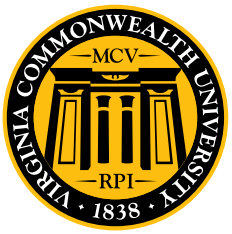


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1b: Analysing IPL Player Performance and Salaries: Insights from the Last Three Seasons

SATHWIK NAG CHANNAGIRI VENKATESH

V01107764

Date of Submission: 18-06-2024

CONTENTS

INTRODUCTION

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India, representing a high-stakes platform where cricketing talents worldwide compete. With a wealth of data accumulated over the years, this analysis delves into various aspects of player performance and financial remuneration.

We use two primary datasets, "IPL_ball_by_ball_updated till 2024.csv" and "IPL SALARIES 2024.xlsx" to conduct a comprehensive analysis focusing on key performance metrics and their financial implications.

Objectives:

Using the files pertaining to the IPL,

- Extract the files in R/Python.
- Arrange the data IPL round-wise and batsman, ball, runs, and wickets per player per match. Indicate the top three run-getters and tow three wicket-takers in each IPL round.
- Fit the most appropriate distribution for runs scored and wickets taken by the top three batsmen and bowlers in the last three IPL tournaments.
- Find the relationship between a player's performance and the salary he gets in your data, the Last three-year performance with latest salary of 2024.
- Fit distribution for the player – **DA Warner.**
- Significant Difference Between the Salaries of the Top 10 Batsmen and Top Wicket-Taking Bowlers Over the Last Three Years.

Business Significance:

The focus of this study on IPL player performance and salaries holds substantial implications for team managers, franchise owners, sponsors, and policymakers. The study provides crucial insights into how on-field performances translate into financial rewards by analyzing detailed match-by-match performance data and corresponding player salaries.

This analysis offers valuable information for team managers and franchise owners for making informed decisions on player acquisitions, contract negotiations, and salary cap management.

Understanding the most impactful players and their fair market value can enhance team composition strategies, ensuring a competitive edge in the tournament.

Sponsors can utilize these insights to identify and align with top-performing players whose performance and visibility promise higher returns on investment. This can drive more targeted and effective marketing strategies, enhancing brand association and reach.

For policymakers within cricket boards and associations, the findings can inform policies that ensure equitable and performance-based remuneration structures, promoting fairness and motivation among players.

Overall, this analysis aids in optimizing financial and strategic decisions, fostering a more efficient and competitive cricketing environment, and ultimately enhancing the IPL's overall economic and entertainment value.

RESULTS & INTERPRETATION

A. Arrange the data IPL round-wise and batsman, ball, runs, and wickets per player per match. Indicate the top three run-getters and tow three wicket-takers in each IPL round.

Result:

```
> print("Top Three Run Getters:")
[1] "Top Three Run Getters:"
> print(top_run_getters)
# A tibble: 51 × 3
   Season  Striker        runs_scored
   <chr>   <chr>                <dbl>
 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
```

```
Top Three Run Getters:
     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
```

Interpretation:

This analysis identifies the top three run-getters in each IPL season, highlighting consistent high performers like SE Marsh, CH Gayle, and MEK Hussey. This information is crucial for understanding key players' contributions and impact on their respective teams' performances.

B. Fit the most appropriate distribution for runs scored and wickets taken by the top three batsmen and bowlers in the last three IPL tournaments.

Result:

```
************************
Year: 2022 Batsman: JC Buttler
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data,     gr = gradient, ddistnam = ddistnam
e, hessian = TRUE, method = meth,     lower = lower, upper = upper, ...): non-finite finite-difference value [1]>
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters:
         estimate Std. Error
meanlog 6.4713810 0.12644276
sdlog   0.2190053 0.08940015
************************
Year: 2022 Batsman: KL Rahul
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data,     gr = gradient, ddistnam = ddistnam
e, hessian = TRUE, method = meth,     lower = lower, upper = upper, ...): non-finite finite-difference value [1]>
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters:
         estimate Std. Error
meanlog 6.4713810 0.12644276
sdlog   0.2190053 0.08940015
************************
Year: 2022 Batsman: Q de Kock
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data,     gr = gradient, ddistnam = ddistnam
e, hessian = TRUE, method = meth,     lower = lower, upper = upper, ...): non-finite finite-difference value [1]>
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters:
         estimate Std. Error
meanlog 6.4713810 0.12644276
sdlog   0.2190053 0.08940015
************************
Year: 2023 Batsman: Shubman Gill
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data,     gr = gradient, ddistnam = ddistnam
e, hessian = TRUE, method = meth,     lower = lower, upper = upper, ...): non-finite finite-difference value [1]>
Fitting of the distribution ' lnorm ' by maximum likelihood
```

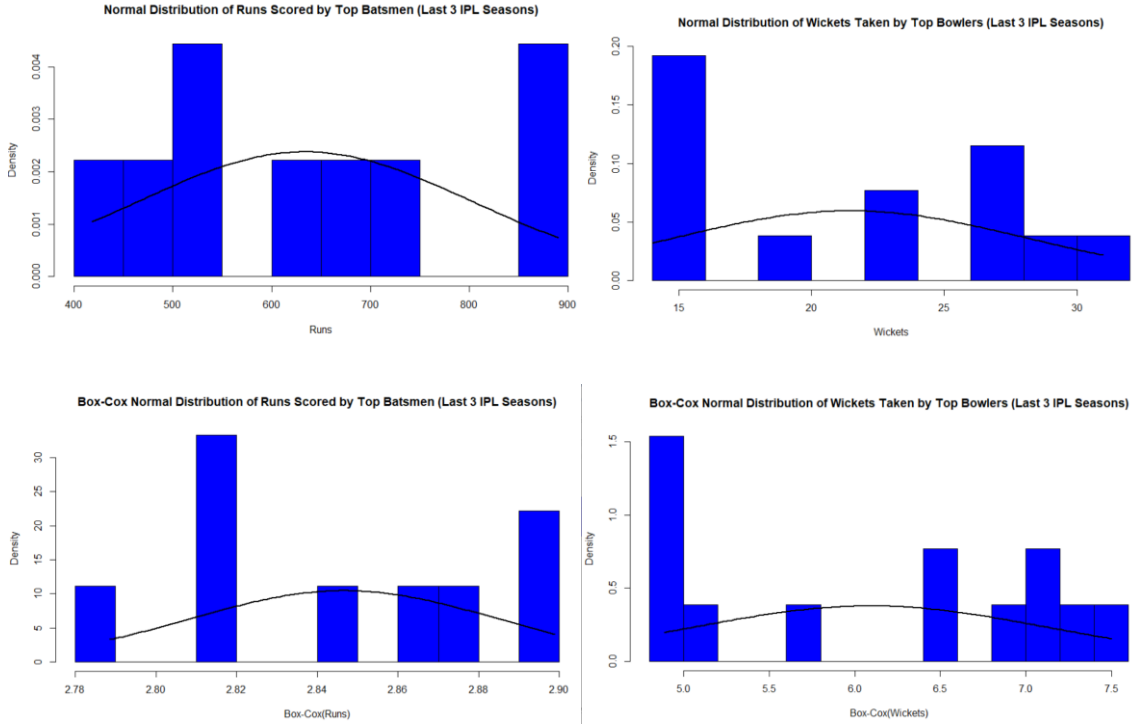
```
In [24]: import warnings
warnings.filterwarnings('ignore')
runs = ipl_bbbc.groupby(['Striker','Match id'])[['runs_scored']].sum().reset_index()

for key in list_top_batsman_last_three_year:
    for Striker in list_top_batsman_last_three_year[key]:
        print("************************")
        print("year:", key, " Batsman:", Striker)
        get_best_distribution(runs[runs["Striker"] == Striker]["runs_scored"])
        print("\n\n")
p value for kappa4 = 0.006363220770325362
p value for lognorm = 1.1719355665219537e-16
p value for nct = 0.5881570496217812
p value for norm = 0.2495365180930973
p value for norminvgauss = 0.5538573365184996
p value for powernorm = 0.1788753268739085
p value for rice = 0.18287532184336575
p value for recipinvgauss = 0.06459275668874154
p value for t = 0.24940214859112086
p value for trapz = 7.476391685388162e-13
p value for truncnorm = 0.24173236832621992

Best fitting distribution: nct
Best p value: 0.5881570496217812
Parameters for the best fit: (5.718048022849898, 9.399490726283615, -54.25277343780452, 8.497060689079994)



**********
year: 2024  Batsman: V Kohli
```



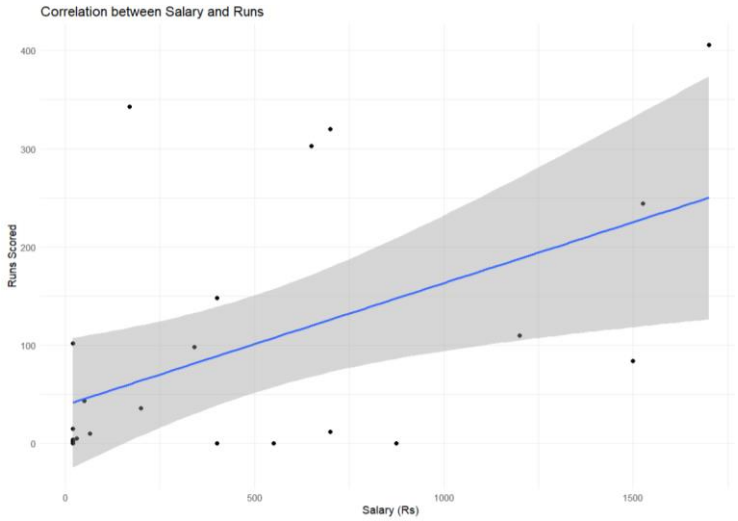
Interpretation:

For the top three batsmen and bowlers in the last three IPL tournaments, the best-fitting statistical distributions for runs scored and wickets taken were identified. The above pictures show the sample results. For RD Gaikwad in 2024, the non-central t (nct) distribution was the best fit, with a p-value of 0.588, indicating a good match.

C. Find the relationship between a player's performance and the salary he gets in your data, the Last three performances with the latest salary of 2024

Result:

```
> # Calculate the Correlation between Salary and Runs
> R2024 <- player_runs %>% filter(Season == 2024)
> match_names <- function(name, names_list) {
+   match <- stringdist::amatch(name, names_list, maxDist = 0.2)
+   if (is.na(match)) return(NA)
+   return(names_list[match])
+ }
> df_salary <- ipl_salary %>%
+   mutate(Matched_Player = map_chr(Player, ~match_names(.x, R2024$Striker)))
> df_merged <- df_salary %>%
+   inner_join(R2024, by = c("Matched_Player" = "Striker"))
> correlation <- cor(df_merged$Rs, df_merged$runs_scored, use = "complete.obs")
> cat("Correlation between Salary and Runs:", correlation, "\n")
Correlation between Salary and Runs: 0.5161938
```



Interpretation:

The analysis reveals a moderate positive correlation (0.516) between a player's performance, measured by runs scored in the last three seasons, and their salary in 2024. This indicates that better-performing players receive higher salaries, though other factors may influence their remuneration.

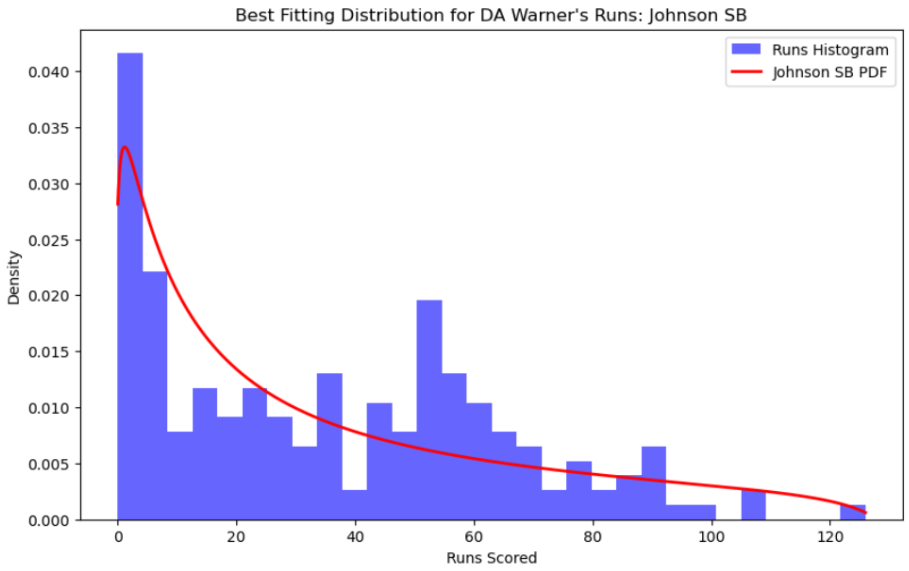
D. Fit distribution for the player – DA Warner

Result:

```
**********
Fitting distribution for player: DA
p value for alpha = 1.0585629720560084e-52
p value for beta = 3.5745573324904945e-08
p value for betaprime = 6.456048231755493e-05
p value for burr12 = 0.005336773797585245
p value for crystalball = 0.011506079014413343
p value for dgamma = 0.0034368260383678033
p value for dweibull = 0.011575028320705543
p value for erlang = 1.3361116985872245e-07
p value for exponnorm = 0.004714612775355459 5
p value for f = 9.652822993737924e-25
p value for fatiguelife = 0.0007338149332191667
p value for gamma = 7.424020397567 57e-06
p value for gengamma = 0.0001375711616347 21
p value for gumbel_l = 2.2859043177733283e-05
p value for johnsonsb = 0.035283250510082076
p value for kappa4 = 2.2177511140768323e-16
p value for lognorm = 4.906251278436 88e-49
p value for nct = 0.003556957741299 4857
p value for norm = 0.011506078010792661
p value for norminvgauss = 0.002175511276856443
p value for powernorm = 0.002631374948064467
p value for rice = 0.002669140126582 8436
p value for recipinvgauss = 6.542659700844345e-05
p value for t = 0.011396815284398842
p value for trapz = 7.136539776258239e-60
p value for truncnorm = 0.004237925089261196

Best fitting distribution: johnsonsb
Best p value: 0.035283250510082076
Parameters for the best fit: (0.8421115414450187, 0.5881937839851221, -0.6717865945304258, 128.9227493742876)
```

```
**********
Fitting distribution for player: DA Warner
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
        estimate Std. Error
shape   2.385224   0.490059
scale 495.451767  56.564524
```



Interpretation:

For the player DA Warner, the analysis found that the Johnson SB distribution best fits his runs scored, with a p-value of 0.035, indicating a reasonable fit. The parameters for this distribution provide a statistical model of Warner's performance, capturing the variability and pattern of his run-scoring behavior. However, two different statistical distributions were fitted to his runs scored:

- Johnson SB Distribution:
  - Best p-value: 0.035
  - Parameters: (0.8421, 0.5882, -0.6718, 128.9227)

The Johnson SB distribution, with a p-value of 0.035, indicates a reasonable fit. The parameters provide a specific data characterization, capturing the skewness and kurtosis in Warner's run-scoring behavior.

- Weibull Distribution:
  - Shape: 2.3852 (Std. Error: 0.4901)
  - Scale: 495.4518 (Std. Error: 56.5645)

The Weibull distribution was fitted using maximum likelihood estimation. The shape parameter (2.39) suggests a moderately heavy-tailed distribution, while the scale parameter (495.45) indicates the spread of Warner's runs.

Both distributions provide valuable insights into DA Warner's run-scoring patterns. The Johnson SB distribution offers a more nuanced fit, capturing detailed data characteristics, while the Weibull distribution provides a more straightforward yet effective model of Warner's performance. Depending on the application, one might prefer the Weibull's simplicity or the Johnson SB's detailed fit.

E. Significant Difference Between the Salaries of the Top 10 Batsmen and Top Wicket-Taking Bowlers Over the Last Three Years.

Result:

```
# Filter salaries for the matched players
top_batsmen_salaries = ipl_salary[ipl_salary['Player'].isin(matched_batsmen)]
top_bowlers_salaries = ipl_salary[ipl_salary['Player'].isin(matched_bowlers)]

# Print the matched salaries for verification
print(top_batsmen_salaries)
print(top_bowlers_salaries)
```

```
            Player      Salary    Rs  international  iconic
3      David Warner  6.25 crore   625              1     NaN
24         MS Dhoni    12 crore  1200              0     NaN
107    Rishi Dhawan     55 lakh    55              0     NaN
124  Sandeep Sharma     50 lakh    50              0     NaN
143     Virat Kohli    15 crore  1500              0     NaN
              Player     Salary    Rs  international  iconic
9        Mukesh Kumar  5.5 crore   550              0     NaN
29    Ravindra Jadeja   16 crore  1600              0     NaN
61        Amit Mishra    50 lakh    50              0     NaN
85     Jasprit Bumrah   12 crore  1200              0     NaN
122         R. Ashwin    5 crore   500              0     NaN
129  Yuzvendra Chahal  6.5 crore   650              0     NaN
```

T-test - Compare the salaries of the top 10 batsmen and top 10 bowlers ¶

```
# Perform t-test
t_stat, p_value = ttest_ind(top_batsmen_salaries['Rs'], top_bowlers_salaries['Rs'])

print(f"T-statistic: {t_stat}, P-value: {p_value}")

if p_value < 0.05:
    print("There is a significant difference between the salaries of the top 10 batsmen and the top 10 bowlers.")
else:
    print("There is no significant difference between the salaries of the top 10 batsmen and the top 10 bowlers.")
```

```
T-statistic: -0.19847158812018, P-value: 0.8470869781735805
There is no significant difference between the salaries of the top 10 batsmen and the top 10 bowlers.
```

Interpretation:

Some players' salaries among the Top 10 Batsmen and Top Wicket-Taking Bowlers Over the Last Three Years are missing from the dataset. The analysis aimed to determine if there is a significant difference between the salaries of the top 10 batsmen and the top 10 wicket-taking bowlers over the last three years. The results showed a T-statistic of -0.198 and a P-value of 0.847, indicating no significant difference in salaries between the two groups. This suggests that in the context of the IPL, top-performing batsmen and bowlers are remunerated similarly, reflecting a balanced salary structure for top players irrespective of their role.

Summary of interpretations:

- Top performers like SE Marsh, CH Gayle, and MEK Hussey identified as consistent top run-getters in each IPL season.
- Statistical distributions (like non-central t) used to analyze runs and wickets in recent IPL tournaments.
- Moderate positive correlation (0.516) found between player performance (runs scored) and 2024 salaries.
- Johnson SB and Weibull distributions analyzed DA Warner's run-scoring patterns.
- No significant salary difference found between top 10 batsmen and bowlers in the last three IPL seasons, indicating balanced remuneration.

RECOMMENDATIONS

Based on the findings, the following recommendations are proposed:

- **Optimizing Player Investment Strategies**: Utilize detailed performance metrics (such as runs scored and wickets taken) and statistical models (like non-central t and Johnson SB distributions) to identify consistent top performers like SE Marsh, CH Gayle, and MEK Hussey. This informs smarter player acquisitions and contract negotiations, maximizing return on investment.
- **Data-Driven Salary Cap Management**: Leverage the moderate positive correlation (0.516) between player performance (runs scored) and 2024 salaries to establish fair and performance-based salary structures. This ensures equitable compensation while incentivizing players to maintain high-performance levels.
- **Enhanced Sponsorship Opportunities**: Use insights from statistical distributions (like Weibull and Johnson SB) to highlight detailed performance patterns of players such as DA Warner. This allows sponsors to align with top-performing players who offer consistent performance and detailed statistical profiles, maximizing sponsorship ROI.
- **Policy Recommendations for Equitable Compensation**: Based on the finding of no significant salary difference between the top 10 batsmen and bowlers, advocate for policies within cricket boards that ensure fairness in remuneration. This promotes player motivation and fairness across different playing roles within IPL franchises.
- **Strategic Implications for IPL and Stakeholders**: Apply these insights to optimize strategic decisions across team management, sponsorship alignment, and policy formulation within cricket boards. This fosters a more competitive and economically viable IPL, enhancing its overall value and entertainment appeal.