Exploratory Data Analysis on Indian Super League

Data analysis basics with Python

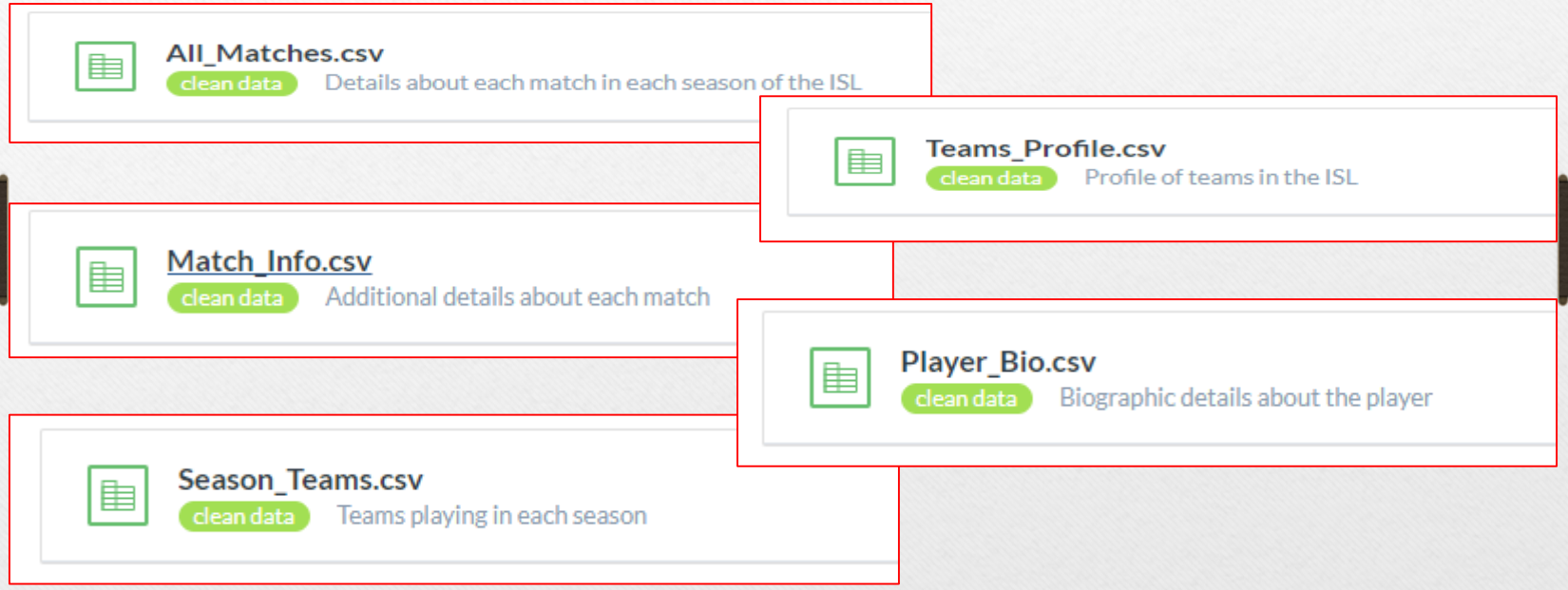
By Chiranjit Pathak

Problem Statement

Based on **quality of games & players, organizational constraints** and **strategic & regional influences** in **India**, how can **ISL expand and resuscitate** its **excitement and popularity** in the **coming seasons**?

Indian Super League Season 2014-2017 data

Data Source:
https://data.world/ajaigovindg/hero-indian-super-league



Based on the above collected data an EDA has been exercised using Numpy, Pandas, Seaborn, Matplotlib, Sklearn, Bokeh and Plotly in Python

Data Processing

Preprocess, Profiling and Post Processing

Data Pre-processing, Profiling and Post-processing

Missing data identification

- This is done to identify the data sets, on which Pandas profiling must be done.

```
def missing_data(data):
    total = data.isnull().sum().sort_values(ascending = False)
    percent = (data.isnull().sum()/data.isnull().count()*100).sort_values(ascending = False)
    return pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data(allmatches_df)
```

	Total	Percent
PS_Team2_Goals	275	98.920863
PS_Team1_Goals	275	98.920863
ET_Team2_Goals	274	98.561151
ET_Team1_Goals	274	98.561151
HT_Team1_Goals	2	0.719424
HT_Team2_Goals	2	0.719424

PS,ET only applies for playoff matches hence replaced by 0;
HT results are replaced by 0.

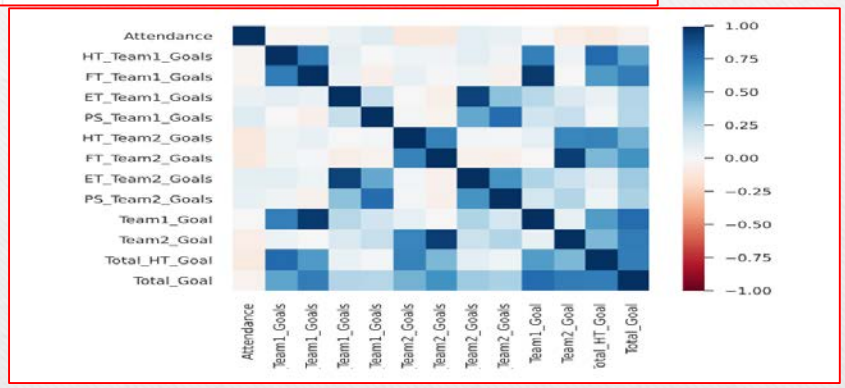
	Total	Percent
Asst_Referee_2	217	78.057554
Asst_Referee_1	217	78.057554
Attendance	3	1.079137

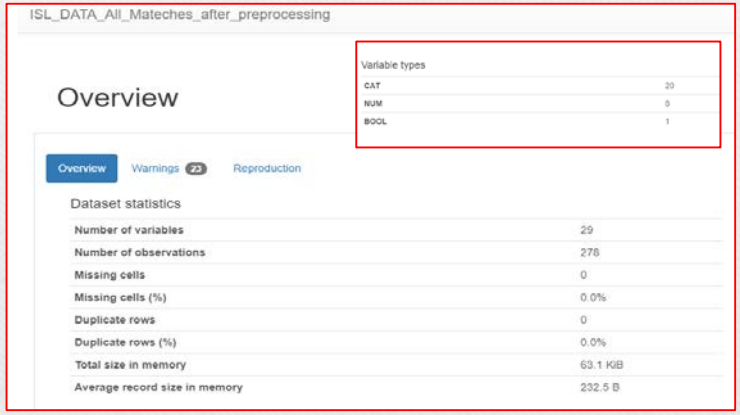
Asst_Referee_1 & 2 have not been used so deleted;
Attendance has been replaced by mean of the respective stadium.

	Total	Percent
height.cm	129	13.767343
dob	9	0.960512

height.cm is replaced by mode (mean, median and mode are very similar).
dob replaced by mode.

```
playerbio_df = playerbio_df.drop_duplicates(subset=['season', 'player'])
```



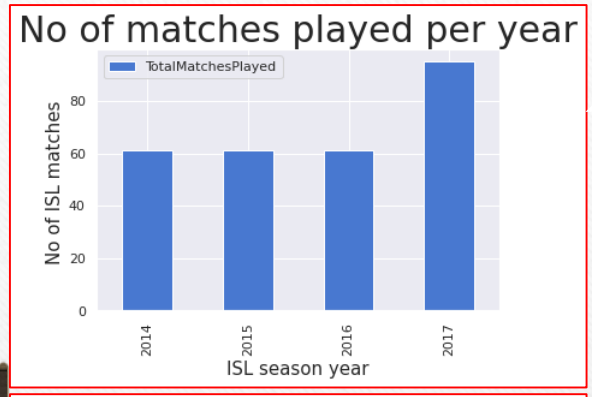


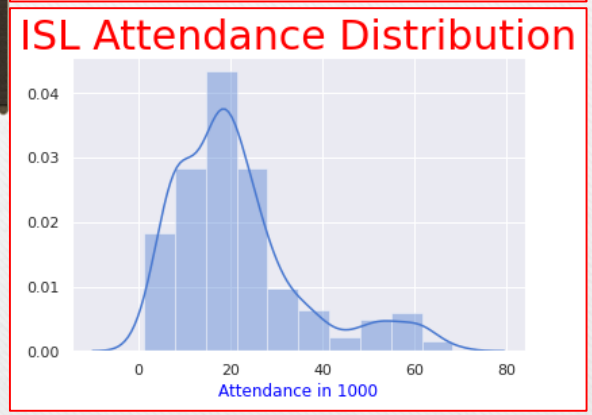
Pandas Profiling has been deployed before and after data processing; Some of the data has been replaced/deleted during pre-processing as indicated.

Identification of challenges

Elaboration on the Problem Statement

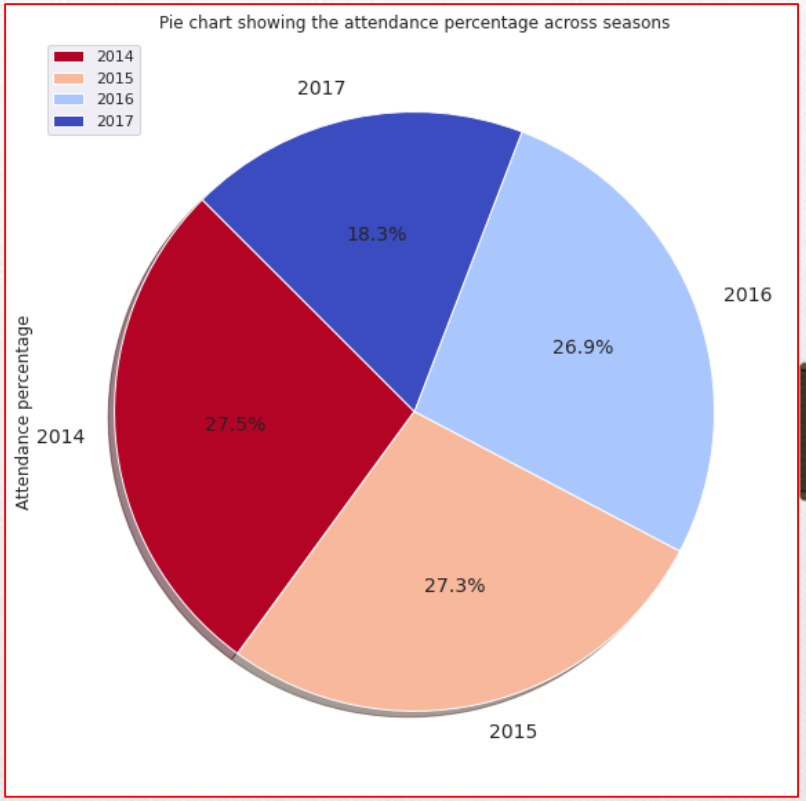
Number of matches played season wise between 2014 to 2017: Match count vs Attendance distribution





More matches

Atten-dance % falling down



Based on the above facts, the key challenge has been understood as Fall in % Attendances which essentially means drop in the popularity of the game.

Basis of the Data Analysis

A detail analysis is being envisaged for the same :

a) Does the Quality of Games are falling down?

b) Does the Player's performances are limiting this drop in popularity?

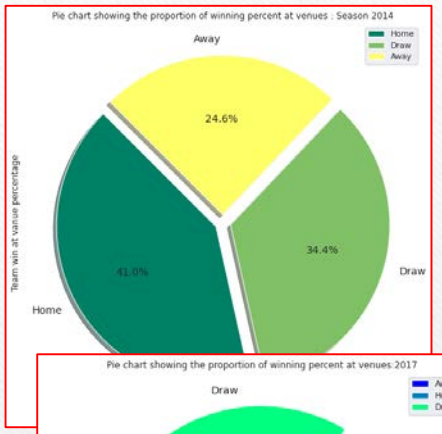
c) Does the infrastructural & organizational bottle-necks are causing these hindrances ?

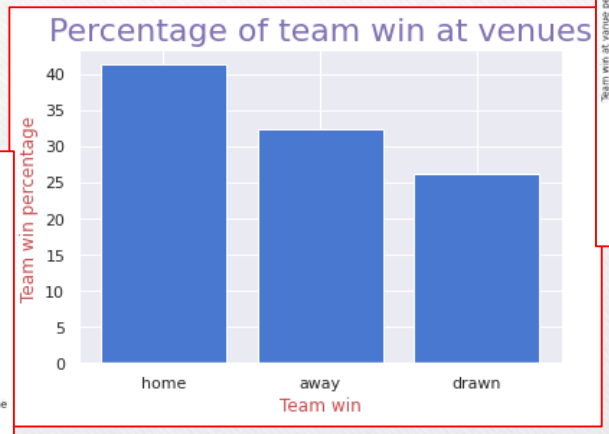
d) Do we need to review some of the strategies based on geographical/regional analysis to resuscitate and expand the excitement?

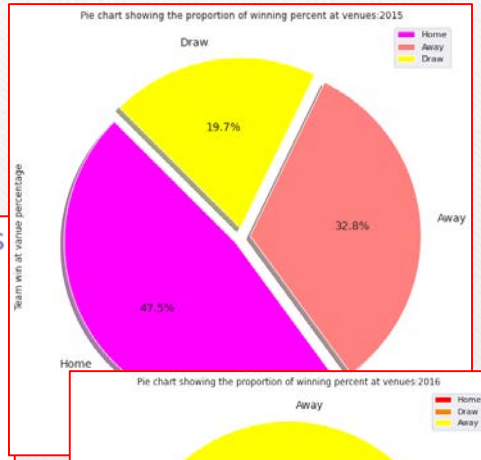
Quality Of the Games

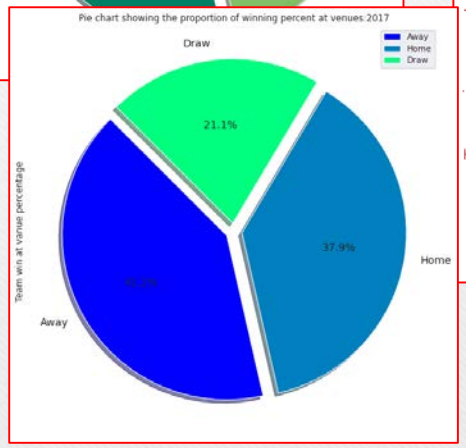
Match/Game wise analysis

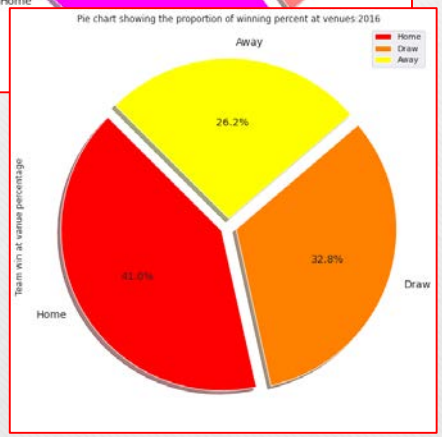
Quality of the Games:
How the Home team winning varies across seasons?





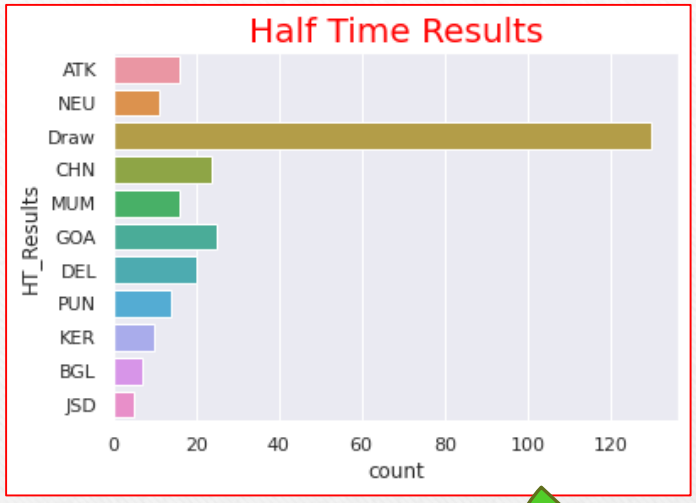


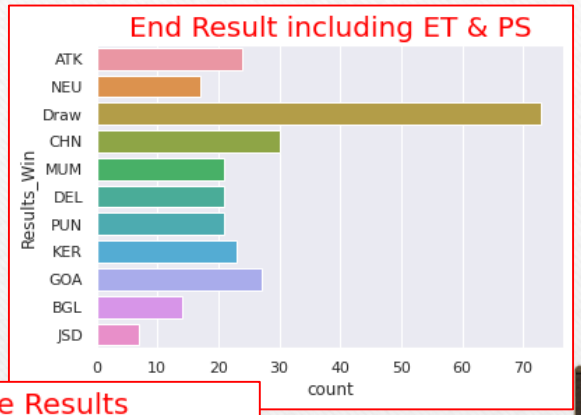


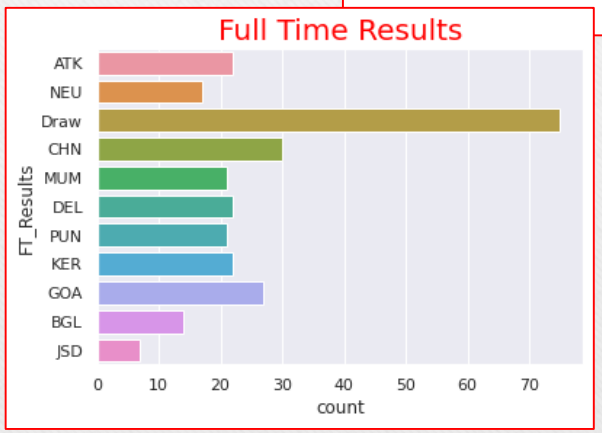


Home teams have not disappointed their fans and the **Away** teams have also challenged well : on an average matches were not being one-sided.

Quality of the Games:
How the match results favoring respective teams?

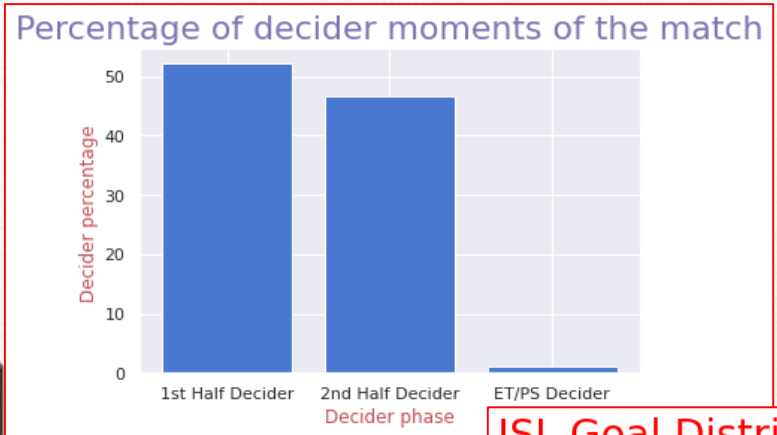


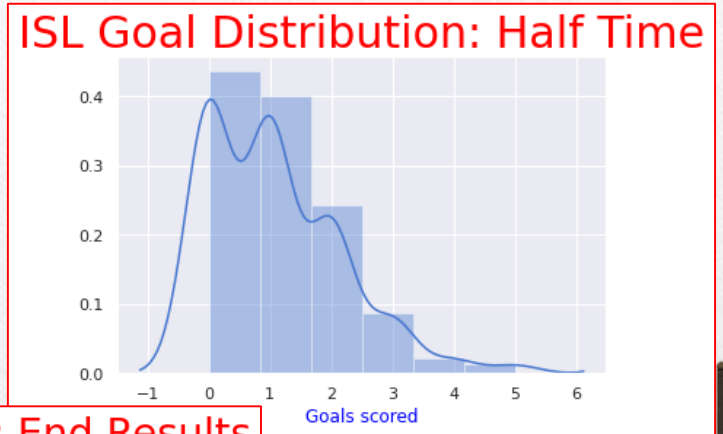


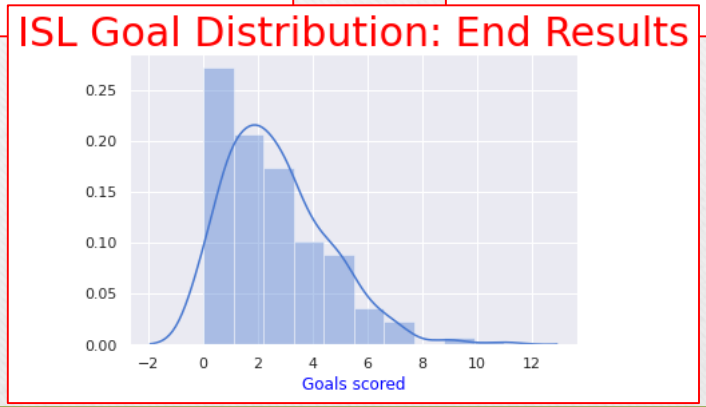


The games brought excitement in the 2nd half; as the number of the draw results (up to half time) reduces at the end of the game.

Quality of the Games:
How the match deciding moments and goals are distributed?



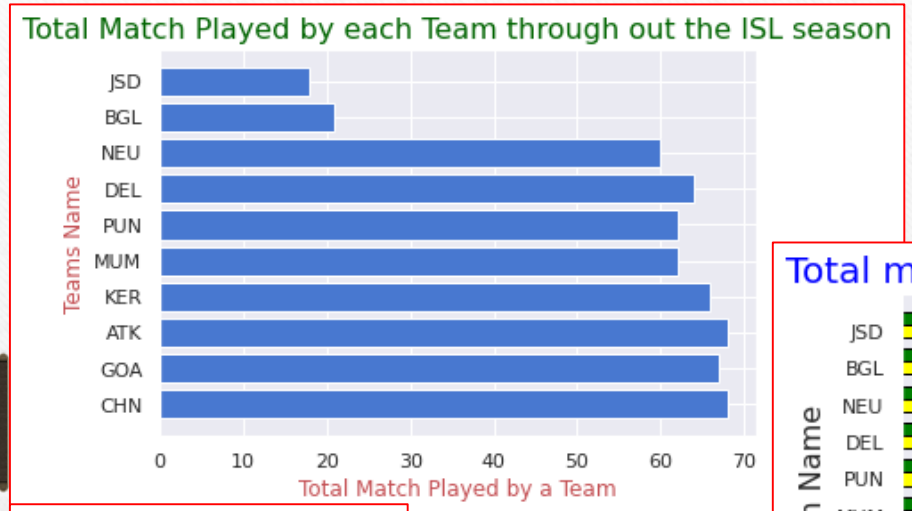


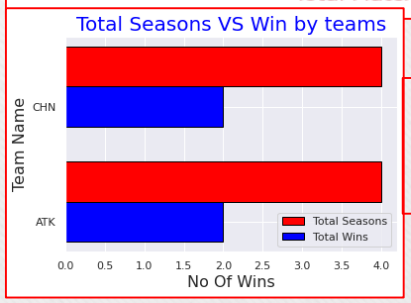


The total goal scored follows normal distribution at the end; whereas total Half Time goals are right skewed.

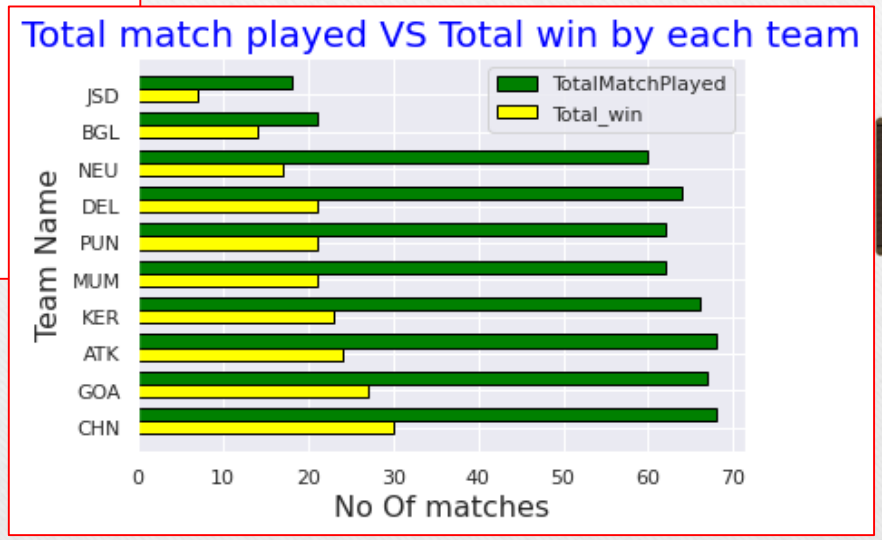
The ratio of the deciding moments of the matches are almost same (~52:47). This shows that the game's excitement are well distributed.

Quality of the Games:
What is the Team wise performances?



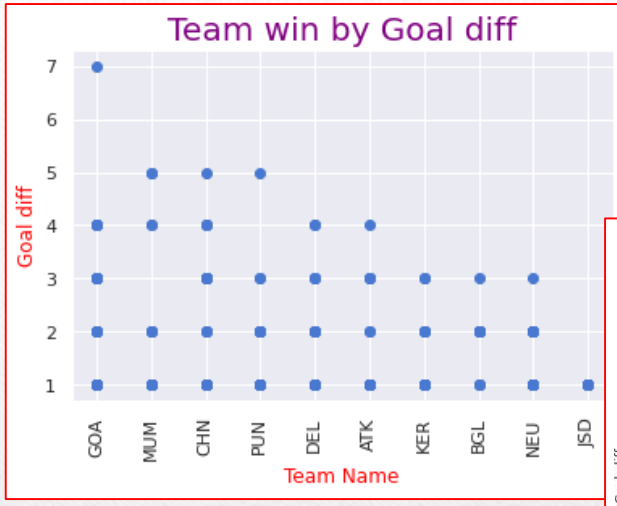


CHN and **ATK** are sharing the title two times each.

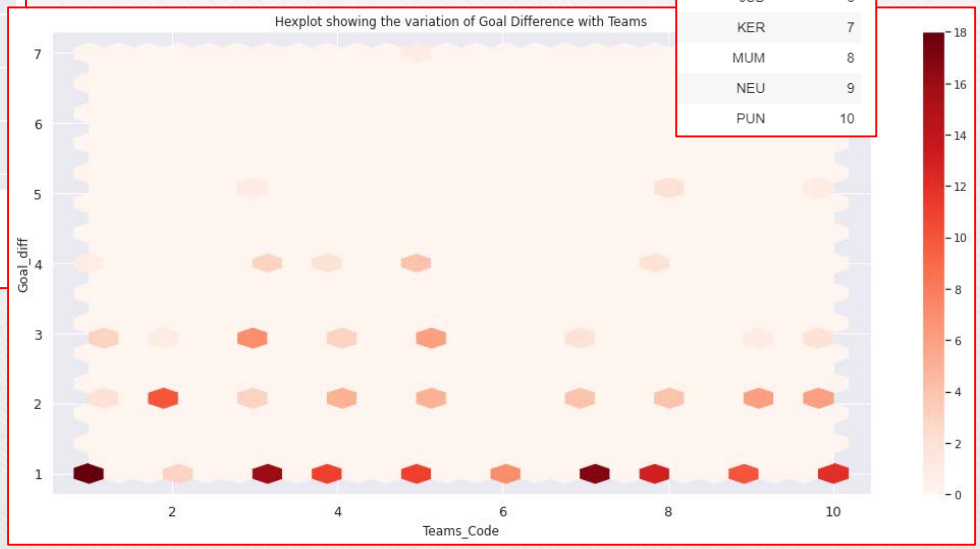


JSD and **BGL** are newly introduced in 2017 season.
CHN, GOA and **ATK** are in leaderboard position respectively.

Quality of the Games:
How the Goal difference occurs throughout the seasons?

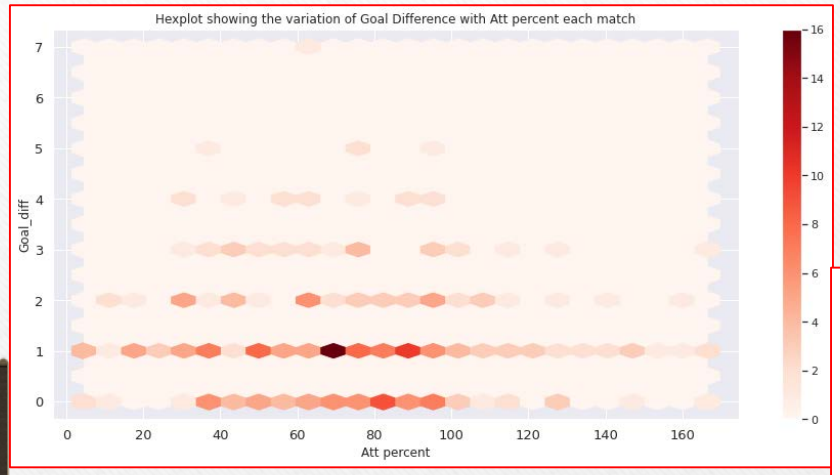


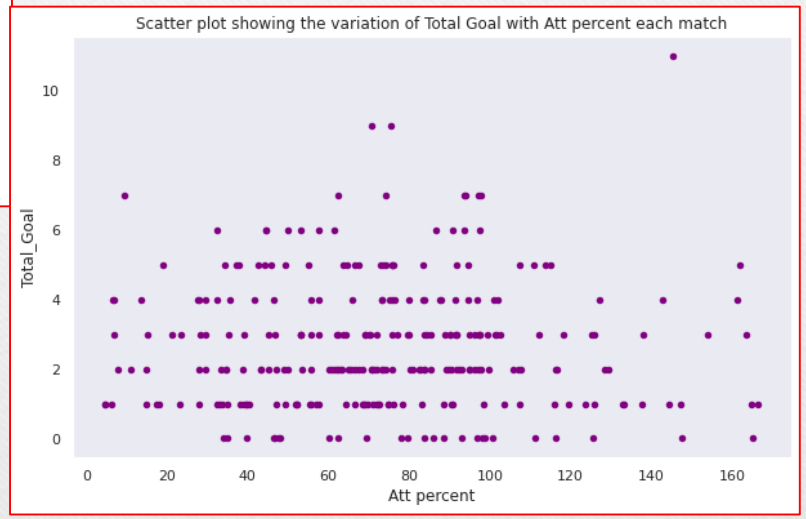
Results_Win	Teams_Code
ATK	1
BGL	2
CHN	3
DEL	4
GOA	5
JSD	6
KER	7
MUM	8
NEU	9
PUN	10



We have seen most matches are with one (1) goal difference i.e. **indication of not being one-sided match**.

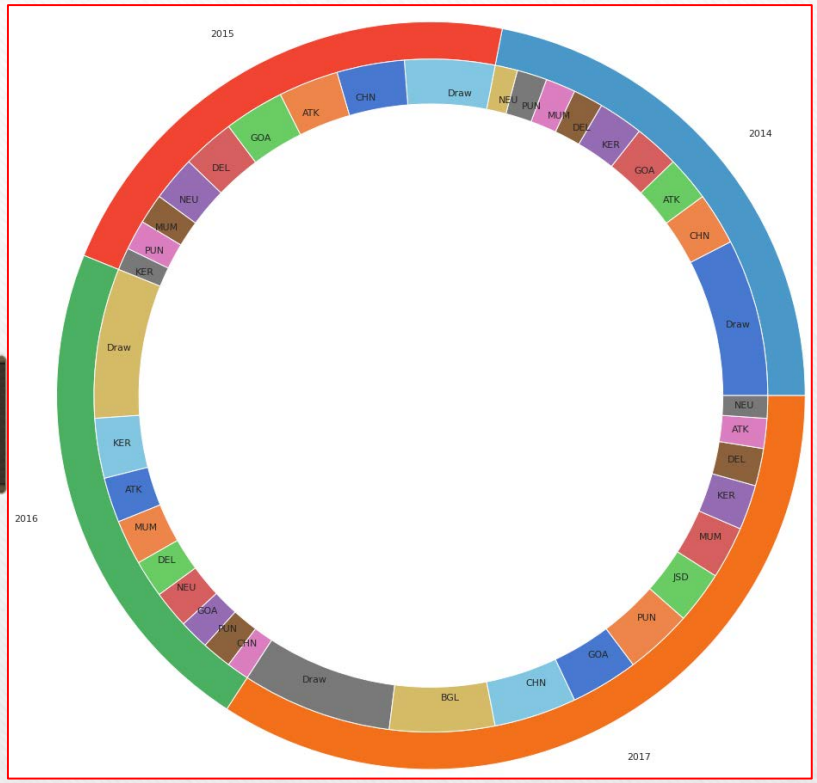
Quality of the Games:
How does the attendance influence the Goal diff., Total goals?

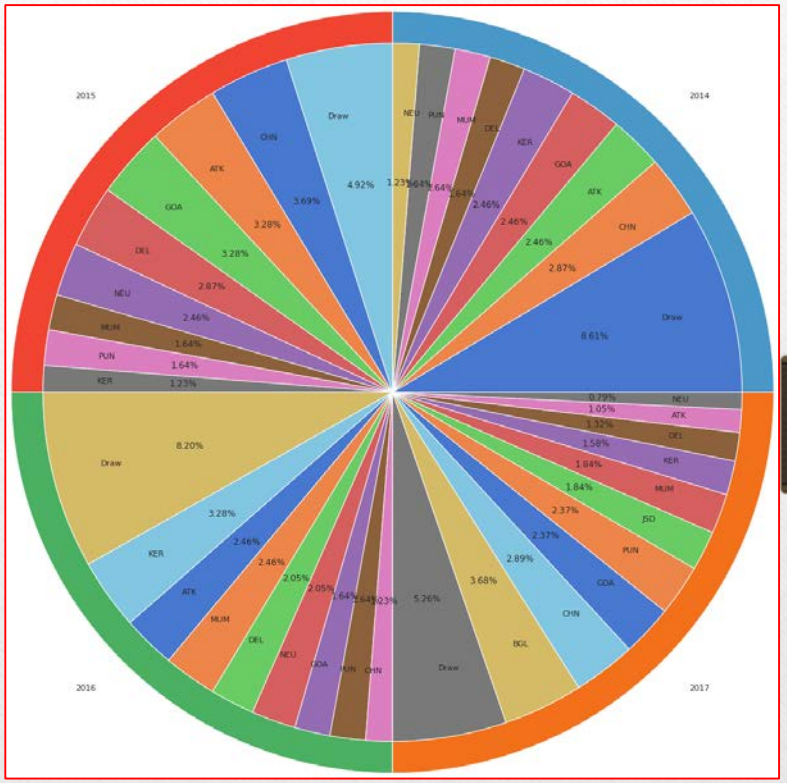




In general, **% attendance** does not show any strong relation with the **total goals and goal differences**.

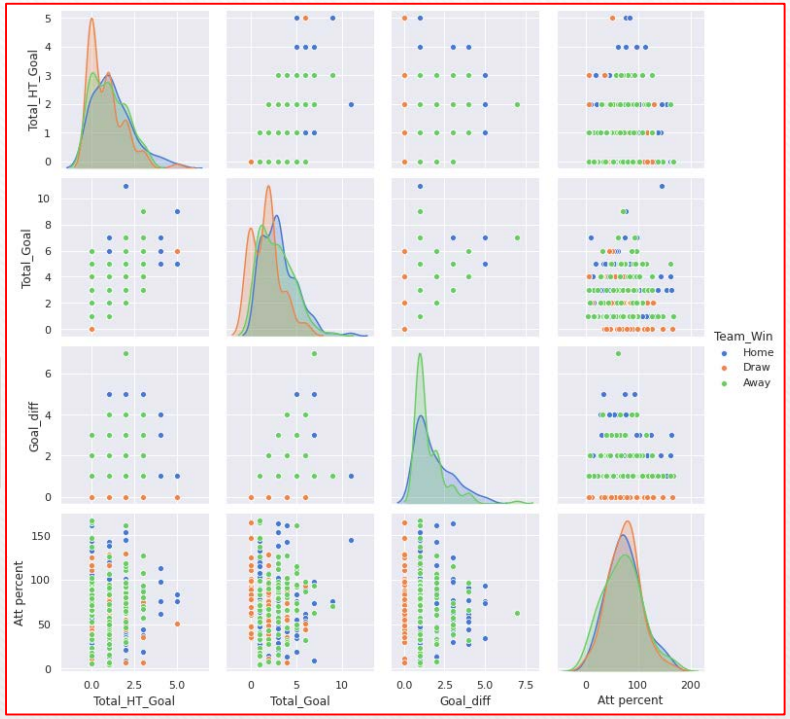
Quality of the Games:
How Teams were performed across all seasons?

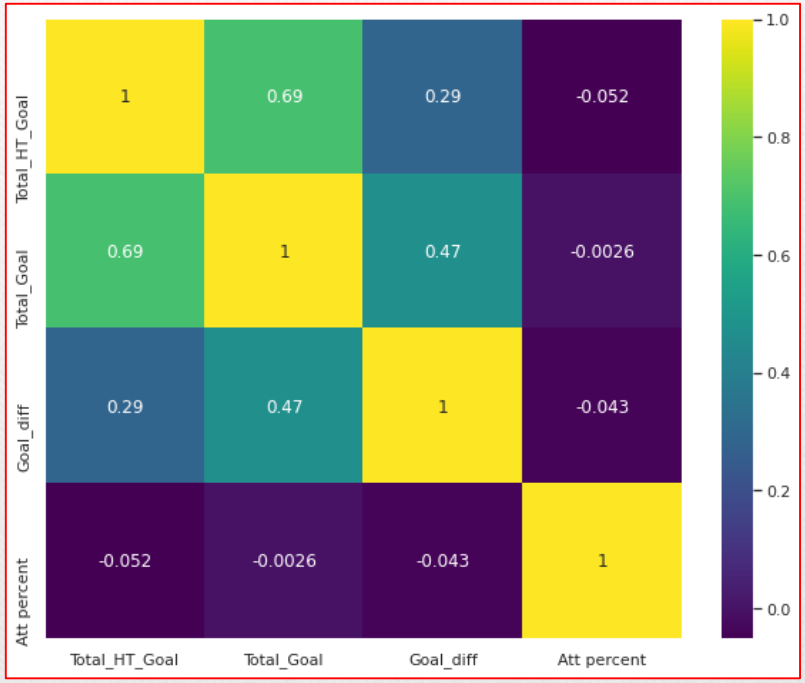




The wins are also well distributed among all the teams : performances seems on similar level ; which also attributed the most draw matches. It means the games are not boring at least.

Quality of the Games:
How does different factors of each matches are correlated ?





Most of the total goals are scored during 1st half (strong correlation +0.69) and also the goal differences are occurred in the 2nd half or later (corr. value +0.47) indicating well distribution of excitement during matches. Attendance % is not well correlated with the match results.

Quality of the Games:
How does all the team strategize their games?

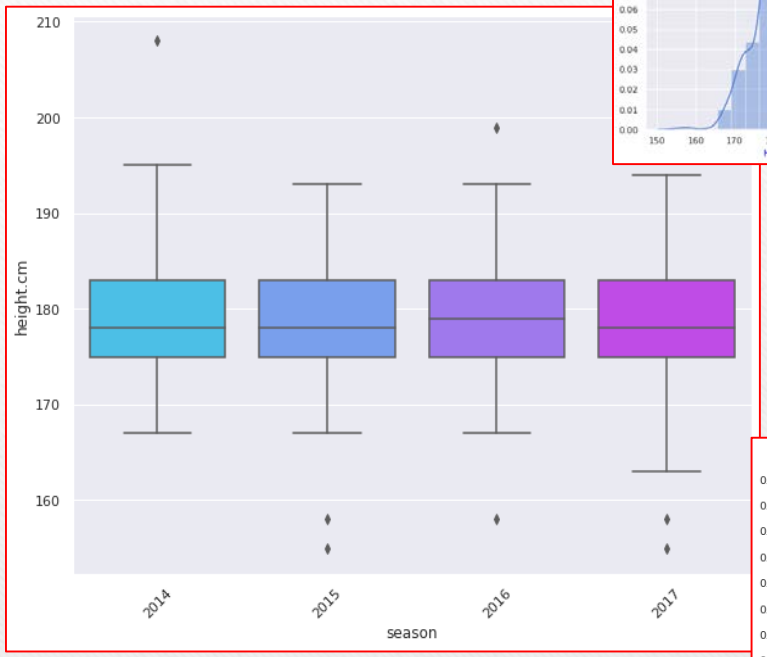
Team_name	win_percentage	HT_ahead_percentage	drawn_percentage	Loss_percentage
CHN	44.117647	35.294118	26.470588	29.411765
GOA	40.298507	37.313433	22.388060	37.313433
ATK	35.294118	23.529412	33.823529	30.882353
KER	34.848485	15.151515	28.787879	36.363636
MUM	33.870968	25.806452	25.806452	40.322581
PUN	33.870968	22.580645	24.193548	41.935484
DEL	32.812500	31.250000	31.250000	35.937500
NEU	28.333333	18.333333	21.666667	50.000000
BGL	66.666667	33.333333	9.523810	23.809524
JSD	38.888889	27.777778	27.777778	33.333333

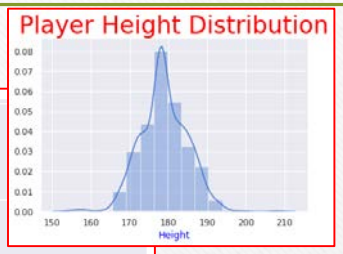
The above analysis depicts that the quality of the game have not been affected much over seasons and there is an optimal balance among teams w.r.t their game strategies (Defensive, Attacking and Underperformer) so there is ample potential for coming seasons to be more exciting and interesting.

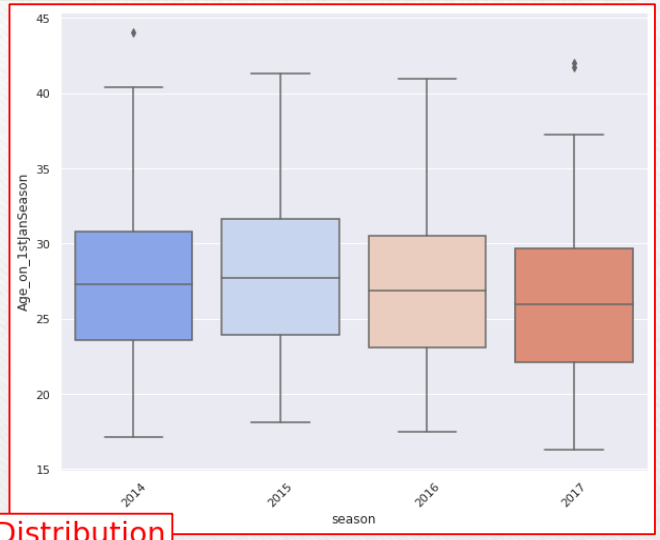
Quality Of the Players

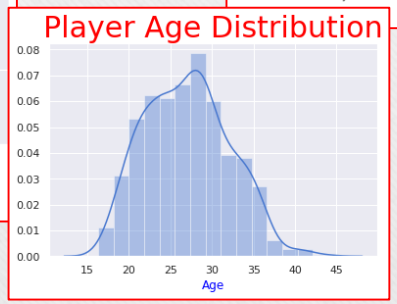
Player bio analysis

Quality of the Players:
How does the Player's anthropometry distributed?



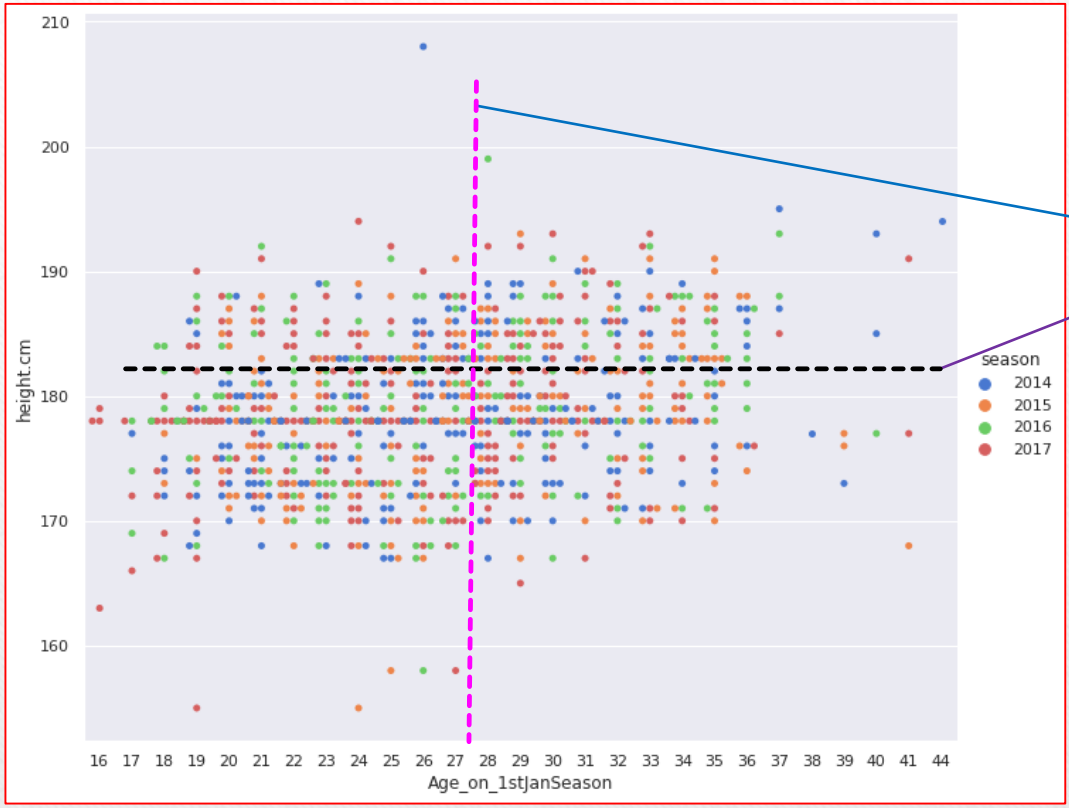






Distribution shows mostly young players have been chosen in 2016 and 2017 so upcoming season could be more exciting.

Quality of the Players:
How does the Player's anthropometry distributed?



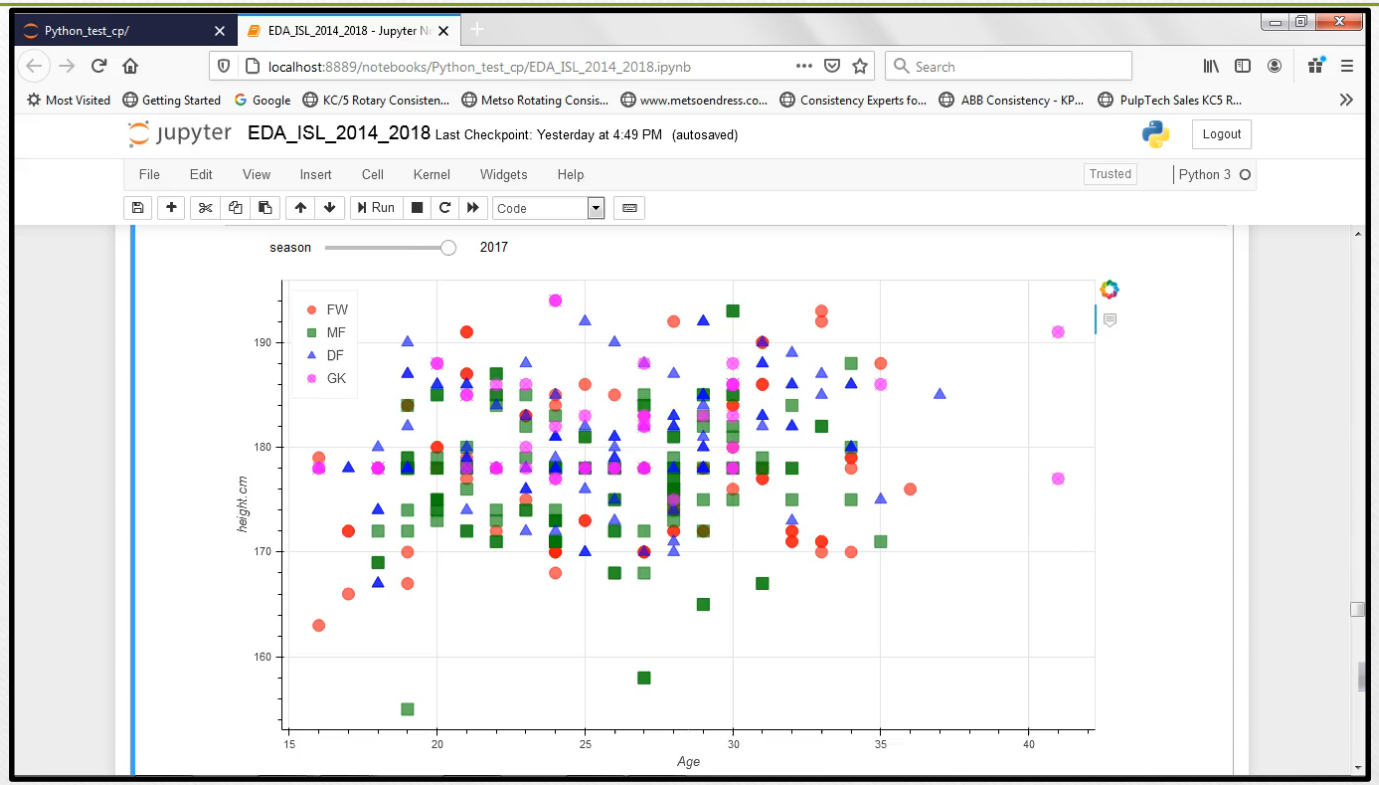
World cup 2018 benchmark
Age (avg.) : 27.4 yrs.
Height (avg.) : 181.7 cm

(# Source : CIES Football Observatory)

	height.cm	Age_on_1stJanSeason
count	937.000000	937.000000
mean	178.897545	26.521878
std	5.996630	5.033414
min	155.000000	16.000000
25%	175.000000	23.000000
50%	178.000000	26.000000
75%	183.000000	30.000000
max	208.000000	44.000000

Distribution depicts that the players selected for the tournament are in line with the global benchmark.

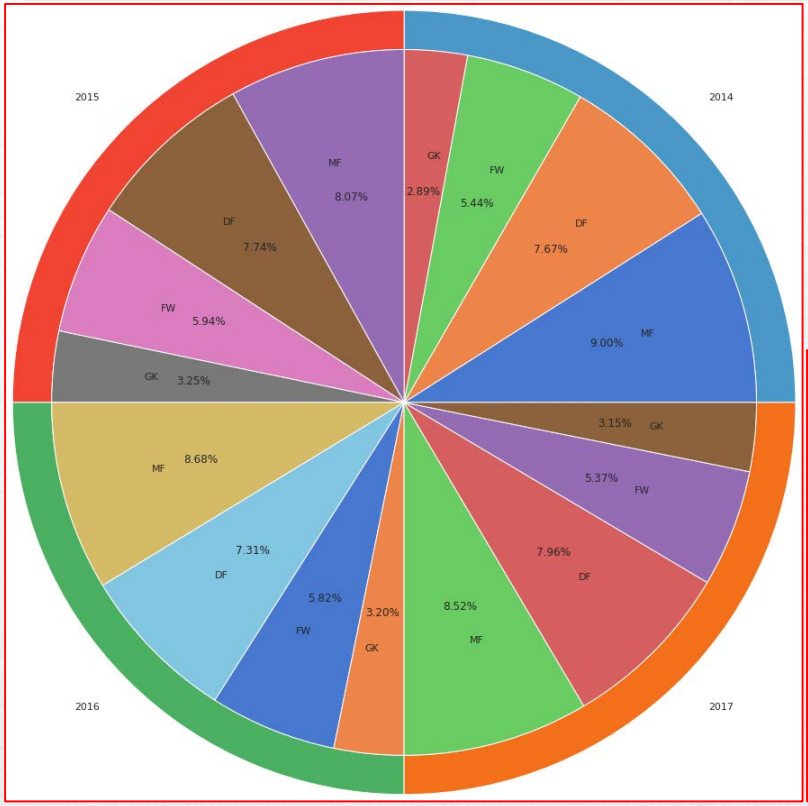
Quality of the Players:
How player's anthropometry are mapped in each season?

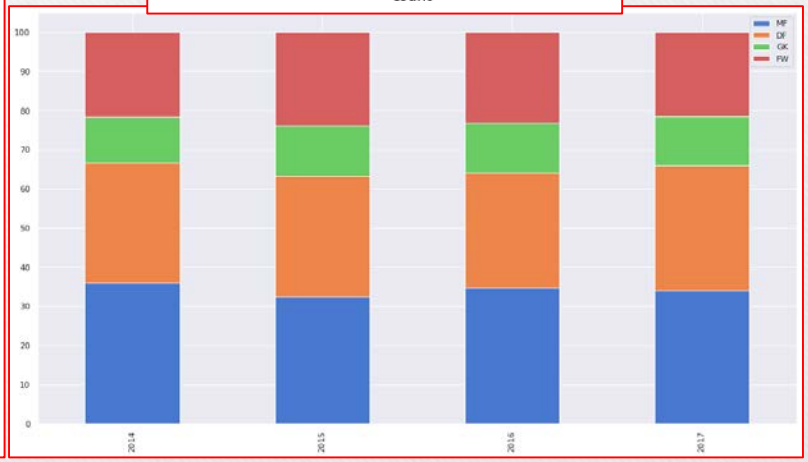


Players are well distributed w.r.t physical parameters/ Strength; so the selection of player does not affecting the quality of the game and hence popularity.

Quality of the Players:

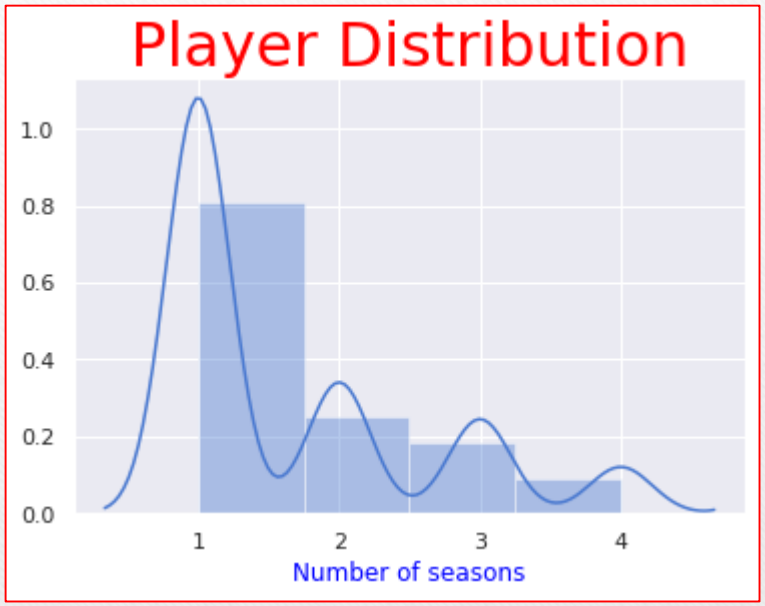
How the players are chosen across positions in each seasons?

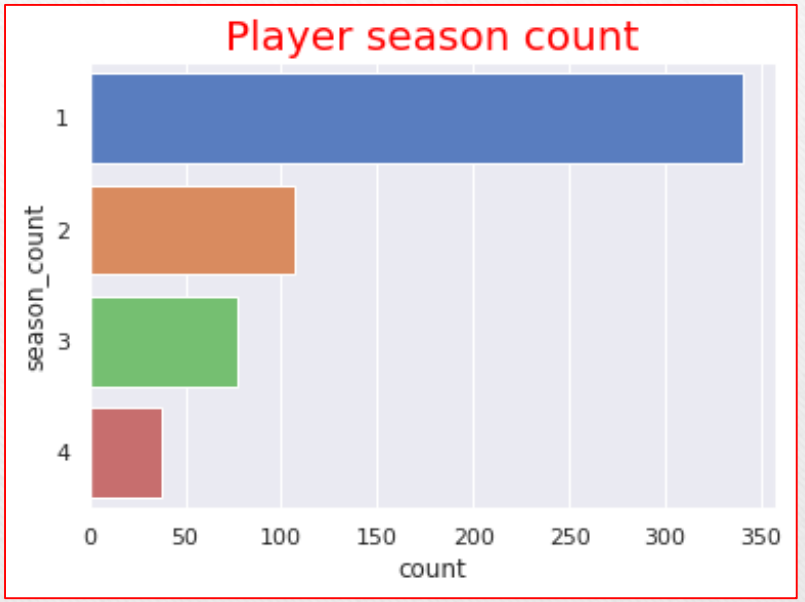




Balanced distribution observed among positions and it is not varying much throughout the seasons.

Quality of the Players:
Whether players are getting replaced in each season?



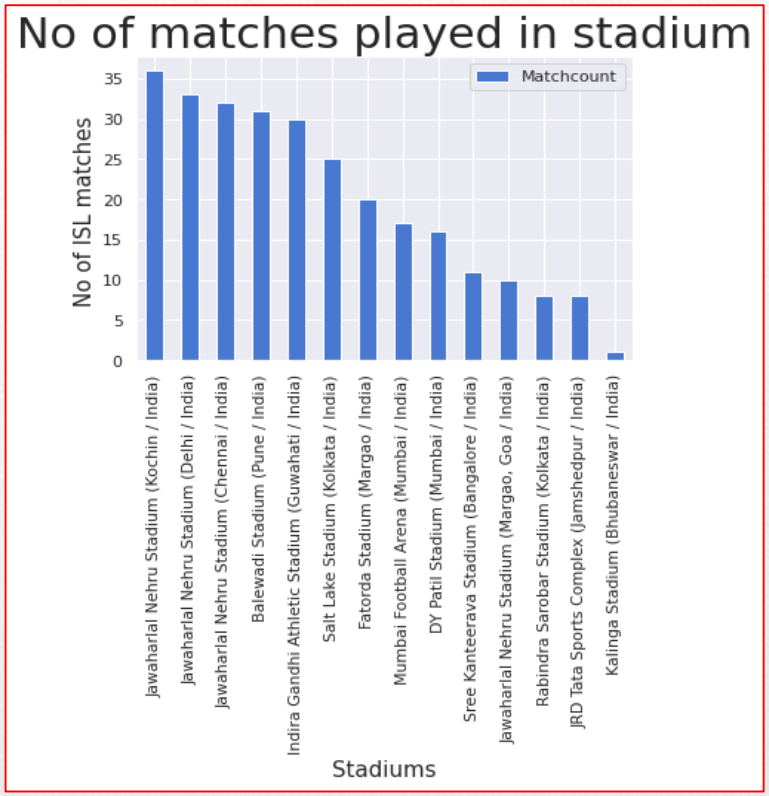


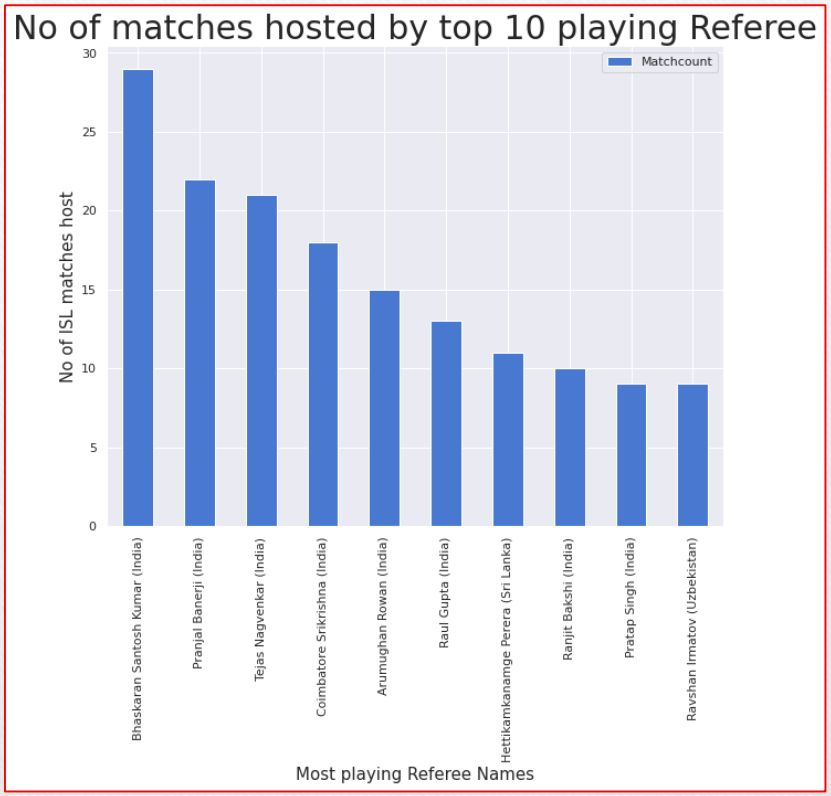
It seems the players are getting replaced each season so the maximum occurrences of players are in one season mostly; may cause confusion over fans and hence the popularity may be declining.

Infrastructural & Organizational aspects

Analysis on the administrational topics

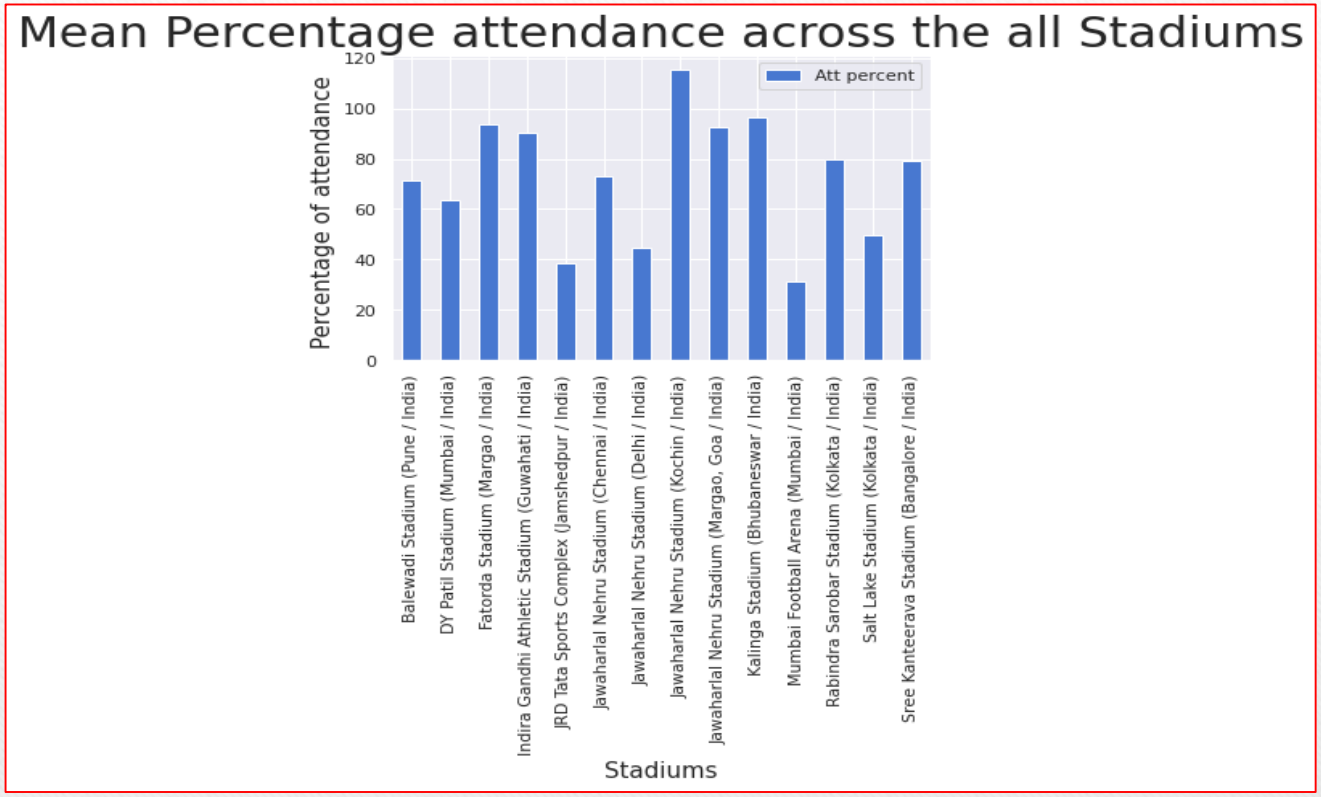
Infrastructural & Organizational aspects:
How the stadiums & referees are being utilized for the games?





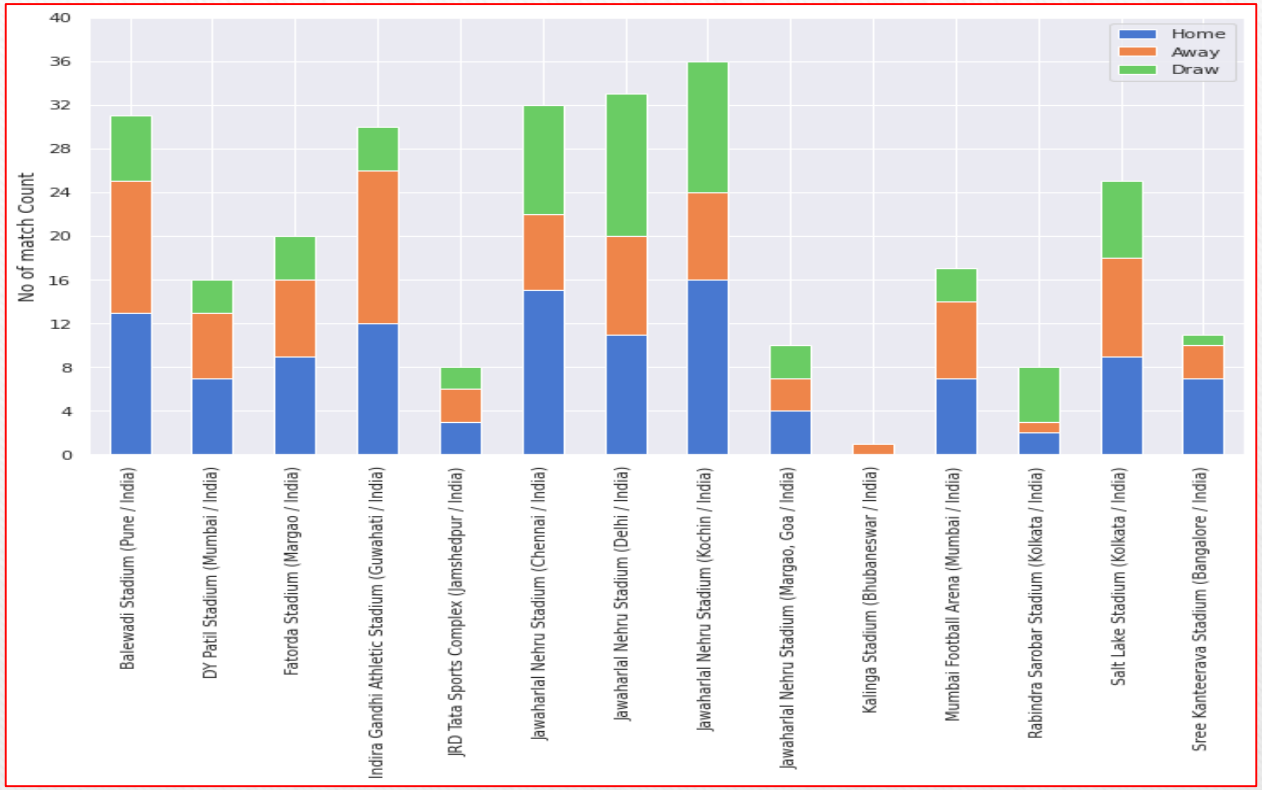
Each teams are well equipped with home stadium.
Referees are well experienced and hence top 10 of them have been selected for at least 8 matches or more.

Infrastructural & Organizational aspects:
How does the attendance varies across all stadiums?



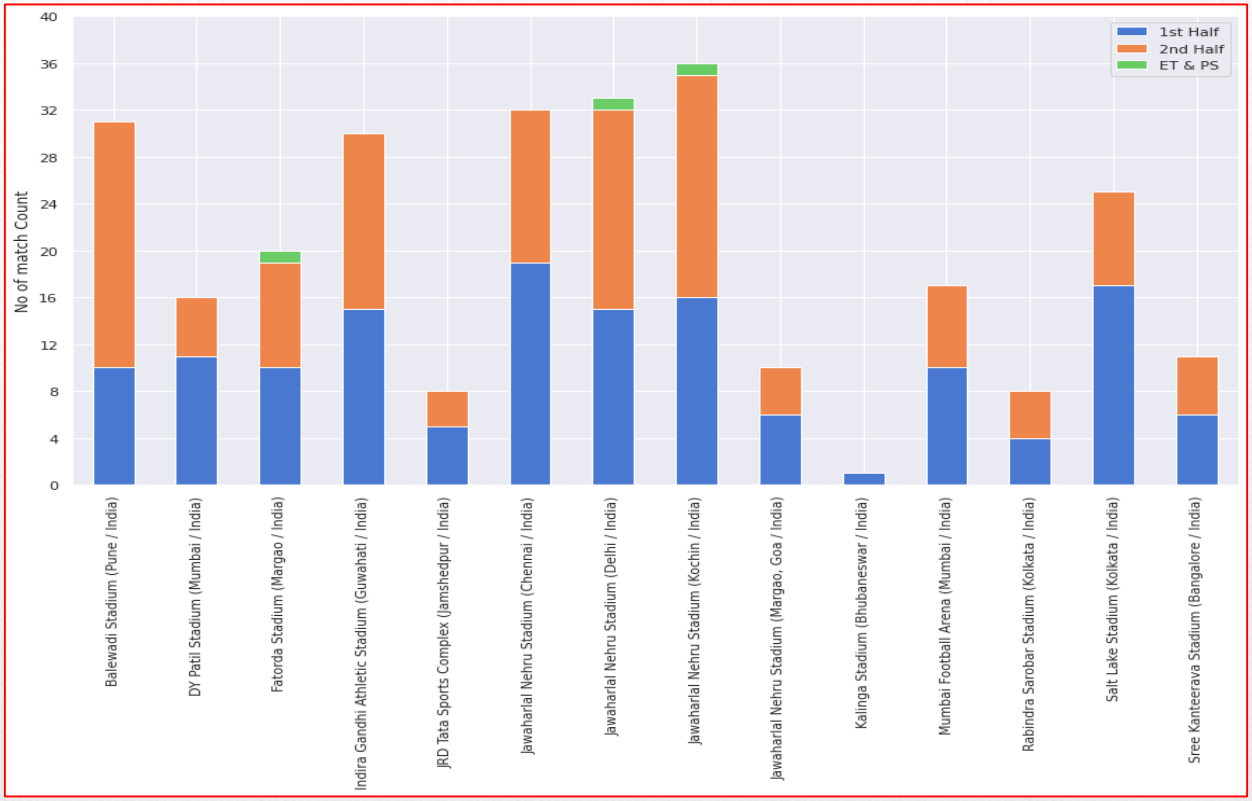
JRD, Jawaharlal Nehru (Delhi), Mumbai Football arena and Salt Lake Kolkata are having **mean attendance of 38%, 44%, 32% and 49% respectively**. Root cause analysis needs to be done further.

Infrastructural & Organizational aspects:
How does the team perform across all the stadiums?



Jawaharlal Nehru (Delhi) and Rabindra Sarobar Stadium Kolkata have been experienced most **Drawn** matches. Balewadi (Pune) and Indira Gandhi Athletic Stadium (Guwahati) have been experienced most **Away team wins**. The same for other stadiums are well balanced.

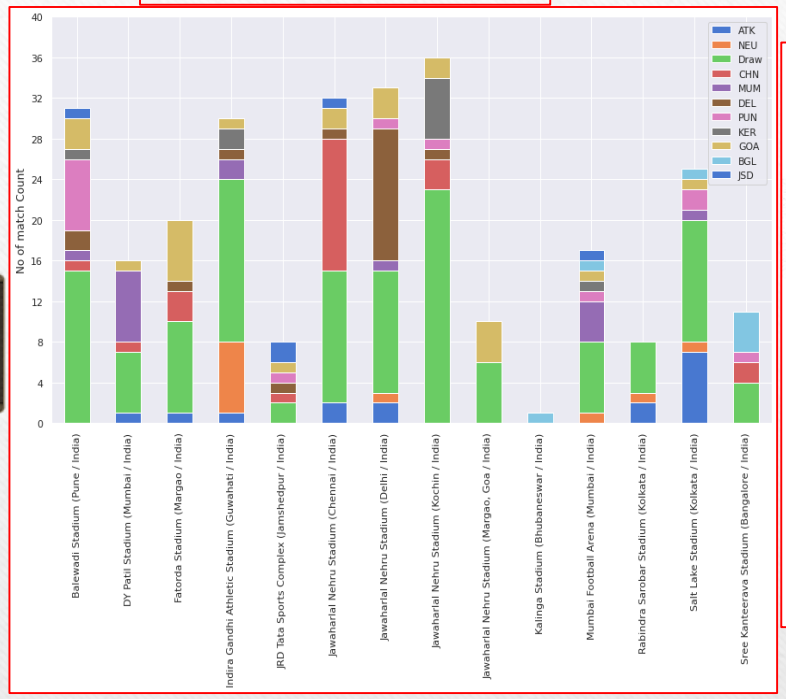
Infrastructural & Organizational aspects:
How does the team perform across all the stadiums?



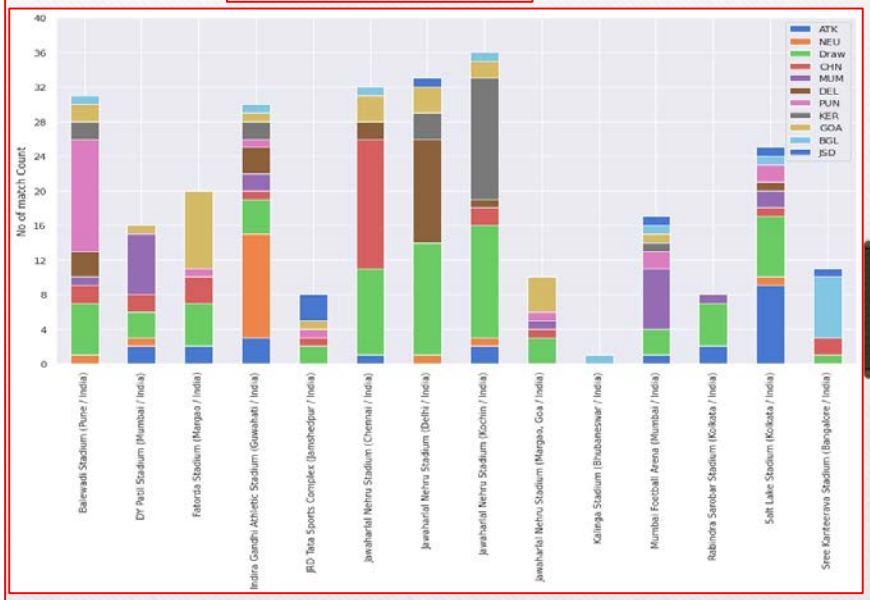
Balewadi (Pune), Jawaharlal Nehru (Kochin) and Jawaharlal Nehru (Delhi) have been experienced most **2nd half decider**. Other stadiums have been experienced with mostly **1st half decider**. Fatorda (Margao), Jawaharlal Nehru (Kochin) and Jawaharlal Nehru (Delhi) have organized the Playoffs.

Infrastructural & Organizational aspects:
How does the team perform across all the stadiums?

Half time Results



End Results

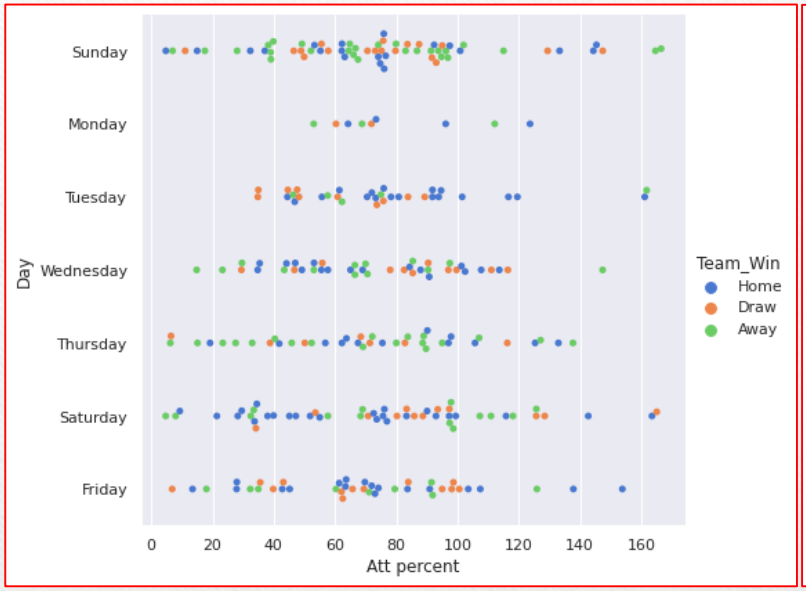


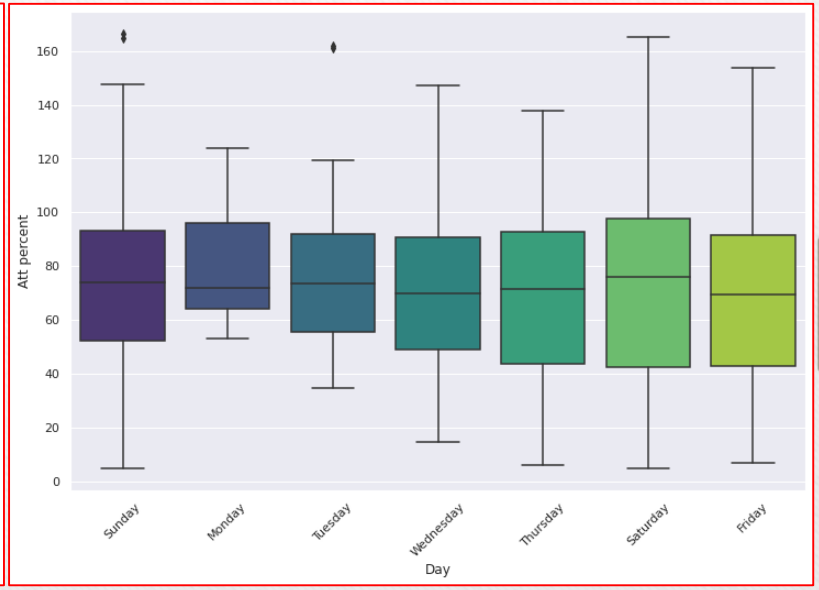
PUN, KER, CHN and NEU are very strong at **Home**. ATK have won most **Away** matches as compared to other teams.

Strategical & Geographical/Regional aspects

Analysis on the Tournament strategies

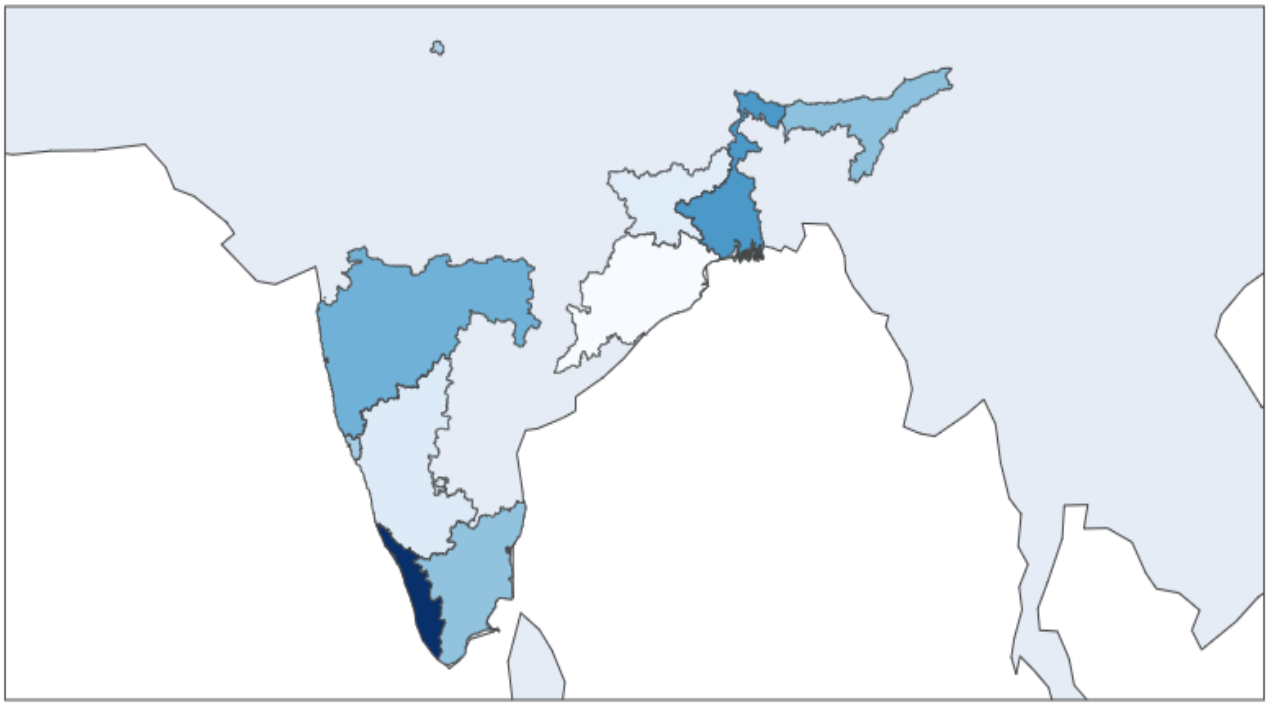
Strategical & Geographical/Regional aspects:
How weekdays/weekend are influencing the attendance?





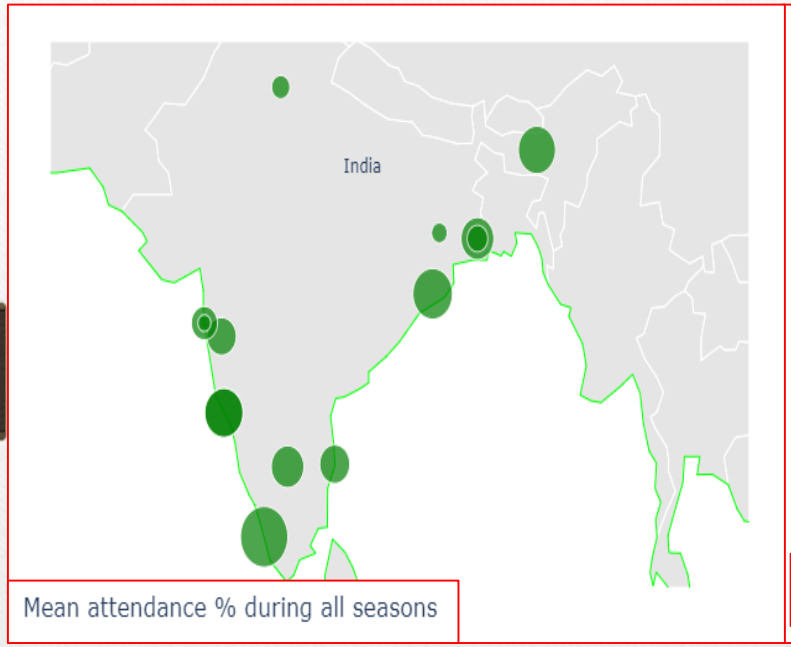
Attendance % have not influenced much neither the Home Team win nor the match day being fallen on a weekday or weekend.

Strategical & Geographical/Regional aspects:
How the geographical factors are influencing the popularity?

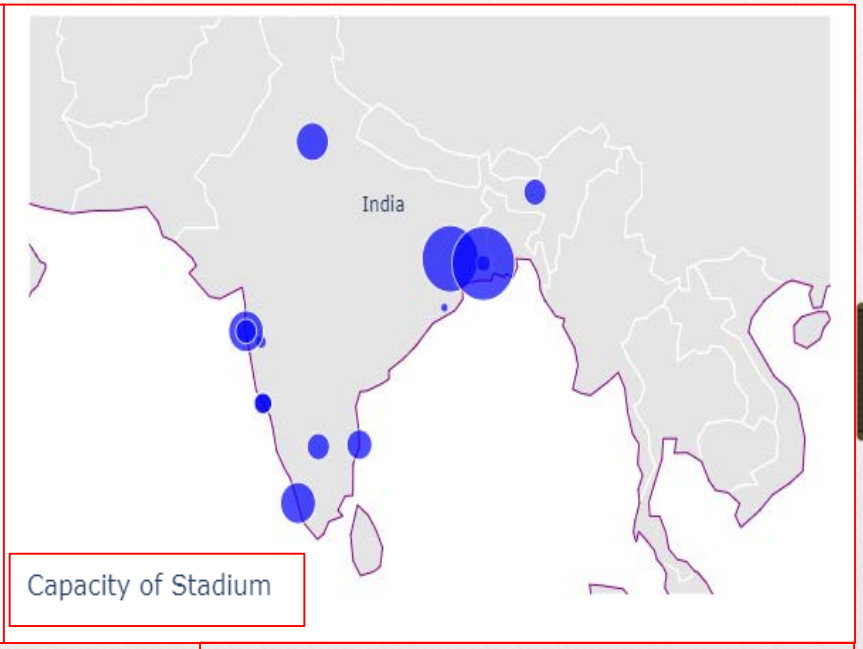


Western & Northern part of India are yet to be **explored**.

Strategical & Geographical/Regional aspects:
How the geographical factors are influencing the popularity?



Mean attendance % during all seasons



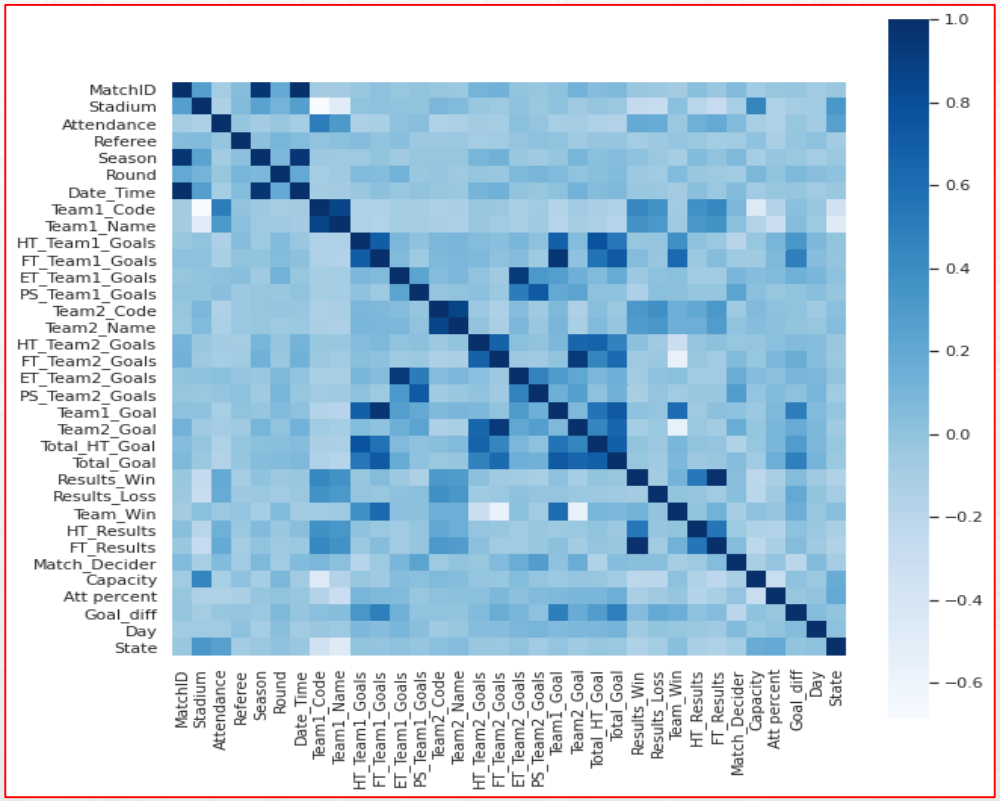
Capacity of Stadium

Kalinga Stadium and Balewadi Stadium are having very low stadium capacity.

JRD, Jawaharlal Nehru (Delhi), Mumbai Football arena and Salt Lake Kolkata are having poor mean attendance %. Root cause analysis needs to be done further.

Summary:

Correlation among all the factors summarized Label-Encoded:



All the discussed topics are being summarized here by means of a **correlation heat-map** for easy reference.

CONCLUSION

- **Game qualities** are at similar & acceptable level throughout all the seasons; so it may not be potentially impactful for the drop in popularity of ISL. Hence the same must be improved or at-least to be maintained in coming seasons.
- **Player distribution** w.r.t position and qualities are also acceptable but retaining of players in same team could be a key challenge in order to re-assure respective fan's emotion hence popularity of the tournament.
- Some of the stadiums are having poor mean attendance (%); more **deep dive needs to be carried** out by the respective State Association in order to have a root cause analysis.
- The tournament is mostly being organized at Eastern and Southern parts of **India**.

Based on the above conclusions we can infer some of the actionable insights further.

ACTIONABLE INSIGHTS

- Some amendments in player's contracts with respective teams may be exercised in order to retain the players at-least up-to two seasons, hence assuring fan's trust.
- **JRD, Jawaharlal Nehru (Delhi), Mumbai Football arena and Salt Lake Kolkata** are having very poor mean attendance % so investigation w.r.t infrastructure, pricing of tickets, logistic/transportation and promotion of events are being recommended and subsequently corrective actions must be taken in order to motivate and attract fans.
- Considering multidimensional regional aspect, **some strategic cities in the Northern and Western** parts of India may be explored for arranging the tournament in coming seasons. However, prior study may be envisaged in order to identify those locations.
- As we have seen the attendance % does not vary much between weekdays and weekends so **OTT platform may be explored** with minimum subscription in order to reach more fans during the emerging digital era; with this more regional teams and more number of games can be planned within same time frame by distributing more weekday matches. However, this might be decided based on prior analysis, which is not been carried out during this study.

The study has been solely carried out over data set of results of all the matches, players & stadium information. Hence, it does not contain any financial analysis of the tournament.