****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical Analysis and Modelling (SCMA 632)**

**A2:**

**NAVIYA MARY VINOD**

**V01108251**

**Date of Submission: 23-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objective | **2** |
| **3.** | Business Significance | **2** |
|  | Results and Interpretation of R and Python | **2** |
|  | Conclusion | **6** |

**INDIAN PREMIER LEAGUE PLAYER DATA ANALYSIS USING PYTHON**

**Introduction**

The Indian Premier League (IPL) is one of the most prominent and widely followed Twenty20 cricket leagues globally. Established in 2008, the IPL features top cricketers from around the world, representing various franchise teams in a highly competitive and entertaining format. This dataset comprises detailed information on IPL matches, providing a granular view of the performance of players and teams over multiple seasons.

#### Dataset Description

The IPL dataset is divided into two primary components:

**Ball-by-Ball Data**: This dataset contains comprehensive information on every ball bowled in IPL matches up until 2024. It includes details such as the match ID, inning number, over number, ball number, batsman, bowler, runs scored, and wicket information. This fine-grained data allows for an in-depth analysis of player performances and match dynamics.

Key Columns:

* 1. match\_id: Unique identifier for each match.
  2. inning: Inning number (1 or 2) of the match.
  3. over: Over number within the inning.
  4. ball: Ball number within the over.
  5. batsman: Name of the batsman facing the ball.
  6. bowler: Name of the bowler delivering the ball.
  7. runs\_off\_bat: Runs scored off the bat on that delivery.
  8. extras: Extra runs awarded (like wides, no-balls).
  9. wicket\_type: Type of dismissal if a wicket falls on that delivery.

**Salary Data**: This dataset includes the salaries of IPL players for the 2024 season. It provides insights into the financial aspects of the league, showing how player salaries correlate with their on-field performance.

Key Columns:

* 1. Player: Name of the player.
  2. Team: The IPL team for which the player is contracted.
  3. Salary (INR): Salary of the player for the 2024 season.

#### Objectives

1. Organize IPL data round-wise to identify the top three run-getters and top three wicket-takers for each round.
2. Fit statistical distributions to the runs scored and wickets taken by top players in the last three IPL tournaments.
3. Examine the relationship between players' performance metrics (runs and wickets) and their salaries, focusing on the latest data from 2024.
4. Perform significance tests to compare the salaries of the top 10 batsmen and top wicket-taking bowlers over the past three years.

**Business Significance**

Analyzing IPL performance data and its correlation with player salaries provides valuable insights for team management and decision-making. Understanding player performance trends helps franchises make informed choices during player auctions and contract negotiations, ensuring optimal investment in talent. This analysis also aids in identifying undervalued players, enabling teams to build competitive squads within budget constraints. Furthermore, such data-driven strategies enhance fan engagement by showcasing the analytical aspects of team building and player performance evaluation.

**Results And Interpretation**

1. Using IPL data, establish the relationship between the player's performance and payment he receives and discuss your findings. Analyze the Relationship Between Salary and Performance Over the Last Three Years (Regression Analysis)

\* Use the data sets [data “IPL\_ball\_by\_ball\_updated till 2024.csv”]

### Relationship Between Player Performance and Payment in IPL

#### Regression Analysis: Player Salary vs. Runs Scored

**Regression Results**:

* **Dependent Variable**: Player Salary (in Rs)
* **Independent Variable**: Runs Scored

**Summary**:

* **R-squared**: 0.094
* **Adjusted R-squared**: 0.089
* **F-statistic**: 18.80, significant at p<0.05p < 0.05p<0.05 (Prob: 2.41e−052.41e-052.41e−05)
* **Coefficients**:
  + Intercept (const): 410.5744
  + Runs Scored: 0.7817

**Interpretation**: The regression model indicates that there is a positive relationship between the number of runs scored by a player and the salary they receive. The coefficient for runs scored (0.7817) suggests that for every additional run scored, the player’s salary increases by 0.7817 units of Rs (currency units in the dataset). The intercept (410.5744) represents the baseline salary when no runs are scored.

The R-squared value of 0.094 indicates that approximately 9.4% of the variability in player salaries can be explained by the number of runs scored. This relatively low R-squared value suggests that while runs scored have a statistically significant impact on salary, many other factors likely influence player salaries.

The Omnibus and Jarque-Bera tests indicate that the residuals are not normally distributed (significant at p<0.05p < 0.05p<0.05), and the Durbin-Watson statistic (2.151) suggests that there is no significant autocorrelation in the residuals.

#### (b)Regression Analysis: Player Salary vs. Wickets Taken

**Regression Results**:

* **Dependent Variable**: Player Salary (in Rs)
* **Independent Variable**: Wickets Taken

**Summary**:

* **R-squared**: 0.057
* **Adjusted R-squared**: 0.037
* **F-statistic**: 2.799, not significant at p<0.05p < 0.05p<0.05 (Prob: 0.101)
* **Coefficients**:
  + Intercept (const): 353.7065
  + Wickets Taken: 13.9594

**Interpretation**: The regression model suggests a positive relationship between the number of wickets taken by a player and the salary they receive. The coefficient for wickets taken (13.9594) indicates that each additional wicket taken increases the player’s salary by 13.9594 units of Rs. The intercept (353.7065) represents the baseline salary when no wickets are taken.

However, the R-squared value of 0.057 shows that only 5.7% of the variability in player salaries can be explained by the number of wickets taken. The relatively low R-squared and the non-significant F-statistic (Prob: 0.101) suggest that wickets taken alone do not have a significant impact on player salaries and that other factors must be influencing salaries.

The Omnibus and Jarque-Bera tests for normality indicate that the residuals are not normally distributed (significant at p<0.05p < 0.05p<0.05), and the Durbin-Watson statistic (1.920) suggests no significant autocorrelation in the residuals.

### Discussion of Findings:

**Player Performance and Salary**:

1. There is a positive relationship between player performance metrics (runs scored and wickets taken) and their salaries. However, the low R-squared values in both models suggest that these performance metrics only explain a small portion of the variance in player salaries.
2. This implies that while performance on the field does contribute to salary determination, many other factors are likely at play. These factors could include player marketability, experience, leadership qualities, team composition needs, and off-field contributions.

**Significance and Model Fit**:

1. The runs scored model is statistically significant, indicating that runs scored are a significant predictor of player salary. However, the wickets taken model is not statistically significant, suggesting that wickets taken alone do not significantly influence salaries.
2. The non-normality of residuals in both models indicates potential issues with model specification or the presence of outliers. This could be addressed by transforming the variables or using robust regression techniques.

**Additional Considerations**:

1. The analysis could be improved by including more predictors such as player experience, age, match impact scores, endorsements, and other performance metrics like strike rate or economy rate.
2. Including interaction terms or using a mixed-effects model to account for team and season variations could also provide a more comprehensive understanding of the factors influencing player salaries.

### Conclusion

While player performance in terms of runs scored and wickets taken has a positive impact on their salaries, these metrics alone do not fully explain the variations in salaries. The low R-squared values indicate the need to consider additional factors and more complex models to accurately capture the determinants of player salaries in the IPL. Further analysis with a broader set of variables and advanced modeling techniques would provide a more detailed understanding of the relationship between player performance and remuneration.

**R**

### (a)Establishing the Relationship Between Player Performance and Salary in the IPL

Using the given IPL data, we conducted regression analyses to establish the relationship between player performance (measured by runs scored and wickets taken) and the salaries they received. We examined data from the last three years (2021-2023) for runs scored and the year 2022 for wickets taken.

#### Analysis of Runs Scored vs. Salary

**Regression Model**:

* **Dependent Variable**: Salary (in Rs)
* **Independent Variable**: Runs Scored

**Results**:

* **Intercept**: 360.666
* **Coefficient for Runs Scored**: 1.087
* **R-squared**: 0.1721
* **Adjusted R-squared**: 0.1694
* **F-statistic**: 63.84 (p-value: 2.752e−142.752e-142.752e−14)

**Interpretation**: The regression model demonstrates a positive relationship between the number of runs scored by a player and their salary. The coefficient of 1.087 indicates that for every additional run scored, the player's salary increases by 1.087 units of Rs. The intercept of 360.666 represents the baseline salary when no runs are scored.

The R-squared value of 0.1721 suggests that 17.21% of the variance in salaries can be explained by the number of runs scored. While this is a significant portion, it indicates that other factors also play a substantial role in determining salaries. The model is statistically significant, as indicated by the p-value of the F-statistic being much less than 0.05.

The residuals indicate some spread around the predicted values, but the overall trend confirms that higher run-scoring players tend to receive higher salaries.

#### Analysis of Wickets Taken vs. Salary

**Regression Model**:

* **Dependent Variable**: Salary (in Rs)
* **Independent Variable**: Wickets Taken

**Results**:

* **Intercept**: 89.94
* **Coefficient for Wickets Taken**: 27.22
* **R-squared**: 0.08356
* **Adjusted R-squared**: 0.03774
* **F-statistic**: 1.824 (p-value: 0.192)

**Interpretation**: The regression model for wickets taken shows a positive but not statistically significant relationship with player salaries. The coefficient of 27.22 indicates that for each additional wicket taken, the player's salary increases by 27.22 units of Rs. The intercept of 89.94 represents the baseline salary when no wickets are taken.

The R-squared value of 0.08356 suggests that only 8.356% of the variance in salaries can be explained by the number of wickets taken. The adjusted R-squared value is even lower at 0.03774, indicating a weak explanatory power. The p-value of the F-statistic (0.192) is greater than 0.05, suggesting that the model is not statistically significant.

The residuals show a wide spread, indicating that many other factors influence player salaries beyond just wickets taken.

### Discussion of Findings

**Runs Scored and Salary**:

1. There is a statistically significant positive relationship between runs scored and player salaries. Players who score more runs tend to receive higher salaries. This relationship is supported by the significant F-statistic and the positive coefficient for runs scored.
2. However, with an R-squared value of 0.1721, runs scored explain only a part of the variability in salaries, suggesting that other factors (e.g., player marketability, experience, team dynamics) also significantly influence salaries.

**Wickets Taken and Salary**:

1. The relationship between wickets taken and player salaries is positive but not statistically significant. This indicates that while players who take more wickets may receive higher salaries, the number of wickets taken alone does not significantly determine salaries.
2. The low R-squared value and non-significant F-statistic imply that many other factors must be considered when assessing the determinants of player salaries for bowlers.

**Additional Factors**:

1. The relatively low explanatory power of the models indicates the importance of other variables not included in these analyses. Factors such as player experience, international reputation, marketability, endorsements, and specific team needs likely play critical roles in determining player salaries.
2. Future analyses could benefit from including these additional variables and exploring interaction effects between different performance metrics.

### Conclusion

While player performance metrics like runs scored and wickets taken do impact salaries, they do not fully explain the variability in salaries. Runs scored have a statistically significant positive relationship with salaries, whereas wickets taken do not show a significant impact. These findings highlight the complexity of salary determination in the IPL, influenced by a myriad of factors beyond on-field performance. Further research incorporating a broader range of variables would provide a more comprehensive understanding of the determinants of player salaries in the IPL.