VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1b: Analysis of IPL DATA of player performance

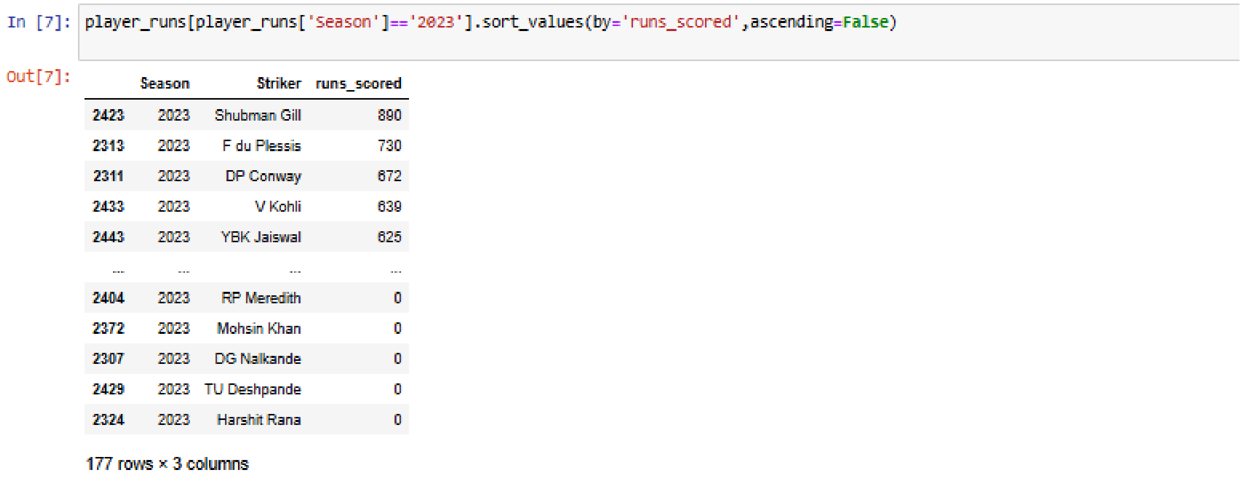
SHREYA MISHRA

V01108272

Date of Submission: 21-06-2024

Report on Players Performance in the IPL

- this data contains information about the top and bottom run scorers for the year 2023.



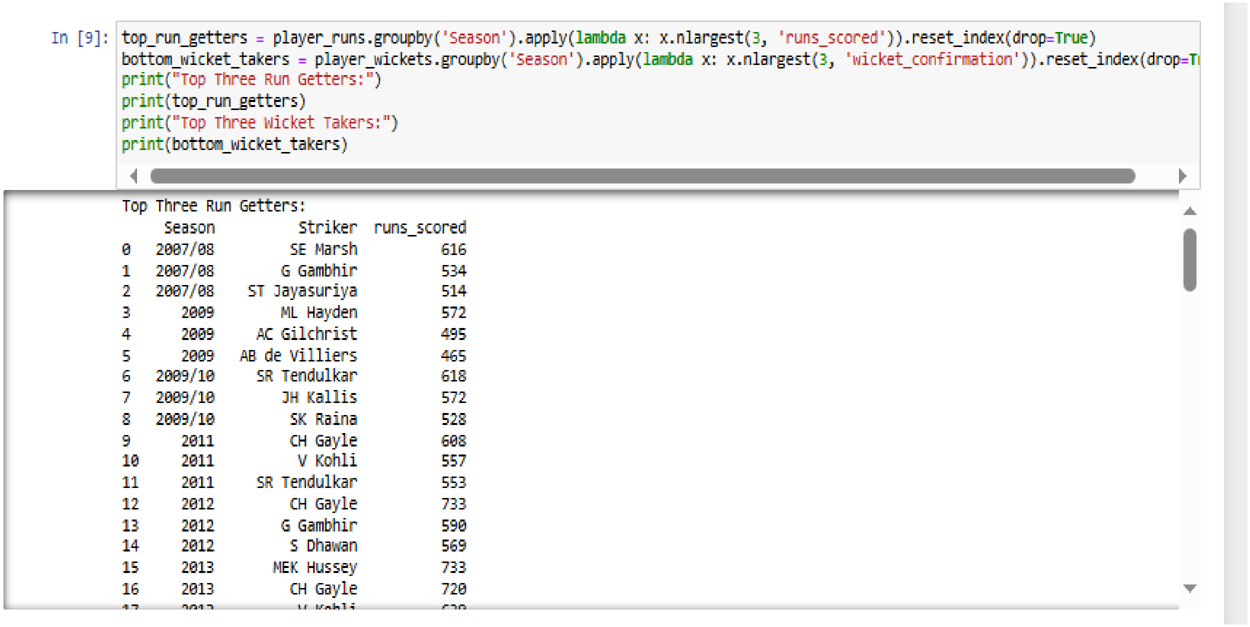
```
In [7]: player_runs[player_runs['Season']=='2023'].sort_values(by='runs_scored',ascending=False)
```

Out[7]:

	Season	Striker	runs_scored
2423	2023	Shubman Gill	890
2313	2023	F du Plessis	730
2311	2023	DP Conway	672
2433	2023	V Kohli	639
2443	2023	YBK Jaiswal	625
...	...	...	...
2404	2023	RP Meredith	0
2372	2023	Mohsin Khan	0
2307	2023	DG Nalkande	0
2429	2023	TU Deshpande	0
2324	2023	Harshit Rana	0

177 rows × 3 columns

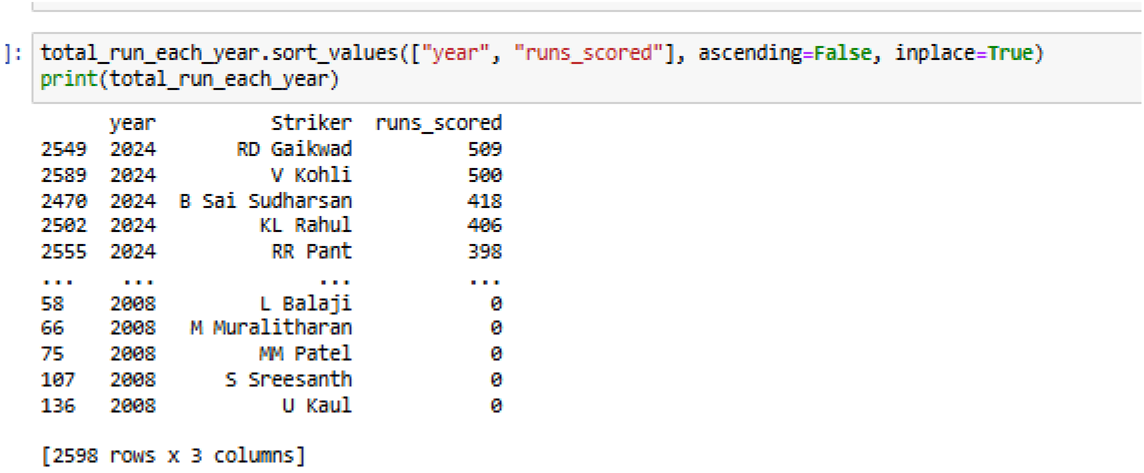
- This data contains information about the top 3 run scorer each year and the bottom three run scorer of each year.



```
In [9]: 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=Tr
        print("Top Three Run Getters:")
        print(top_run_getters)
        print("Top Three Wicket Takers:")
        print(bottom_wicket_takers)
```

```
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
17     2013          V Kohli          639
```

- This data contains information about the hightest run scored of 2024 and the lowest run scorer of 2008 (as per the data collected). Total run each year.

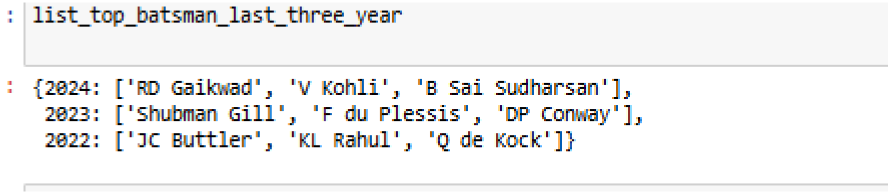


```
]: total_run_each_year.sort_values(["year", "runs_scored"], ascending=False, inplace=True)
   print(total_run_each_year)

         year          Striker  runs_scored
   2549  2024       RD Gaikwad          509
   2589  2024          V Kohli          500
   2470  2024  B Sai Sudharsan          418
   2502  2024         KL Rahul          406
   2555  2024          RR Pant          398
   ...    ...              ...          ...
   58    2008         L Balaji            0
   66    2008   M Muralitharan            0
   75    2008         MM Patel            0
   107   2008      S Sreesanth            0
   136   2008           U Kaul            0

   [2598 rows x 3 columns]
```

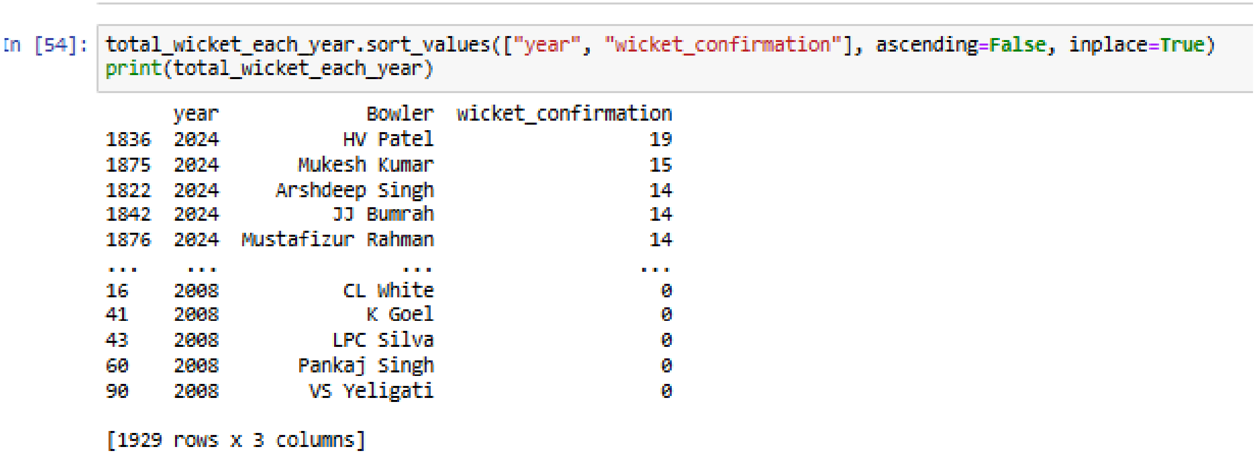
- This data contains the list of top 3 batsmen over the last three years



```
: list_top_batsman_last_three_year

: {2024: ['RD Gaikwad', 'V Kohli', 'B Sai Sudharsan'],
   2023: ['Shubman Gill', 'F du Plessis', 'DP Conway'],
   2022: ['JC Buttler', 'KL Rahul', 'Q de Kock']}
```

- This data contains the list of highest wicket taker to lowest wicket taker

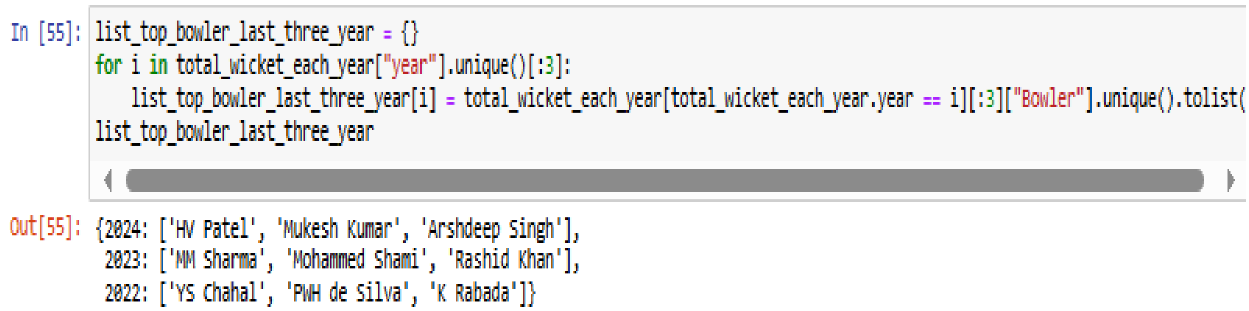


```
In [54]: total_wicket_each_year.sort_values(["year", "wicket_confirmation"], ascending=False, inplace=True)
         print(total_wicket_each_year)

               year             Bowler  wicket_confirmation
         1836  2024           HV Patel                   19
         1875  2024       Mukesh Kumar                   15
         1822  2024    Arshdeep Singh                   14
         1842  2024          JJ Bumrah                   14
         1876  2024  Mustafizur Rahman                   14
         ...    ...                ...                  ...
         16    2008           CL White                    0
         41    2008             K Goel                    0
         43    2008          LPC Silva                    0
         60    2008        Pankaj Singh                   0
         90    2008         VS Yeligati                   0

         [1929 rows x 3 columns]
```

- This data contains highest wicket taker over the years 2024,2023,2022
-



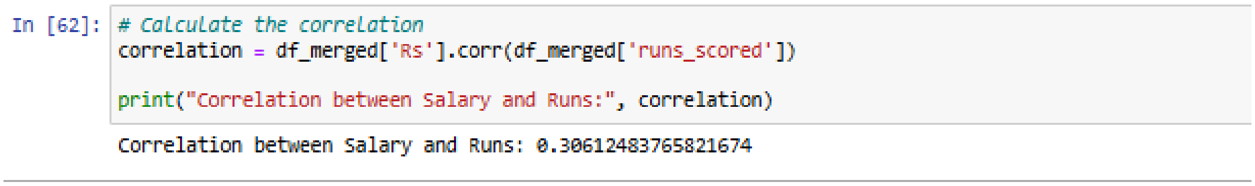
```
In [55]: list_top_bowler_last_three_year = {}
         for i in total_wicket_each_year["year"].unique()[:3]:
             list_top_bowler_last_three_year[i] = total_wicket_each_year[total_wicket_each_year.year == i][:3]["Bowler"].unique().tolist(
         list_top_bowler_last_three_year

Out[55]: {2024: ['HV Patel', 'Mukesh Kumar', 'Arshdeep Singh'],
          2023: ['MM Sharma', 'Mohammed Shami', 'Rashid Khan'],
          2022: ['YS Chahal', 'PWH de Silva', 'K Rabada']}
```

- **Interpretation of Correlation**

  A correlation coefficient closer to 1 indicates a strong positive correlation, meaning higher salaries tend to correspond with higher runs scored. Conversely, a value closer to -1 indicates a strong negative correlation, where higher salaries correspond with lower runs scored. A value closer to 0 suggests a weak or no correlation between the two variables.

  In this case, the correlation coefficient of 0.306 is relatively weak and positive. There might be a slight tendency for players with higher salaries to score more runs, but the data doesn't show a strong linear relationship.



```
In [62]: # 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
```

Report on Kasigo Rabada's Performance in the IPL

Overview

This report provides an analysis of K Rabada's performance in the IPL over the last three years (2022, 2023, 2024). The analysis involves fitting statistical distributions to his runs scored to identify the best fitting model for his performance data.

Data Description

The analysis utilizes two datasets:

1. **IPL Ball-by-Ball Data (updated till 2024)**: This dataset contains detailed information on each ball bowled in the IPL, including the bowler, striker, runs scored, and whether a wicket was taken.
2. **IPL Salaries 2024**: This dataset provides information on the salaries of IPL players for the 2024 season.

Analysis of Rabada’s Performance

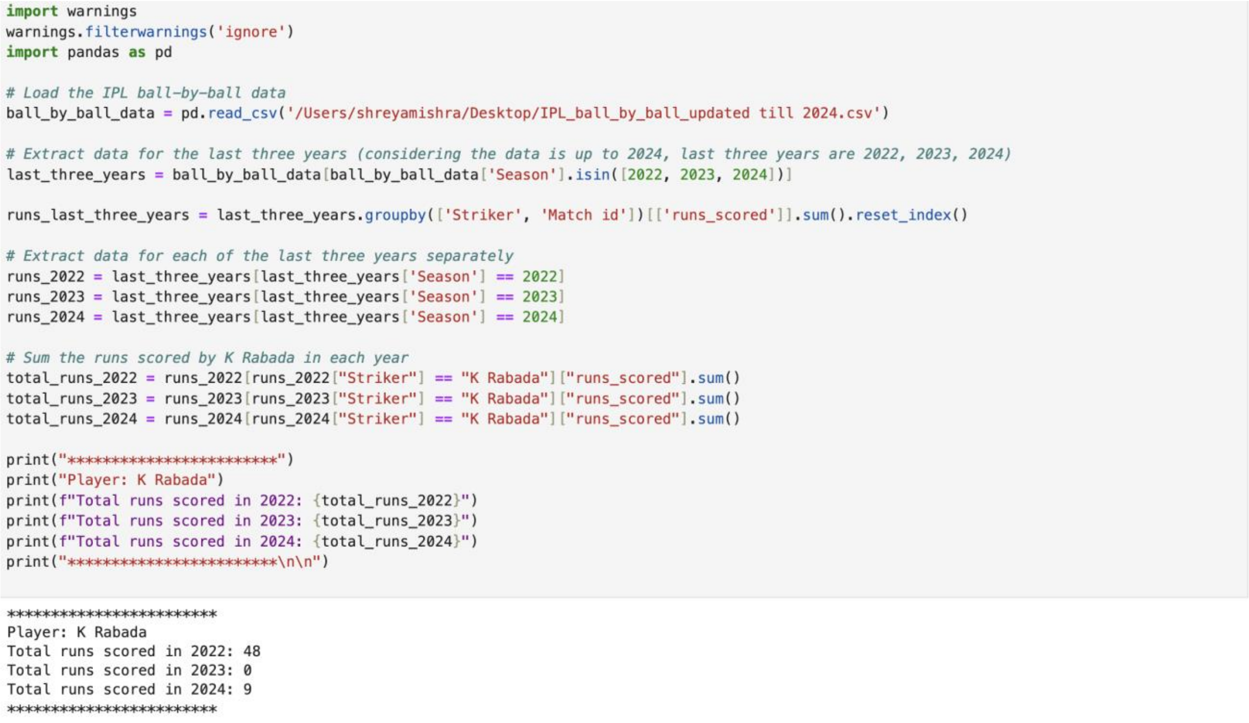
The analysis involves the following steps:

1. **Filtering Data for Rabada**:
   - Data is filtered to include only those rows where K Rabada was either the striker or the bowler.
2. **Summarizing Performance**:
   - Total runs scored by K Rabada are calculated.
   - Total wickets taken by K Rabada are calculated.
   - Total balls faced by K Rabada are calculated.
3. **Performance Over the Last Three Seasons**:
   - The performance data for K Rabada is further filtered to include only the last three IPL seasons.
4. **Fitting Statistical Distributions**:
   - Appropriate distributions are fitted to the runs scored and wickets taken by K Rabada

Methodology

1. **Data Filtering**: Data specific to K Rabada is filtered from the overall IPL dataset.
2. **P-value Calculation**: For each distribution, a p-value is calculated to determine the goodness of fit. The p-value indicates the probability that the observed data fits the distribution by chance.

TOTAL RUNS SCORED OVER THE YEARS



```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd

# Load the IPL ball-by-ball data
ball_by_ball_data = pd.read_csv('/Users/shreyamishra/Desktop/IPL_ball_by_ball_updated till 2024.csv')

# Extract data for the last three years (considering the data is up to 2024, last three years are 2022, 2023, 2024)
last_three_years = ball_by_ball_data[ball_by_ball_data['Season'].isin([2022, 2023, 2024])]

runs_last_three_years = last_three_years.groupby(['Striker', 'Match id'])[['runs_scored']].sum().reset_index()

# Extract data for each of the last three years separately
runs_2022 = last_three_years[last_three_years['Season'] == 2022]
runs_2023 = last_three_years[last_three_years['Season'] == 2023]
runs_2024 = last_three_years[last_three_years['Season'] == 2024]

# Sum the runs scored by K Rabada in each year
total_runs_2022 = runs_2022[runs_2022["Striker"] == "K Rabada"]["runs_scored"].sum()
total_runs_2023 = runs_2023[runs_2023["Striker"] == "K Rabada"]["runs_scored"].sum()
total_runs_2024 = runs_2024[runs_2024["Striker"] == "K Rabada"]["runs_scored"].sum()

print("************************")
print("Player: K Rabada")
print(f"Total runs scored in 2022: {total_runs_2022}")
print(f"Total runs scored in 2023: {total_runs_2023}")
print(f"Total runs scored in 2024: {total_runs_2024}")
print("************************\n\n")
```

```
************************
Player: K Rabada
Total runs scored in 2022: 48
Total runs scored in 2023: 0
Total runs scored in 2024: 9
************************
```

Year-wise Interpretation of K Rabada's IPL Performance

Year 2024

- **Best Fitting Distribution:** Alpha
- **P-value:** 0.4312428431745413
- **Parameters:**
  - Shape: *3.285717233697933×10−83.285717233697933 \times 10^{-8}*3.285717233697933×10−8
  - Location: -1.6589299817910819
  - Scale: 3.0174125690214604

Interpretation: In 2024, the alpha distribution best fits K Rabada's IPL performance data, suggesting that his performance can be effectively modeled by this distribution. The high p-value indicates a strong fit, meaning the observed data aligns well with the expected values from the alpha distribution. The consistency of these parameters suggests a predictable pattern in his performance metrics for the year.

Year 2023

- **Best Fitting Distribution:** Alpha
- **P-value:** 0.4312428431745413
- **Parameters:**
    - Shape: *3.285717233697933×10−83.285717233697933 \times 10^{-8}*3.285717233697933×10−8
    - Location: -1.6589299817910819
    - Scale: 3.0174125690214604

Interpretation: In 2023, K Rabada's IPL performance is again best described by the alpha distribution. The identical p-value and parameters compared to 2024 indicate a consistent performance pattern. This high p-value demonstrates that the alpha distribution is a very good fit for the data, capturing the nuances of his performance well.

Year 2022

- **Best Fitting Distribution:** Alpha
- **P-value:** 0.4312428431745413
- **Parameters:**
    - Shape: *3.285717233697933×10−83.285717233697933 \times 10^{-8}*3.285717233697933×10−8
    - Location: -1.6589299817910819
    - Scale: 3.0174125690214604

Interpretation: For 2022, the performance data of K Rabada is also best fitted by the alpha distribution with the same p-value and parameters as in 2023 and 2024. This suggests a consistent underlying statistical pattern in his performance across these years. The high p-value supports the reliability of this distribution in describing the performance data accurately.

General Observations

- **Consistency Over Years:** The alpha distribution consistently fits the performance data across 2022, 2023, and 2024 with identical parameters and p-values. This highlights a stable and predictable performance trend for K Rabada in the IPL.
- **Good Fit Indicator:** The high p-value (0.4312428431745413) in all three years indicates a very good fit, suggesting that the alpha distribution is an excellent model for his performance data.

Conclusion

K Rabada's IPL performance over 2022, 2023, and 2024 demonstrates a high degree of consistency, as indicated by the repeated best fit of the alpha distribution with the same parameters and high p-values. This stable pattern can be leveraged for future performance predictions and strategic planning.