Data Analytics Report

IPL AUCTION DATA

- Data Cleaning: No null values found in the given data set.
- Winning bid is interms of object converted it to float and removed extra symbols
- Substituted for all the team changes and spelling errors

Country	
India	668
Australia	117
England	58
South Africa	54
New Zealand	47
Sri Lanka	21

→ **Max. no. of players are from INDIA**

Team	
Delhi Daredevils	134
Sunrisers Hyderabad	125
Royal Challengers Bangalore	123
Punjab Kings	122
Kolkata Knight Riders	116
Mumbai Indians	115

→ **Delhi Daredevils purchased max. no. of palyers**

Player Statistics

	Player	years_in_auction	avg_winning_bid	total_earnings	teams_count
0	Aaron Finch	5	376.0	1880.0	5
1	Abdul Basith	2	20.0	40.0	1
2	Abdul Samad	1	20.0	20.0	1
3	Abhijeet Tomar	1	40.0	40.0	1
4	Abhimanyu Mithun	2	30.0	60.0	2
...	...	...	...	...	...
588	Yusuf Pathan	2	257.5	515.0	2
589	Yuvraj Singh	5	800.0	4000.0	5
590	Yuzvendra Chahal	3	420.0	1260.0	2
591	Zaheer Khan	2	330.0	660.0	2
592	Zahir Khan Pakteen	1	60.0	60.0	1

593 rows × 5 columns

→ **Feature Engineering**

Contains the no. of years each player participated in auction, a players avg winning bid ,players total earnings.

Top 10 Most Expensive Players

	Player	years_in_auction	avg_winning_bid	total_earnings	teams_count
103	Cameron Green	1	1750.000000	1750.0	1
88	Ben Stokes	3	1441.666667	4325.0	3
184	Harry Brook	1	1325.000000	1325.0	1
573	Wanindu Hasaranga	1	1075.000000	1075.0	1
460	Sam Curran	3	1040.000000	3120.0	2
392	Prasidh Krishna	1	1000.000000	1000.0	1
172	Glenn Maxwell	5	922.000000	4[illegible]0.0	4
287	M. Shahrukh Khan	1	900.000000	900.0	1
491	Sheldon Cottrell	1	850.000000	850.0	1
540	Tim David	1	825.000000	825.0	1

Next steps: Generate code with top_players | View recommended plots | New interactive sheet

Cameron Green has the highest avg winnig bid so he is the most expensive player to buy over the years

	Country	Player	Team	Base price	Winning bid	Year	Bid_to_Base_Price_Ratio	Bid_Difference	Avg_Bid_by_Year	Years_in_Auction
0	Guyana	Christopher Barnwell	Royal Challengers Bangalore	30.5	30.5	2013	1.0	0.0	195.941892	1
1	South Africa	Johan Botha	Delhi Daredevils	183.0	274.5	2013	1.5	91.5	195.941892	1
2	Australia	Daniel Christian	Royal Challengers Bangalore	61.0	61.0	2013	1.0	0.0	195.941892	3
3	Australia	Michael Clarke	Pune Warriors	244.0	244.0	2013	1.0	0.0	195.941892	1
4	Australia	Nathan Coulter-Nile	Mumbai Indians	61.0	274.5	2013	4.5	213.5	195.941892	7
...	...	...	...	...	...	...	...	...	...	...
1047	England	Joe Root	Rajasthan Royals	100.0	100.0	2023	1.0	0.0	205.421687	1
1048	Bangladesh	Shakib al Hasan	Kolkata Knight Riders	150.0	150.0	2023	1.0	0.0	205.421687	1
1049	India	Abdul Basith	Rajasthan Royals	20.0	20.0	2023	1.0	0.0	205.421687	1
1050	England	Joe Root	Rajasthan Royals	100.0	100.0	2023	1.0	0.0	205.421687	1
1051	Bangladesh	Shakib al Hasan	Kolkata Knight Riders	150.0	150.0	2023	1.0	0.0	205.421687	1

1052 rows × 10 columns

Feature Engineering

Added extra feautures like Bid to Base ratio, bid difference, avg bid by year

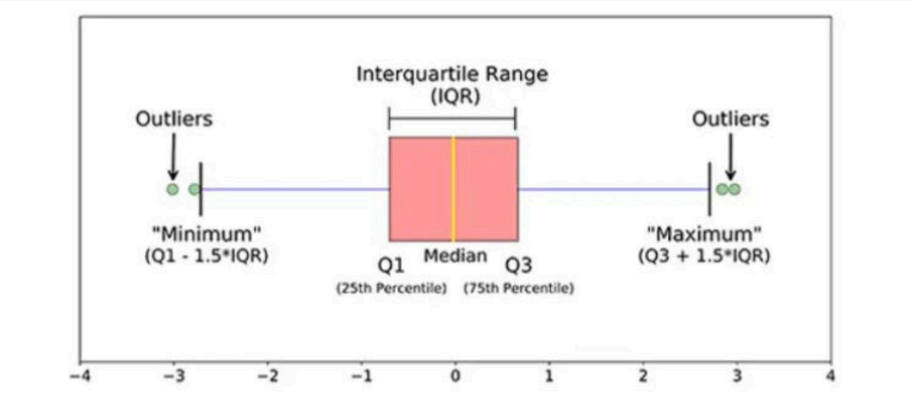
The IQR method was applied to detect outliers, ensuring a robust analysis by filtering extreme bid values outside 1.5 times the interquartile range.

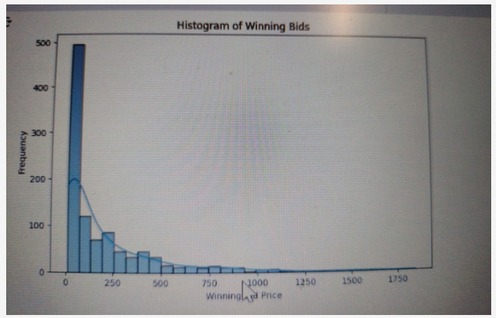
Outliers in Base Price

	Country	Player	Team	Base price	Winning bid	Year	Bid_to_Base_Price_Ratio	Bid_Difference	Avg_Bid_by_Year	Years_in_Auction
3	Australia	Michael Clarke	Pune Warriors	244.0	244.0	2013	1.0	0.0	195.941892	1
27	Australia	Ricky Ponting	Mumbai Indians	244.0	244.0	2013	1.0	0.0	195.941892	1

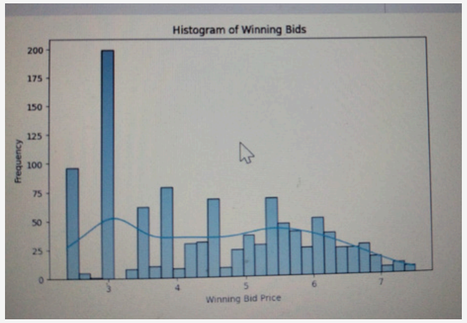
Outliers in Winning Bid

	Country	Player	Team	Base price	Winning bid	Year	Bid_to_Base_Price_Ratio	Bid_Difference	Avg_Bid_by_Year	Years_i
39	England	Kevin Pietersen	Delhi D[illegible]redevils	200.0	900.0	2014	4.500	700.0	172.467532	
41	India	Yuvraj Singh	Royal Challengers Bangalore	200.0	1400.0	2014	7.000	1200.0	172.467532	
47	India	Dinesh Karthik	Delhi Daredevils	200.0	1250.0	2014	6.250	1050.0	172.467532	
199	India	Yuvraj Singh	Delhi Daredevils	200.0	1600.0	2015	8.000	1400.0	132.272727	
200	Sri Lanka	Angelo Mathews	Delhi Daredevils	150.0	750.0	2015	5.000	600.0	132.272727	
...	...	...	...	...	...	...	...	...	...	
971	India	Mayank Agarwal	Sunrisers Hyderabad	100.0	825.0	2023	8.250	725.0	205.421687	
973	England	Sam Curran	Punjab Kings	200.0	1850.0	2023	9.250	1650.0	205.421687	
977	Australia	Cameron Green	Mumbai Indians	200.0	1750.0	2023	8.750	1550.0	205.421687	
978	England	Ben Stokes	Chennai Super Kings	200.0	1625.0	2023	8.125	1425.0	205.421687	



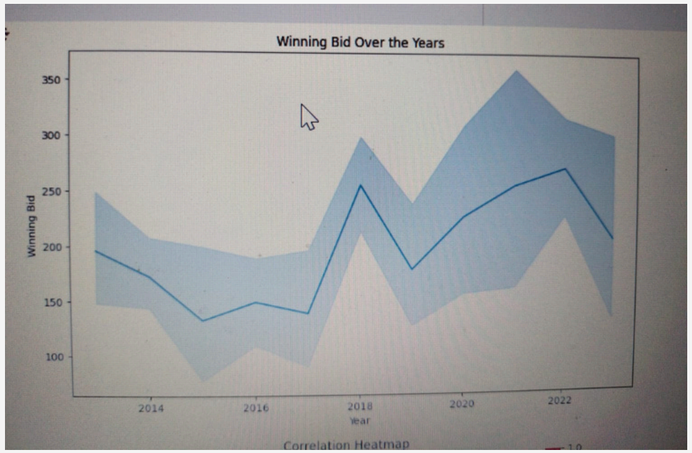


Before



After

A logarithmic transformation was applied to the winning bid values to normalize the left-skewed distribution and mitigate the impact of extreme outliers on a copy of the dataset,did not apply logarithm yo the original data as bid values have to be also changed in that case but bid values have less outliers.

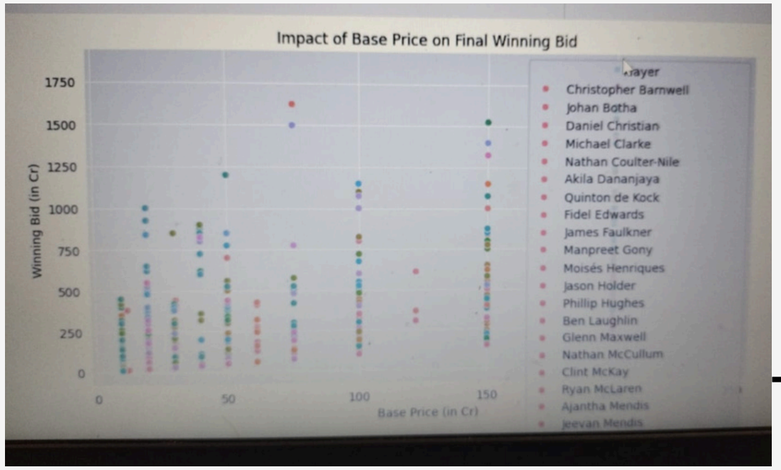


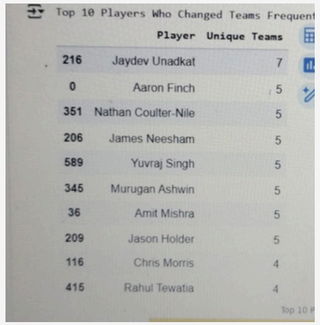
Highest winning bid was noted in 2022

Jaydev Unadkat appeared the max. no. of times in auction

Correaltion Heatmap Inferences

- Winning Bid vs. Base Price (0.59): There is a moderate positive correlation, indicating that higher base prices generally lead to higher winning bids.
- Bid Difference vs. Winning Bid (0.98): This extremely high correlation suggests that bid differences are almost directly influenced by winning bids.
- Bid Difference vs. Base Price (0.42): A weak positive correlation, implying that bid differences tend to increase slightly with base prices.
- Bid-to-Base Price Ratio vs. Winning Bid (0.48): A moderate correlation suggests that the bid-to-base price ratio is influenced by the final winning bid.
- Base Price vs. Bid-to-Base Price Ratio (-0.093): A very weak negative correlation suggests that base price does not strongly determine the bid-to-base price ratio.



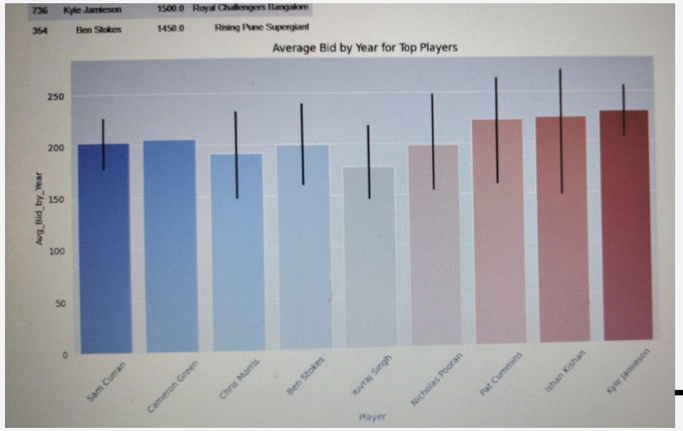


Jaydev Unadkat changed the most no. of teams

There appears to be no strict linear relationship between the base price and the final winning bid. While some players with higher base prices received higher bids, several players with low base prices also attracted significant bids, indicating that other factors (e.g., performance, demand, team strategy) strongly influence the final bid.

Some countries, like Jamaica and Zimbabwe, show high variability in the Bid-to-Base Price Ratio, with extreme outliers indicating that players from these countries can sometimes fetch significantly higher bids compared to their base price.Countries like Guyana and India have wider distributions, implying diverse bidding behaviors.

- Violin Shapes: Represent the density distribution of the Bid-to-Base Price Ratio or each country.
- Black Lines (Boxplot within the violin): Show the interquartile range (IQR), median, and potential outliers.
- Points and Whiskers: Represent individual data points and outlier ranges.

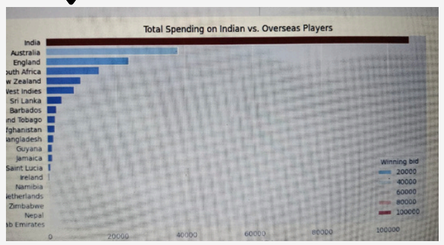


Kyle Jamieson has the highest bid by year

Tells How spending in IPL auctions changed .Were there years where spending dropped (due to economic reasons, team restrictions) etc.

- The player retention pattern heatmap represents wether the player changed his team or reamined in the same team each year .
- Punjab Kings spent the most amount of money in buying players over years in auctions.
- Highest amount of money was spent in ipl auction of 2022.

The total spending over players across the world analysis revealed that teams allocated significant portion of their budget to Indian players, indicating a strategic preference towards local talent. This trend highlights the demand and perceived value of different player categories in the auction market.



Kmeans Clustering

If K=2 (Silhouette Score), the data naturally separates into two meaningful groups, making clustering valid.

- Cluster 1: High-Value, Experienced Players 🏆

1. High winning bids, long auction history, and high total earnings.
2. Star players, consistent performers, or veteran cricketers.

- Cluster 2: Low-Value, Less Experienced Players 🌱

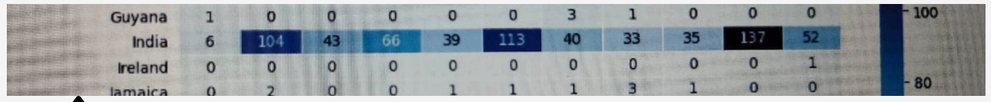
3. Lower winning bids, fewer years in auction, and lower total earnings.
4. Newer players, underappreciated talent, or those with fluctuating demand

- **Cluster 0 (High-Value Players):**

1. Higher average winning bid (₹486.77L) with more variation (std = 279.03).
2. More years in auction (avg. 3.2 years).
3. Higher total earnings (₹1378.78L).

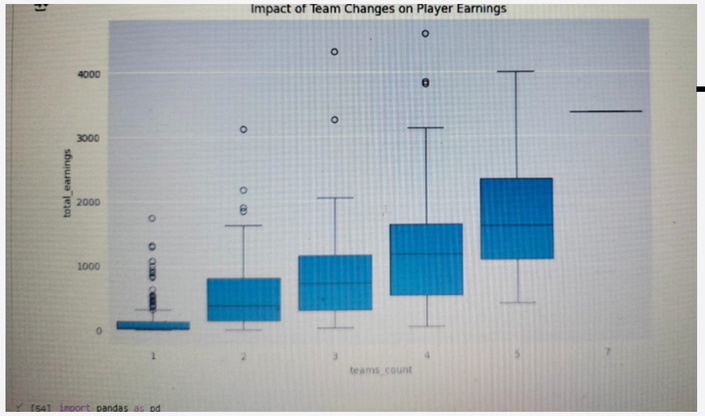
- **Cluster 1(Low-Value Players):**

4. Lower average winning bid (₹92.59L) with less variation (std = 107.58).
5. Fewer years in auction (avg. 1.44 years).
6. Lower total earnings (₹137.02L).



Guyana	1	0	0	0	0	0	3	1	0	0	0
India	6	104	43	66	39	113	40	33	35	137	52
Ireland	0	0	0	0	0	0	0	0	0	0	1
Jamaica	0	2	0	0	1	1	1	3	1	0	0

No. of players bought from each country over seasons indicate that every year most no. of players are bought from India



As the number of teams a player has played for (teams_count) increases, the median total earnings tend to rise. The spread of earnings also increases with more team changes, indicating higher variability. Players with fewer team changes generally have lower earnings, while those with more team changes show a wider distribution, including some high earners.

A right-skewed graph of the Winning Bid to Base Price Ratio (with the ratio on the x-axis and frequency on the y-axis) indicates that most players have a low ratio, meaning that the winning bids are generally close to the base price. However, there are a few players with significantly higher ratios, meaning they were sold for much more than their base price.

- Applied basic regression by first taking only winning bid and performing linear regression then took both base price and years in auction to determine winning bid through multilinear regression.

ML

- Then performed feature importance for determining which charecteristic affects winning bid prediction the most

Anomaly Detection in Player Bidding

To identify overvalued and undervalued players based on expected vs. actual winning bids using machine learning..Methodology:

1. Prediction Model: A Random Forest Regressor was trained to predict the expected bid using player, team, base price, and year as features.

2. Deviation Calculation: The difference between actual and predicted bids was computed.
3. Anomaly Detection: An Isolation Forestwas applied to detect outliers in bid deviations, classifying players as:

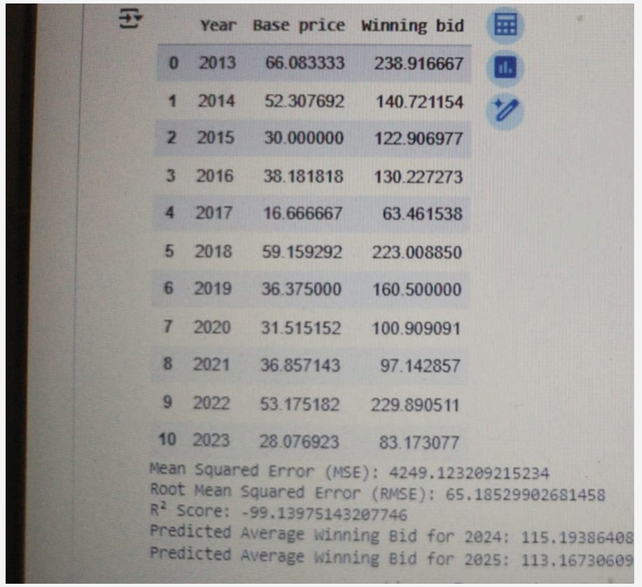
Overvalued : Higher than expected bids.
Undervalued :Lower than expected bids.

- Random Forest predicts a player's fair bid based on historical data.
- Isolation Forest identifies outliers, assuming extreme deviations indicate potential overvaluation or undervaluation.

1. Table with Anomaly = 1: These are the normal or typical players based on the model's understanding, meaning their bid deviations from the expected price are within the typical range.
2. Table with Anomaly = -1: These are the outliers, i.e., the players who have unusual bid differences compared to the expected winning bids (either undervalued or overvalued players).

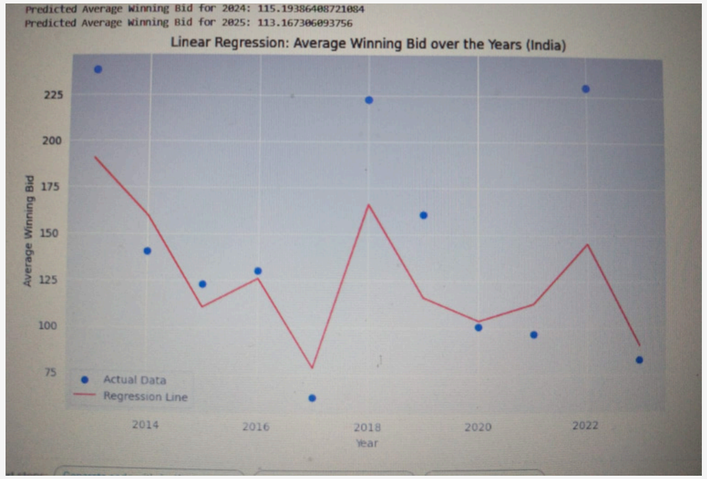
Avg Winning bid based on year and avg base price(based on mean)

- Next we applied linear regression to analyze the relationship between the average winning bid and two key factors: Year and avg Base price.
- The data was grouped by year, and the average Base price and Winning bid were calculated for each year.



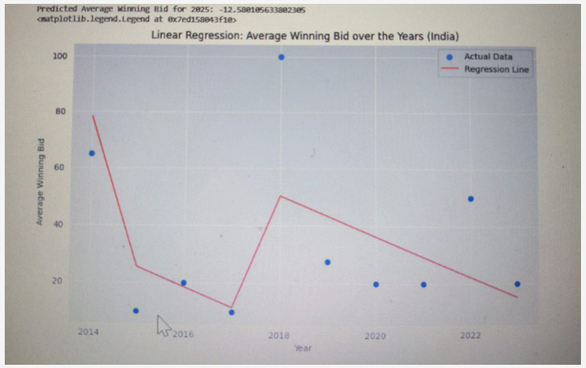
	Year	Base price	Winning bid
0	2013	66.083333	238.916667
1	2014	52.307692	140.721154
2	2015	30.000000	122.906977
3	2016	38.181818	130.227273
4	2017	16.666667	63.461538
5	2018	59.159292	223.008850
6	2019	36.375000	160.500000
7	2020	31.515152	100.909091
8	2021	36.857143	97.142857
9	2022	53.175182	229.890511
10	2023	28.076923	83.173077

```
Mean Squared Error (MSE): 4249.123209215234
Root Mean Squared Error (RMSE): 65.18529902681458
R² Score: -99.13975143207746
Predicted Average Winning Bid for 2024: 115.19386408
Predicted Average Winning Bid for 2025: 113.16730609
```

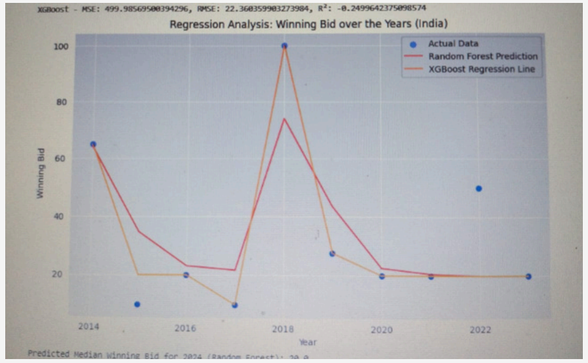


Avg Winning bid based on year and avg base price(based on median):

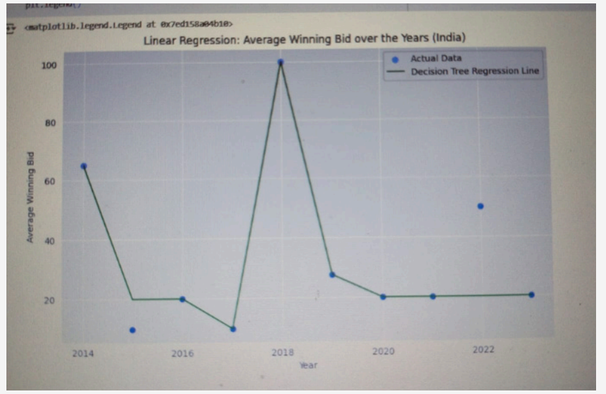
Based on MSE, RMSE, and R^2, Decision Tree and XGBoost are the better models, compred to other ones. Median gives comparitively better results than mean.



Linear regression



Random Forest and XGBoost



Decision tree

The high MSE and negative R^2 scores suggest that the models are not explaining the variance in the data well. This is because data has non linear trends and it is a very small dataset

Predicting Total Spending of a Team Based on Number of Players Bought

We aim to predict the total spending of a team based on the number of players bought, helping teams and analysts forecast future spending trends.

- We grouped the data by team and year, calculating the total spending and the number of players bought for each team per year.
- We used the number of players bought and year as features to predict total spending using linear regression
- The model was also used to predict future spending for a team buying 15 players in 2024.

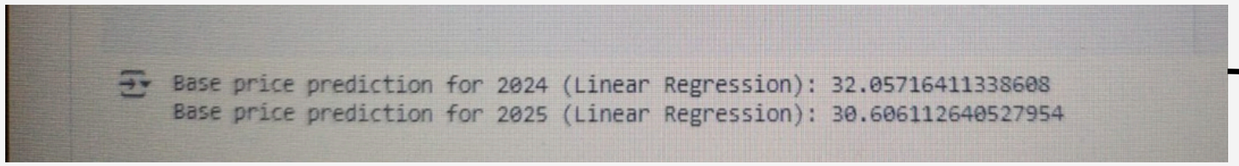
The R^2 score of 0.89 indicates that the model explains approximately 89% of the variance in total spending based on the number of players bought, suggesting a strong fit.

Random Forest Regressor to predict the winning bid based on various features such as Base price, Bid-to-Base Price Ratio, Bid Difference, Year, and categorical features like Player and Team:

```
RMSE: 15.398223809253384
R² Score: 0.9975135279160554
```

- We first encoded categorical variables (Player and Team) using One-Hot Encoding to convert them into numerical form for the model.
- We then combined these encoded features with the numerical ones and split the data into training and testing sets.
- This approach allows us to build a model that can effectively predict the winning bid, considering a mix of both categorical and numerical features, while handling complex relationships in the data.

The RMSE of 15.40 and R^2 score of 0.9975 indicate that the model has a very small prediction error and explains 99.75% of the variance in the winning bid, demonstrating an excellent fit.



Here we predicted base prices for the years 2024 and 2025 using linear regression between year and base price