I P L A U C T I O N D A T A S E T

EXPLORATORY DATA ANALYSIS

Arham Aneeq / 240005009
Department of Metallurgical Engineering & Material Sciences

INTRODUCTION

The initial received dataset, `final_dataset.csv` contained 1052 entries with 7 columns. Cleaning involved the dropping of one redundant column, seven invalid, and three duplicate entries. This was followed by rudimentary univariate analysis. Data was then refactored into three tables aggregating data according to individual Player statistics, individual Team statistics, and yearly statistics. This was followed by correlation analysis.

CLEANING & MANIPULATION

Importing final_dataset.csv into the dataframe df, the initial dataframe is summarised by

```
# df.info()
RangeIndex: 1052 entries, 0 to 1051
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   1052 non-null   int64
 1   Country      1052 non-null   object
 2   Player       1052 non-null   object
 3   Team         1052 non-null   object
 4   Base price   1052 non-null   float64
 5   Winning bid  1052 non-null   object
 6   Year         1052 non-null   int64
dtypes: float64(1), int64(2), object(4)
memory usage: 57.7+ KB

#df.nunique()
Unnamed: 0     1052
Country          29
Player          593
Team             17
Base price       15
Winning bid     128
Year             11
dtype: int64
```

We perform the following to clean the datafram df:

- We drop the redundant column 'Unnamed: 0'
- We strip entries in the column Winning bid of non-numeric characters (including commas), and convert the dtype of the column to float64.
- We strip entries in the Country column of trailing and leading whitespaces
- We drop entries in which the Winning bid is priced lower than the Base price, since such auctions are invalid under IPL rules (we also create a column 'Change' representing this data)

We further notice the following:

- There are multiple typos in the 'Team' column, these are corrected
- There are references to older names of teams, these have been replaced by their latest name
- There are references to both the 'West Indies' and to the individual countries that comprise it. These entries have been replaced by the 'West Indies'
- We remove all entries referring to players who are accounted for by more than one country

Finally, we drop duplicates from the table. These yield

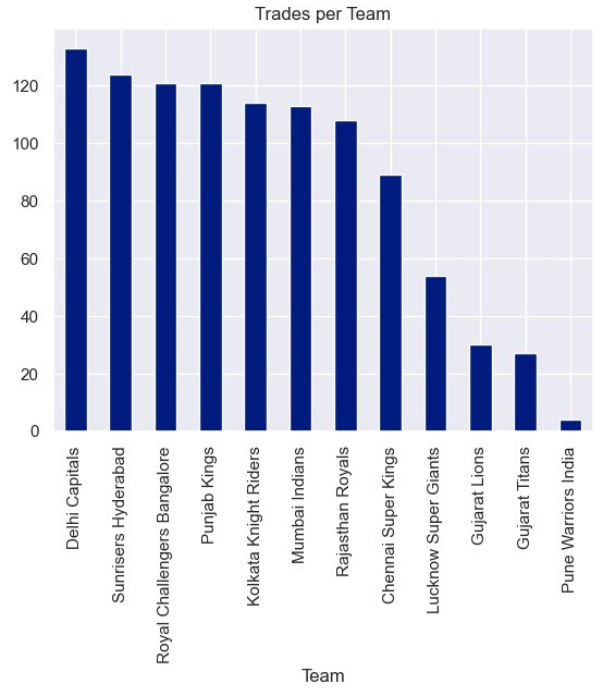
```
# df.info()
RangeIndex: 1038 entries, 0 to 1037
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Country      1038 non-null   object
 1   Player       1038 non-null   object
 2   Team         1038 non-null   object
 3   Base price   1038 non-null   float64
 4   Winning bid  1038 non-null   float64
 5   Year         1038 non-null   int64
 6   Change       1038 non-null   float64
dtypes: float64(3), int64(1), object(3)
memory usage: 56.9+ KB

# df.nunique()
Country         15
Player         589
Team            12
Base price      15
Winning bid    121
Year            11
Change         138
dtype: int64
```

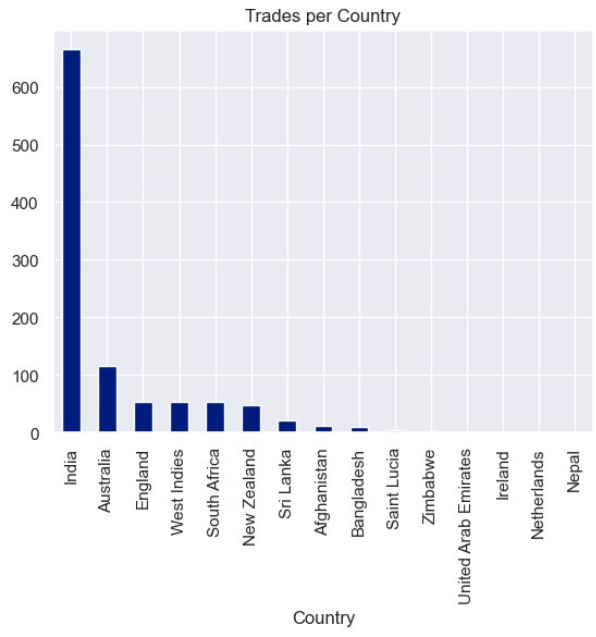
We also create three new tables, one grouped by individual players, by individual teams, and by year.

BASIC ANALYSIS

A total of 12 different teams participated in IPL auctions in the ten year period between 2013 and 2023. Most teams traded roughly a similar number of players in total, with some exceptions.

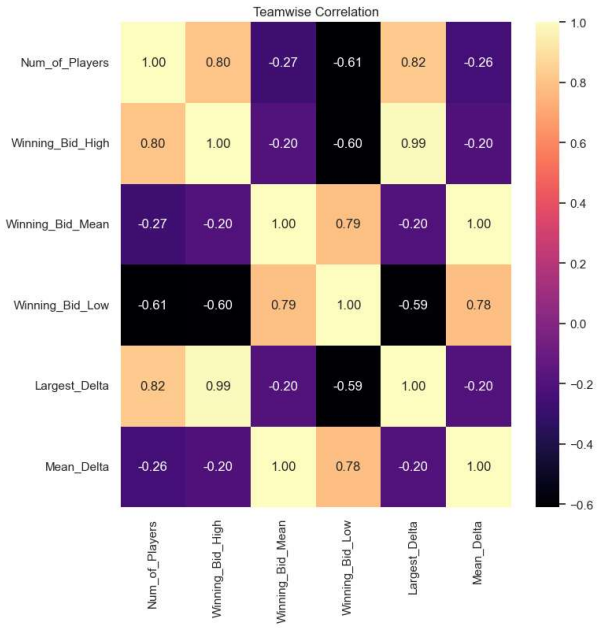


The 589 distinct players themselves are traded mostly only once in the ten-year period (Q1, Q2, Q3 = 1), with the upper quartile (Q4) being traded only twice. These players represent 15 different countries; however, Indian players account for the vast majority of these trades, with more than 60% of traded players being Indian.

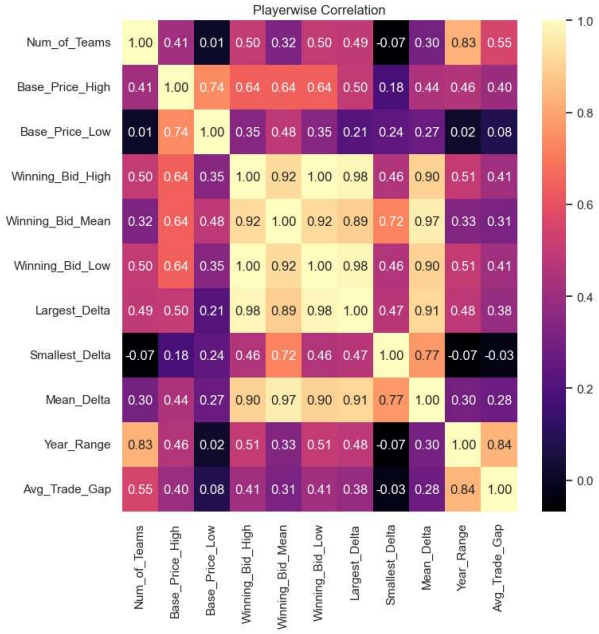


Individual players are purchased, on average, 1.59 times, with most players (up to Q3) being accounted for by only one team. The maximum number of distinct teams that a player is accounted for by is six. For players that have been traded more than once, they remain in their team for an average of 2.62 years before being put up for auction again. The bottom 50% of base prices are between 10 and 30, however, the top 25% in this field command base prices starting at 100 and even go as far up as 244. For 25% of trades, there is no change between the base price and the winning bid, and the median relative increase is only 8%. However, at best, the winning bid was 49 times the original base price. Before analysing particular multivariate relationships, we assess heatmaps representing a basic linear correlation matrix for each of the three aggregate tables.

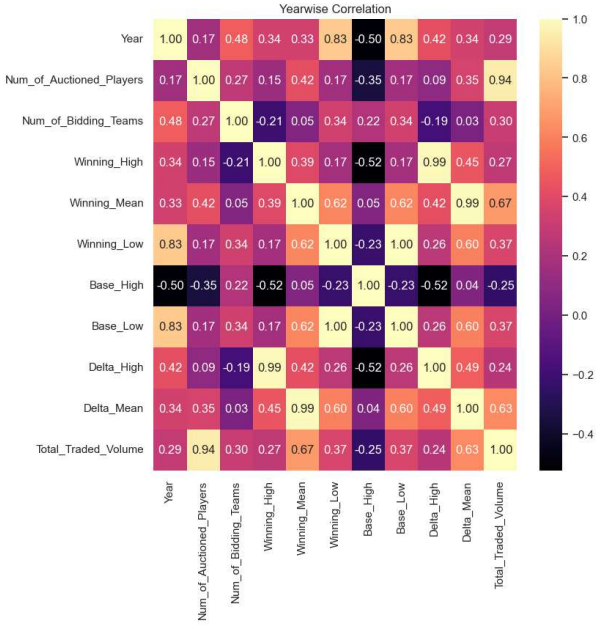
Besides obvious correlates, for example, such as that between mins, means, and maxes, or between a higher low bid implying a higher winning bid, we find several significant correlations.



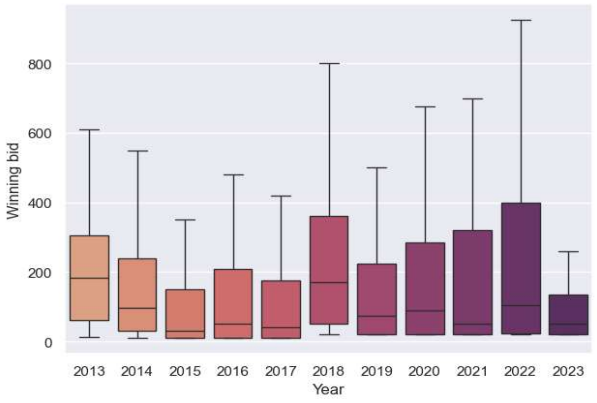
While there are several negative correlations, there are a surprising number of perfect or near perfect correlates, even for statistically related metrics. We observe that teams with a high number of total players are in general less willing to increase past the base price, and that the player count correlates negatively with lower winning bids; it does however correlate highly with a team's highest winning bid.

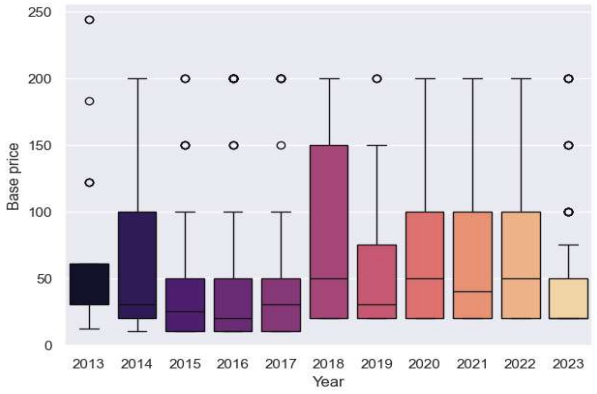


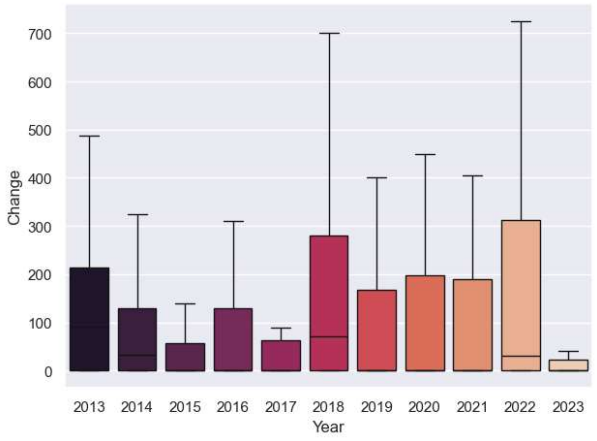
From a player wise perspective mostly weak positive correlations, though it is worth noting that there exists a weak negative correlation between the number of teams that a player has been in, and the smallest increase between the winning bid and base price.



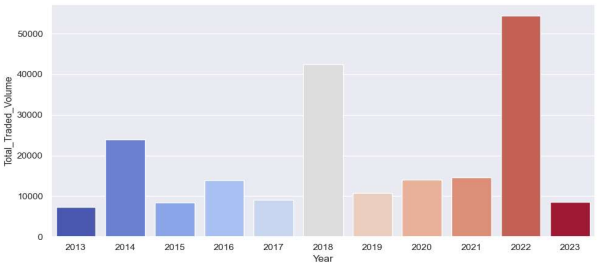
The year correlates strongly with both the minimum base price and the winning bid, both of which correlate perfectly with each other. The maximum base price correlates negatively with the number of players up for auction but increases with the number of teams. Surprisingly, the highest winning prices also negatively correlate with the number of teams participating. The time-wise relations are scrutinised further in the following boxplots.



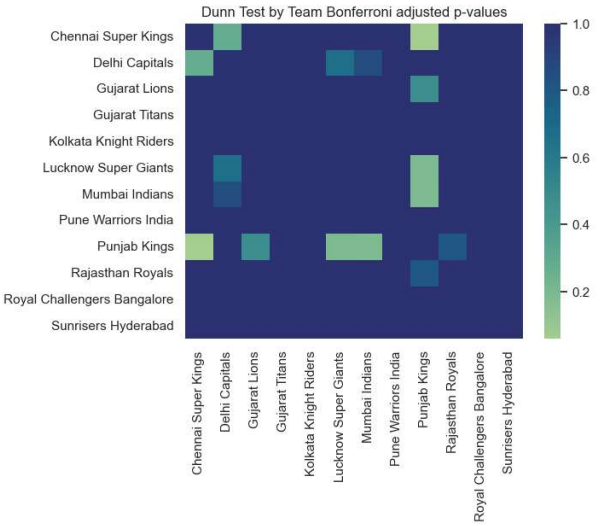




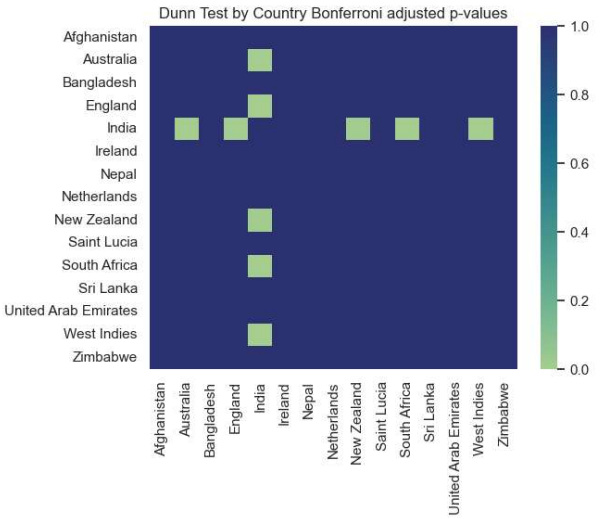
Fliers have been removed from most of these boxplots since their presence compresses the IQR and whiskers to an unreadable size. We also see that the traded volume spikes every four years.



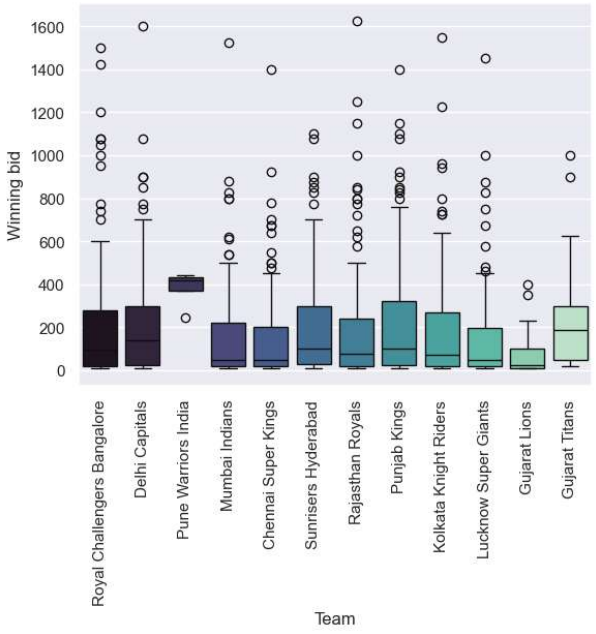
To determine whether different teams have significantly different willingness to bid higher than the base price, we perform a Kruskal-Wallis test with a confidence level $p = 0.05$. The H-Statistic and p-value returned are respectively, 30.34 and 0.0014, which is indicative of significant differences between teams. We perform a post hoc Dunn test, with a Bonferroni p-adjust which generates the following heatmap, based on median values for the Winning bid.

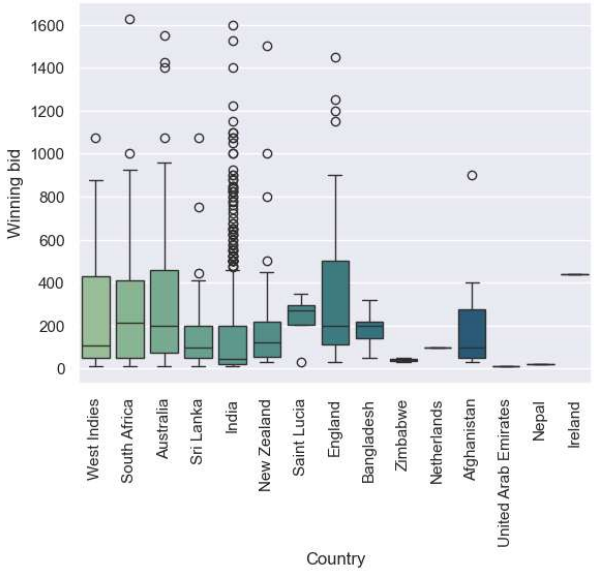


The Kruskal-Wallis test as applied to identifying significant differences in the Winning Bid by country yield even stronger results, with an H-statistic of 141 and a staggeringly small p-value of 3.96×10^{-23}.



The heatmap generated by a Dunn test shows conclusively that a few countries have significantly different distributions. While Kruskal-Wallis and Dunn indicate the existence of qualitative differences between teams and countries, we can gain further insight by consulting boxplots showing the winning bid against categorical data like the country or bidding team.



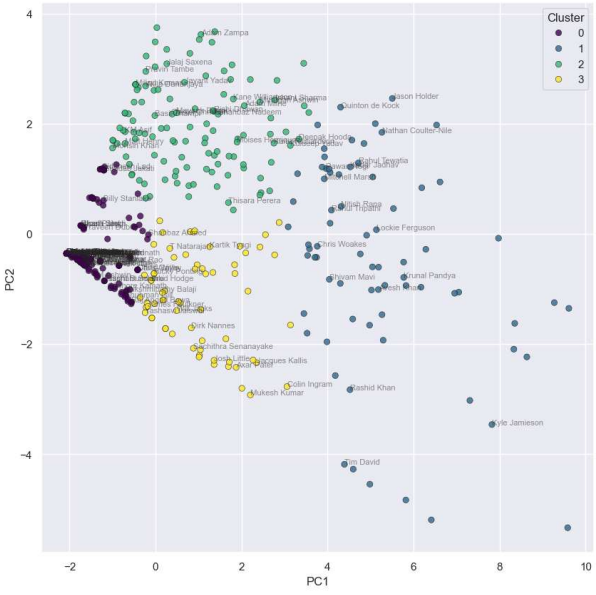


One can easily surmise the relative power differentials that players from different countries have from this graph. While India's fliers command most high value trades, keeping in mind that India accounts for over 60% of trades, we see that the IQR falls far below strong foreign countries like England, Australia, South Africa, and the aggregated West Indies. It is also interesting note that the highest winning bids for each time are

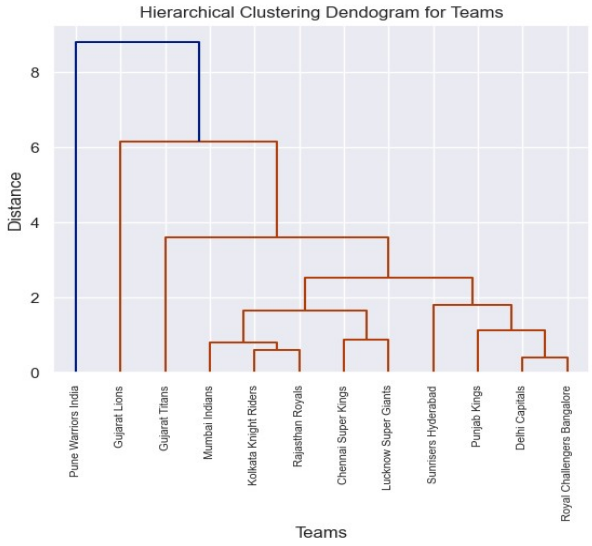
FURTHER ANALYSIS

We build a K Means Clustering model using the columns various features from the player_df graph.

While the clusters look undifferentiated due to the input dimension being greater than displayed, a Silhouette score of 0.504 indicates acceptably demarcated clusters. We can however, apply Principal Component Analysis followed by K-Means Clustering to get a slightly better separated graph, with a Silhouette of 0.578.

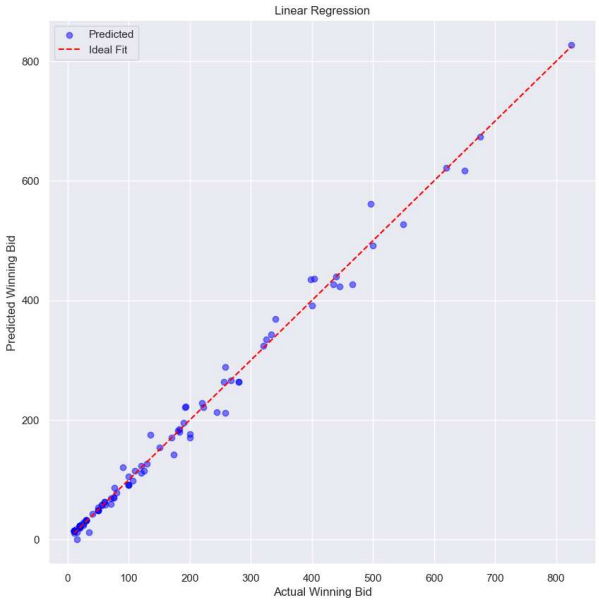


We can use a clustering method on the teams dataset to create a graph known as a dendogram.



The dendogram reinforces the earlier noted disparity between Pune Warriors India and all other teams. It also identifies high performing teams and groups clusters them, such as KKR, MI, and RR.

We attempted regression with two models, a linear regression model and a polynomial regression model, however, the performance increase in the polynomial model was negligible. The regression model took in features related to the player to predict the mean winning bid.



With a R^2-score of 0.9926, and an RMSE of 14.60, this model is quite accurate.