****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modeling (SCMA 632)**

**A1b: Preliminary preparation and analysis of data- Descriptive statistics**

**SAMPREETH S SHETTY**

**V01107634**

**Date of Submission: 18-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **3** |
| **4.** | Recommendations | **-** |
| **5.** | Codes | **-** |
| **6.** | References | **-** |

**Introduction**

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India, which is usually held between March and May every year by teams representing eight cities or states of India. The club was founded by the Board of Control for Cricket in India (BCCI) in 2008 and has since become one of the most popular and competitive cricket leagues in the world.

The IPL is known for its high-scoring matches, thrilling finishes, and the involvement of cricketing stars from around the globe. It has been a platform for young Indian cricketers to showcase their talents and rub shoulders with some of the best in the game. The league follows a double round-robin format followed by playoffs and the final.

The IPL also has a massive fan following and is known for its entertainment value, with cheerleaders and music adding to the festive atmosphere of the matches. It has significantly contributed to the global popularity of cricket, bringing in significant revenues through broadcasting rights, sponsorships, and ticket sales.

The data sets used in this project are titled ‘IPL\_ball\_by\_ball\_updated till 2024’ and ‘IPL SALARIES 2024’. The former contains details about the ball-by-ball account of all the matches from the tournament from 2008 to 2024, while the latter has information regarding the salaries earned by the players over the seasons, also from 2008 to 2024.

The player assigned to this project work was named ‘RD Gaikwad,’ one of the top-performing batsmen. The project aims to fit the most suitable probability distribution for the top three batsmen and bowlers and the one player assigned using Python and R.

After obtaining the probability distribution, the relationship between the runs scored by a player and the salary of said player was also determined by computing the coefficient of correlation.

**Results**

The top three batsmen and bowlers were fitted with probability distributions and the best fits are as follows.

Year 2024, Batsman: RD Gaikwad

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: Virat Kohli

Best fitting distribution: beta

Best p-value: 0.7807091136830002

Parameters for the best fit: (0.816277299300862, 2.3391761669196907, -3.0251144495756596e-31, 130.79371484721577)

Year 2024, Batsman: B Sai Sudharsan

Best fitting distribution: f

Best p-value: 0.9743698554720728

Parameters for the best fit: (7.230079711691059, 94.80999484543659, -0.46870159044880233, 39.84202109781083)

Year 2024, Bowler: HV Patel

Best fitting distribution: alpha

Best p-value: 0.0002993252328930706

Parameters for the best fit: (5.200800514990576, -4.106246473111661, 27.580368990504883)

Year 2024, Bowler: Mukesh Kumar

Best fitting distribution: alpha

Best p-value: 0.6028771589628603

Parameters for the best fit: (6.113363581345144, -5.245777123804531, 39.57745263632695)

Year 2024, Bowler: Arshdeep Singh

Best fitting distribution: t

Best p-value: 0.004473243416689088

Parameters for the best fit: (4.822497644715119, 1.1162819391895469, 0.9153269129308039)

Correlation Coefficient for the correlation between Salary and runs

Correlation between Salary and Runs: 0.30612483765821674

**Interpretation**

**H\_0:** The distribution is a fit for the data.

**H\_1:** The distribution is not a fit for the data.

We identified the best-fitting distribution, a non-central t-distribution, with a p-value of 0.558 for the probability distribution of RD Gaikwad. This p-value suggests that the model is a decent fit to the data as a p-value higher than 0.05 indicates that the null hypothesis cannot be rejected. Hence, the model is fit for the distribution.

The first parameter (5.718) is the degree of freedom, the second (9.399) is the non-centrality parameter, and the third (-54.253) and the fourth (8.497) are the location and parameters scale, respectively.

The rest of the batsmen and bowlers have been fitted with the best-suited distribution with p-values greater than 0.05, thus rejecting the null hypothesis and the listed parameters, respectively.

The **Correlation Coefficient** of 0.306 indicates a positive but weak linear relationship between Salary and Runs Scored. This suggests that as the Salary increases, there is a chance that the number of runs might also increase.