1. **Important Questions**

i) Which team had weak distributions and were highly dependent on a few players?

ii) What is the ratio of average to Strike Rate?

iii) Do teams prefer SR over average while picking their players?

iv) Which player was the most important for a team?

v) Was there any team with a significantly higher number of better batsmen than other teams?

vi) Is Runs an Important Factor while measuring a batter's performance?

vii) Which players were underused by a team?

viii) Which team had a higher concentration of overpaid players?

ix) Do players with more sixes have lower average?

x) What was the most important factor while picking a player?

xi) Do team owners look at hundreds and fifties of players?

xii) Can players be effective even with low average but high strike rates?

xiii) Which players were cheaper based on their General Batting Performance Index (2*SR+1.9*Average+1*Runs+1.5*Fours+1.75*Sixes+20*Fifties+25*Hundreds)?

Which players justified their price tag?

xiv) Which performance measure is preferable for a team owner?

xv) Which team had the best ratio of salary to performance?

2. **Performance Measure**

i) General Batting Performance Index

Comparison on (2*SR+1.9*Average+1*Runs+1.5*Fours+1.75*Sixes+20*Fifties+25*Hundreds)

Based on Salary:

Comparison on (2*SR+1.9*Average+1*Runs+1.5*Fours+1.75*Sixes+20*Fifties+25*Hundreds)/Salary

ii) Impact Factor

Ratio of Fours+Sixes(Boundaries) vs Runs

iii) Consistent Performance

Comparison on Avg vs SR

iv) Fast Milestone Scoring

Fifties+Hundreds(Milestones) vs Boundaries(Fours+Sixes)

v) Heavy Hitters

Ratio of Sixes vs SR

3. **Hypothesis and Experiments**

i) Teams prefer Higher SR over Average and pay more for them

ii) Runs are no longer an important factor
iii) high strike rate usually leads to lower average.
iv) Teams don't value Hundreds or Fifties
v) Boundaries are an important factor while picking a player

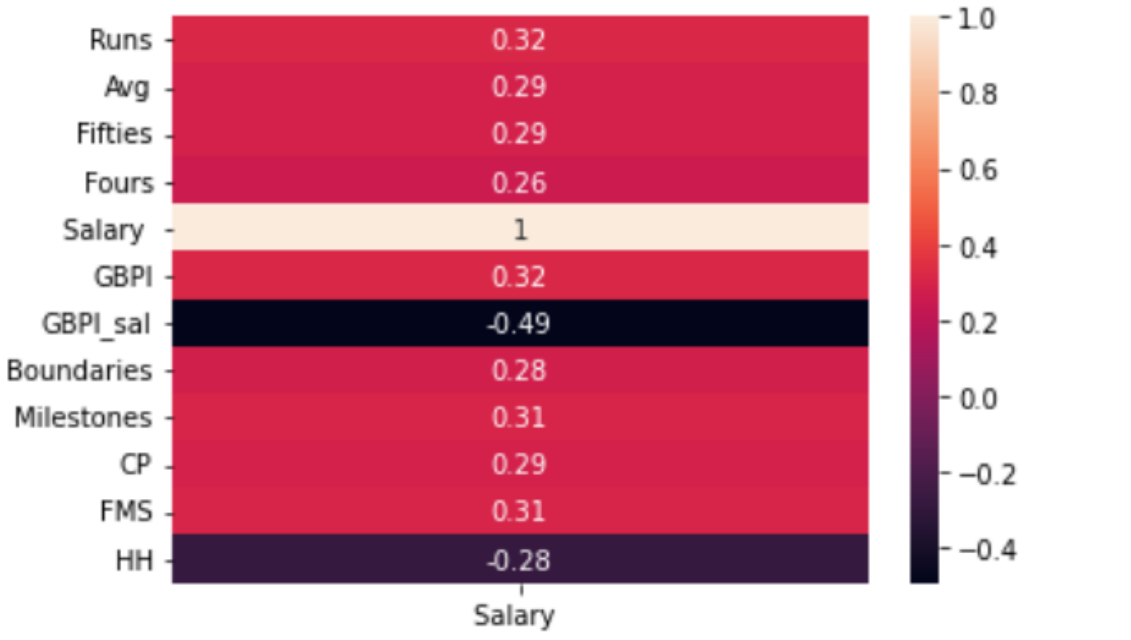
```
5]  players_df[['SR','Salary ']].corr()
```

	SR	Salary
SR	1.000000	0.105567
Salary	0.105567	1.000000

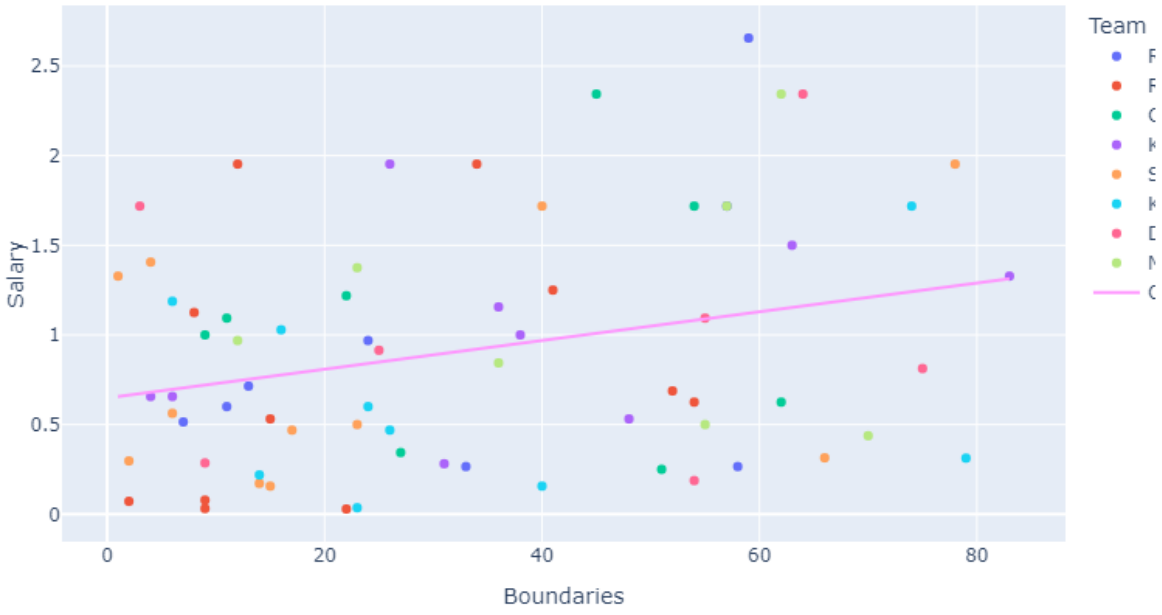
```
5] players_df[['Avg','Salary ']].corr()
```

	Avg	Salary
Avg	1.000000	0.291995
Salary	0.291995	1.000000

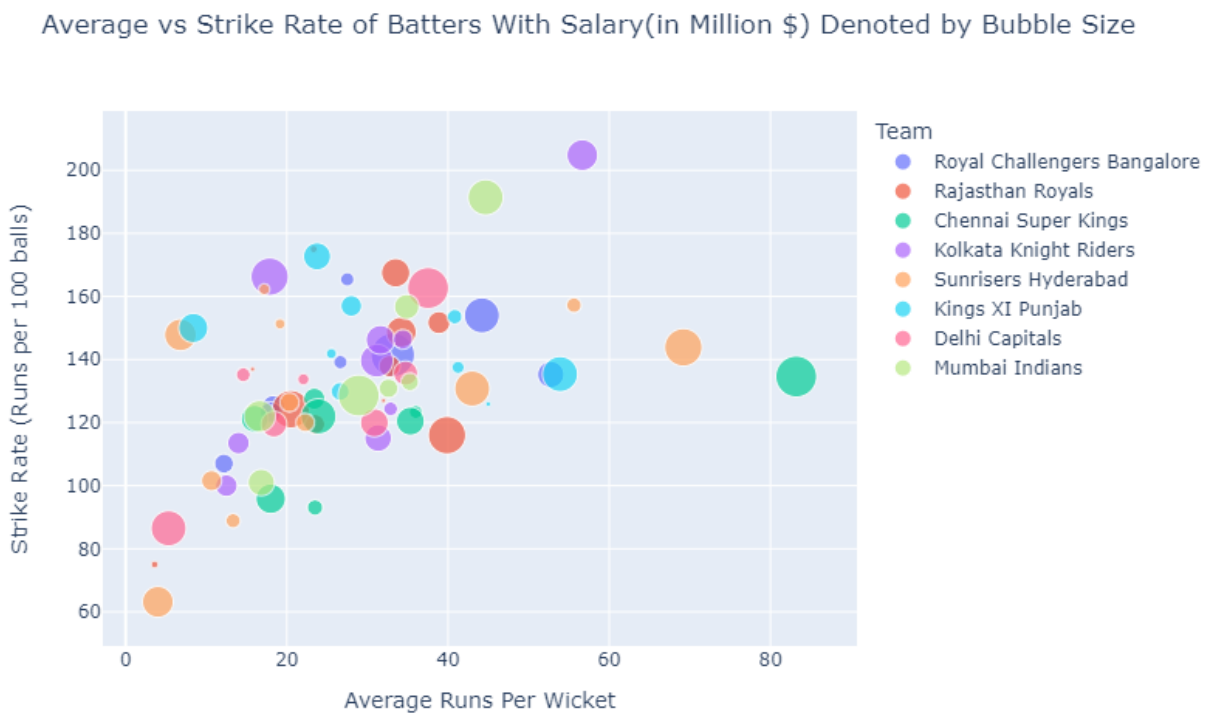
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f97b64f8790>
```



Boundaries vs Salary (in Million $) by team



Milestones vs Salary (in Million $) divided by General Batting Performance Index



i) H1 has been disproved using fig 1 and fig 2 as they show more correlation between average and salary compared to Strike Rate and Salary. Teams are surprisingly still valuing players with higher averages compared to players with higher Strike Rates.

ii) H2 stands true as runs is the highest correlated primitive variable with salary of a player at 0.32. Teams still value runs even if they come at a slower pace.

iii) H3 is inconclusive as the general trend in fig 5 shows us that with the increase in average, there is also an increase in Strike Rate. However, players with one of the highest averages tend to have an Strike Rate which is around the mean.

iv) H4 can be considered as true because fig 2 shows that milestones and fms are highly valued by teams at 0.31. Fig 4 also shows that players with more milestones generally are paid more as their GBPI is high too.

v) H5 is proven wrong because fig 2 shows a weak correlation between salary and boundaries. More importantly, fig 3 shows a trendline which is more or less straight. Hence, Teams aren't that interested in the number of boundaries hit by a player.

4. Predictive Modeling

Random Forest Regressor Model is used to predict the continuous variable which is the salary of a player. It gives a high accuracy of 90.12% which is very impressive. This model uses the variables which have a correlation of greater than 0.25 with the salary. The fig2 in the Hypothesis Section illustrates those variables.

```
rain_test_split(scaled_normal_players_df, normal_players_df_target, test_size=0.3,random_state=21)
```

```
09] model= RandomForestRegressor(max_depth =27, random_state =21,n_estimators=38,criterion='squared_error')
    #model=RandomForestRegressor()
```

```
10] model.fit(x_train,y_train)

    RandomForestRegressor(max_depth=27, n_estimators=38, random_state=21)
```

```
11] predictions=model.predict(x_test)
```

```
12] mse = np.mean((predictions - y_test)**2)
    score=model.score(x_test,y_test)
    print("MSE: ",mse)
    print("Score: ",score)

    MSE:  0.027232706578657166
    Score:  0.9012908231739506
```

Random Forest Classifier Model is used to predict the salary class of a player which is a discrete variable. A good accuracy of 76.1% is given by it. It uses all the continuous variables.

```
[118] from sklearn.ensemble import RandomForestClassifier
```

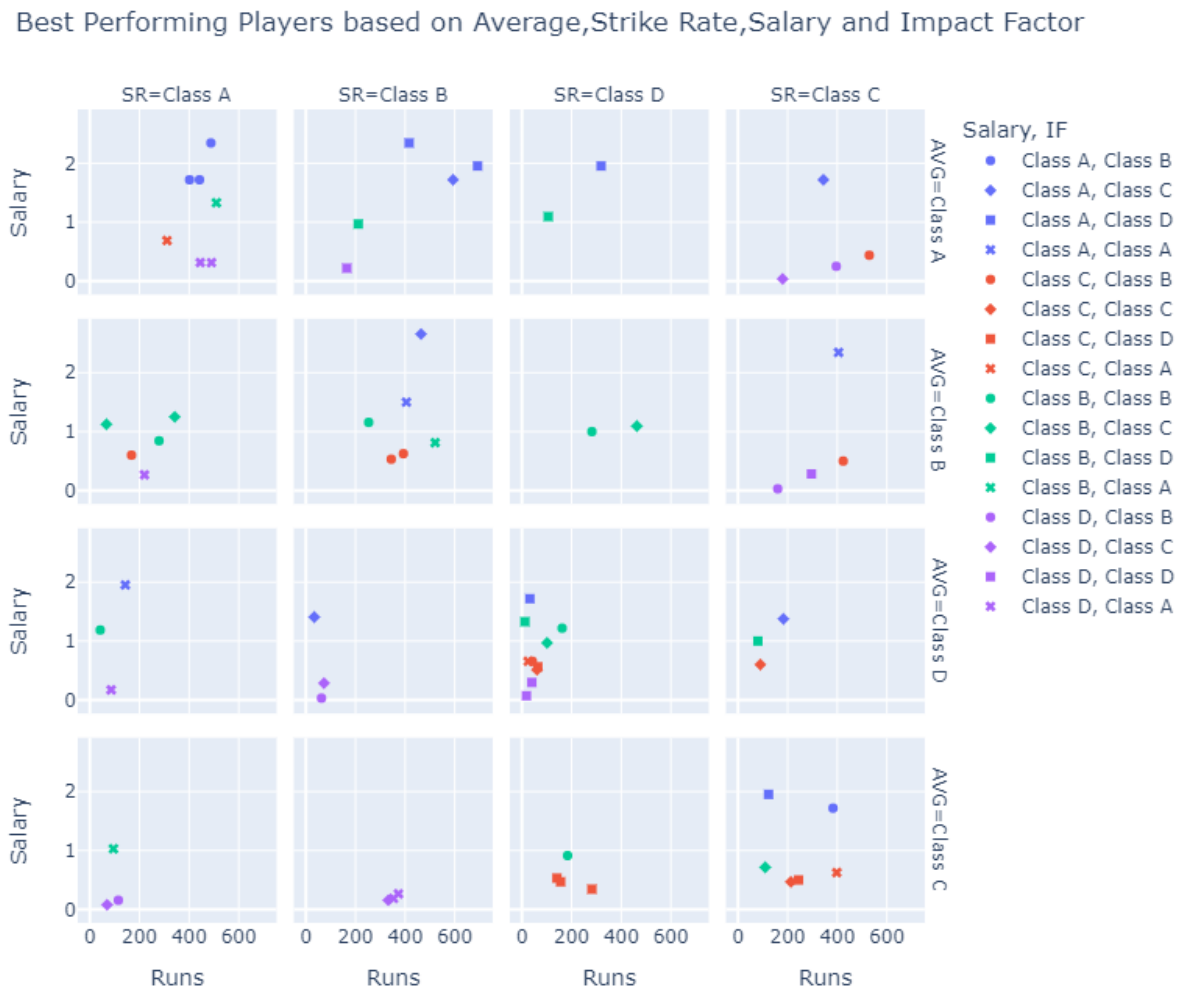
```
[162] ain, y_test = train_test_split(players_df.iloc[:,2:17], players_df['Salary_Encoded'], test_size=0.3,ra
```

```
[163] model=RandomForestClassifier(max_depth =27, random_state =21,n_estimators=38,criterion='entropy')
```

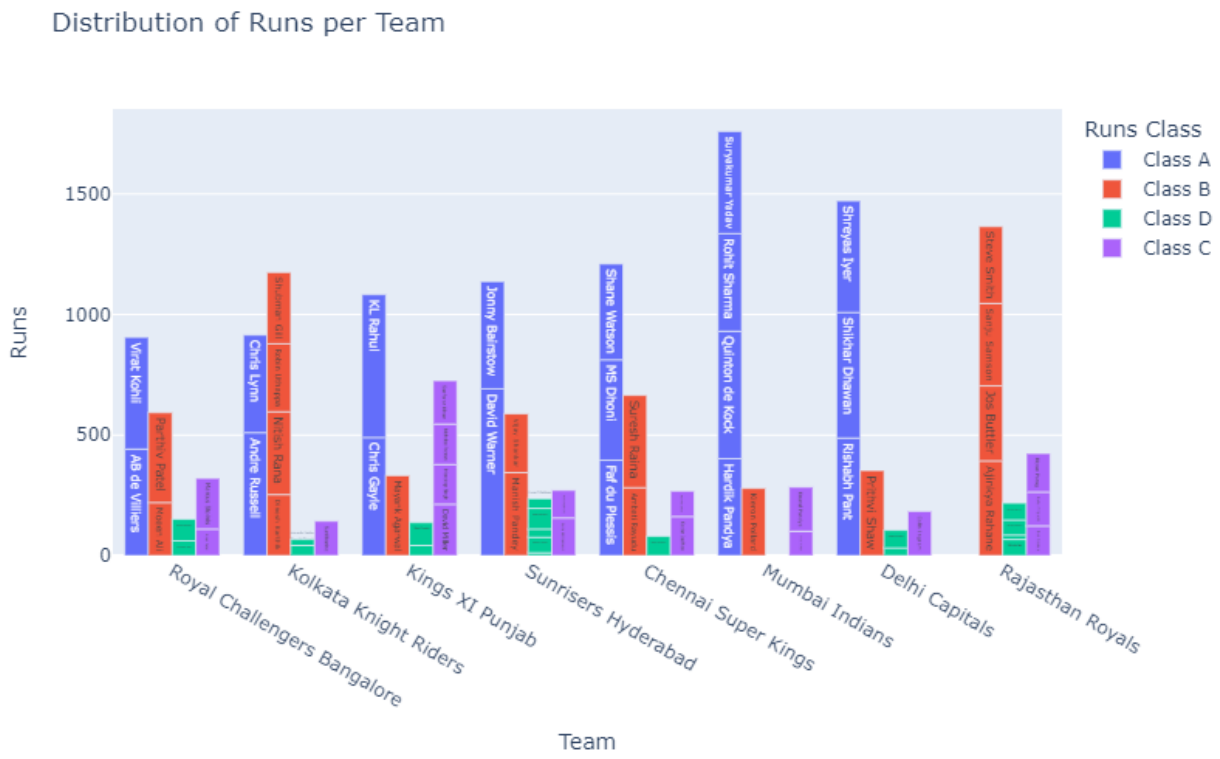
```
[164] model.fit(x_train,y_train)
      predictions=model.predict(x_test)
      score=model.score(x_test,y_test)
      print(score)

      0.7619047619047619
```

5. Plots and Visualization

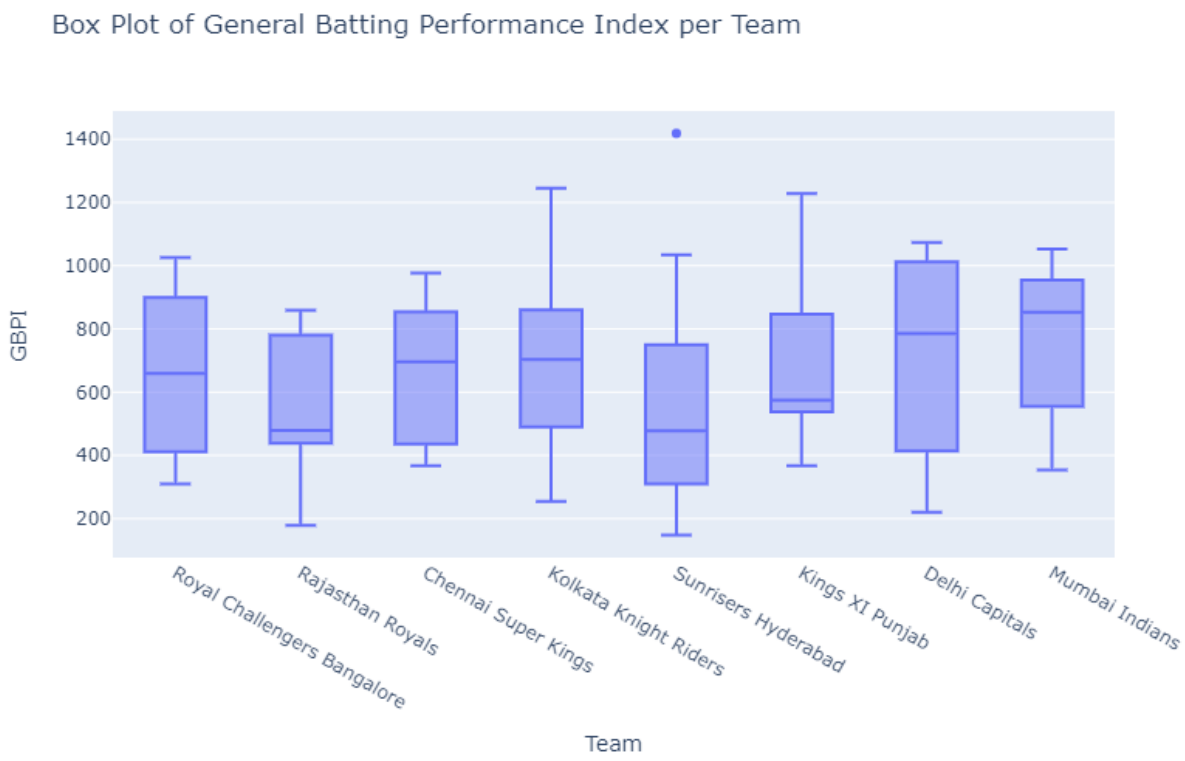
Plots



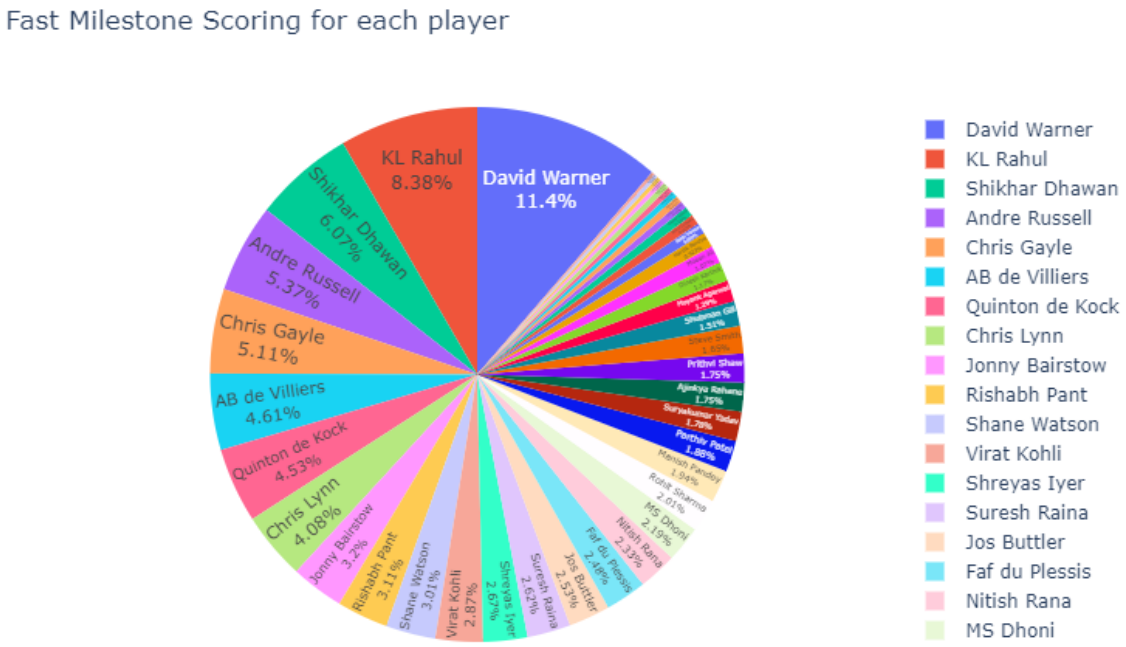
Johnny Bairstow and Chris Gayle provide the best in everything and are also the cheapest salary class. Andre Russell is slightly more expensive, but can probably be considered as the best batsman in the whole league. These three players should be most sought after in the next edition of the league.



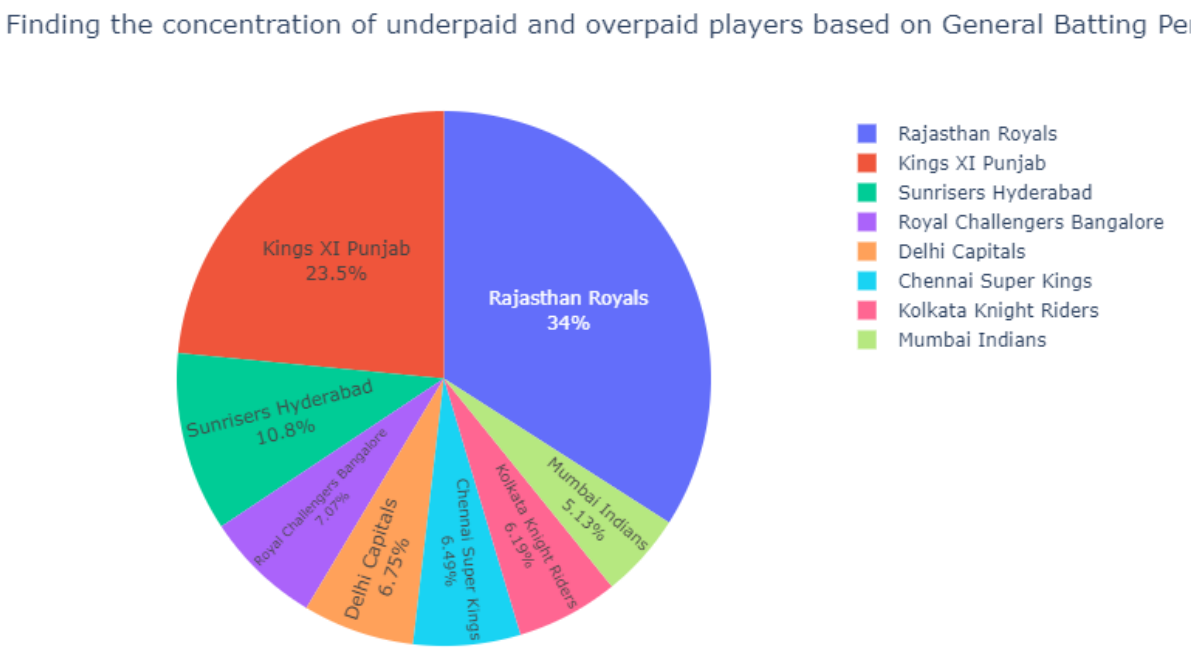
MI and DC have a good distribution of runs with more players in class A. RR has no player in class A which tends to tell us that they may have performed poorly in the season. In the next edition, RR must buy one of the players in class A category belonging to other teams if they are to improve. MI came in 5th and this may have been because of the lack of players in class D. If a couple of fringe players would have added a few runs here and there, they may have qualified for the playoffs.



One of the reasons for the terrible performance of the SRH in the 2021 edition could be the lowest mean for GBPI with the lone spark being David Warner who also had the highest GBPI out of all players. It can be observed that poor performing teams have a high difference between q1 and mean or a low mean. Moreover, this plot shows us that GBPI can be considered a good performance measure. Its correlation with the salary was the highest too.



David Warner has more than 10% of Fast Milestone Scoring out of all players. SRH can improve in the next edition if they improve their other players while retaining David Warner. Retaining him is a must. Moreover KL Rahul and Shikhar Dhawan are also great openers and teams who can splash the cash must go for them as they can score big at a fast pace.



RR and KXIP had underpaid players performing more which gives us a sense that their main batsmen who got high salaries were unable to perform nicely. It also shows us that MI had overpaid too many players. If they would have wisely used their funds, some more runs could have been enough for them to clinch a top 4 spot. For a good season, a team must have a fair balance of GBPI per salary. Teams with too high a percentage are going to perform the poorest and too low are gonna miss that extra factor.