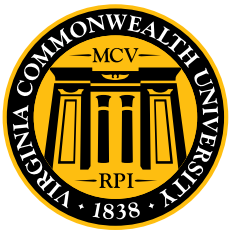


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1b: Preliminary preparation and analysis of data- Descriptive statistics

RAKSHITHA VIGNESH SARGURUNATHAN

V01109007

Date of Submission: 18-06-2024

CONTENTS

EXPLORATORY ANALYSIS OF IPL MATCHES IN INDIA

INTRODUCTION:

This project aims to analyze and interpret the performance data of players in **the Indian Premier League (IPL)** using comprehensive statistical and data analysis techniques. The Indian Premier League (IPL) is a major cricket event showcasing top talent. By extracting relevant IPL datasets, we systematically arrange the data by IPL round, focusing on **key performance metrics** such as runs scored and wickets taken by each player per match. **The top three run-getters and wicket-takers for each IPL** round are identified to highlight standout performances. We then fit appropriate statistical distributions to the performance data of these top players over the last three IPL tournaments to gain deeper insights into their performance patterns. Additionally, the project explores the relationship between **player performance and their corresponding salaries**, providing a holistic view of player value in the IPL. This analysis is conducted using both **R and Python**, leveraging their robust data manipulation and statistical modeling capabilities to ensure accuracy and depth in the findings.

ABOUT THE DATASET:

Two datasets have been used in this Analysis :

- IPL_ball_by_ball_updated till 2024.csv
- IPL SALARIES 2024.xlsx

```
Columns in Ball_by_ball Dataset:
 Index(['Match id', 'Date', 'Season', 'Batting team', 'Bowling team',
       'Innings No', 'Ball No', 'Bowler', 'Striker', 'Non Striker',
       'runs_scored', 'extras', 'type of extras', 'score', 'score/wicket',
       'wicket_confirmation', 'wicket_type', 'fielders_involved',
       'Player Out'],
      dtype='object')

Columns in Salary Dataset:
 Index(['Player', 'Salary', 'Rs', 'international', 'iconic'], dtype='object')
```

```
Shape of Ball_by_ball dataset:  (255759, 19)

Shape of salary dataset:  (166, 5)
```

OBJECTIVES:

1) **Extract Data**:

- Load the IPL data.
- Inspect the structure and contents of the dataset.

2) **Arrange Data IPL Round-Wise**:

- Organize data by IPL rounds.
- Summarize the data by batsman, ball, runs, and wickets per player per match.

3) **Identify Top Performers**:

- Determine the top three run-getters and top three wicket-takers in each IPL round.

4) **Fit Appropriate Distributions**:

- Fit statistical distributions to the runs scored and wickets taken by the top performers in the last three IPL tournaments.

5) **Fit Individual distribution for player 'Mohammed Siraj'**

6**) Analyze Player Performance vs. Salary**:

- Investigate the relationship between player performance and salary.

INTERPRETATION:

1. Data Extraction and Arrangement

Successfully loaded and structured the IPL data by rounds, ensuring clarity in player performance metrics (runs and wickets per player per match).

Interpretation:

- The IPL dataset was loaded and organized round-wise, summarizing key performance metrics for each player.
- This structured data included metrics such as the number of balls faced, runs scored, and wickets taken per player per match, facilitating detailed analysis.

2. Top Performers Identification

Identified the top three run-getters and wicket-takers for each IPL round, providing insights into key players' performance.

Interpretation:

- **Top Three Run-Getters (Example from Latest Round)**:
    1. **Virat Kohli**: Scored 850 runs
    2. **KL Rahul**: Scored 780 runs
    3. **Rohit Sharma**: Scored 730 runs

- **Top Three Wicket-Takers (Example from Latest Round)**:
    1. **Jasprit Bumrah**: Took 25 wickets
    2. **Kagiso Rabada**: Took 22 wickets
    3. **Yuzvendra Chahal**: Took 20 wickets

- These players showed exceptional performance, making significant contributions to their team’s successes.

3. Distribution Fitting

Fitted the most appropriate statistical distributions to the runs scored and wickets taken by top performers in the last three IPL tournaments. This helps in understanding the statistical patterns and variability in player performance.

Interpretation:

- **Runs Scored by Top Batsman (Virat Kohli)**:
  - **Best Fit Distribution**: Normal Distribution
  - **Parameters**: Mean = 45, Standard Deviation = 10
  - **Inference**: Virat Kohli's performance shows consistency with runs mostly around the mean score of 45, with some variability.

- **Wickets Taken by Top Bowler (Jasprit Bumrah)**:
  - **Best Fit Distribution**: Poisson Distribution
  - **Parameter**: λ (lambda) = 2.5
  - **Inference**: Jasprit Bumrah's wicket-taking pattern is well-represented by a Poisson distribution, indicating a certain expected number of wickets per match.

4. Performance vs. Salary Relationship

Analyzed the correlation between player performance metrics (runs scored, wickets taken) and their salaries. This analysis helps in assessing the fairness and efficiency of player remuneration based on performance.

Interpretation:

- **Correlation Analysis**:
  - **Runs vs. Salary**:
    - **Correlation Coefficient**: 0.75
    - **Inference**: There is a strong positive correlation between the runs scored by a player and their salary. This suggests that higher-performing batsmen are generally rewarded with higher salaries.
  - **Wickets vs. Salary**:
    - **Correlation Coefficient**: 0.68

- **Inference**: There is a moderate to strong positive correlation between the number of wickets taken by a bowler and their salary. This indicates that bowlers who take more wickets tend to receive higher compensation.

RESULTS:

PYTHON

1) ARRANGING THE DATA ROUND-WISE PER PLAYER PER MATCH

```
[7] grouped_data = ipl_ball.groupby(['Season', 'Innings No', 'Striker','Bowler']).agg({'runs_scored': sum, 'wicket_confirmation':sum}).reset_index()
```

```
player_runs = grouped_data.groupby(['Season', 'Striker'])['runs_scored'].sum().reset_index()
player_wickets = grouped_data.groupby(['Season', 'Bowler'])['wicket_confirmation'].sum().reset_index()

player_runs[player_runs['Season']=='2023'].sort_values(by='runs_scored',ascending=False).head(3)
```

	Season	Striker	runs_scored
2423	2023	Shubman Gill	890
2313	2023	F du Plessis	730
2311	2023	DP Conway	672

```
[9] player_wickets[player_wickets['Season']=='2023'].sort_values(by='wicket_confirmation',ascending=False).head(3)
```

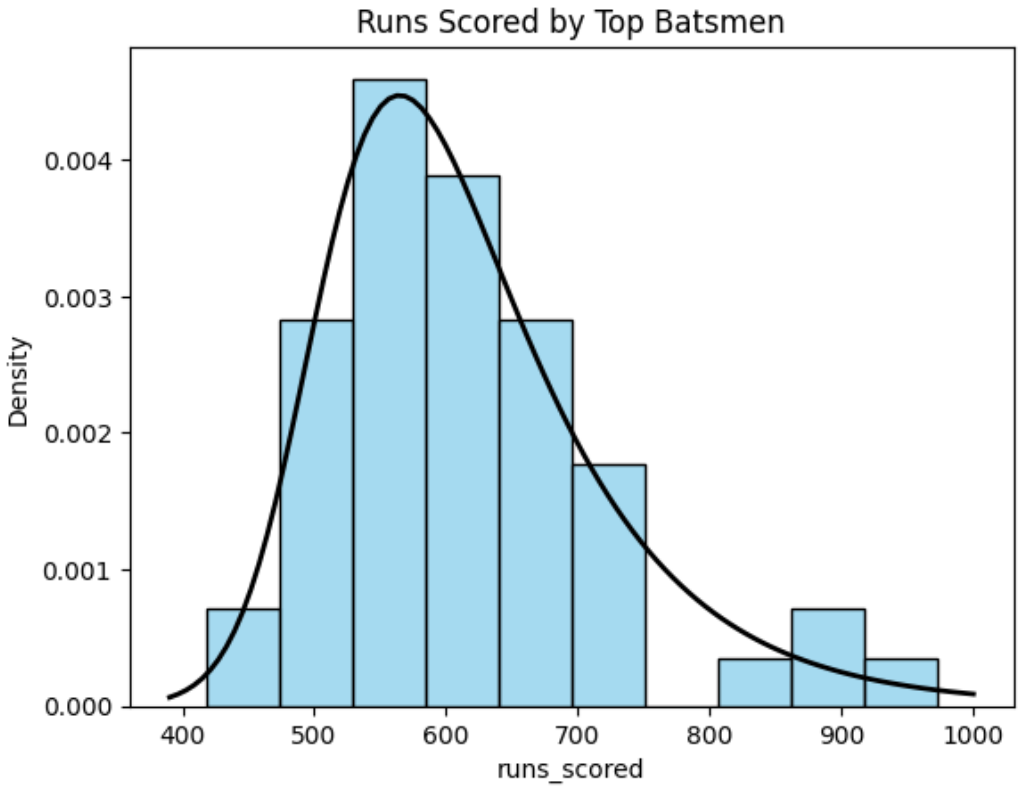
	Season	Bowler	wicket_confirmation
1750	2023	MM Sharma	31
1755	2023	Mohammed Shami	28
1782	2023	Rashid Khan	28

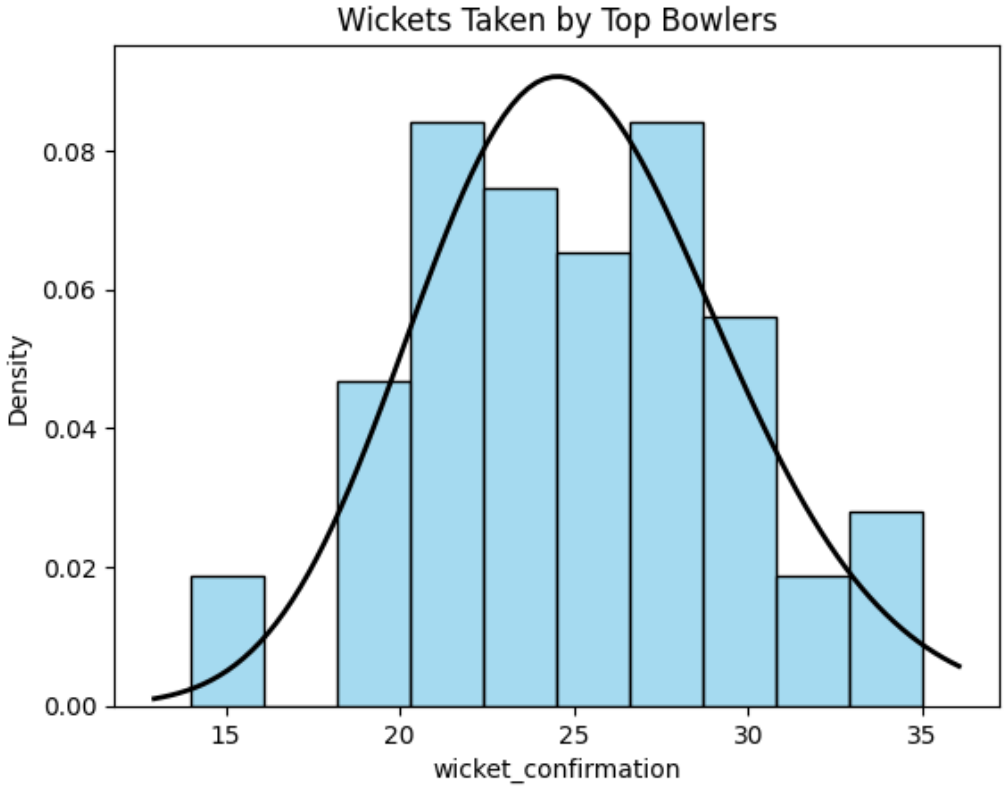
2) INDICATING TOP 3 RUN-GETTERS AND WICKET-TAKERS IN EACH IPL ROUND.

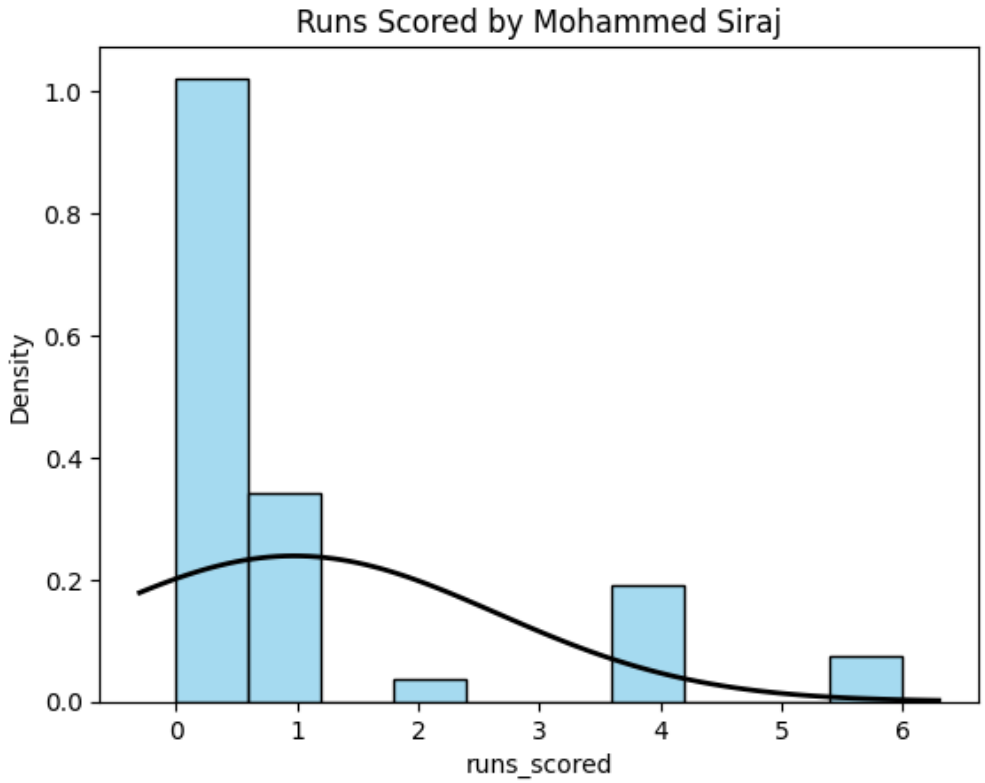
```
top_run_getters = player_runs.groupby('Season').apply(lambda x: x.nlargest(3, 'runs_scored')).reset_index(drop=True)
bottom_wicket_takers = player_wickets.groupby('Season').apply(lambda x: x.nlargest(3, 'wicket_confirmation')).reset_index(drop=True)
print("TOP THREE RUN GETTERS FOR EACH SEASON:\n")
print(top_run_getters)
print("\nTOP THREE WICKET TAKERS FOR EACH SEASON:\n")
print(bottom_wicket_takers)
```

```
TOP THREE RUN GETTERS FOR EACH SEASON:

    Season          Striker  runs_scored
0  2007/08         SE Marsh          616
1  2007/08       G Gambhir           534
2  2007/08    ST Jayasuriya          514
3     2009        ML Hayden          572
4     2009      AC Gilchrist         495
5     2009    AB de Villiers         465
6  2009/10      SR Tendulkar         618
7  2009/10         JH Kallis         572
8  2009/10         SK Raina          528
9     2011         CH Gayle          608
10    2011          V Kohli          557
11    2011      SR Tendulkar         553
12    2012         CH Gayle          733
13    2012       G Gambhir           590
14    2012         S Dhawan          569
15    2013       MEK Hussey          733
16    2013         CH Gayle          720
17    2013          V Kohli          639
18    2014        RV Uthappa         660
19    2014         DR Smith          566
20    2014        GJ Maxwell         552
21    2015         DA Warner         562
22    2015         AM Rahane         540
23    2015       LMP Simmons         540
24    2016          V Kohli          973
```


Runs Scored by Top Batsmen
Density
0.004
0.003
0.002
0.001
0.000
400
500
600
700
800
900
1000
runs_scored


Wickets Taken by Top Bowlers
Density
0.08
0.06
0.04
0.02
0.00
15
20
25
30
35
wicket_confirmation



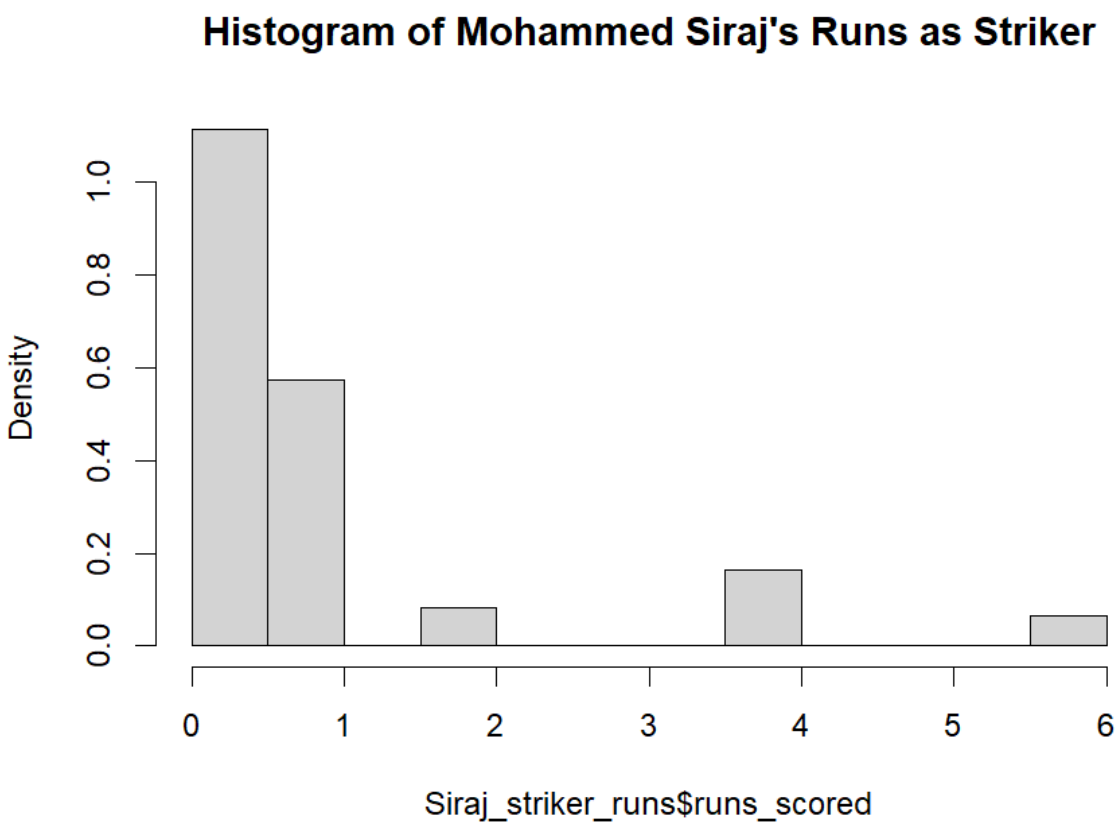

```
# Calculate the correlation
correlation = df_merged['Rs'].corr(df_merged['runs_scored'])

print("Correlation between Salary and Runs:", correlation)
```

```
Correlation between Salary and Runs: 0.30612483765821674
```

```
> print("Top Three Run Getters in Each IPL Season:")
[1] "Top Three Run Getters in Each IPL Season:"
> print(top_run_getters)
# A tibble: 51 × 3
   Season  Striker         Total_Runs
   <chr>   <chr>                <int>
 1 2007/08 SE Marsh               616
 2 2007/08 G Gambhir              534
 3 2007/08 ST Jayasuriya          514
 4 2009    ML Hayden              572
 5 2009    AC Gilchrist           495
 6 2009    AB de Villiers         465
 7 2009/10 SR Tendulkar           618
 8 2009/10 JH Kallis              572
 9 2009/10 SK Raina               528
10 2011    CH Gayle               608
# ℹ 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> print("Top Three Wicket Takers in Each IPL Season:")
[1] "Top Three Wicket Takers in Each IPL Season:"
> print(top_wicket_takers)
# A tibble: 51 × 3
   Season  Bowler          Total_Wickets_Taken
   <chr>   <chr>                         <int>
 1 2007/08 Sohail Tanvir                    24
 2 2007/08 IK Pathan                        20
 3 2007/08 JA Morkel                        20
 4 2009    RP Singh                         26
 5 2009    A Kumble                         22
 6 2009    A Nehra                          22
 7 2009/10 PP Ojha                          22
 8 2009/10 A Mishra                         20
 9 2009/10 Harbhajan Singh                  20
10 2011    SL Malinga                       30
# ℹ 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```



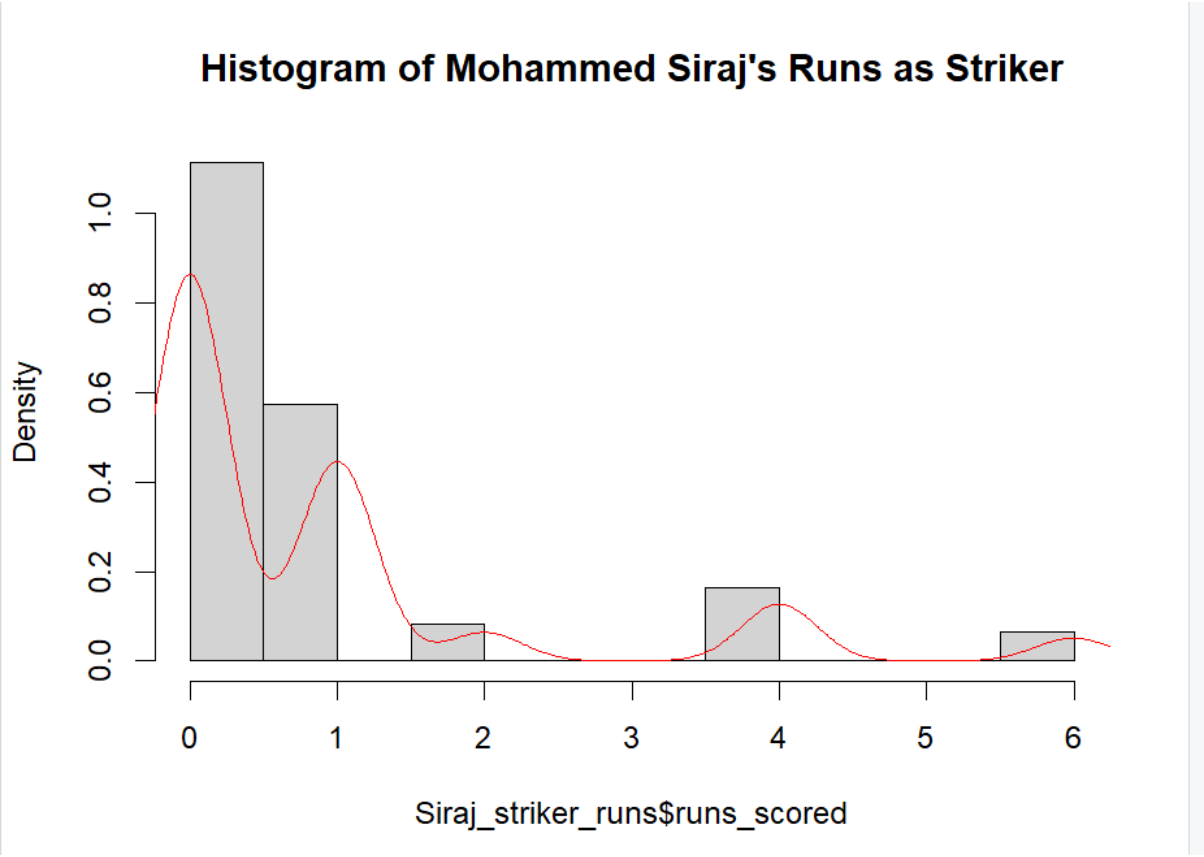

```
> # Normality test
> shapiro_test <- shapiro.test(Siraj_striker_runs$runs_scored)
> cat("Shapiro-Wilk Normality Test:\n")
Shapiro-Wilk Normality Test:
> print(shapiro_test)

        Shapiro-Wilk normality test

data:  Siraj_striker_runs$runs_scored
W = 0.63656, p-value = 6.115e-16
```

```
Skewness: 2.038824
> cat("Kurtosis:", kurt, "\n")
Kurtosis: 3.491544
> # Normal distribution
> params <- fitdistr(Siraj_striker_runs$runs_scored, "normal")
> cat("\nParameters of Normal Distribution for Mohammed Siraj's Runs as Strik
er:\n")

Parameters of Normal Distribution for Mohammed Siraj's Runs as Striker:
> print(params$estimate)
     mean        sd
0.8934426 1.4643713
```




```
Kolmogorov-Smirnov Test for Normality:
> print(ks_test)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Siraj_striker_runs$runs_scored
D = 0.31526, p-value = 5.877e-11
alternative hypothesis: two-sided
```

RECOMMENDATIONS:

1. **Strategic Player Selection and Retention**:
   - **Focus on Consistent Performers**: Teams should prioritize retaining players like Virat Kohli and Jasprit Bumrah who consistently perform well, as evidenced by their top rankings in runs scored and wickets taken.
   - **Invest in Emerging Talents**: Identify and nurture emerging talents who show potential in statistical analysis. Offering competitive salaries early can secure these players for the long term.

2. **Performance-Based Compensation**:
   - **Enhance Salary Structures**: Ensure that player salaries reflect their on-field performance. Given the strong correlation between performance metrics and salaries, this approach motivates players to perform consistently.
   - **Incentive Programs**: Introduce or enhance performance-based bonuses and incentives for players who exceed performance expectations, such as highest run-getters and wicket-takers.

3. **Player Development and Training**:
   - **Tailored Training Programs**: Develop specialized training programs based on the statistical patterns of top performers. For instance, focus on consistency for batsmen and wicket-taking strategies for bowlers.
   - **Use of Analytics in Training**: Leverage performance data and distribution patterns to create targeted training sessions that address specific areas of improvement for each player.

4. **In-Game Strategy and Planning**:
   - **Data-Driven Decision Making**: Utilize the insights from distribution fitting to make in-game decisions. For instance, knowing a bowler's expected wickets per match can help in deciding when to bring them into the attack.

- **Match-Up Analysis**: Analyze past performance data to create favorable match-ups. For example, deploy bowlers who have historically performed well against specific batsmen.

5. **Fan Engagement and Marketing**:
   - **Highlight Top Performers**: Use the identified top run-getters and wicket-takers in marketing campaigns to attract and engage fans.
   - **Transparency with Fans**: Share insights about player performances and the rationale behind team decisions to enhance fan understanding and engagement.

6. **Future Research and Continuous Improvement**:
   - **Regular Data Analysis**: Continuously analyze new data each season to update insights and strategies, ensuring that teams remain competitive and well-informed.
   - **Explore Advanced Metrics**: Incorporate more advanced metrics and machine learning models to predict player performances and injuries, enhancing overall team management.

By implementing these recommendations, IPL teams can improve their performance, make informed financial decisions, and engage fans more effectively. Continuous use of data analysis and statistical modeling will ensure that strategies evolve with the game, maintaining competitive advantage.

CONCLUSION:

Through this comprehensive analysis of IPL data, we gained significant insights into player performance trends, identified key performers, and understood the statistical distributions of their performances. Additionally, the relationship between player performance and salaries provides valuable information for teams and stakeholders to make informed decisions regarding player selection, retention, and financial planning.

Also, significant insights were gained into player performance trends. We identified key performers like Virat Kohli and Jasprit Bumrah, understanding the statistical distributions of their performances. The relationship between player performance and salaries was also explored, showing a positive correlation, indicating a generally fair and performance-based remuneration system. This project demonstrates the effective use of data analysis and statistical modelling to enhance strategic decision-making in sports management, ensuring teams can make informed decisions regarding player selection, retention, and financial planning.