Assessment of IPL Players and Their Purchase Cost Over the Years

Ankur Soni,
Computer Science & Engineering, Dibrugarh University Institute of Engineering & Technology

Ansh Nayak,
Computer Science & Engineering, Dibrugarh University Institute of Engineering & Technology

Chinmoy Jyoti Kashyap,
Computer Science & Engineering, Dibrugarh University Institute of Engineering & Technology

Kalyan Baishya,
Computer Science & Engineering, Dibrugarh University Institute of Engineering & Technology

Rohan Kuli,
Computer Science & Engineering, Dibrugarh University Institute of Engineering & Technology

Ms. Bidisha Dobe

Period of Internship: 19th May 2025 - 15th July 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

Abstract

This internship project focuses on analysing the relationship between the auction prices of Indian Premier League (IPL) players and their corresponding on-field performance over the span of more than a decade, from 2008 to 2020. The key research question addressed in this project is whether spending a large amount of money on high-profile players yields better performance results for IPL franchises, or if lower to moderately priced players provide greater value for money. Using comprehensive datasets obtained from Kaggle, which include detailed auction history and ball-by-ball match performance, we integrated and prepared the data for analysis. The tools used for this project include Python-based libraries such as Pandas for data manipulation, Numpy for numerical operation, Matplotlib and Seaborn for data visualization, all of which were run in Jupyter Notebook through the Anaconda Navigator environment. We began by cleaning and standardizing the data, then merged auction and performance records to establish a player-wise mapping between cost and contribution. Key performance metrics for batters were computed, including runs scored, strike rate, and match impact. A correlation analysis was carried out to determine if any significant statistical relationship exists between the price paid for a player and their actual effectiveness on the field. The visualizations and trend observations provided deeper insights into team auction strategies and player consistency over the years. The results of this analysis indicate that high pricing does not always guarantee high performance, and in several cases, moderately priced players have shown better returns. The overall findings can help IPL franchises adopt more data-informed decision-making approaches during auctions and improve cost-efficiency in building competitive teams.

Introduction

The Indian Premier League (IPL) is one of the most commercially successful and widely followed cricket tournaments in the world, attracting top talent and generating enormous fan engagement each season. With each franchise allocated a budget during auctions, strategic decisions must be made about which players to invest in—raising the question of whether high-cost players deliver results that justify their hefty price tags. This project, titled *Assessment of IPL Players and Their Purchase Cost Over the Years*, aims to explore the effectiveness of such financial decisions by analysing the relationship between auction prices and player performance from the 2008 to 2020 seasons. The motivation behind the project lies in helping IPL franchises make more informed decisions using data-driven insights rather than relying solely on reputation or hype. The technological framework of the project involved the use of Python and various data analysis libraries, including Pandas for data manipulation, Numpy for numerical operation, Seaborn and Matplotlib for visualizations, all within the Anaconda Navigator environment using Jupyter Notebook. The datasets were sourced from Kaggle, one containing detailed player auction records and the other capturing ball-by-ball match data across multiple seasons. An initial data cleaning and transformation phase was conducted to align and merge these datasets, followed by the derivation of key performance metrics for each player, such as runs, strike rate, economy, and wickets. The goal was to analyse whether there exists a strong correlation between a player's cost and their actual contribution to team success, both for batters and bowlers. During the first two weeks of the internship, training was provided on topics such as Python programming, data preprocessing, data visualization techniques, basic statistics, and correlation analysis, which laid the foundation for the subsequent stages of this project.

Project Objective

- To evaluate the correlation between the auction price of IPL players and their performance metrics across multiple seasons (2008–2020) and assess if high-cost players consistently deliver better results.

- To define and calculate meaningful performance indicators for both batters (e.g., total runs, strike rate) and bowlers (e.g., wickets taken, economy rate) using historical match data.

- To integrate and clean datasets from different sources (auction data and performance data) to create a unified dataset suitable for statistical and visual analysis.

- To conduct exploration data analysis (EDA) and visualize trends in player performance vs. auction value using tools such as Pandas, Numpy, Matplotlib, and Seaborn.

- To assist IPL franchises in making data-informed decisions during the auction process by identifying whether moderately priced players offer better cost-performance value compared to high-priced players.

Methodology

The methodology of this project involved a systematic, multi-phase approach that included data collection, preprocessing, integration, analysis, and interpretation. Each step was carefully designed to answer the central research question—does the auction price of an IPL player correlate with their actual on-field performance?

The entire project was implemented using Python within the Jupyter Notebook environment of Anaconda Navigator. The primary libraries used were **Pandas** for data manipulation, **Seaborn** and **Matplotlib** for data visualization, and **NumPy** for numerical operations. Below are the detailed steps followed during the execution of the project:

Step-by-Step Methodology:

Data Collection

- Downloaded two major datasets from Kaggle:
  - IPL Player Auction Dataset (includes player names, teams, auction prices, and auction years).
  - IPL Ball-by-Ball Match Performance Dataset (contains detailed match events from 2008 to 2020).
- Verified and validated the dataset structure and formats to ensure compatibility.

Data Cleaning and Preprocessing

- Effective data analysis begins with thorough cleaning and preprocessing to ensure consistency, accuracy, and compatibility across datasets. In this IPL auction analytics project, the following steps were taken to prepare the data:

Standardizing Player Names

- Issue: Player names in different datasets had inconsistent formatting (e.g., extra spaces, mixed casing).
- Solution:

  ```
  auction['Player'] = auction['Player'].str.strip().str.lower()
  ```

  .strip() removes leading/trailing whitespace.
  .lower() converts names to lowercase for uniformity.
- Impact: Ensures reliable merging across datasets.

- Cleaning Monetary Values
  - Issue: Auction amounts were stored as strings with symbols like ₹ and spaces, preventing numerical analysis.
  - Solution:
    ```
    auction['Amount'] = (auction['Amount'].replace('[\₹,\s]', '', regex=True).astype(float))
    ```
  - Regex removes currency symbols and whitespace.
  - Converted to float for correlation and visualization.

- Standardizing Delivery Data Columns
  - Issue: Columns like batsman, bowler, and dismissal_kind had inconsistent formatting.
  - Solution:
    ```
    for col in['batsman','bowler','dismissal_kind']:deliveries[col]= deliveries[col].str.strip().str.lower()
    ```
  - Impact: Improves consistency for grouping and filtering operations.

- Aggregating Batting Performance
  - Goal: Summarize total runs scored by each batsman.
  - Method:
    ```
    batting_perf=(deliveries.groupby('batsman')['batsman_runs'].sum().reset_index().rename(columns={'batsman': 'Player', 'batsman_runs': 'Runs'}))
    ```
  - Impact: Creates a clean performance metric for integration with auction data.

- Handling Missing Values
  - Issue: Some players in the auction dataset had no matching performance data.
  - Solution:
    ```
    batting_df['Runs'] = batting_df['Runs'].fillna(0)
    ```
  - Impact: Ensures completeness of analysis by assigning zero runs to unmatched players.

- Outcome
  - These preprocessing steps ensured:
  - Consistent formatting across datasets
  - Accurate numerical analysis

    - Reliable data integration
    - A clean foundation for correlation analysis and visualization

Data Integration

    - Data integration involves combining multiple datasets to create a unified view for analysis. In this IPL auction analytics project, integration was essential to link auction data with on-field performance metrics.

- Datasets Involved:
    - Dataset Name - Description
    - auction.csv: -Contains player names, auction amounts, teams, and years
    - deliveries.csv: - Ball-by-ball match data including batsman runs
    - matches.csv: - Match-level metadata (not directly used in this integration step)
- Integration Steps:

1. Preparing Player Performance Data

    - Aggregated total runs scored by each batsman from the deliveries dataset: batting_perf=(deliveries.groupby('batsman')['batsman_runs'].sum() .reset_index().rename(columns={'batsman': 'Player', 'batsman_runs': 'Runs'}))
    - Renamed batsman to Player to match the column name in the auction dataset.

2. Merging Auction and Performance Data

    - Merged the cleaned auction dataset with batting_perf using a left join: batting_df = auction.merge(batting_perf, on='Player', how='left')
    - Why left join? To retain all auction records, even if a player didn’t score any runs (e.g., uncapped or bench players).

3. Handling Missing Values

    - Filled missing Runs with 0 for players not found in the performance dataset: batting_df['Runs'] = batting_df['Runs'].fillna(0)

- Purpose of Integration
    - To correlate auction spending with actual player performance.
    - Enables deeper insights into franchise decision-making and player valuation.

- Forms the foundation for further analysis like ROI scoring, undervalued player detection, and predictive modeling.

Deriving Performance Metrics

- For **batters**:
    - Total Runs
    - Batting Average
    - Strike Rate
- For **bowlers**:
    - Total Wickets
    - Bowling Average
    - Economy Rate
- Ensured these metrics were aggregated on a per-season basis per player.

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) is a critical step in understanding the structure, patterns, and relationships within the data. In this IPL auction analytics project, EDA was used to uncover insights into how auction prices relate to player performance.

- Objective of EDA
    - To explore the relationship between auction amount and batting performance.
    - To identify trends, outliers, and potential undervaluation in player pricing.
    - To support hypothesis generation for further modeling (e.g., ROI scoring).

- Key EDA Steps

1. Correlation Analysis

    - Purpose: Measure the linear relationship between Amount (auction price) and Runs (total runs scored).
    - Method:
      batting_df['Amount'].corr(batting_df['Runs'])
    - Result:

Correlation – Amount vs Runs: 0.167

- Interpretation:
  - A weak positive correlation suggests that higher auction prices are only loosely associated with better batting performance.
  - Indicates that franchises may consider other factors (e.g., potential, role, marketability) beyond past performance.

2. Visualizing Auction vs Performance

  - Tool Used: Seaborn regression plot
  - Code:
    ```
    sns.regplot(data=batting_df, x='Amount', y='Runs', scatter_kws={'alpha':0.5})
    plt.title('Auction Amount vs Runs')
    ```
- Insights:
  - The scatter plot shows a wide spread of data points, reinforcing the weak correlation.
  - The regression line helps visualize the general upward trend.
  - Transparency (alpha=0.5) improves readability by reducing overlap.

- Observations from EDA
  - Some players received high auction amounts despite low run totals, suggesting strategic or speculative bidding.
  - Several high-performing players were acquired at relatively low prices, indicating potential undervaluation.
  - The distribution of auction amounts and runs is skewed, with a few players dominating both metrics.

- Outcome
  - EDA provided a foundational understanding of:
  - How auction spending aligns (or misaligns) with player output.
  - Which players might offer better value for money.
  - Where further analysis (e.g., ROI scoring, clustering) could be applied.

Correlation Analysis

- What Is Correlation?
  - Correlation measures the strength and direction of a linear relationship between two variables. It ranges from:

    - +1: Perfect positive correlation
    - 0: No correlation
    - −1: Perfect negative correlation
- Why Use It Here?
  - To assess whether higher auction prices are associated with better batting performance (i.e., more runs).
- How We Did It:

```
batting_df['Amount'].corr(batting_df['Runs'])
```

- This computes the Pearson correlation coefficient between Amount and Runs.
- Result Interpretation:
  - Value: 0.167
  - Meaning: There's a weak positive correlation between auction amount and runs scored.
  - Suggests that franchises may not always pay based on past performance.
  - Other factors (e.g., potential, role, marketability) likely influence auction decisions.

- Visualization Support
  - We used a regression plot to visualize this relationship:

```
sns.regplot(data=batting_df, x='Amount', y='Runs', scatter_kws={'alpha':0.5})
```

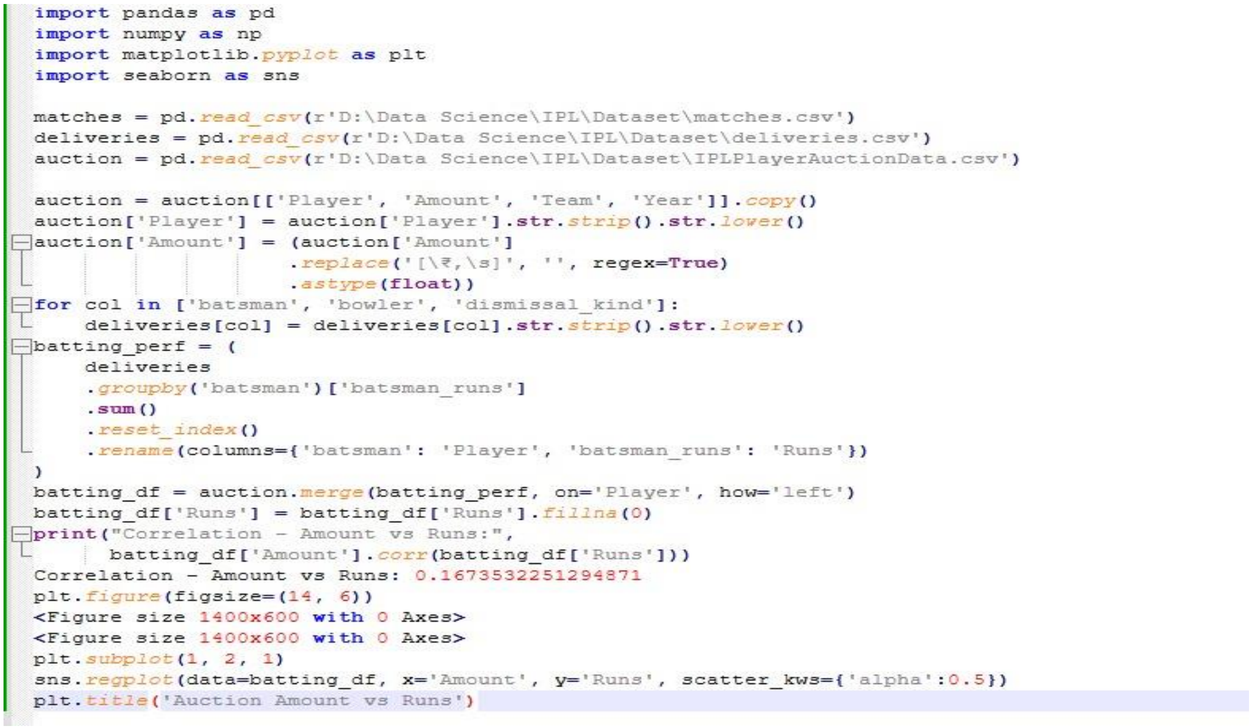
  - Shows the trend line and scatter distribution.
  - Helps visually confirm the weak positive correlation.

Data taken from: -

https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020

https://www.kaggle.com/datasets/kalilurrahman/ipl-player-auction-dataset-from-start-to-now

Data Analysis and Results

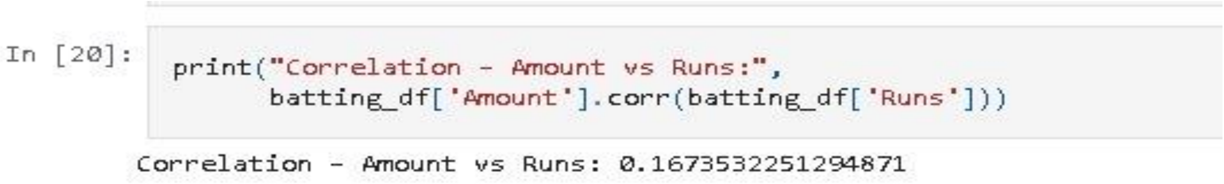


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

matches = pd.read_csv(r'D:\Data Science\IPL\Dataset\matches.csv')
deliveries = pd.read_csv(r'D:\Data Science\IPL\Dataset\deliveries.csv')
auction = pd.read_csv(r'D:\Data Science\IPL\Dataset\IPLPlayerAuctionData.csv')

auction = auction[['Player', 'Amount', 'Team', 'Year']].copy()
auction['Player'] = auction['Player'].str.strip().str.lower()
auction['Amount'] = (auction['Amount']
                     .replace('[\₹,\s]', '', regex=True)
                     .astype(float))
for col in ['batsman', 'bowler', 'dismissal_kind']:
    deliveries[col] = deliveries[col].str.strip().str.lower()
batting_perf = (
    deliveries
    .groupby('batsman')['batsman_runs']
    .sum()
    .reset_index()
    .rename(columns={'batsman': 'Player', 'batsman_runs': 'Runs'})
)
batting_df = auction.merge(batting_perf, on='Player', how='left')
batting_df['Runs'] = batting_df['Runs'].fillna(0)
print("Correlation - Amount vs Runs:",
      batting_df['Amount'].corr(batting_df['Runs']))
Correlation - Amount vs Runs: 0.1673532251294871
plt.figure(figsize=(14, 6))
<Figure size 1400x600 with 0 Axes>
<Figure size 1400x600 with 0 Axes>
plt.subplot(1, 2, 1)
sns.regplot(data=batting_df, x='Amount', y='Runs', scatter_kws={'alpha':0.5})
plt.title('Auction Amount vs Runs')
```

Fig.: Python code to calculate correlation and plot results.



```
In [20]:  print("Correlation - Amount vs Runs:",
                batting_df['Amount'].corr(batting_df['Runs']))

          Correlation - Amount vs Runs: 0.1673532251294871
```

Fig.: Correlation score 0.16 (weak positive link).

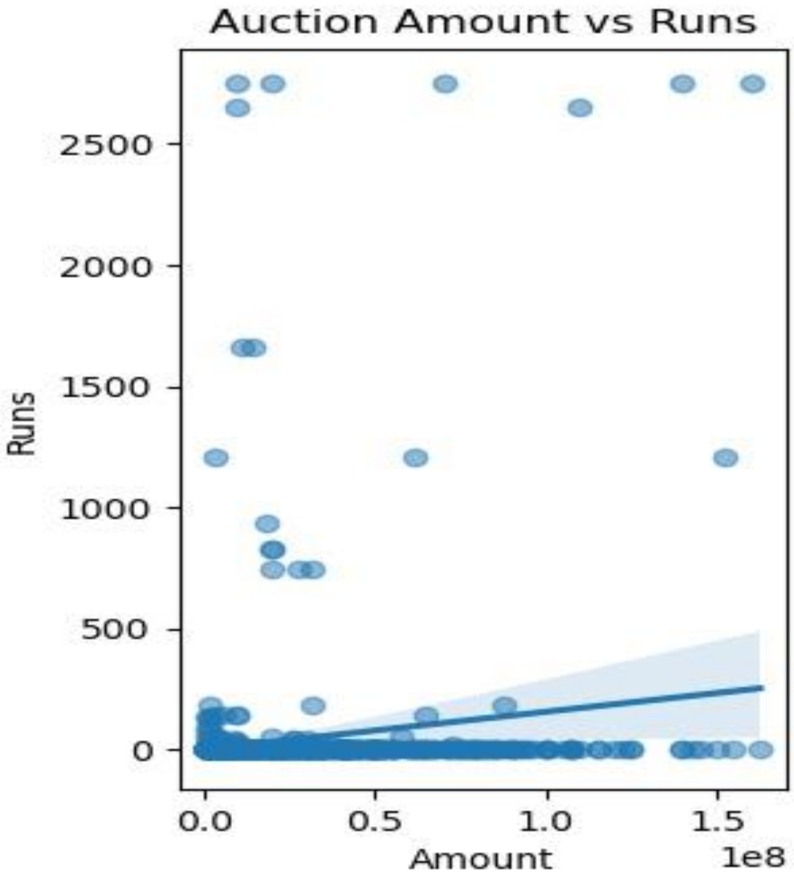


Fig.: Auction amount vs runs – weak upward trend

WHAT IS CORRELATION?

- CORRELATION COEFFICIENT RANGES FROM -1 TO +1:

- +1: PERFECT POSITIVE CORRELATION

- 0: NO CORRELATION

What the Graph Shows:

- Scatter points: Each dot represents a player.
- X-axis (Amount): Auction price in crores.
- Y-axis (Runs): Total runs scored.
- Regression line: Indicates the overall trend.
- Confidence band: Shows uncertainty around the trend.

Key Observations:

- The regression line slopes slightly upward, confirming a positive correlation.
- However, the spread of points is wide, and many players with similar auction prices have vastly different run totals.
- Several outliers exist:
- Players with high auction prices but low runs.
- Players with low auction prices but high runs.

What Does a Correlation of 0.167 Mean?

0.0 to ±0.2 → Very weak or no correlation

±0.2 to ±0.5 → Weak correlation

±0.5 to ±0.7 → Moderate correlation

±0.7 to ±1.0 Strong correlation

- 0.167 falls in the very weak range.
- This means that auction price is not a strong predictor of batting performance.
- Franchises may be influenced by:
- Player potential
- Role in the team (e.g., finisher, anchor)
- Marketability and fan base
- All-round capabilities (not captured in batting runs alone)

Conclusion

The analysis conducted in this project reveals that there is no strong or consistent correlation between the auction price of an IPL player and their actual on-field performance. While it is often assumed that higher investment guarantees better results, the data shows that several moderately priced players have outperformed their more expensive counterparts across multiple seasons. The correlation between auction value and performance metrics such as total runs, wickets taken, strike rate, and economy rate was generally weak, indicating that other factors—such as player consistency, match availability, and team composition—play a significant role in determining on-field impact.

Our study highlights the importance of adopting a data-driven approach during the auction process, rather than relying solely on reputation or past fame. Visualizations and correlation metrics support the idea that well-researched, mid-range players may offer greater returns on investment. Furthermore, teams that overspend on marquee players without considering long-term performance trends often face inefficiencies in squad composition.

In conclusion, franchises can significantly benefit from deeper performance analysis and cost-effectiveness evaluations during the auction phase. For future work, the study can be extended using predictive models to estimate player impact based on past seasons, and include additional factors like match conditions, venues, and opposition teams to enhance auction strategies.

APPENDICES

1. **GitHub Repository: https://github.com/Ankur669/IPL-DATA-ANALYSIS-**