# Cricket: Indian Premier League

Erica Cau
e.cau@studenti.unipi.it
Student ID: 545126

Simona Mazzarino
s.mazzarino@studenti.unipi.it
Student ID: 620022

Federico Mazzoni
f.mazzoni6@studenti.unipi.it
Student ID: 524324

## ABSTRACT

In this project, the IPL dataset was analysed, aiming to extract some interesting information about the data. In order to do so, we focused on the preliminary step of the Knowledge Discovery Process, and, in particular, on data understanding, data correlation and outlier detection [1]

## 1 INTRODUCTION

Nowadays, data can be generated by anything, especially on a large scale. In fact, thanks to the continuous development of technologies, everything is a potential source of data, from temperature sensors to social networks: everything can be used to gain more knowledge and unexpected insights about a specific field. This is the reason why it is important to know how to study and analyse them to gain some useful insights to better understand the world surrounding us. Those kind of analysis are divided in several steps, which allow to explore the data in many different ways using basic or advanced methods. In this project, we focused on the exploration and study of the IPL dataset, trying to understand the nature of the data, their distributions and the correlation between their features.

## 2 DATA UNDERSTANDING

The dataset is composed by two different subsets: *matches* and *deliveries*, the former containing data about all the matches of IPL played between 2008 and 2017, while the later containing statistics about all the played balls in the matches. First of all, it was necessary to explore those two subsets in order to gain a better insight about their features and records.

To do so, basic informations about the two datasets have been visualized. One problem that didn't go unnoticed was that in the *matches* dataset, the feature *Umpire3* was only composed by missing values and the column was therefore dropped. Then, it was also noticed that other features of *matches* had a very small number of missing values. The affected rows were therefore deleted. On the other hand, most of the values of *player_dismissed, dismissal_kind* and *fielder* features (from the *Deliveries* dataset) are missing values: however, from a semantic point of view the lack of those records is significant. Those values aren't actually missing, but their lack means that there wasn't a dismissed player in most situations. Thus, we didn't fill or drop the rows with NaN values because they gave us some kind of information about every single match.

An overview of the total set of features can be seen in Table 1.

Then, we deepened the general analysis of the numeric features by calculating the mean, the standard deviation, the minimum and maximum value, the first, the second and the third percentile of the numerical features. While for the categorical ones we visualized the total number of values, the number of unique values, the mode and its frequency. Subsequently, we calculated some synthetic measures about won matches, matches played in each season, the role of *player of match* and the role of *batsman*. We reported the results in Table 2.

| | |
|---|---|
| **Matches** | id, season, city, date, team1, team2, toss_winner, toss_decision, result, dl_applied, winner, win_by_runs, win_by_wickets, player_of_match, venue, umpire1, umpire2, umpire3 |
| **Deliveries** | match_id, inning, batting_team, bowling_team, over, ball, batsman, non_striker, bowler, is_super_over, wide_runs, bye_runs, legbye_runs, noball_runs, penalty_runs, batsman_runs, extra_runs, total_runs, player_dismissed, dismissal_kind, fielder |

**Table 1: List of all the features in the two subsets**

| | Won Matches | Matches played in each Season | Player of Match | Batsman |
|---|---|---|---|---|
| **mean** | 45.21 | 63.60 | 3.15 | 326.38 |
| **std** | 31.78 | 7.50 | 3.37 | 617.31 |
| **min** | 5.00 | 57.00 | 1.00 | 1.00 |
| **25%** | 12.25 | 59.00 | 1.00 | 18.00 |
| **50%** | 52.00 | 60.00 | 2.00 | 71.00 |
| **75%** | 72.25 | 69.75 | 4.00 | 285.00 |
| **max** | 92.00 | 76.00 | 18.00 | 3494.00 |

**Table 2: Synthetic Measures of some interesting features**

Even at a first glance, from the table can be noticed that the values of the mean in the first two columns are pretty significative because they are not far from the value of its 50th percentile, i.e. the median. Moreover, looking at the value of their standard deviation, we can easily say that the distribution of won matches and of matches played in each season was quite around the median value. On the other hand, the distribution of the features *Player of Match* and *Batsman* is pretty unbalanced (both of them have high values of Kurtosis and Skewness, which means that the distribution isn't normal and very skewed on the left). In the feature *Player of Match*, we see that the title was gained several times by a small number of players. On the other hand, a significatly higher amount of player won the title a small amount of times. Same thing happens with the role of *batsman*: some players cover that role a lot (up to 3494 times!). Thus, it can be said that the two features have a Zipf's distribution.

Finally, we also decided to observe whether the distribution of the two numeric features of *matches* was normal. In order to do so, we calculated calculating the *Kurtosis* coefficient of *win_by_runs* and *win_by_wickets*. As it can be seen in Table 3, the Kurtosis value

[1] **Dataset**: https://www.kaggle.com/manasgarg/ipl

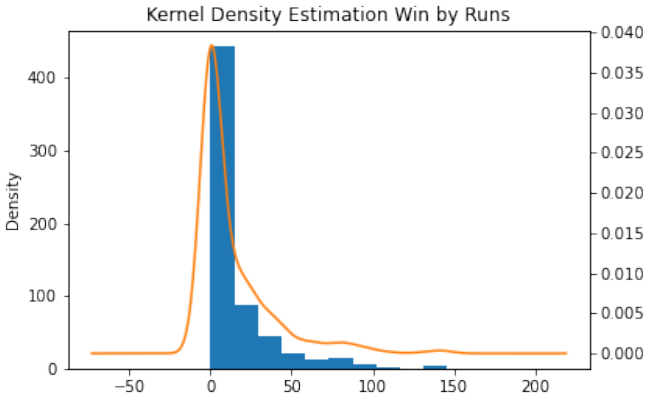


**Figure 1: Kernel Density Estimation of *win_by_runs***

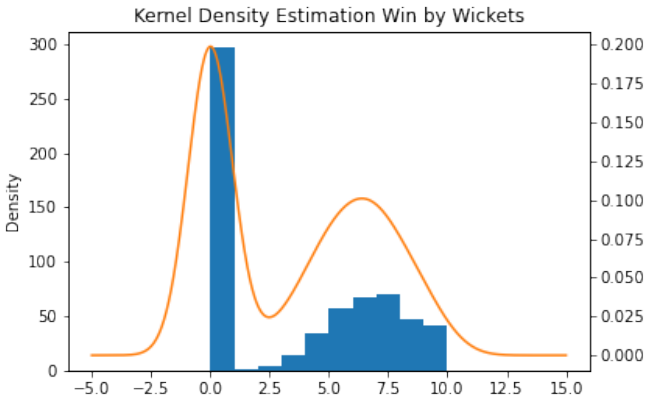


**Figure 2: Kernel Density Estimation of *win_by_wickets***

| | Win by Runs | Win by Wickets |
|---|---|---|
| **Kurtosis** | 7.45817 | -1.53173 |
| **Skewness** | 2.52206 | 0.26756 |

**Table 3: Values of Kurtosis and Skewness coefficient for the numeric features of *Matches***

of both features wasn't around zero, so we rejected the hypothesis of normal distribution. That's why we measured the *Skewness* coefficient that shows the asymmetry of the distributions.

The second step of the work was the visual analysis of the distribution of the most interesting features. The first one was the number of matches won by each team of IPL, that showed us the most winning team, *Mumbai Indians*, that won the 14.7% of matches, followed by *Kolkata Knight Riders* and *Chennai Super Kings* (Fig. 3). Therefore, we further explored the feature, by plotting the number of matches won by each team year by year (Fig. 4). This showed also that the year with fewer number of matches was 2014, as it can also be seen in Fig. 5 ; moreover, we can also see that the team

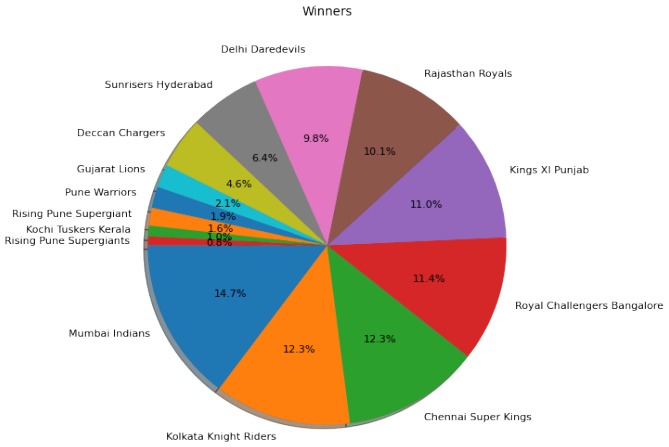


**Figure 3: Matches won by team**

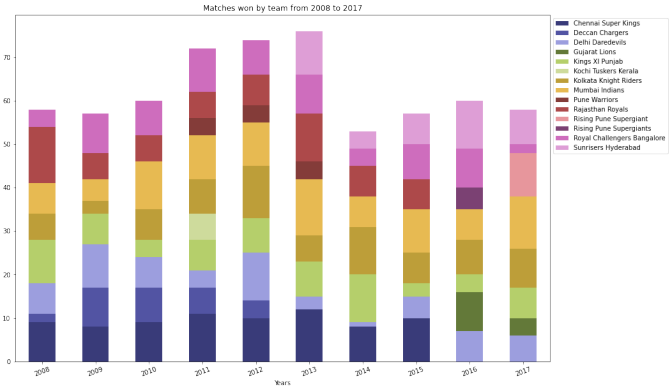


**Figure 4: Number of match won by team for each season**

*Sunrisers Hyderabad* started to play in IPL only since 2013, and that the name of *Rising Pune Supergiant* changed between 2016 and 2017, before being definitely suspended from IPL. It was also visualized the number of matches by city, that allow to discover that Mumbai hosted more than 80 matches, followed by Bangalore and Kolcata (Fig. 6).

Furthermore, we observed whether the toss decision (bat or field) influenced the result of the match. Comparing the plots of *Matches Won by Team* (Fig. 3) with the following one, which shows the toss decision by winner teams (Fig. 7), it can be seen that the two teams, *Kochi Knight Riders* and *Rising Prune Supergiant*, which always chose *field* as toss decision, are the last ones considering the number of matches won by team. So, we can say that there could be a mild relation between the toss decision and the final result of the match. Additionally, the analysis focused on discovering which was the most common dismissal kind in the 10 years-span time covered by the IPL dataset (Fig. 8). Apparently, 7438 cases of player dismissed from the match are registered in the dataset: most of them were batsman dismissed by being *caught*, *bowled* and *run out*. The less common dismissal methods, instead, were *obstructing the*

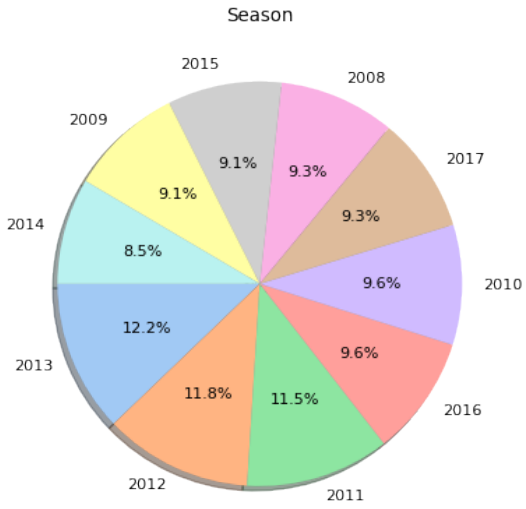


**Figure 5: Number of match for each season**

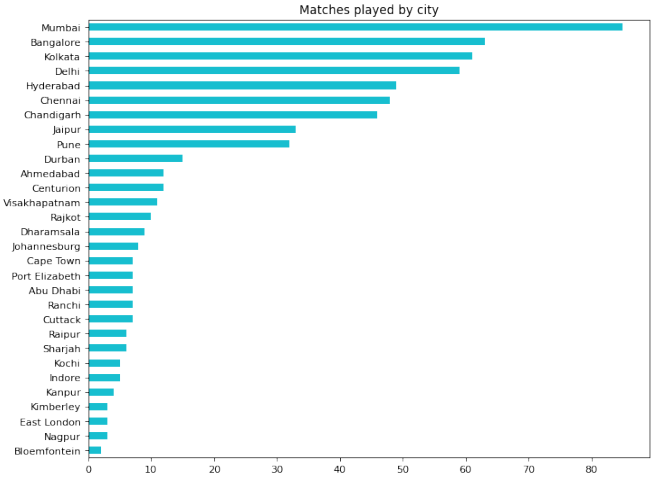


**Figure 6: Match played by city**

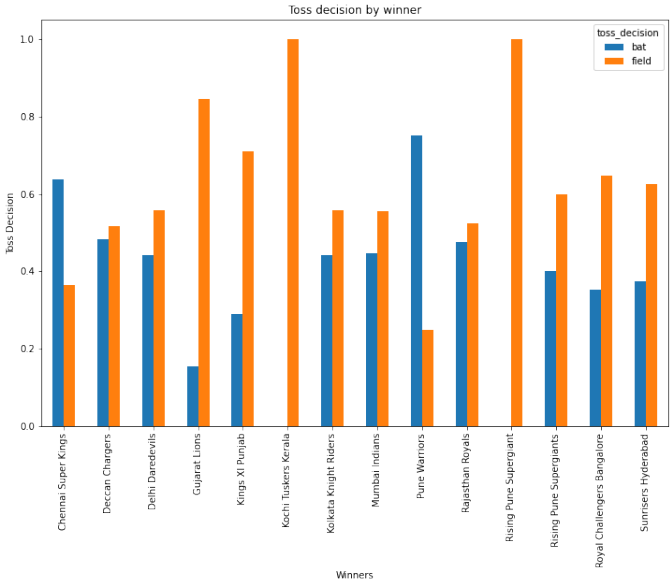


**Figure 7: Toss Decision by Winner**

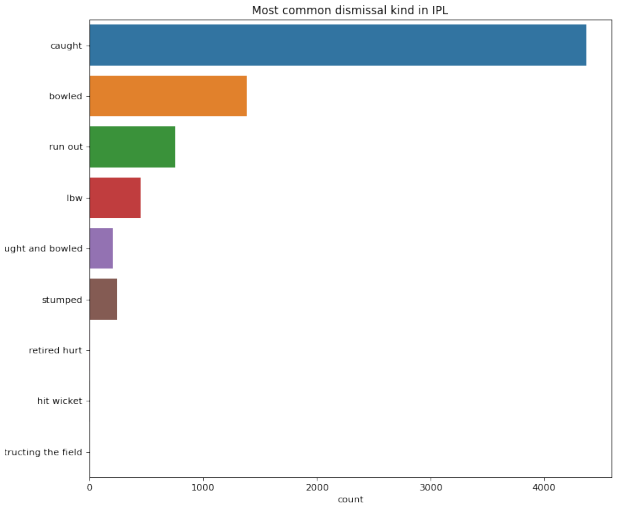


**Figure 8: Most common dismissal kind**

*field* (which occured one time) *hit wicket* and *retired hurt* (both occuring nine times).

Ultimately, we visualized the top 5 batsman and the top 10 players of the match (Fig. 9, 10): as we can see, the best batsmans, except for the second one (*V Kohli*), are also among the best 10 players of the match.

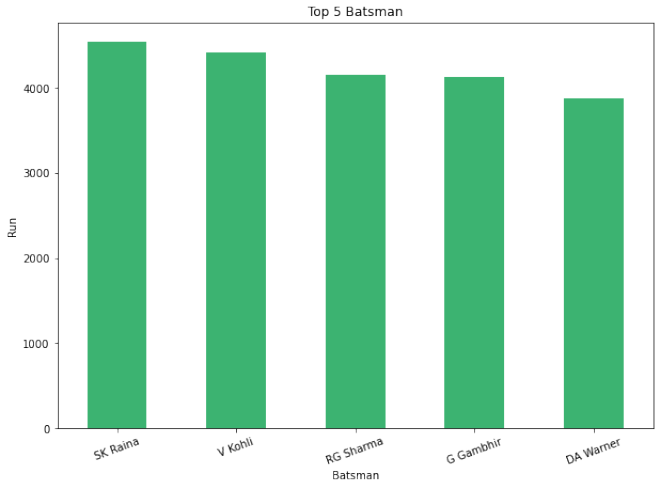


**Figure 9: Top 5 Batsman**

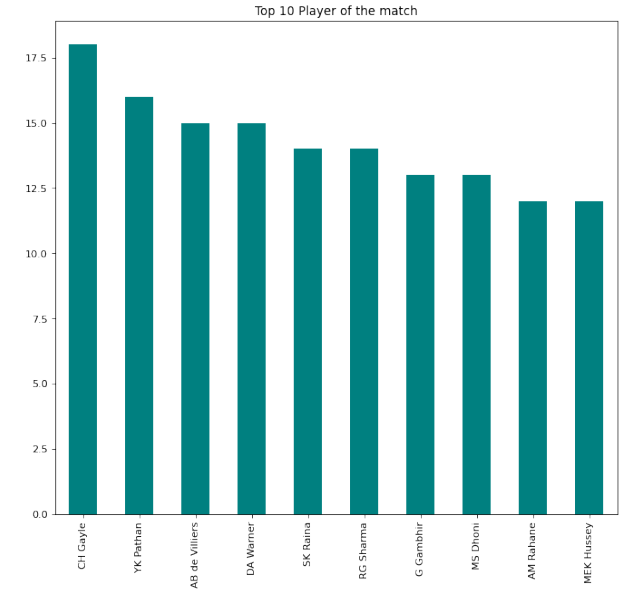


**Figure 10: Top 10 player of the match**

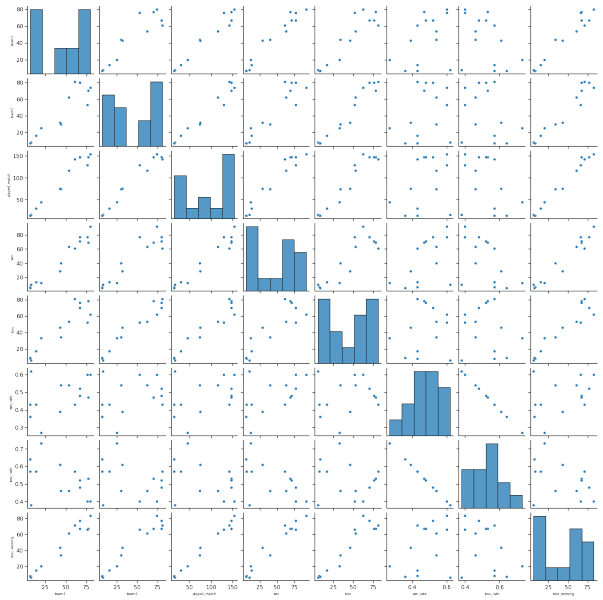

**Figure 11: Pairplot of the *Match_Stats* subset**

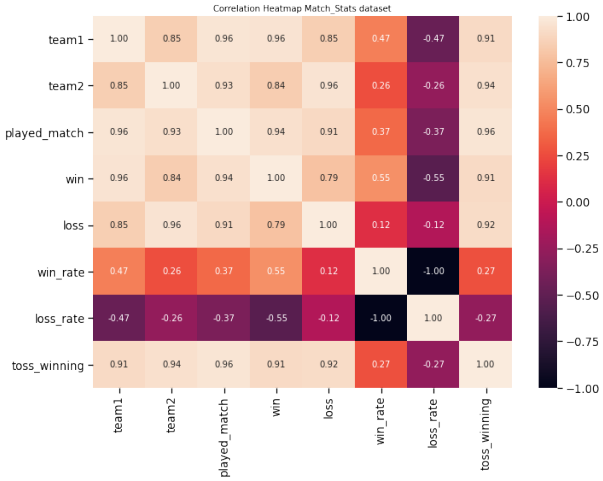


**Figure 12: Spearman's rank correlation of *Match_Stats dataset***

## 3 DATA CORRELATION

We tried to calculate the correlation in both dataset. The main problem was the nature of the features: in all the two datasets most of the variables were categorical or, while numerical, discrete in nature, so it was impossible to obtain meaningful (and useful) results. For this reason we created a new subset, called *Match_Stats*, in which we aggregated numeric values about the number of wins and losses of each IPL team. We defined some new features as well, such as *win_rate* and *loss_rate*, obtained by dividing the number of won (or lost) matches by the total number of played matches.
We proceeded by plotting the pairplot of the *Match_Stats* dataset, with whom we observed the distribution of the features, seeing that just *win_rate* and *loss_rate* presented an approximately normal distribution (Fig. 11). Thus, we considered to use the *Spearman's rank correlation coefficient*, instead of Pearson. The results obtained are reported below in Fig. 12. For the purposes of our exploration, we focused on the values of the correlation between *played_match* and *win*, and the one between *win* and *toss_winning*. By observing the positive correlation between the two pairs above, we plotted their linear regression, considering as independent variables *played_match*, in the first case, and *toss_winning* in the second one, and as dependent variables *win* in the first plot and *win* in the second case (Fig. 13, 14).

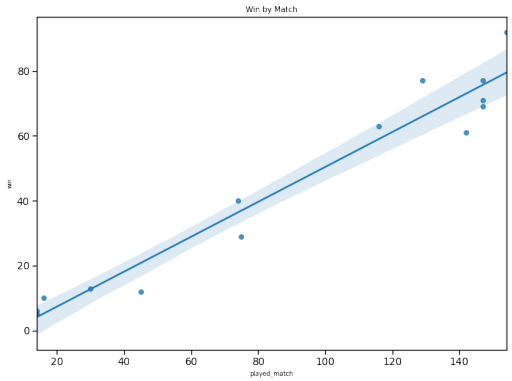


**Figure 13: Linear regression *win* by *played_match***

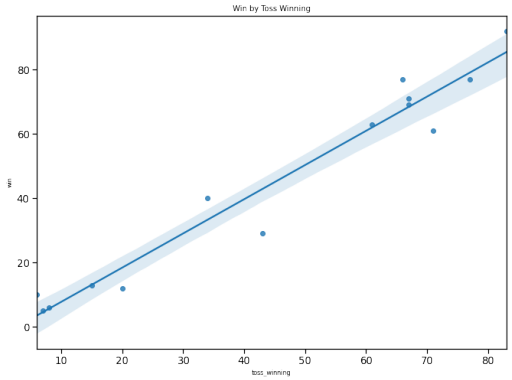


**Figure 14: Linear regression *win* by *toss_winning***

## 4 OUTLIER ANALYSIS

The phase of outlier detection was exploited by using visual methods such as boxplots.

One example of plotted boxplot is the one in Fig. 15, that shows the presence of some "outliers". However, by deepening the analysis, it was proved that the points were not proper outliers, but real rare observations appearing in the dataset with lesser frequency than the values in the range {0,2}. Subsequently, by considering the observation done in the *Data Understanding* paragraph, where we saw that the distribution of *Player of Match* and *Batsman* was unbalanced, we tried to plot the boxplots of these two variables to observe whether there were some outliers. As we can see in Fig. 16 and Fig. 17, both features present some outliers (18 in *Player of Match* and 67 in *Batsman*), therefore we deleted them and calculated again the synthetic measures. In Table 4 we can see the standard deviation value strongly decreasing, compacting a bit the distribution of both features.

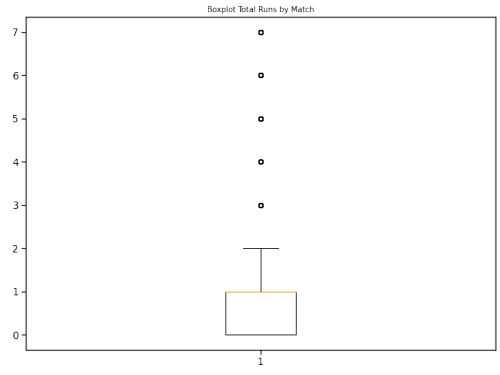


**Figure 15: Boxplot of the variable *total_runs***

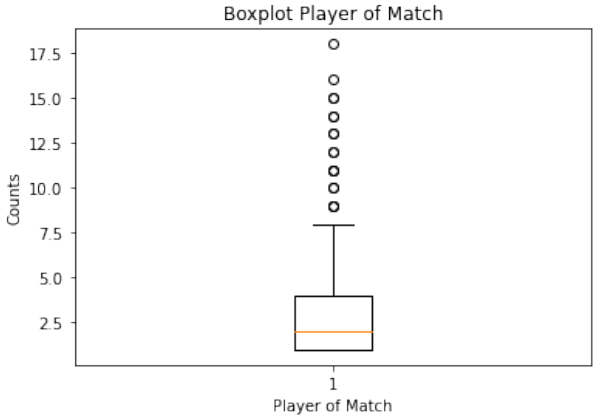


**Figure 16: Boxplot of the variable *Player of Match***

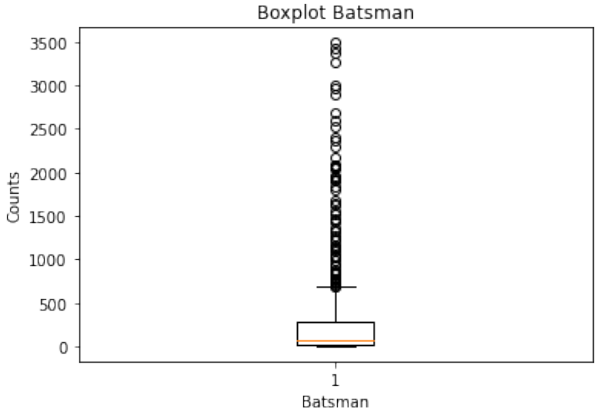


**Figure 17: Boxplot of the variable *Batsman***

| | Player of Match | Batsman |
|---|---|---|
| **mean** | 2.25 | 109.62 |
| **std** | 1.64 | 141.48 |
| **min** | 1.00 | 1.00 |
| **25%** | 1.00 | 13.25 |
| **50%** | 2.00 | 46.50 |
| **75%** | 3.00 | 143.00 |
| **max** | 8.00 | 684.00 |

**Table 4: Syntetic Measures of the features *Player of Match* and *Batsman* without outliers**

## 5 CONCLUSIONS

In these previous paragraphs we obtained several information about the IPL matches as the number of matches played in each season, the team which won the most and the top 5 batsman. Moreover, we observed the correlation between pairs of features and we studied the outliers of the distributions. So, it can be noticed that this kind of analysis was about the descriptive characterics of features. A way to improve this study could be using some techniques, such as clustering methods, to see if there are some hidden connections between data, or classification methods, to do some kind of predictions for the future seasons of the Indian Premiere League.