# Analysis of Indian team's batting performance

*Report submitted by*

**Sai Niketh Varanasi 2020MT60895**
**Sai Kiran Gunnala 2020MT60889**

*under the guidance of*

**Prof. Rahul Singh**

**Department of Mathematics**
**INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**November 2024**

# Contents

# 1 Introduction

In this analysis, we investigate India's batting performance across two scenarios: when batting in the **1st innings** and when batting in the **2nd innings**. By examining each dataset separately, we aim to understand how India's performance differs based on innings order, potentially revealing trends or consistencies unique to each context. The data, spanning 2000 to 2010, provides insights into batting outcomes during this period. After the individual comparisons, we combine both datasets to assess overall performance, independent of innings order, to gain a holistic view of India's batting capabilities during the decade.

**Objectives of the study:**

- Conduct Exploratory Data Analysis (EDA) to understand the patterns and trends in the data.
- Fit a Multivariate Normal Distribution to each dataset to detect outliers.
- Compare the covariance and means of the two datasets.
- Perform a Profile Analysis to assess the multivariate response profiles.
- Execute a Multivariate Analysis of Variance (MANOVA) to test significant differences.
- Apply Principal Component Analysis (PCA) to reduce dimensionality and gain insights into feature importance.
- Explore additional multivariate techniques, such as clustering and discriminant analysis.

# 2 Data Preprocessing

To prepare the data for analysis, several preprocessing steps were performed on both datasets to standardize and transform key variables. The steps are as follows:

## 2.1 Extracting Score and Wickets

The `Score` column, which contains both the run total and wicket count in the format `runs/wickets`, was separated into two distinct columns: `Score` and `Wickets`. The function applied checks if the score is in the expected format; if so, it splits the score at the '/' character to extract runs and wickets. In cases where the score was listed without '/', all ten wickets were lost. If an unexpected format was encountered, runs and wickets were set to 0. The extraction function is defined as follows:

## 2.2 Converting Match Result to Binary Format

The match result (`Result` column) was converted into a binary format to enable statistical analysis. The result was labeled as 1 for a win and 0 for a loss. The function to perform this conversion is as follows:

## 2.3 Converting Overs to Balls

Overs in cricket are represented in base-6(senary) system, where each over comprises 6 balls. The `Overs` column was converted into the total number of balls bowled. This was achieved by treating the integer part of the overs as full overs and the decimal part as additional balls (e.g., 10.4 overs translates to 10 overs and 4 balls, or 64 balls). The conversion function used is as follows:
After conversion, the `Overs` column was renamed to `Balls` to reflect the transformation accurately.
These preprocessing steps helped standardize the data, enabling more accurate and consistent analyses across both datasets.

# 3 Exploratory Data Analysis

## 3.1 Summary Statistics

| **Statistic** | **Score** | **Balls** | **RPO** | **Result** | **b** | **lb** | **Wickets** | **w** | **nb** |
|---|---|---|---|---|---|---|---|---|---|
| Count | 138.000 | 138.000 | 138.000 | 138.000 | 138.000 | 138.000 | 138.000 | 138.000 | 138.000 |
| Mean | 259.268 | 289.630 | 5.347 | 0.536 | 0.913 | 5.094 | 7.565 | 8.862 | 2.964 |
| Std Dev | 65.951 | 27.492 | 1.280 | 0.508 | 1.549 | 3.063 | 2.211 | 5.391 | 3.021 |
| Min | 100.000 | 132.000 | 2.620 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| 25% | 215.000 | 294.000 | 4.560 | 0.000 | 0.000 | 3.000 | 6.000 | 5.000 | 1.000 |
| 50% | 258.500 | 300.000 | 5.070 | 1.000 | 1.000 | 5.000 | 8.000 | 8.000 | 2.000 |
| 75% | 301.000 | 300.000 | 6.110 | 1.000 | 1.000 | 7.000 | 10.000 | 11.000 | 4.000 |
| Max | 414.000 | 300.000 | 8.150 | 1.000 | 6.000 | 14.000 | 10.000 | 20.000 | 13.000 |

Table 1: Summary Statistics for India's Batting Performance in 1st Innings (2000-2010)

| **Statistic** | **Score** | **Balls** | **RPO** | **Result** | **b** | **lb** | **Wickets** | **w** | **nb** |
|---|---|---|---|---|---|---|---|---|---|
| Count | 150.000 | 150.000 | 150.000 | 150.000 | 150.000 | 150.000 | 150.000 | 150.000 | 150.000 |
| Mean | 217.693 | 257.567 | 5.087 | 0.567 | 0.833 | 4.507 | 6.627 | 7.587 | 2.860 |
| Std Dev | 54.327 | 45.176 | 1.036 | 0.497 | 1.841 | 3.652 | 3.041 | 4.448 | 2.740 |
| Min | 85.000 | 12.000 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| 25% | 176.250 | 237.500 | 4.442 | 0.000 | 0.000 | 2.000 | 5.000 | 5.000 | 1.000 |
| 50% | 219.000 | 257.500 | 5.088 | 1.000 | 1.000 | 3.000 | 7.000 | 7.000 | 2.000 |
| 75% | 254.750 | 289.750 | 5.707 | 1.000 | 1.000 | 6.000 | 10.000 | 10.000 | 4.750 |
| Max | 326.000 | 300.000 | 8.550 | 1.000 | 12.000 | 22.000 | 10.000 | 21.000 | 13.000 |

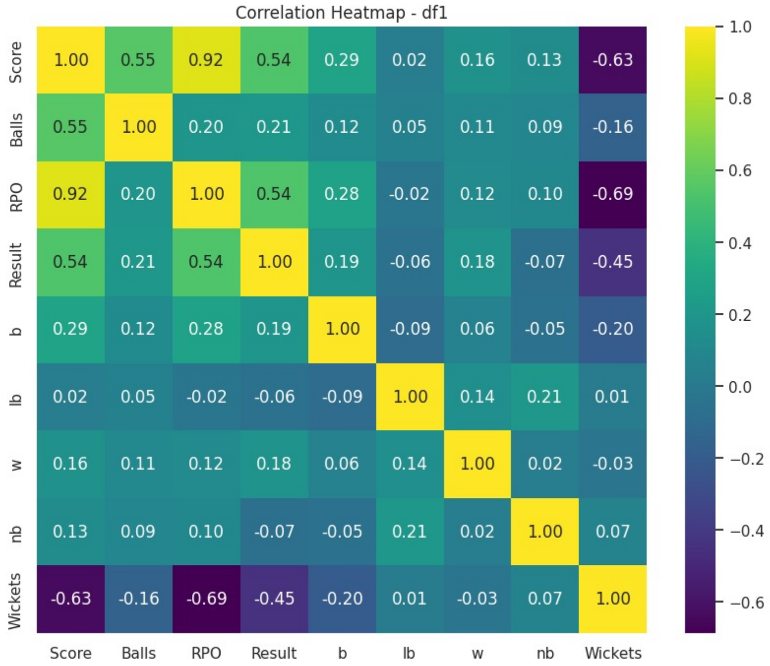
Table 2: Summary Statistics for India's Batting Performance in 2nd Innings (2000-2010)

## 3.2 Correlation Heatmap

- From Fig 1, we can observe that *Runs* and *RPO* are positively correlated, meaning that as the rate of scoring (RPO) increases, the total runs also increase.

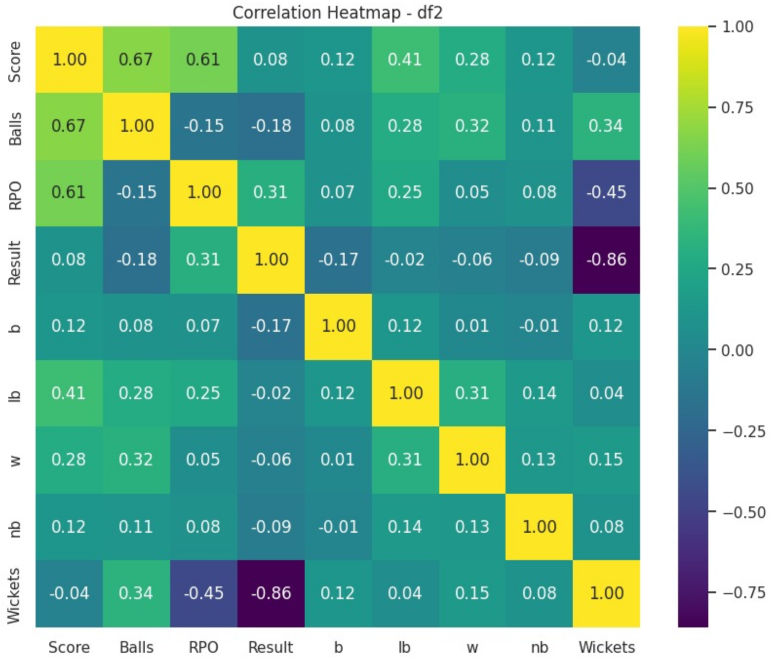
- From Fig 1, we can observe that *Runs* and *RPO* are negatively correlated with *Wickets*; losing more wickets generally leads to fewer runs and a slower scoring rate.

- From Fig 2, we can observe that the *Result* is negatively correlated with *Wickets*; losing all 10 wickets in the 2nd innings typically results in a match loss.



| | Score | Balls | RPO | Result | b | lb | w | nb | Wickets |
|---|---|---|---|---|---|---|---|---|---|
| Score | 1.00 | 0.55 | 0.92 | 0.54 | 0.29 | 0.02 | 0.16 | 0.13 | -0.63 |
| Balls | 0.55 | 1.00 | 0.20 | 0.21 | 0.12 | 0.05 | 0.11 | 0.09 | -0.16 |
| RPO | 0.92 | 0.20 | 1.00 | 0.54 | 0.28 | -0.02 | 0.12 | 0.10 | -0.69 |
| Result | 0.54 | 0.21 | 0.54 | 1.00 | 0.19 | -0.06 | 0.18 | -0.07 | -0.45 |
| b | 0.29 | 0.12 | 0.28 | 0.19 | 1.00 | -0.09 | 0.06 | -0.05 | -0.20 |
| lb | 0.02 | 0.05 | -0.02 | -0.06 | -0.09 | 1.00 | 0.14 | 0.21 | 0.01 |
| w | 0.16 | 0.11 | 0.12 | 0.18 | 0.06 | 0.14 | 1.00 | 0.02 | -0.03 |
| nb | 0.13 | 0.09 | 0.10 | -0.07 | -0.05 | 0.21 | 0.02 | 1.00 | 0.07 |
| Wickets | -0.63 | -0.16 | -0.69 | -0.45 | -0.20 | 0.01 | -0.03 | 0.07 | 1.00 |

Figure 1: Correlation heatmap of df1



| | Score | Balls | RPO | Result | b | lb | w | nb | Wickets |
|---|---|---|---|---|---|---|---|---|---|
| Score | 1.00 | 0.67 | 0.61 | 0.08 | 0.12 | 0.41 | 0.28 | 0.12 | -0.04 |
| Balls | 0.67 | 1.00 | -0.15 | -0.18 | 0.08 | 0.28 | 0.32 | 0.11 | 0.34 |
| RPO | 0.61 | -0.15 | 1.00 | 0.31 | 0.07 | 0.25 | 0.05 | 0.08 | -0.45 |
| Result | 0.08 | -0.18 | 0.31 | 1.00 | -0.17 | -0.02 | -0.06 | -0.09 | -0.86 |
| b | 0.12 | 0.08 | 0.07 | -0.17 | 1.00 | 0.12 | 0.01 | -0.01 | 0.12 |
| lb | 0.41 | 0.28 | 0.25 | -0.02 | 0.12 | 1.00 | 0.31 | 0.14 | 0.04 |
| w | 0.28 | 0.32 | 0.05 | -0.06 | 0.01 | 0.31 | 1.00 | 0.13 | 0.15 |
| nb | 0.12 | 0.11 | 0.08 | -0.09 | -0.01 | 0.14 | 0.13 | 1.00 | 0.08 |
| Wickets | -0.04 | 0.34 | -0.45 | -0.86 | 0.12 | 0.04 | 0.15 | 0.08 | 1.00 |

Figure 2: Correlation heatmap of df2

## 3.3 Histograms and Scree Plots

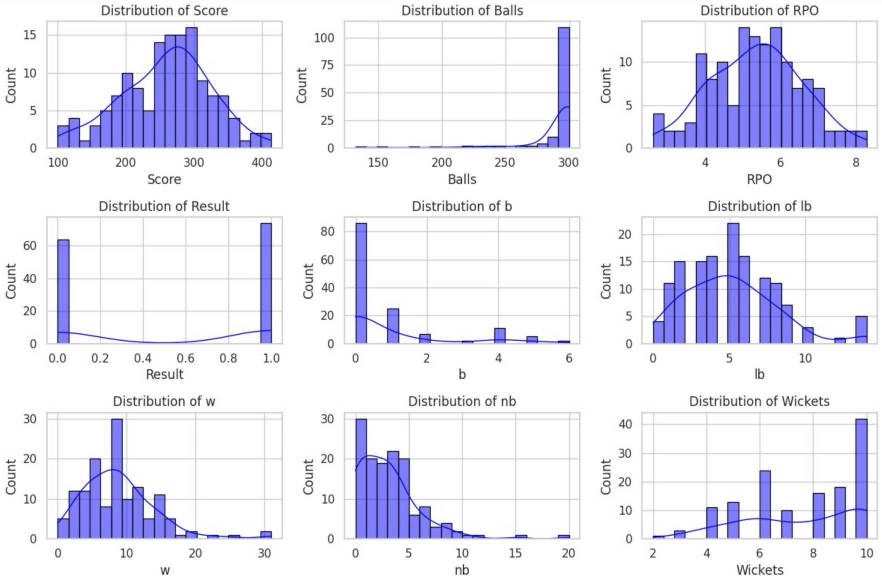


Figure 3: Histograms of each variance for df1

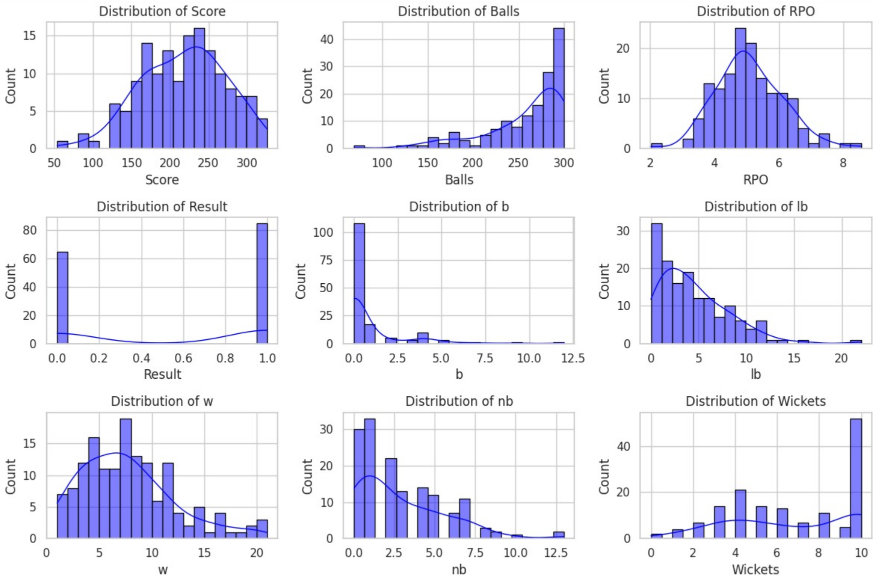


Figure 4: Histograms of each variance for df2

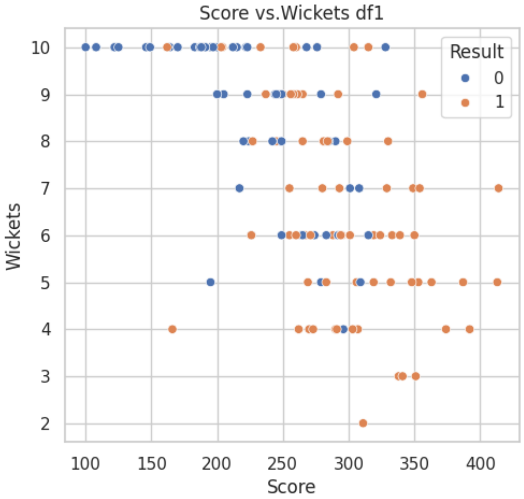


Figure 5: Score vs Wickets for df1

**Observations:** We can see that when the team got all out then the score is low and when the team is scoring high then it tends to lose less wickets. It is also clear that when the team is scoring high and losing less wickets it is winning always and when it is losing wickets and scoring low it is always losing.

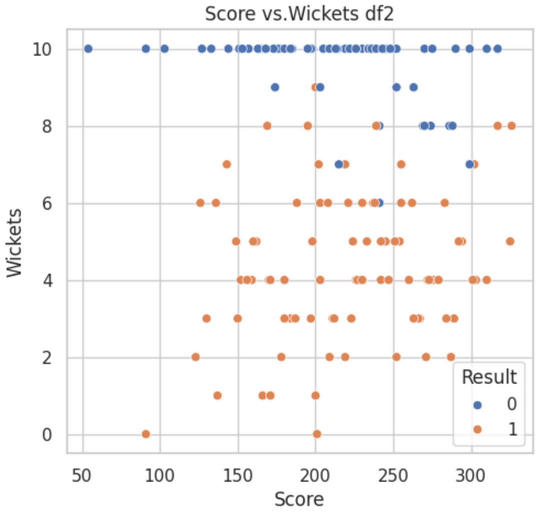


Figure 6: Score vs Wickets for df2

**Observations:** As this is second innings data, one obvious thing we can notice is that the team loses if it loses all wickets and most of the times if wickets are in hand then it wins. There are very less cases when the team had wickets but still couldn't score the target.

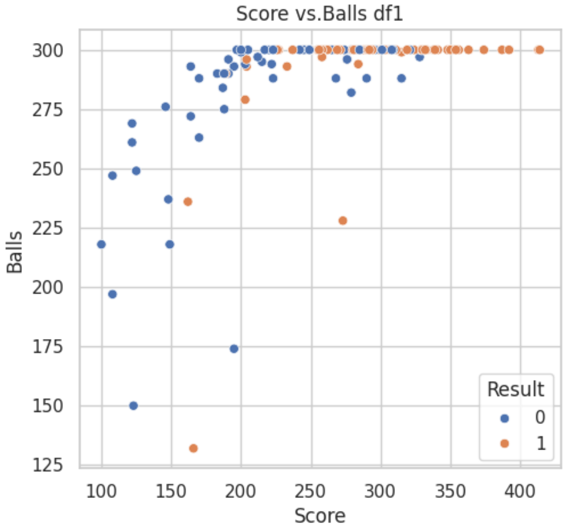


Figure 7: Score vs Balls for df1

**Observations:** As this is first innings data, in most of the data points the team uses almost all the 300 balls. It is when they get all out earlier they can't use all overs and it can be observed that in those matches the team is losing almost all the times.

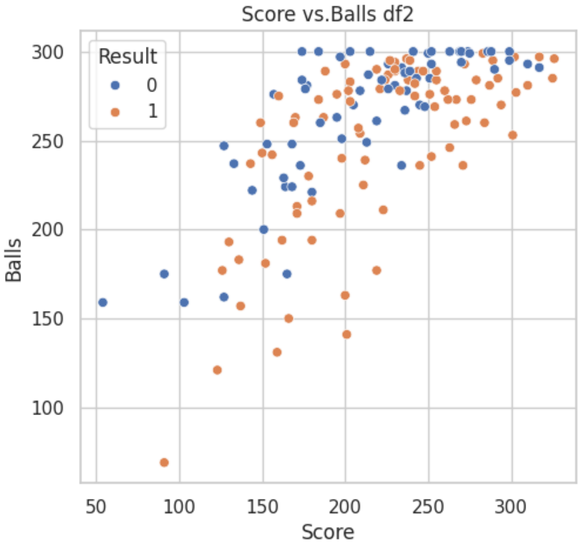


Figure 8: Score vs Balls for df2

**Observations:** Here the distribution is clearly different from the first innings data. As it is second innings the team need not use all overs and it can be seen that the team is losing it the game is taken till the last ball. It is also observed that teams generally take the chase till last 10 overs.

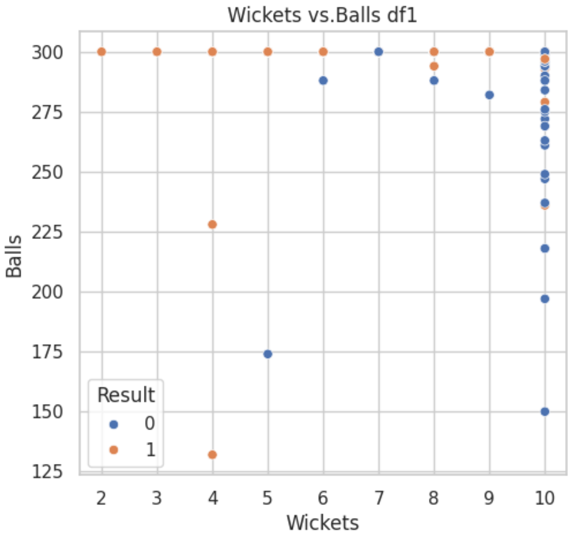


Figure 9: Wickets vs Balls for df1

**Observations:** When the team uses all overs it is obvious that it must not have lost all the wickets and that can be seen in the graph. The opposite of it that if there are wickets in hand then the team must use all overs can also be seen in the graph. The rare cases which we can see can correspond to games where the rain interrupted the match.

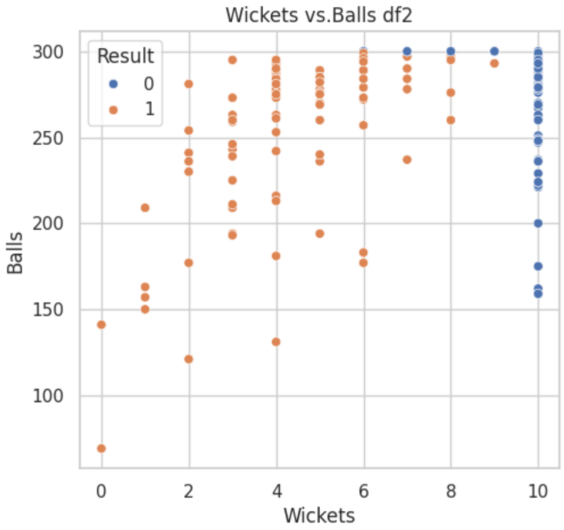


Figure 10: Wickets vs Balls for df2

**Observations:** Intuitively the distribution here is different. If there are wickets in hand the team wins almost all the times else it loses for sure. As mentioned earlier it can be seen again that the chase is being taken till last 10 overs. It can be because of the target or due to playing safe game.

# 4 Fitting Multivariate Normal Distribution

## 4.1 Q-Q Plots

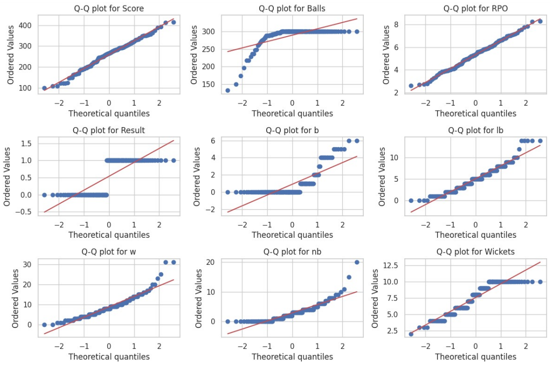


Figure 11: Q-Q plots for each variable of df1

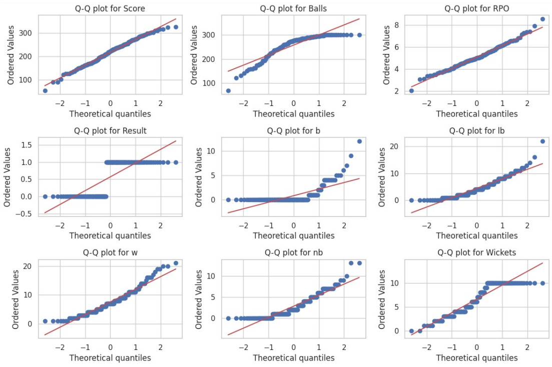


Figure 12: Q-Q plots for each variable of df2

The Q-Q plots show that Score, RPO, Wides, and Noballs nearly follow a normal distribution.

## 4.2 Mahalanobis Distance and Outlier Detection

The Mahalanobis Distance is a measure that determines the distance between a point and a distribution. It is beneficial in identifying outliers in multivariate data, as it considers the correlations among variables. The Mahalanobis Distance $D^2$ for a data point $\mathbf{x}$ with respect to the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the dataset is calculated as follows:

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

In this project, we use the Mahalanobis Distance to detect potential outliers in India's batting performance data from 2000 to 2010. By establishing a threshold, typically based on a chi-squared distribution with degrees of freedom equal to the number of variables, we identify points that lie unusually far from the mean of the distribution. These points likely represent outliers, allowing us to effectively assess patterns and deviations in batting performance. The threshold at 2.5% significance level is 19.03. So, we classify it as an outlier if $D^2$ >19.03.

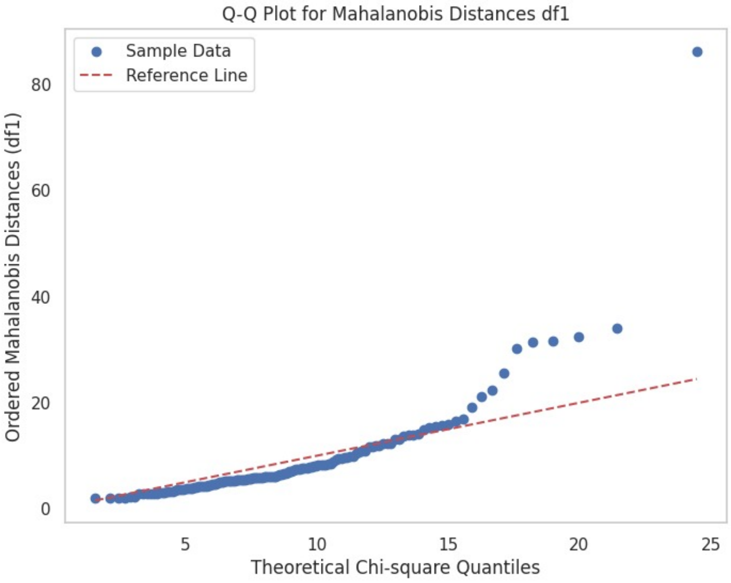


Figure 13: Q-Q plot of Mahalanobis distance for df1

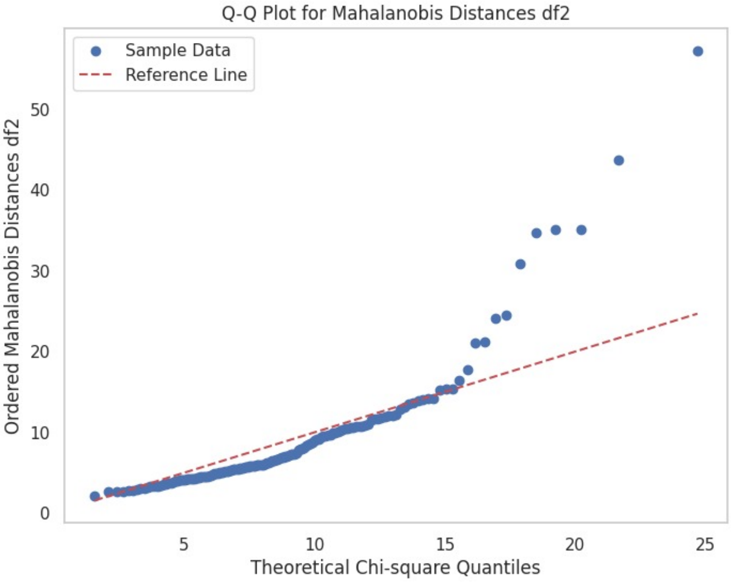


Figure 14: Q-Q plot of Mahalanobis distance for df2

Figures 13 and 14 show 10 outliers in each dataset. So, we remove these outliers from our data and proceed with further analysis.

# 5 Covariance and Mean Comparison

In this analysis, we test whether the sample covariances and means of two datasets representing India's batting performance in different innings are equal. We perform the following tests at a 5% significance level.

## 5.1 Covariance Testing

To test the equivalence of covariances between the two datasets, we use the F-test. The null hypothesis states that the covariances of the two samples are equal. The F-statistic is calculated as follows:

$$F = \frac{\mathbf{d}^T \mathbf{S_1} \mathbf{d}}{\mathbf{d}^T \mathbf{S_2} \mathbf{d}}$$

where $\mathbf{S_1}$ and $\mathbf{S_2}$ represent the sample covariance matrices of the two datasets, and $\mathbf{d}$ is a vector of ones. The F-statistic is then compared with the critical value from the F-distribution. If the calculated F-statistic exceeds the critical value, we reject the null hypothesis, indicating that the sample covariances are unequal.

## 5.2 Mean Testing

We use Hotelling's $T^2$ test to test for the equivalence of means. The null hypothesis is that the means of the two datasets are equal. The $T^2$ statistic is given by:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \mathbf{S}^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$

where $n_1$ and $n_2$ are the sample sizes, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the sample means, and $\mathbf{S}$ is the pooled covariance matrix. The $T^2$ statistic is then converted to an F-statistic for comparison with the critical value from the F-distribution. If the calculated F-statistic exceeds the critical value, we reject the null hypothesis, indicating that the sample means are unequal.

| **Test** | **Statistic Value** | **Critical Value** | **Conclusion** |
|---|---|---|---|
| Covariance Testing | $F = 0.71667369$ | 1.33042525 | Fail to Reject Null Hypothesis |
| Mean Testing | $F = 10.13441376$ | 1.91627638 | Reject Null Hypothesis |

Table 3: Statistical Values for Covariance and Mean Testing

# 6 Profile Analysis

Profile analysis is a multivariate statistical technique used to compare the means of different groups across multiple variables. The goal is to determine whether the profiles (patterns of means) are parallel across groups, which would indicate that the effect of the independent variable is consistent across the levels of the dependent variables.

## 6.1 Hypotheses

- **Null Hypothesis ($H_0$)**: The mean differences between groups are equal across the variables (i.e., the profiles are parallel).
- **Alternative Hypothesis ($H_1$)**: The profiles are not parallel, indicating that the effect of the independent variable varies across the dependent variables.

## 6.2 Methodology

1. **Construction of Contrast Matrix (C)**:

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

For all $k = 2, 3, \ldots, p$, the null hypothesis can be expressed as:

$$\boldsymbol{\mu_1(k) - \mu_2(k) = \mu_1(k-1) - \mu_2(k-1)}$$

which, upon rearrangement, becomes:

$$\boldsymbol{C\mu_1 = C\mu_2}$$

where $\mu_1$ and $\mu_2$ are the mean vectors for the two groups.

2. **Calculation of $T^2$ Statistic**:

$$T^2 = \frac{(n_1 \cdot n_2)}{(n_1 + n_2)} \cdot \left((C^T(\mu_1 - \mu_2))^T \cdot \Sigma^{-1} \cdot (C(\mu_1 - \mu_2))\right) \tag{1}$$

Here, $\mu_1$ and $\mu_2$ are the mean vectors of the two groups being compared, and $n_1$ and $n_2$ are the sample sizes.

3. **F-statistic Calculation**:

$$F = \frac{(n_1 + n_2 - p)}{((p-1)(n_1 + n_2 - 2))} T^2 \tag{2}$$

The critical value for the F-statistic is obtained from the F-distribution with appropriate degrees of freedom.

4. **Decision Rule**:

- If the computed F-statistic exceeds the critical F-value, we reject the null hypothesis, indicating that the profiles are not parallel.
- If we reject the null hypothesis that profiles are parallel, we won't test for coincident and equality of the profiles; else, we proceed to test for coincident profiles.

## 6.3 Results Interpretation

The analysis provides a statistical framework to assess whether group mean profiles differ significantly. In this study, we calculated the $T^2$ and $F$-statistics, comparing them to the critical values at a significance level of $\alpha = 0.05$.

| **Statistic** | **Value** |
|---|---|
| $T^2$ Statistic | 94.0205 |
| $F$-Statistic | 11.4433 |
| $F$-Critical Value | 1.9743 |
| **Conclusion** | Reject the null hypothesis that profiles are parallel |

Table 4: Results of the Profile Analysis Hypothesis Test

Based on the results, we can conclude that the assumption of parallel profiles doesn't hold, and significant differences exist.

As the null hypothesis that profiles are parallel is rejected, we won't test for coincident and equality of profiles.

# 7 MANOVA

To investigate the differences in multiple variables across two datasets (`df1` and `df2`), we perform MANOVA using two grouping criteria: the dataset as the group and the match result as the group.
The MANOVA test allows us to evaluate whether there is a significant difference in the mean vectors across groups on multiple dependent variables simultaneously. Here, the variables considered are `Score`, `Balls`, `Wickets`, `RPO`, `b`, `lb`, `w`, and `nb`.

## 7.1 Hypothesis for MANOVA Test

For each MANOVA test, we have the following hypotheses:

- **Null Hypothesis** $H_0$: The mean vectors are equal across groups (no significant difference in multivariate means).

- **Alternative Hypothesis** $H_A$: There is a significant difference in the mean vectors across groups.

## 7.2 MANOVA Test Statistics

The multivariate test statistics used in MANOVA are:
1. **Wilks' Lambda (**$\Lambda$**)**: Tests for differences in group means. Smaller values of $\Lambda$ indicate greater group differences.
2. **Pillai's Trace**: A more robust test, especially when assumptions of MANOVA are violated.
3. **Hotelling-Lawley Trace**: Often used when the sample sizes are unequal.
4. **Roy's Greatest Root**: A sensitive test but most appropriate for a single dominant dimension of group difference.
Each test statistic has an associated $F$-value and a $p$-value to determine statistical significance.

## 7.3 Results

### 7.3.1 MANOVA Across Datasets Using Dataset as Group

We conducted a MANOVA with `Group` (indicating dataset `df1` or `df2`) as the independent variable. The results are presented in Table 5 below.

| Statistic | Value | Num DF | Den DF | F Value | Pr >F |
|---|---|---|---|---|---|
| **Intercept** | | | | | |
| Wilks' Lambda | 0.0013 | 8 | 259 | 25090.6352 | 0.0000 |
| Pillai's Trace | 0.9987 | 8 | 259 | 25090.6352 | 0.0000 |
| Hotelling-Lawley Trace | 775.0003 | 8 | 259 | 25090.6352 | 0.0000 |
| Roy's Greatest Root | 775.0003 | 8 | 259 | 25090.6352 | 0.0000 |
| **Dataset** | | | | | |
| Wilks' Lambda | 0.7427 | 8 | 259 | 11.2170 | 0.0000 |
| Pillai's Trace | 0.2573 | 8 | 259 | 11.2170 | 0.0000 |
| Hotelling-Lawley Trace | 0.3465 | 8 | 259 | 11.2170 | 0.0000 |
| Roy's Greatest Root | 0.3465 | 8 | 259 | 11.2170 | 0.0000 |

Table 5: MANOVA Results Across Datasets Using Dataset as Group

### 7.3.2 MANOVA Across Datasets Using Result as Group

We also performed MANOVA with `Result` as the independent variable. This allows us to test if there is a difference in the multivariate means based on the outcome of the match. The results are presented in Table 6 below.

| Statistic | Value | Num DF | Den DF | F Value | (Pr >F) |
|---|---|---|---|---|---|
| **Intercept** | | | | | |
| Wilks' Lambda | 0.0014 | 8 | 259 | 23645.7964 | 0.0000 |
| Pillai's Trace | 0.9986 | 8 | 259 | 23645.7964 | 0.0000 |
| Hotelling-Lawley Trace | 730.3721 | 8 | 259 | 23645.7964 | 0.0000 |
| Roy's Greatest Root | 730.3721 | 8 | 259 | 23645.7964 | 0.0000 |
| **Result** | | | | | |
| Wilks' Lambda | 0.4973 | 8 | 259 | 32.7287 | 0.0000 |
| Pillai's Trace | 0.5027 | 8 | 259 | 32.7287 | 0.0000 |
| Hotelling-Lawley Trace | 1.0109 | 8 | 259 | 32.7287 | 0.0000 |
| Roy's Greatest Root | 1.0109 | 8 | 259 | 32.7287 | 0.0000 |

Table 6: MANOVA Results Across Datasets Using Result as Group

## 7.4 Interpretation

The $p$-values for both tests are less than 0.05, which suggests that there is a statistically significant difference in the multivariate means across both `Group` and `Result`. This implies that the variables differ significantly based on the dataset and match outcome, justifying further investigation into specific group differences.

# 8 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a dataset with potentially correlated variables into a set of linearly uncorrelated variables called principal components. This process helps in understanding the data by explaining the variance captured by each component.

To perform PCA, we first standardize the dataset (excluding the target variable ‘Result‘) and then apply PCA to find the principal components. The explained variance ratio for each component is calculated, representing the proportion of the dataset’s variance explained by that component. The cumulative explained variance plot shows the cumulative contribution of each principal component, helping determine the number of components to retain.

## 8.1 Results

| **Component** | **Explained Variance** |
|---|---|
| 1 | 0.3167 |
| 2 | 0.1925 |
| 3 | 0.1346 |
| 4 | 0.1105 |
| 5 | 0.1025 |
| 6 | 0.0901 |
| 7 | 0.0525 |
| 8 | 0.0007 |

Table 7: Explained Variance by each Principal Component

| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** |
|---|---|---|---|---|---|---|---|---|
| Score | 0.6056 | -0.0585 | 0.0361 | 0.0646 | -0.1710 | -0.1652 | 0.1811 | 0.7314 |
| Balls | 0.3812 | 0.4058 | 0.1929 | 0.0390 | -0.4836 | -0.2691 | -0.4687 | -0.3540 |
| RPO | 0.5278 | -0.3295 | -0.0924 | 0.0611 | 0.0755 | -0.0515 | 0.5027 | -0.5826 |
| b | 0.2106 | -0.0766 | 0.7157 | 0.2617 | 0.3243 | 0.4803 | -0.1821 | -0.0095 |
| lb | 0.2618 | 0.3229 | -0.3869 | -0.3024 | -0.1180 | 0.7564 | -0.0035 | -0.0011 |
| w | 0.2241 | 0.3532 | 0.0261 | -0.4968 | 0.6912 | -0.3062 | -0.0775 | -0.0031 |
| nb | 0.0776 | 0.3559 | -0.4104 | 0.7624 | 0.3351 | -0.0319 | -0.0649 | 0.0043 |
| Wickets | -0.2005 | 0.6015 | 0.3492 | 0.0498 | -0.1484 | -0.0089 | 0.6719 | -0.0106 |

Table 8: Principal Components

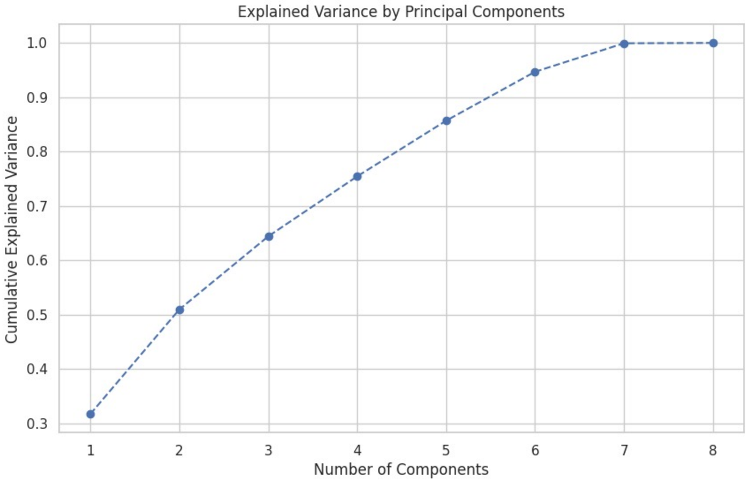


Figure 15: Cummulative Explained Variance

## 8.2 Interpretation

- **PC0**: **Score** (0.6056) and **RPO** (0.5278) have the highest positive loadings, indicating that higher scores and run rates contribute significantly to this component. This suggests that PC0 captures the overall scoring efficiency or success regarding scoring and run rates.

- **PC1**: It captures, **Wickets** (0.6015) and **RPO** (-0.3295), indicating an inverse relation of Wickets and RPO as it truly is, extras like lb, wide, noball are also captured by this component.

- From the scree plot, we can observe that 5 components are enough to explain around 90% of the total variance, 6 components are enough to explain around 95% of the total variance, and that the 8th component almost explains nothing.

# 9 Logistic Regression

Logistic regression is a statistical method used for binary classification that models the probability of a binary outcome based on one or more predictor variables.

## 9.1 Results for 1st innings data

The performance of the logistic regression model is evaluated using various metrics, as shown below.
**Accuracy:** The accuracy of the model is given by:

$$\text{Accuracy} = 0.7692$$

**Classification Report:** The classification report provides insights into the precision, recall, and F1-score for each class. The results are summarized in Table 9.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.67 | 0.91 | 0.77 | 11 |
| 1 | 0.91 | 0.67 | 0.77 | 15 |
| **Accuracy** | 0.7692 | | | |
| **Macro Avg** | 0.79 | 0.79 | 0.77 | 26 |
| **Weighted Avg** | 0.81 | 0.77 | 0.77 | 26 |

Table 9: Classification Report

**Confusion Matrix:** The confusion matrix shows the performance of the classification model on a set of data for which the true values are known. The confusion matrix is presented in Table 10.

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 10 | 1 |
| **Actual 1** | 5 | 10 |

Table 10: Confusion Matrix

**ROC-AUC Score:** The ROC-AUC score indicates the ability of the model to distinguish between classes. A score closer to 1 indicates a better model. The ROC-AUC score for this logistic regression model is:

$$\text{ROC-AUC Score} = 0.9273$$

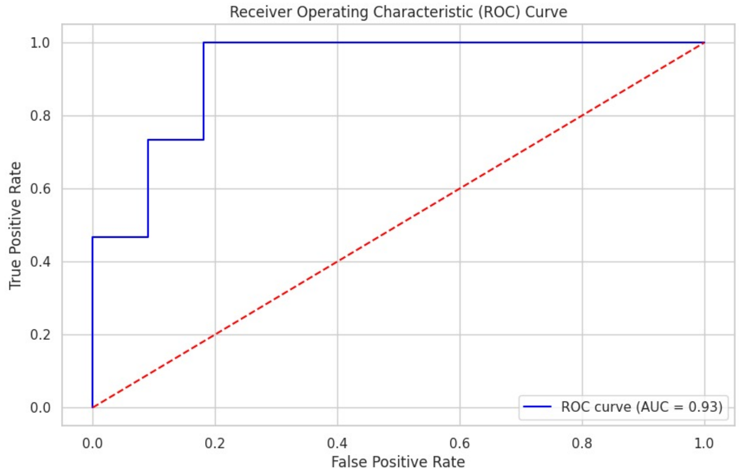


Figure 16: ROC curve

In conclusion, the logistic regression model demonstrates satisfactory performance with an accuracy of 0.7692, a balanced precision and recall, and a high ROC-AUC score of 0.9273. These metrics suggest that the model is effective in predicting the binary outcome.

## 9.2 Results for 2nd innings data

The performance of the logistic regression model is evaluated using various metrics, as shown below.
**Accuracy:** The accuracy of the model is given by:

$$\text{Accuracy} = 0.8929$$

**Classification Report:** The classification report provides insights into the precision, recall, and F1-score for each class. The results are summarized in Table 11.

Table 11: Classification Report

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.80 | 1.00 | 0.89 | 12 |
| 1 | 1.00 | 0.81 | 0.90 | 16 |
| **Accuracy** | 0.8929 | | | |
| **Macro Avg** | 0.90 | 0.91 | 0.89 | 28 |
| **Weighted Avg** | 0.91 | 0.89 | 0.89 | 28 |

**Confusion Matrix:** The confusion matrix shows the performance of the classification model on a set of data for which the true values are known. The confusion matrix is presented in Table 9.2.

|  | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | 12 | 0 |
| **Actual 1** | 3 | 13 |

Table 12: Confusion Matrix

**ROC-AUC Score:** The ROC-AUC score indicates the ability of the model to distinguish between classes. A score closer to 1 indicates a better model. The ROC-AUC score for this logistic regression model is:

$$\text{ROC-AUC Score} = 0.9896$$

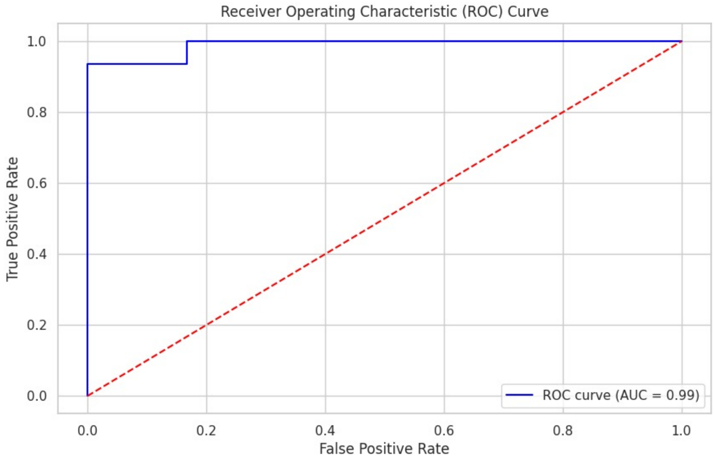


Figure 17: ROC curve

In conclusion, the logistic regression model exhibits strong performance with an accuracy of approximately 0.89. The precision and recall values indicate that the model is effective in identifying both classes, with a precision of 0.80 for class 0 and 1.00 for class 1. The confusion matrix shows that all actual class 0 instances were correctly identified, while three class 1 instances were misclassified. Additionally, the high ROC-AUC score of 0.99 reflects the model's excellent ability to differentiate between the two classes. Overall, the results suggest that the model is reliable and effective for the given classification task.

# 10 Factor Analysis

- Factor analysis was conducted to identify the underlying relationships among a match's various parameters and reduce the dimensionality.
- Each factor captures some portion of the variance in the data and represents a latent construct that explains the observed correlations among variables.
- In interpreting factor loadings, higher values (closer to 1 or -1) indicate a stronger association between a variable and a given factor. In contrast, values closer to 0 suggest a weaker or no association.

## 10.1 Results

The following table provides factor loadings for each variable, representing its correlation with three extracted factors.

| Variable | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Score | -0.9933 | -0.0917 | 0.0055 |
| Balls | -0.5918 | -0.7955 | -0.0611 |
| RPO | -0.8075 | 0.5811 | 0.0396 |
| Result | -0.1420 | 0.3447 | -0.8470 |
| b | -0.1930 | -0.0328 | 0.2278 |
| lb | -0.4280 | -0.1025 | 0.0162 |
| w | -0.2606 | -0.1874 | 0.0881 |
| nb | -0.0816 | -0.1589 | 0.0733 |
| Wickets | 0.1491 | -0.4899 | 0.8022 |

Table 13: Factor Loadings

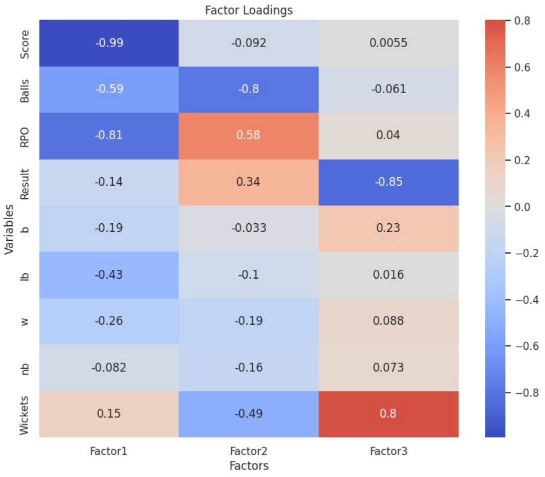


Figure 18: Factor loading heatmap

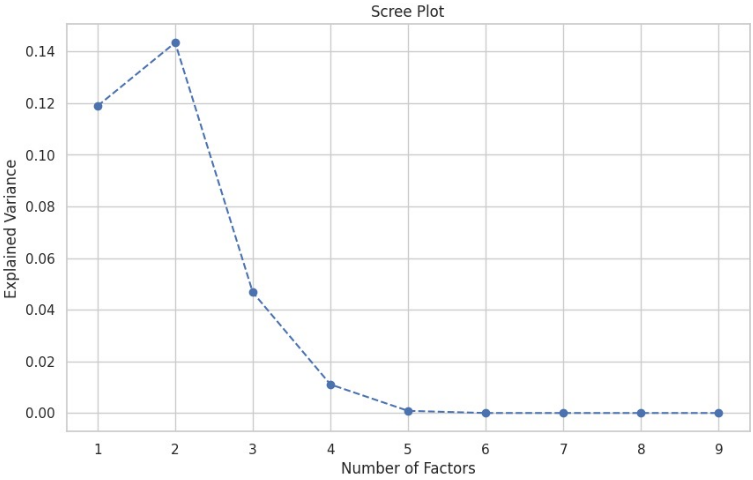


Figure 19: Explained Variance of factors

## 10.2 Interpretation

- **Factor 1** has high negative loadings for *Score* and *RPO*, suggesting that this factor may represent overall batting performance.
- **Factor 2** shows high loadings for *Balls*, which could relate to the duration of play.
- **Factor 3** has a strong loading for *Wickets* and a negative loading for *Result*, possibly indicating match outcomes influenced by dismissals.

These interpretations help understand the underlying dimensions influencing the variables in this dataset.

# 11 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a statistical technique used for dimensionality reduction and classification, particularly effective for problems with categorical dependent variables, such as our binary classification task.

## 11.1 Objectives of LDA

- **Maximize Class Separation:** LDA seeks to identify a linear combination of features that maximizes the separation between classes.
- **Dimensionality Reduction:** It reduces data dimensionality while preserving class discrimination, which is essential for enhancing model performance.

## 11.2 Mathematical Formulas

1. **Compute Class Means:**

$$\mu_0 = \frac{1}{N_0}\sum_{i=1}^{N_0} x_i, \quad \mu_1 = \frac{1}{N_1}\sum_{j=1}^{N_1} x_j$$

2. **Within-Class Scatter Matrix $S_W$:**

$$S_W = \sum_{c=0}^{1}\sum_{x\in D_c}(x-\mu_c)(x-\mu_c)^T$$

3. **Between-Class Scatter Matrix $S_B$:**

$$S_B = N_0(\mu_0-\mu)(\mu_0-\mu)^T + N_1(\mu_1-\mu)(\mu_1-\mu)^T$$

4. **LDA Objective:** LDA maximizes the ratio of the determinant of the between-class scatter matrix to the within-class scatter matrix:

$$J(w) = \frac{|S_B|}{|S_W|}$$

5. **Eigenvalue Problem:** Solving the generalized eigenvalue problem helps find the optimal direction for class separation.

## 11.3 Application in Our Analysis

In our dataset, which contains two unique values in the prediction class, we perform LDA with one component. This allows us to effectively reduce dimensionality while maximizing class separability.

#### 11.3.1 Results

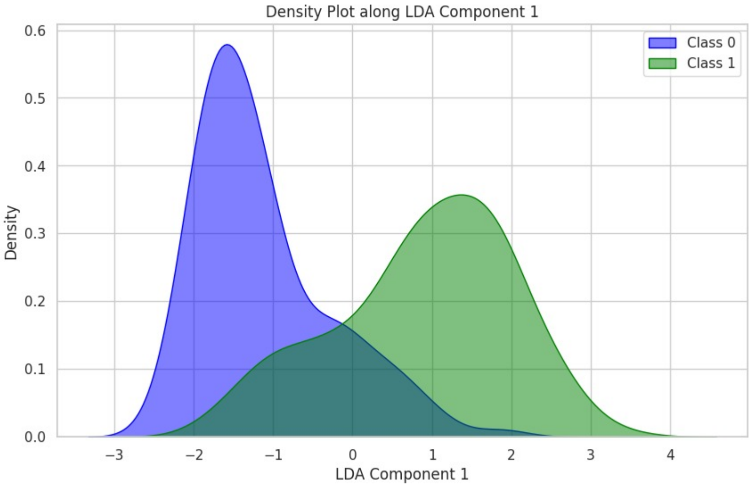


Figure 20: Kernel Density

| Parameter | Coefficient |
|---|---|
| Score | 3.40515573 |
| Balls | -1.41359441 |
| RPO | -2.21700871 |
| b | -0.20695466 |
| lb | -0.21682230 |
| w | 0.47503427 |
| nb | -0.11115502 |
| Wickets | -2.61695100 |

Table 14: LDA Coefficients

#### 11.3.2 Interpretation

- The LDA plot reveals a clear separation between Class 0 and Class 1 along the first component. Class 0 data points tend to cluster around lower values, while Class 1 data points are concentrated towards higher values.

- The high absolute values of *Score*, *RPO*, and *Wickets* indicate strong influence of these values on *Result*. The negative coefficient of wickets indicate that if the team lose more wickets they tend to lose matches, and the net effect of score and RPO is positive which says that if the team scores more they tend to win the match.

- The negative coefficient of *Balls* is may be due the data of 2nd innings as team wins matches if they chase faster hence less number of balls.

# 12 Conclusion

From the tests we have conducted we can conclude that the batting approach of team India is different w.r.t to 1st and 2nd innings. The important factors for the match results are mainly *Score*, *RPO*, *Wickets*, and *Balls* and we have found them to be important factors while performing Factor Analysis and Linear Discriminant Analysis. *Extras (b,lb,w,nb)* have less contribution towards the match result.