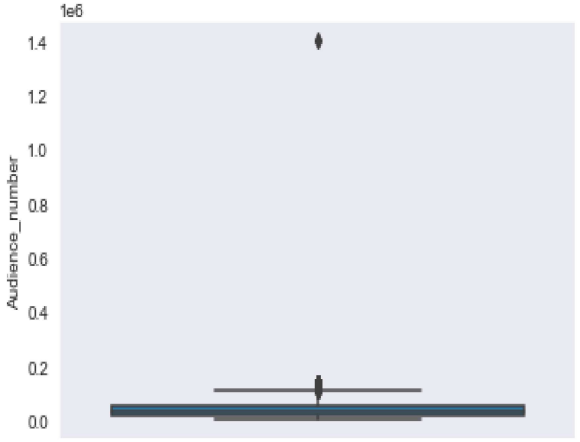
```
for i,col_val in enumerate(list(df.columns)):
    if df[col_val].isnull().sum()>0 :
        df[col_val]=imputer.fit_transform(df[col_val].values.reshape(-1,1))[:,0]
```

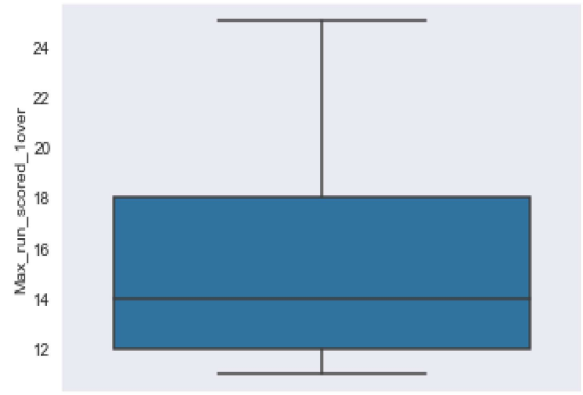
4.2 Outlier treatment (if required)

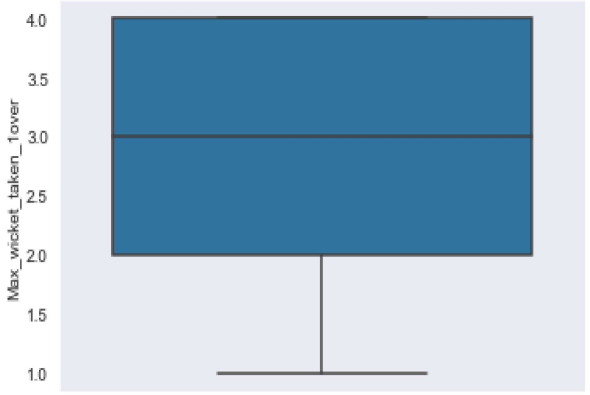


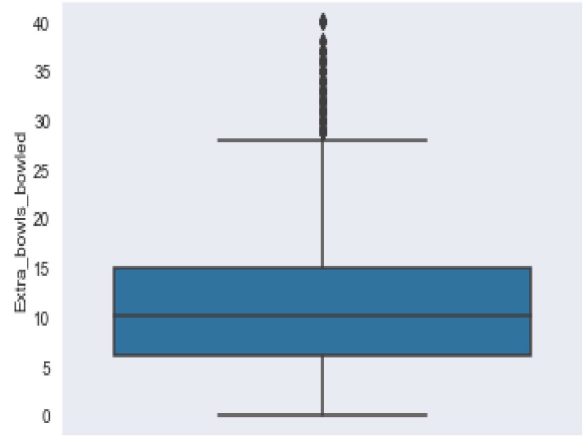


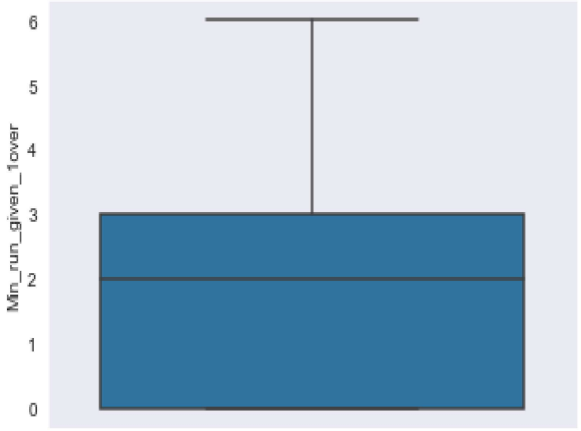

All_rounder_in_team
4.0
3.5
3.0
2.5
2.0
1.5
1.0

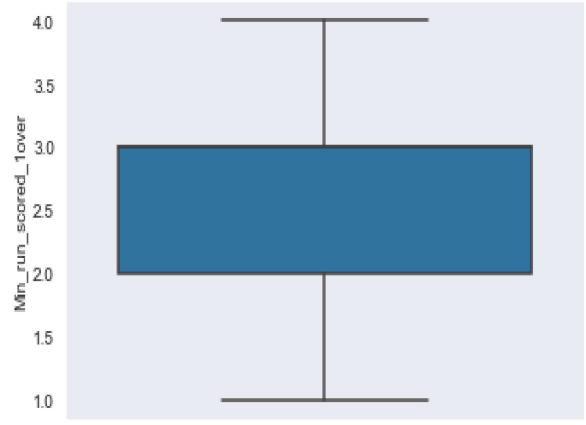

1e6
Audience_number
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

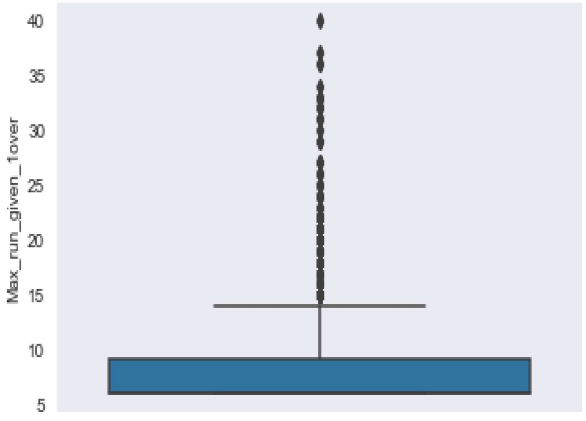

Max_run_scored_1over
24
22
20
18
16
14
12

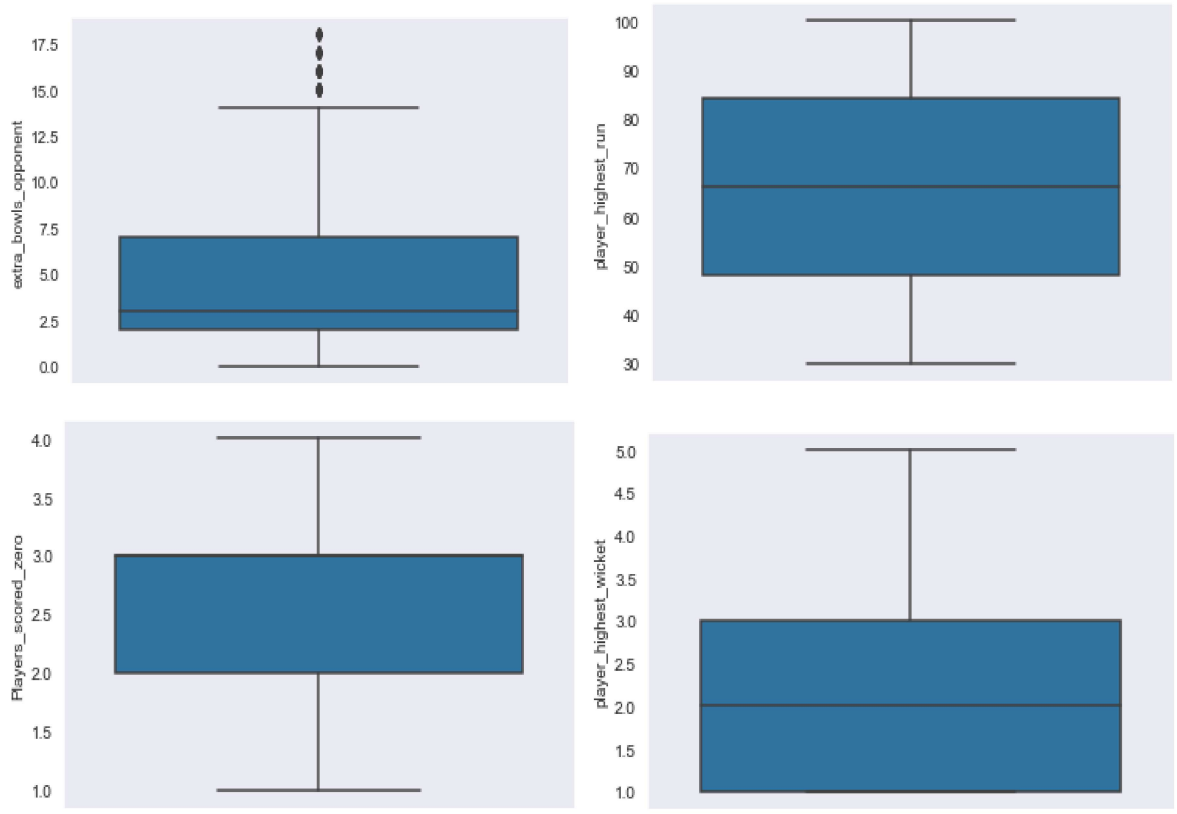

Max_wicket_taken_1over
4.0
3.5
3.0
2.5
2.0
1.5
1.0


Extra_bowls_bowled
40
35
30
25
20
15
10
5
0


Min_run_given_1over
6
5
4
3
2
1
0


Min_run_scored_1over
4.0
3.5
3.0
2.5
2.0
1.5
1.0


Max_run_given_1over
40
35
30
25
20
15
10
5



Not in a favour of doing any outlier treatment as most of the numeric data here is discrete and hence the outliers might be able add value to the model . More so the numeric data which is continuous has minimal outliers . Like the audience number has one observation which stands out and most of the others are in the right range.

5. Feature Engineering

5.1 Removal of unwanted variables (if applicable)

Game_number and Wicket_keeper_in_team are both redundant columns and have been removed. Chose not to remove any other columns and left to the model phase where the variable importance would be judged.

```
In [99]:  1  df.drop(['Game_number','Wicket_keeper_in_team'],axis=1,inplace=True)
```

5.2 Variable transformation (if applicable)

```
Match_light_type :  3
Night              296
Day and Night      541
Day               2041
Name: Match_light_type, dtype: int64

Match_format :  4
20-20         6
Test        125
T20         864
ODI        1865
Name: Match_format, dtype: int64

First_selection :  3
Bat           11
Batting     1138
```

```
Name: First_selection, dtype: int64

Players_scored_zero :  5
Three        5
1          166
4          285
2          744
3         1730
Name: Players_scored_zero, dtype: int64


player_highest_wicket :  6
Three        7
5          138
4          211
3          427
2         1063
1         1084
```

The highlighted data seems to be recorded incorrectly and required replacement and this was done to ensure the right categories are picked up by the model

```
df['player_highest_wicket']=df['player_highest_wicket'].replace(to_replace='Three',value=3)
```

```
df['Players_scored_zero'] = df['Players_scored_zero'].replace(to_replace='Three',value=3)
```

```
df['First_selection'] = df['First_selection'].replace(to_replace='Bat',value='Batting')
```

```
df['Match_light_type'] = df['Match_light_type'].replace(to_replace='Night',value='Day and Night')
```

```
df['Match_format'] = df['Match_format'].replace(to_replace='20-20',value='T20')
```

The variables has been encoded to numeric values for the following variables

```
df['Result'] = df['Result'].replace(to_replace='Win',value=1)
df['Result'] = df['Result'].replace(to_replace='Loss',value=0)
```

```
df['Offshore'] = df['Offshore'].replace(to_replace='Yes',value=1)
df['Offshore'] = df['Offshore'].replace(to_replace='No',value=0)
```

```
df['Match_light_type'] = df['Match_light_type'].replace(to_replace='Day',value=1)
df['Match_light_type'] = df['Match_light_type'].replace(to_replace='Day and Night',value=0)
```

```
df['First_selection'] = df['First_selection'].replace(to_replace='Bowling',value=1)
df['First_selection'] = df['First_selection'].replace(to_replace='Batting',value=0)
```

5.3 Addition of new variables (if required)

No new variables were added at this stage . But before proceeding with the model one hot encoding would be required on few categories which would increase the number of column not essentially the number of variables.

6. Business Insights from EDA

6.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business

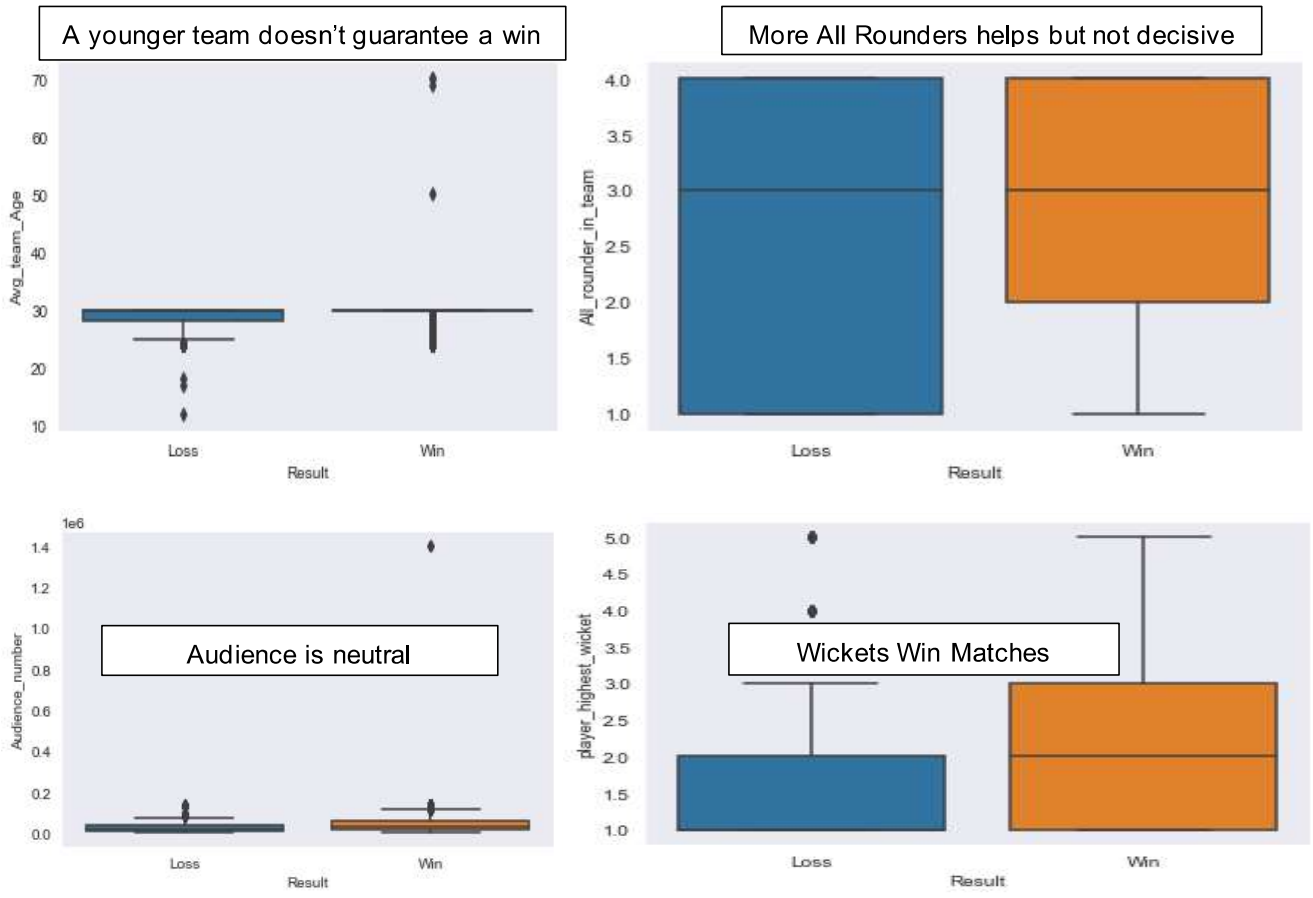
```
Result :  2
Loss     473
Win     2457
```

Data is not balanced with more wins than losses but that would be the nature of the sports as every teams' endeavour is to win more and more matches and hence this should be the way the data is expected . Don't see any treatment on this would be needed .

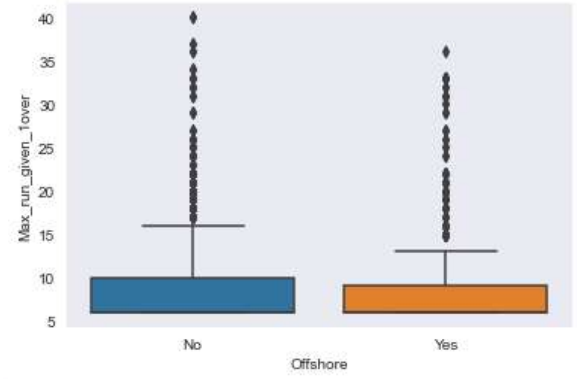
More so the problem statement is also to find the wining strategy for the team

6.2 Any business insights using clustering (if applicable)

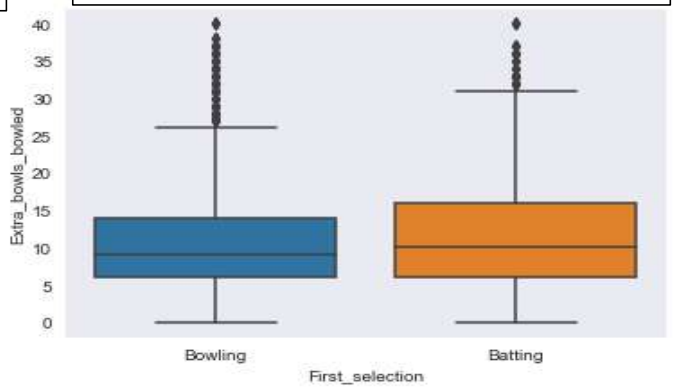
Variable plotted against Match Result



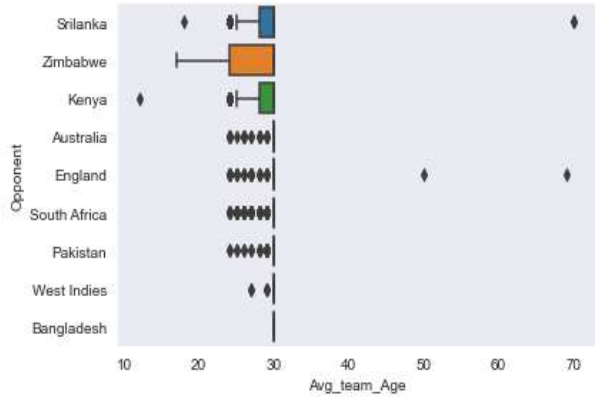
Home condition makes us more expensive



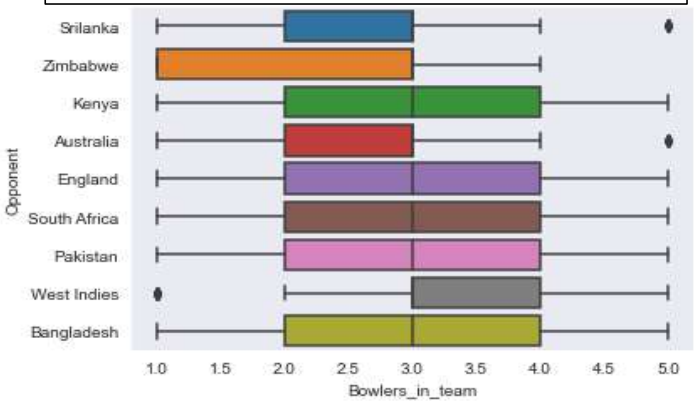
Batting First extra balls bowled are higher



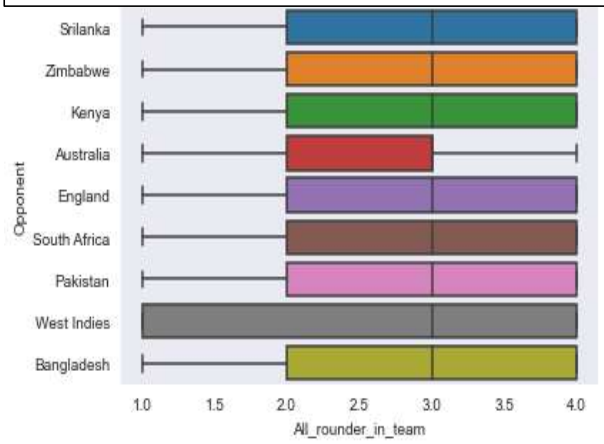
Average Age of opponents is 30



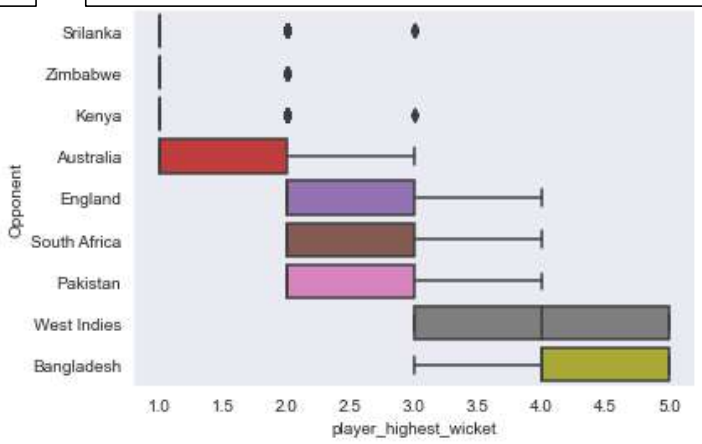
Opponents play with 3 bowlers on average



Opponents play with 3 All Rounders on average

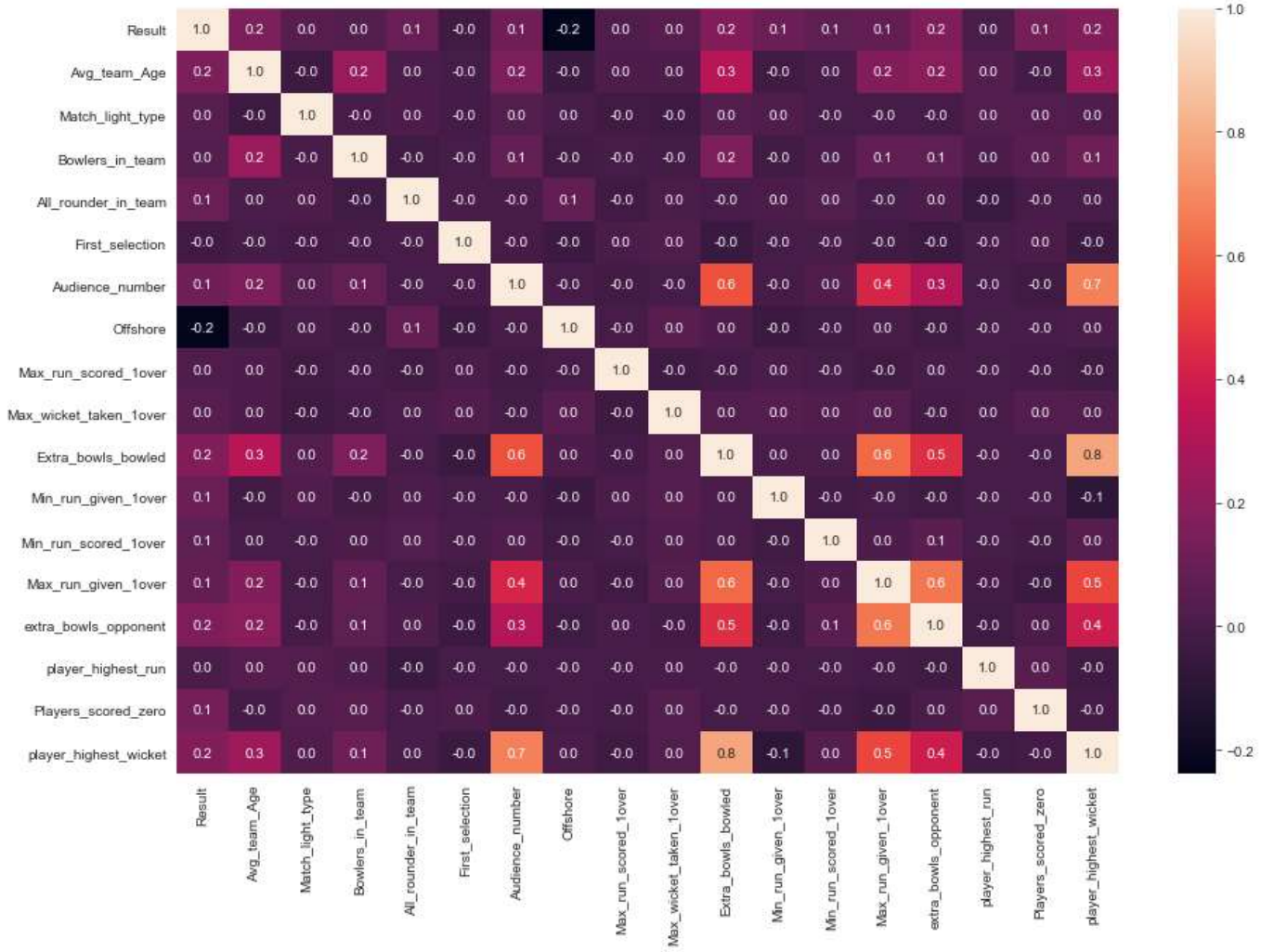


Bangladesh has player with highest wicket



Seasons have less or no impact on team performance

6.3 Any other business insights



- Audience numbers have no correlation with result but seems to be correlated with players_highest_wicket which means Audience brings the best in a player but may not be true for the entire team
- Audience seems to be a distraction as well . extra_balls_bowled and extra_ball_bowled_opponent both are positively correlated with Audience
- Extra_balls_bowled is positively correlated with player_highest_wicket
- Max_run_given_1over is positively correlated with extra_bowls_bowled