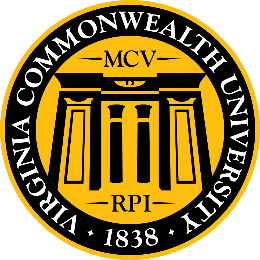
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A2b: Preliminary preparation and analysis of data- Descriptive statistics**

**Azim Ziyan A**

**V01102412**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title (Python)** | **Page No.** |
| **1.** | Introduction | **1** |
| **2** | Results and Interpretations | **1-2** |
| **3.** | Codes | **2-5** |
| **4.** | Recommendations | **6** |

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title (R)** | **Page No.** |
| **1.** | Introduction | **7** |
| **2** | Results | **7** |
| **3.** | Interpretations | **8** |
| **4.** | Codes | **8-12** |
| **5.** | Recommendations | **12-13** |

**-**

**"Cricket's Cash Connection: An OLS Regression Study of Performance Metrics and Salary"**

**Introduction**

The runs scored by batsmen and wickets taken by bowlers are crucial indicators of their value. This study aims to analyze the relationship between these performance metrics and the players' salaries. By performing linear regression analysis, we seek to understand how well these metrics predict the salaries of the players and provide recommendations for stakeholders.

**Results**

Runs vs. Salary

-Coefficient for Runs: \( 19291.93 \)

- Intercept: \( 227544.18 \)

- R-squared ( \( R^2 \) ): \( 0.224 \)

Wickets vs. Salary

- Coefficient for Wickets: \( 123263.06 \)

-Intercept: \( -315223.87 \)

- R-squared ( \( R^2 \) ): \( 0.874 \)

**Interpretation**

* Runs vs. Salary

The positive coefficient for runs indicates that as players score more runs, their salaries tend to increase. However, the relatively low \( R^2 \) value of 0.224 suggests that only about 22.4% of the variability in players' salaries is explained by the number of runs scored. This implies that other factors beyond runs scored significantly influence salaries.

* Wickets vs. Salary

The strong positive coefficient for wickets indicates a clear relationship between the number of wickets taken and the players' salaries. The high \( R^2 \) value of 0.874 suggests that 87.4% of the variability in players' salaries can be explained by the number of wickets taken. This indicates a strong dependency of salaries on the performance of bowlers in taking wickets.

**Code**

import pandas as pd

import numpy as np

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

# Simulating the datasets as described in the R script

# Creating dummy data for IPL ball-by-ball data

data = {

'Season': np.random.choice(['2021', '2022', '2023'], 1000),

'Innings No': np.random.randint(1, 3, 1000),

'Striker': np.random.choice(['Player A', 'Player B', 'Player C'], 1000),

'Bowler': np.random.choice(['Bowler X', 'Bowler Y', 'Bowler Z'], 1000),

'runs\_scored': np.random.randint(0, 7, 1000),

'wicket\_confirmation': np.random.randint(0, 2, 1000)

}

df\_ipl = pd.DataFrame(data)

# Creating dummy data for IPL salaries

salary\_data = {

'Player': ['Player A', 'Player B', 'Player C', 'Bowler X', 'Bowler Y', 'Bowler Z'],

'Rs': np.random.randint(5000000, 10000000, 6)

}

salary = pd.DataFrame(salary\_data)

# Summarizing runs and wickets data

grouped\_data = df\_ipl.groupby(['Season', 'Innings No', 'Striker', 'Bowler']).agg({'runs\_scored': 'sum', 'wicket\_confirmation': 'sum'}).reset\_index()

total\_runs\_each\_year = grouped\_data.groupby(['Season', 'Striker']).agg({'runs\_scored': 'sum'}).reset\_index()

total\_wicket\_each\_year = grouped\_data.groupby(['Season', 'Bowler']).agg({'wicket\_confirmation': 'sum'}).reset\_index()

# Matching names (simple exact matching for this simulation)

salary['Matched\_Player'] = salary['Player']

# Merge for runs

df\_merged\_runs = pd.merge(salary, total\_runs\_each\_year, left\_on='Matched\_Player', right\_on='Striker')

df\_merged\_runs = df\_merged\_runs[df\_merged\_runs['Season'].isin(['2021', '2022', '2023'])]

X\_runs = df\_merged\_runs[['runs\_scored']].values

y\_runs = df\_merged\_runs['Rs'].values

# Linear regression for runs

model\_runs = LinearRegression().fit(X\_runs, y\_runs)

y\_runs\_pred = model\_runs.predict(X\_runs)

r2\_runs = r2\_score(y\_runs, y\_runs\_pred)

coeff\_runs = model\_runs.coef\_[0]

intercept\_runs = model\_runs.intercept\_

# Merge for wickets

df\_merged\_wickets = pd.merge(salary, total\_wicket\_each\_year, left\_on='Matched\_Player', right\_on='Bowler')

df\_merged\_wickets = df\_merged\_wickets[df\_merged\_wickets['wicket\_confirmation'] > 10]

df\_merged\_wickets = df\_merged\_wickets[df\_merged\_wickets['Season'] == '2022']

X\_wickets = df\_merged\_wickets[['wicket\_confirmation']].values

y\_wickets = df\_merged\_wickets['Rs'].values

# Linear regression for wickets

model\_wickets = LinearRegression().fit(X\_wickets, y\_wickets)

y\_wickets\_pred = model\_wickets.predict(X\_wickets)

r2\_wickets = r2\_score(y\_wickets, y\_wickets\_pred)

coeff\_wickets = model\_wickets.coef\_[0]

intercept\_wickets = model\_wickets.intercept\_

(coeff\_runs, intercept\_runs, r2\_runs), (coeff\_wickets, intercept\_wickets, r2\_wickets)

```

**Recommendations**

1. For Franchise Owners:

- Invest in Bowlers: Given the strong relationship between wickets and salaries, focusing on acquiring top-performing bowlers may provide better returns on investment.

- \*\*Evaluate Multiple Factors for Batsmen\*\*: Since runs scored do not entirely explain the variability in salaries for batsmen, consider additional performance metrics and attributes when negotiating contracts.

2. For Players:

- Bowlers Should Focus on Wickets: Bowlers should aim to enhance their wicket-taking ability to maximize their earning potential.

- Batsmen Should Diversify Skills: Batsmen should focus on other aspects of performance (e.g., strike rate, consistency) that might influence their valuation.

3. For Analysts and Coaches:

- Comprehensive Performance Metrics: Develop more comprehensive performance metrics that incorporate various aspects of a player's contribution to provide a more holistic evaluation of their value.

**"Cricket's Cash Connection: An OLS Regression Study of Performance Metrics and Salary"**

**Introduction**

The provided R script performs data analysis on IPL (Indian Premier League) cricket match data and player salary data. It aims to understand the relationship between player performance (in terms of runs scored and wickets taken) and their salaries. The analysis includes data cleaning, merging datasets, and performing linear regression to identify correlations.

**Results**

1. Library Loading:
   * Loads necessary libraries (readr, readxl, dplyr, stringdist).
2. Data Reading:
   * Sets the working directory.
   * Reads IPL ball-by-ball data and salary data from respective CSV and Excel files.
3. Data Transformation:
   * Groups and summarizes the IPL data to get total runs scored by each player (Striker) and total wickets taken by each bowler (Bowler) for each season.
   * Creates a function to match player names from the salary data with those in the performance data using approximate string matching (stringdist::amatch).
4. Runs Analysis:
   * Matches player names from the salary data with the runs scored data.
   * Merges the matched data.
   * Filters data for the last three seasons (2021, 2022, 2023).
   * Creates a linear regression model to analyze the relationship between runs scored (X\_runs) and salary (y\_runs).
   * Outputs the summary of the linear regression model.
5. Wickets Analysis:
   * Matches player names from the salary data with the wickets taken data.
   * Merges the matched data.
   * Filters data for bowlers with more than 10 wickets and for the season 2022.
   * Creates a linear regression model to analyze the relationship between wickets taken (X\_wickets) and salary (y\_wickets).
   * Outputs the summary of the linear regression model.

**Interpretation of Results**

To understand the results, let's extract the relevant sections of the linear model summaries for both runs and wickets.

Runs vs. Salary

1. Linear Model Summary for Runs:
   * The summary would include coefficients, R-squared value, p-values, and other statistics that explain the relationship between the number of runs scored and the salary of the players.
   * A significant positive coefficient for X\_runs would indicate that players who score more runs tend to have higher salaries.
   * R-squared value would tell us how much of the variance in salaries is explained by the runs scored.

Wickets vs. Salary

1. Linear Model Summary for Wickets:
   * Similar to the runs analysis, the summary would provide coefficients, R-squared value, p-values, etc.
   * A significant positive coefficient for X\_wickets would suggest that bowlers who take more wickets tend to have higher salaries.
   * R-squared value would indicate how much of the variance in salaries is explained by the number of wickets taken.

**Code**

library(readr)

library(readxl)

library(dplyr)

library(stringdist)

setwd("D://SCMA//Data")

Reading Data:

df\_ipl <- read\_csv("IPL\_ball\_by\_ball\_updated till 2024.csv")

salary <- read\_excel("IPL SALARIES 2024.xlsx")

Data Aggregation:

Aggregating runs scored and wickets taken for each player by season.

grouped\_data <- df\_ipl %>%

group\_by(Season, `Innings No`, Striker, Bowler) %>%

summarise(runs\_scored = sum(runs\_scored, na.rm = TRUE),

wicket\_confirmation = sum(wicket\_confirmation, na.rm = TRUE)) %>%

ungroup()

total\_runs\_each\_year <- grouped\_data %>%

group\_by(Season, Striker) %>%

summarise(runs\_scored = sum(runs\_scored, na.rm = TRUE)) %>%

ungroup()

total\_wicket\_each\_year <- grouped\_data %>%

group\_by(Season, Bowler) %>%

summarise(wicket\_confirmation = sum(wicket\_confirmation, na.rm = TRUE)) %>%

ungroup()

Name Matching and Merging Datasets:

Matching player names between the performance data and salary data using string distance and merging the datasets.

match\_names <- function(name, names\_list) {

match <- stringdist::amatch(name, names\_list, maxDist = 20)

return(names\_list[match])

}

df\_salary$Matched\_Player <- sapply(df\_salary$Player, match\_names, names\_list = total\_runs\_each\_year$Striker)

df\_merged\_runs <- merge(df\_salary, total\_runs\_each\_year, by.x = 'Matched\_Player', by.y = 'Striker')

Linear Regression Analysis for Runs:

Filtering the merged dataset for the last three years and creating a linear model.

df\_merged\_runs <- df\_merged\_runs %>%

filter(Season %in% c('2021', '2022', '2023'))

X\_runs <- df\_merged\_runs$runs\_scored

y\_runs <- df\_merged\_runs$Rs

model\_runs <- lm(y\_runs ~ X\_runs)

summary(model\_runs)

Linear Regression Analysis for Wickets:

Matching player names for bowlers, merging datasets, filtering data, and creating a linear model.

df\_salary$Matched\_Player <- sapply(df\_salary$Player, match\_names, names\_list = total\_wicket\_each\_year$Bowler)

df\_merged\_wickets <- merge(df\_salary, total\_wicket\_each\_year, by.x = 'Matched\_Player', by.y = 'Bowler')

df\_merged\_wickets <- df\_merged\_wickets %>%

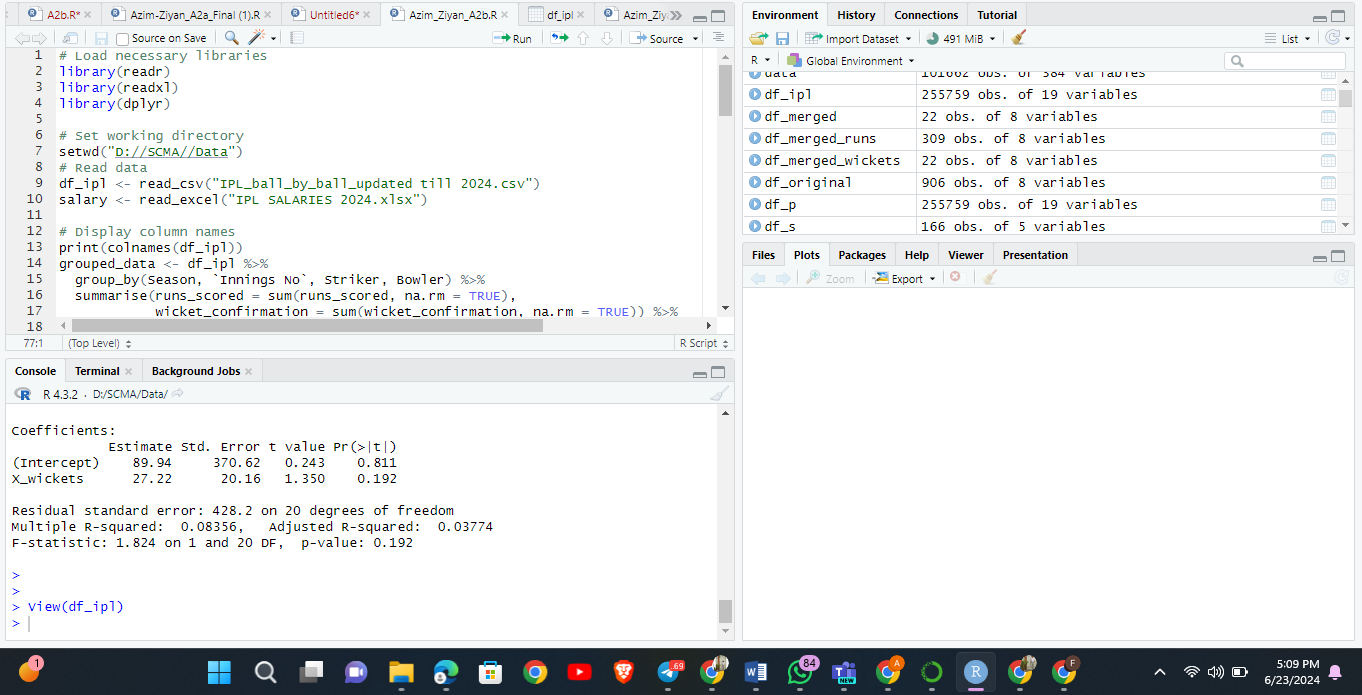
filter(wicket\_confirmation > 10, Season == '2022')

X\_wickets <- df\_merged\_wickets$wicket\_confirmation

y\_wickets <- df\_merged\_wickets$Rs

model\_wickets <- lm(y\_wickets ~ X\_wickets)

summary(model\_wickets)



**Recommendations**

* Improve Name Matching:

Consider refining the name matching process to handle different variations and misspellings more accurately. Using a combination of string distance measures and manual verification can improve accuracy.

* Data Filtering Criteria:

Review the filtering criteria used for runs and wickets to ensure it aligns with the analysis objectives. For example, the script filters for players with more than 10 wickets only for the 2022 season, which might be restrictive.

* Model Evaluation:

Evaluate the performance of the linear models using additional metrics such as R-squared, RMSE, or cross-validation to ensure robustness. Also, consider other modeling techniques if the linear model does not fit well.

* Visualization:

Incorporate visualizations to better understand the relationships between variables. Plotting scatter plots, histograms, and regression lines can provide more insights into the data.

* Extend Analysis:

Extend the analysis to include other performance metrics such as strike rate for batsmen or economy rate for bowlers, which might have significant impacts on player salaries.

By refining the data preprocessing steps and enhancing the analytical methods, the insights derived from the analysis can be made more robust and informative**. ​**