**Multivariate Analysis on CPL – 2020**

**Variables of Cricket that influence Batting & Bowling:**

Caribbean Premier League (abbreviated to CPL or CPLT20) is a professional Twenty20 championship

cricket league in West-Indies that has become very popular among cricket fans worldwide. Twenty20

is the latest format of cricket, and a typical match lasts about 3.5 hours. When compared to One Day

International (ODI) or Test Cricket games, its fast-paced style and shortened duration are two key

reasons for the increasing popularity of Twenty20 in recent years. Beginning in 2013, CPL had

completed its eighth consecutive season by September 2020. For the 2020 competition there were 6

competing teams: Jamaica Tallawahs, St Lucia Zouks, Guyana Amazon Warriors, Barbados Tridents,

St Kitts & Nevis Patriots, Trinbago Knight Riders and Trinbago Knight Riders winning most times (4)

and in 2020 also.

Franchise can acquire players by several ways. One of these methods is to buy players

at player auctions and also mid-season transfers. Obviously, these players were paid huge amount

by the franchises based on their performances. After tournament completion, it is always interesting

“How did these quality players actually perform?” ‘’Had they done value to their price tag?”

Analysis on particular players will help the franchises to buy them in auction and have some value to

their name in cart in auction.

In limited-overs cricket, like Twenty20, the batsmen goal is to score as many as runs as possible

for few balls. On the other hand, the key goal in test cricket is building longer innings instead of

trying to score runs for each ball is. For a batting analysis, there is set of widely-recognized variables

that can be used to measure the quality of each batsman. These variables are commonly used by

cricket commentators and sports authorities, and are also shown in scoreboards to describe player

profiles.

The variables used here are:

**Runs**: Total Runs scored by a batsman in CPL 2020. Higher the runs scored stronger the impact.

**Average**: Number of Runs scored by a player per total number of dismissals in CPL 2020. Also, there

are loopholes like “Not-outs” were not consider where some batsman has high average with less

number of times dismissed.

**Strike-Rate**: Number of Runs scored by player per 100 balls faced by him in CPL 2020.Higher the

Strike-Rate more impact the Batsman have in the match.

**Fours:** Total Number of 4s scored by a player in CPL 2020.Batsman scores boundaries to improve

run-rate and also his own strike-rate. Higher the boundaries have high impact.

**Sixes:** Total Number of 6s scored by a player in CPL 2020. Batsman scores sixes to improve

run-rate and also his own strike-rate. Higher the boundaries have high impact.

**HF:** Number of 50s + (2\* Number of 100s) scored by a player in CPL 2020. Mainly 50s and 100s

shows how long the player has impact in the match and also a greater number of HF says how many

matches does a player has impact in the tournament (CPL-2020).

Figure 1 shows the Individual Histograms for Batting Variables based on the performance of 88

players with least being ‘0’ runs who dismissed once in CPL 2020. Figure 1.a shows the Histogram of

Runs scored by 88 players in CPL 2020. Figure 1.b shows the Histogram of Individuals Average of the

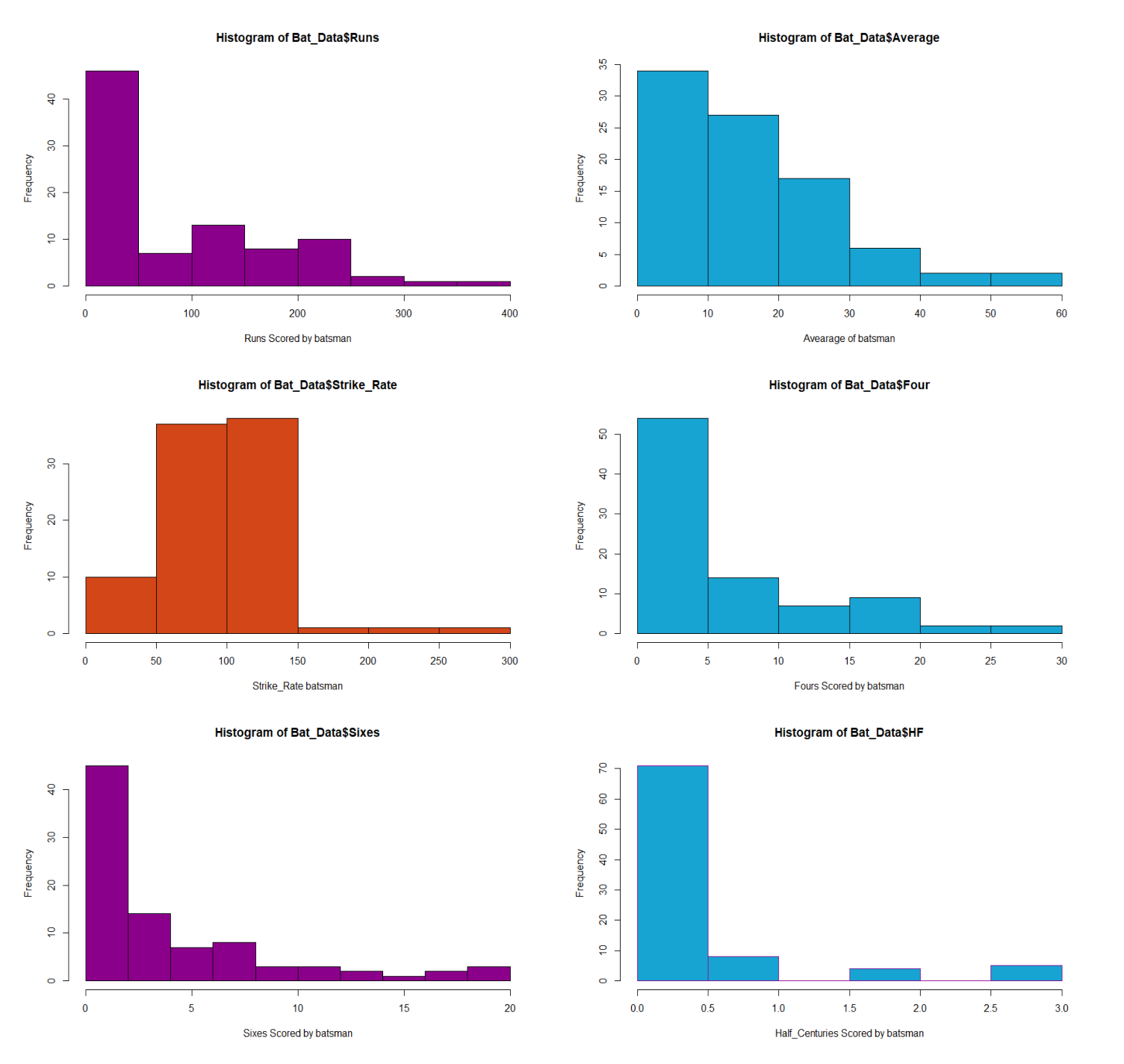
88 players in CPL 2020. Figure 1.c shows the Histogram of Individuals Strike-Rate of the 88 players in

CPL 2020. Figure 1.d shows the Histogram of Individuals Fours Scored by the 88 players in CPL 2020.

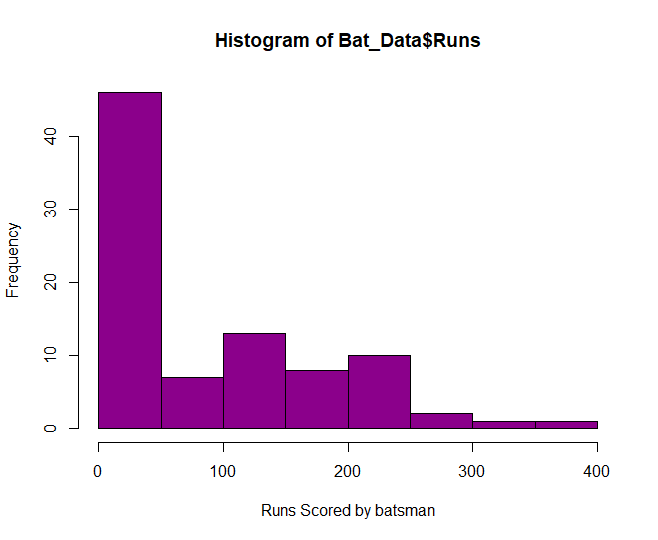
Figure 1.e shows the Histogram of Individuals Sixes scored by the 88 players in CPL 2020. Figure 1.e

shows the Histogram of Individuals HF (# of 50s + 2\*# of 100s) scored of the 88 players in CPL 2020.

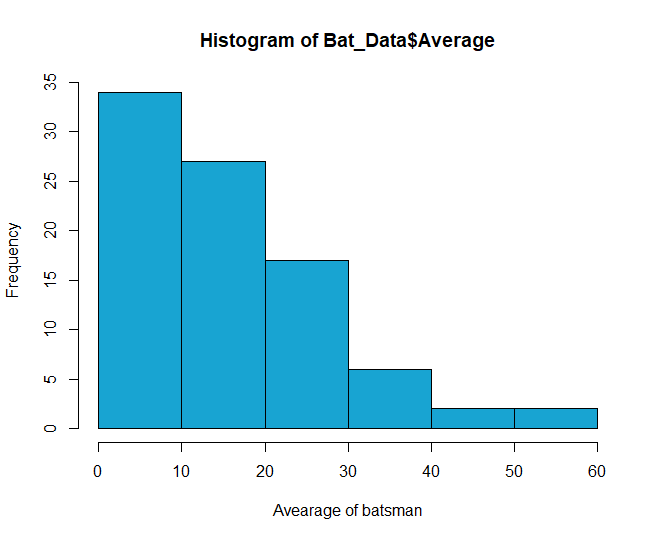
**Figure 1. Individual Histograms for Batting Variables**



**Figure 1- a The Histogram of Individual Runs scored in CPL 2020**

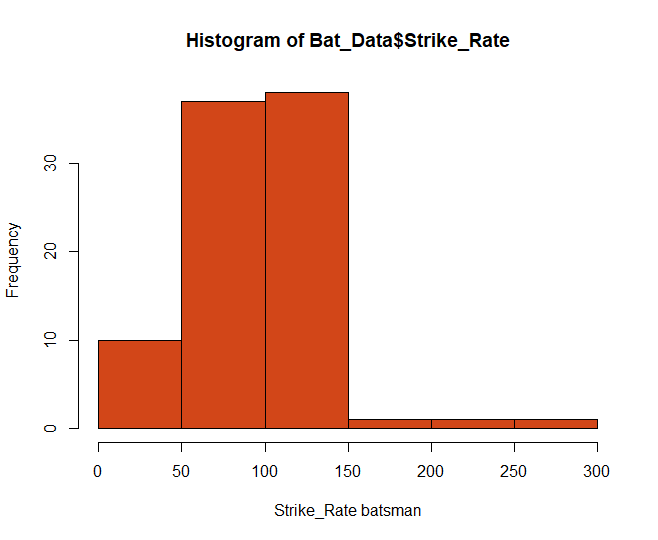


**Figure 1- b The Histogram of Individual Average of 88 players in CPL 2020**

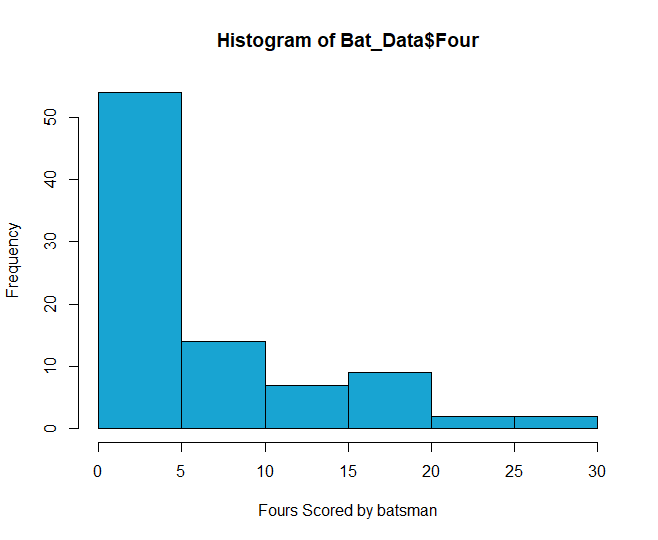


.

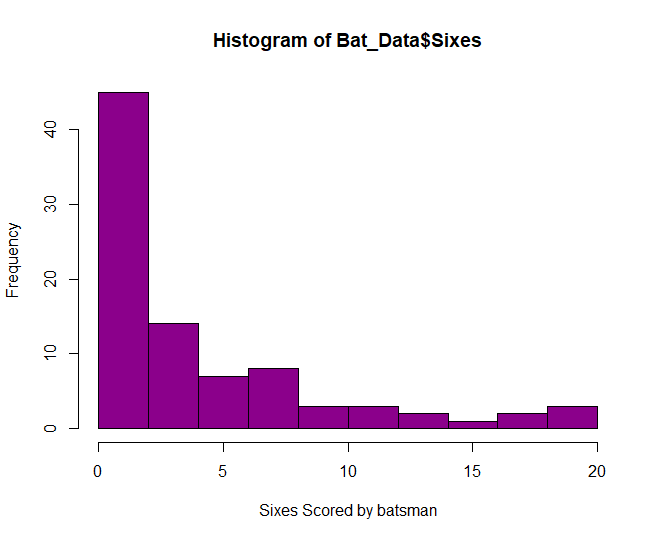
**Figure 1- c The Histogram of Individual Strike-Rate of 88 players in CPL 2020**



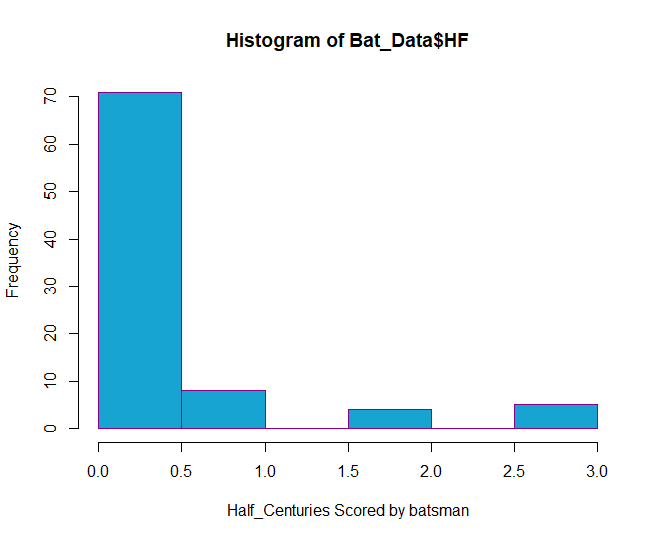
**1- d The Histogram of Individual 4s scored in CPL 2020**



**Figure 1- e The Histogram of Individual 6s scored in CPL 2020**



**Figure 1-f The Histogram of Individual HF (# of 50s +2\*# of 100s) scored in CPL 2020**



**Figure 2 Correlation Structure of Batting Variables**

Matrix Plot of Runs, Average, Strike-Rate, Fours, Sixes, HF

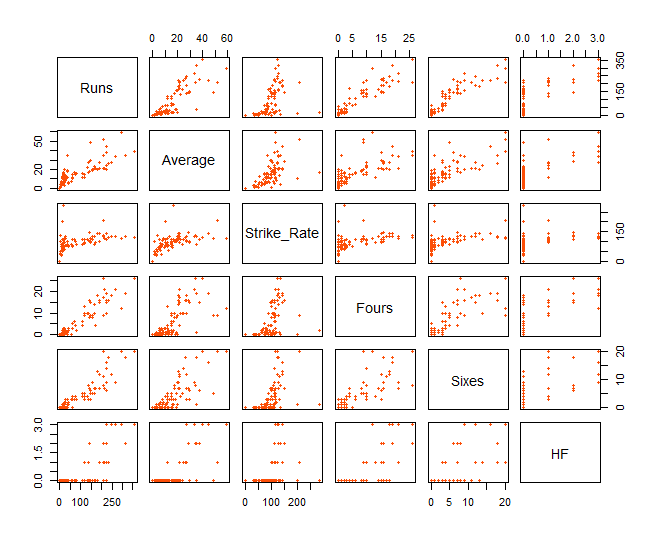


Figure 2 shows matrix plot between batting variables where we can see some correlation between

them. The plot shows that Runs and Fours, Runs and Sixes, and Ave and Runs are considerably

correlated, as might be anticipated.

Now, for measuring or for doing analysis on bowling performance in the tournament (CPL 2020)

there are other variables to be consider. The variables are wickets, strike-rate, economy and

average (bowling different from batting average).

The variables used for measuring bowling performance are:

**Wickets:** Total number of wickets taken by a player (bowler) in CPL 2020. Higher the number of

wickets say higher the impact that the player has in CPL 2020.

**Strike-Rate:** Total number of balls per total number of wickets taken by a player in CPL 2020. Lower

the Strike-Rate says that the players take very less balls to take wickets in CPL 2020. Often, captains

go to their strike bowler (Bowler with less strike-rate) to take wickets and break the partnership.

**Economy:** Total number of runs conceded per total number of overs bowled by a player in CPL 2020.

Low Economy ensures the bowler and bowling team to concede less number of runs which in return

allows the team to chase or defend low targets. Having low economy in Limited overs cricket like T-

20 ensures a good impact as many batsmen go after bowler from ball # 1.

**Average (Bowling Average):** Total number of runs conceded per total number of wickets taken by a

player in CPL 2020. Lower the average higher the impact in the match in CPL 2020. Often, Captains

go to bowlers with not only low strike-rate but also with low average in which taking wickets and

conceding less runs per wicket taken has major impact in limited overs cricket like T-20.

Figure 3 shows the Individual Histograms for Bowing Variables based on the performance of 56

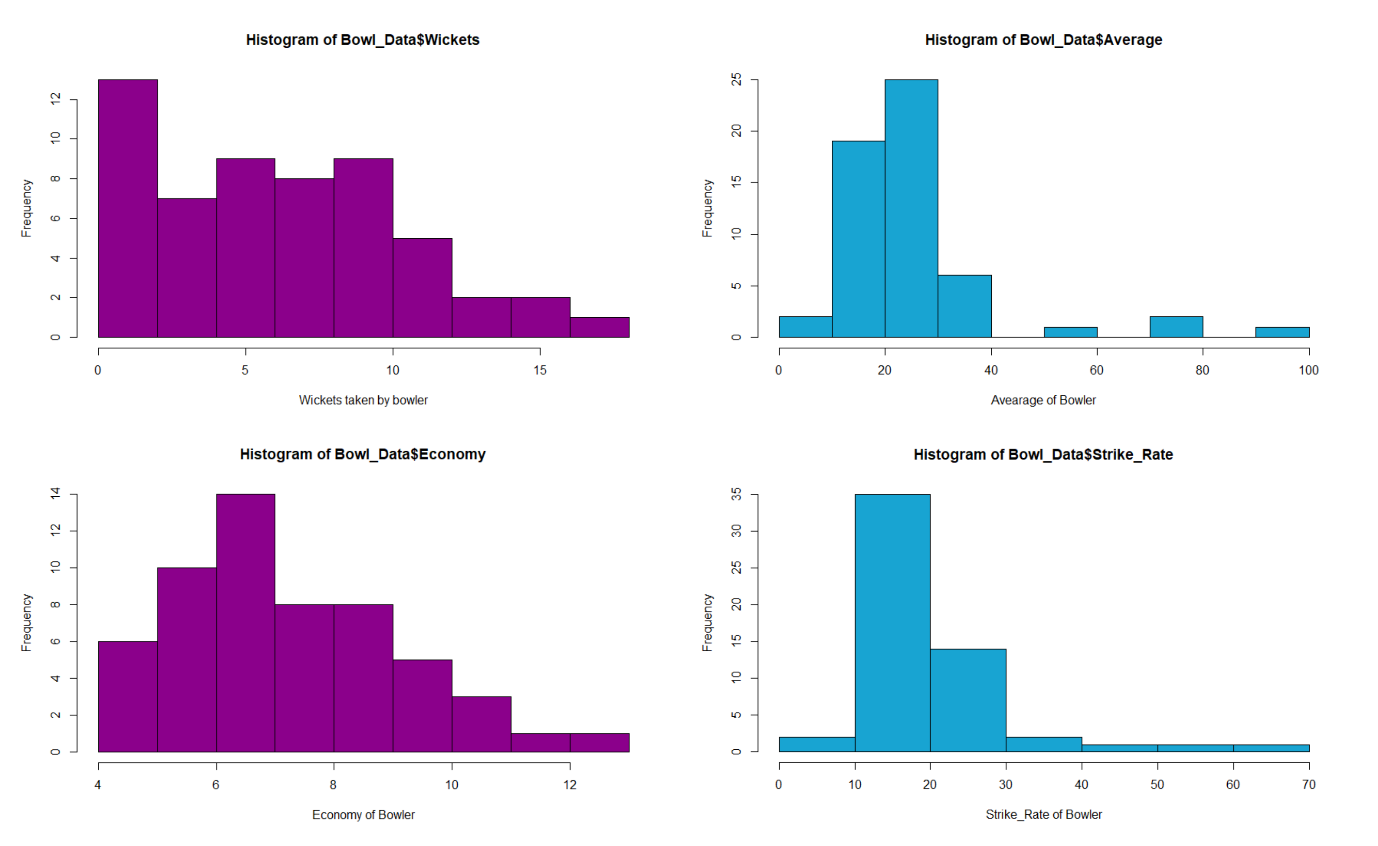
players with least being ‘1’ wicket taken by them in the CPL 2020. Figure 3.a shows the Histogram of

wickets taken by players in CPL 2020. Figure 3.b shows the Histogram of Individuals Average of the

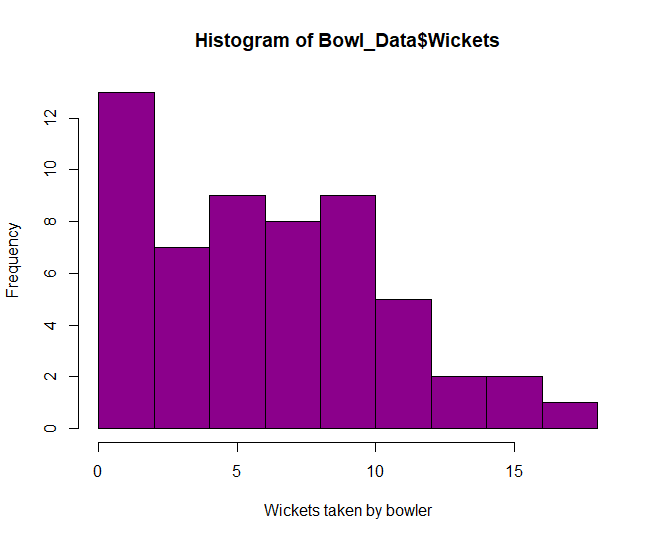
56 players in CPL 2020. Figure 3.c shows the Histogram of Individuals Economy of the 56 players in

CPL 2020. Figure 3.d shows the Histogram of Individuals Strike-Rate of the 56 players in CPL 2020.

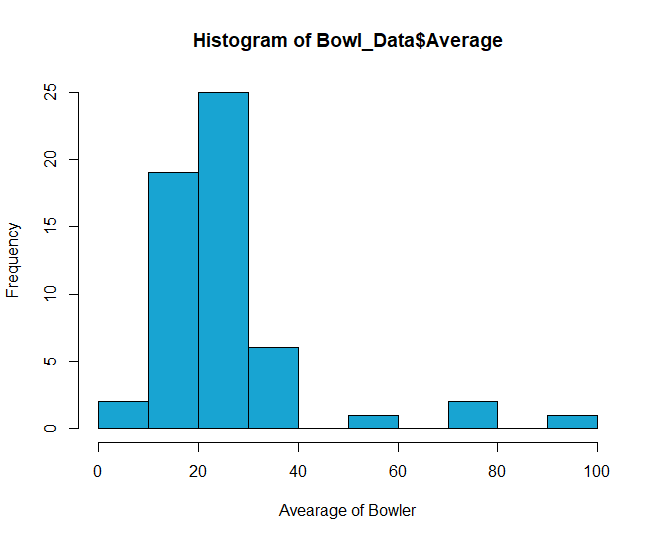
**Figure 3 Individual Histograms for Bowling Variables**



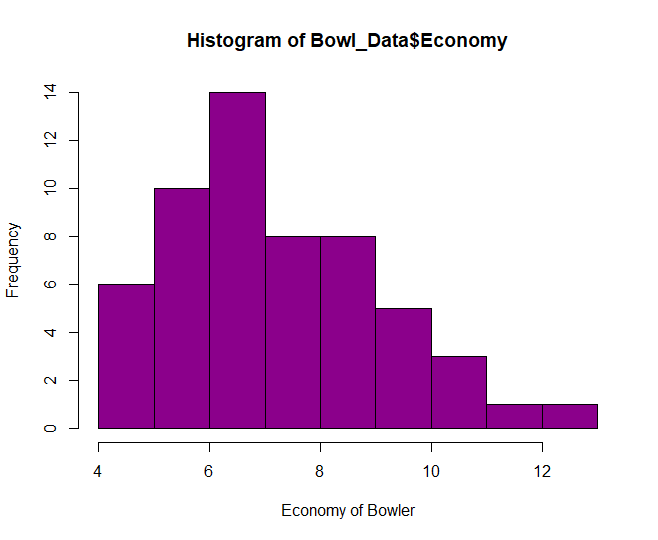
**Figure 3.a The Histogram of Individual Wickets Taken in CPL 2020**



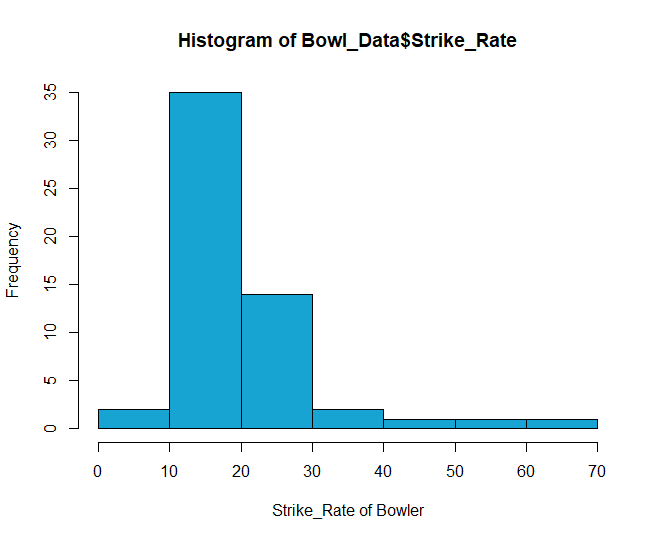
**Figure 3.b The Histogram of Individual Bowing Average in CPL 2020**



**Figure 3.c The Histogram of Individual Bowling Economy in CPL 2020**



**Figure 3.d The Histogram of Individual Bowling Strike-Rate in CPL 2020**



**Matrix Plot of Wickets, Bowl. Average, Economy and Bowl. Strike-Rate**

Figure 4. Sample Correlation Matrix of Bowling Variables

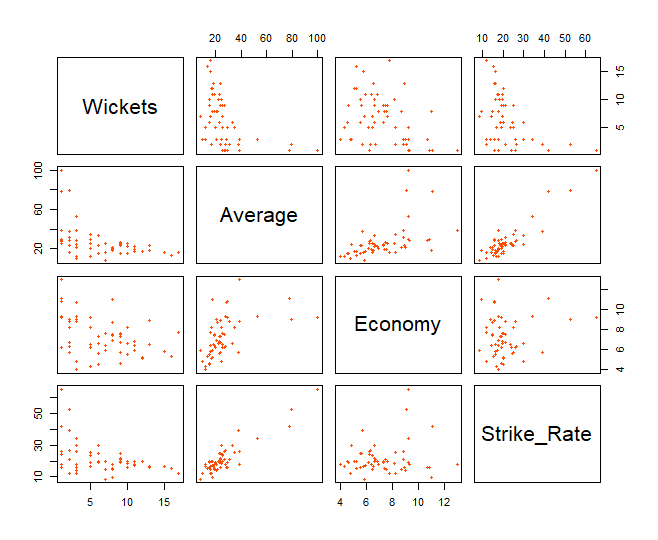


Figure 4 shows matrix plot between bowling variables where we can see some correlation between

them. The plot shows that Wickets and Average, Wickets and Economy, and Wickets and Strike-Rate

are considerably correlated, as might be anticipated. We can see, Bowling Average and the Strike

Rate are highly positively correlated. All the other variables are somewhat negatively correlated with

the number of wickets. However, each one of these variables measures a different quality of a

bowler, even though they are correlated. Constructing an overall measure of performance by using

some kind of weighted averaging would be the ideal way to handle this situation. This suggests the

potential usefulness of the Principal Component Analysis technique for bowler performance as well.

**Ranking using the First Principal Component (components with S.D. >=1):**

Readers can find excellent introductions to Principal Component Analysis (PCA) in the works of

Johnson and Wichern (2007), Dawkins (1989), and Watnik and Levine(2001). Nalik & Khattree (1999)

provide an example in which principal component analysis is used with sports data. Principal

Component Analysis (PCA) is a nonparametric variable reduction technique wellsuited for correlated

data that can be effectively used in our context. One objective of principal component analysis is to

collapse a set of correlated variables into fewer uncorrelated variables as linear combinations of the

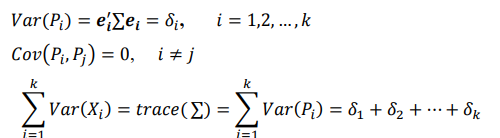
original variables.

Briefly, if  is a k-vector of random variables with variance-covariance matrix

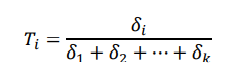
and corresponding eigenvalue-eigenvector pairs  , where  , then the principal components  are defined by



Furthermore, it can be shown that



Consequently, the proportion of total variability due to the principal component is given by



If the first principle component captures a substantial percentage of the total variation in the

observations, it can possibly be used to discriminate between the kvectors. Indeed, if accounts for

most of the variation seen in the data, then there is good reason to believe that it can successfully

be used for ranking purposes. For this reason, we call this technique the First Principal Component

(FPC) ranking method. In practice, it is customary to use the correlation matrix instead of the

variance-covariance matrix when the measurement units for the components of the  data vector

are largely dissimilar. For this reason, the correlation matrix is used in this analysis.

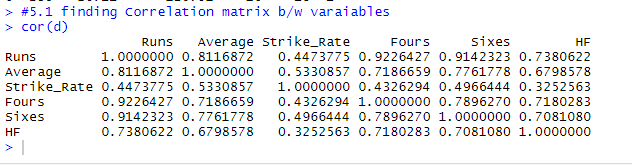
**Ranking Batsmen in CPL 2020:**

This analysis includes the batting variables Runs, Average (Batting), Strike-Rate (Batting), Fours,

Sixes and HF for all the 88 batsmen with least being ‘0’ runs scored in CPL 2020. Figure 5 shows the

correlation matrix for batting variables of 88 batsmen in CPL 2020.

**Figure-5 correlation matrix for 88 batsmen**



Values for each of these variables were collected together into a (6 x 1) t column vector of the

form (Runs, Ave, SR, Fours, Sixes, HF) t for each of 88 batsmen. These we call the batting vectors.

Once data have been obtained, the (6 x 6) sample correlation matrix associated with the sample

batting vectors may be examined for the correlation structure inherent in these cricket variables.

Figure 6 shows the standard deviation of 6 components (Batting Variables) and cumulative variability

or Proportion of 6 components (Batting variables).

**Figure 6 Ordered component with respected S.D. and Cumulative Proportion**

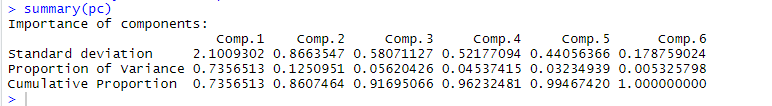


Figure 7. show the eigen values of corresponding batting variables (6 components) like Runs,

Average (Batting), Strike-Rate (Batting), Fours, Sixes and HF scored in CPL 2020.

**Figure 7. Eigen values of corresponding batting variables (6 Components)**

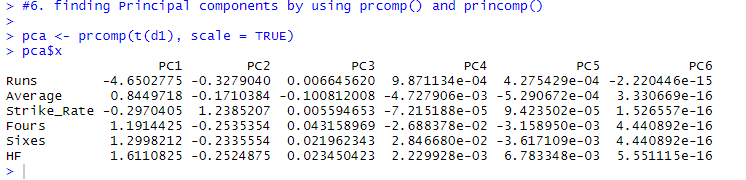
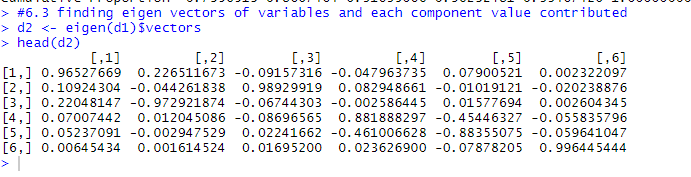


Figure 8. shows the Corresponding Eigen vector pairs for sample correlation matrix i.e., batting

variables (6 Components) like Runs, Average (Batting), Strike-Rate (Batting), Fours, Sixes and HF

scored in CPL 2020.

**Figure 8. Eigen vector pairs of batting variables (6 Components)**



From figure 6 and figure 9, We can clearly see that only Component [1] has Standard Deviation

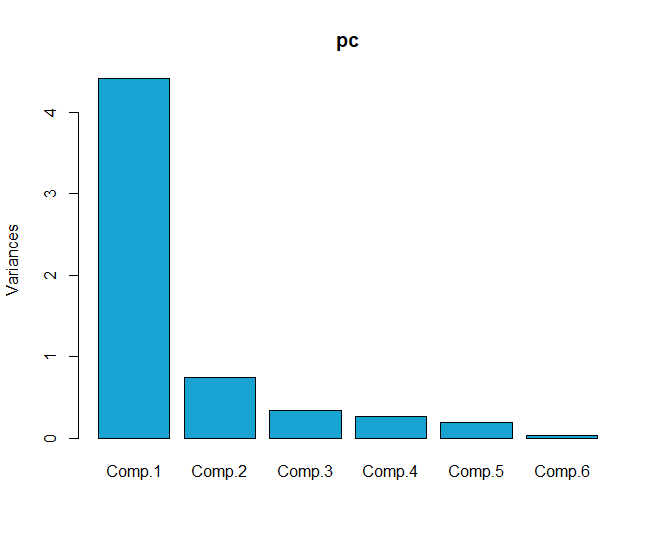
(2.1009302) > 1 and variance (4.42) > 1 and also, it explains variability of 73% of total variability.

Hence the corresponding eigen value of Component [1] are column [1] are only consider and the

first principal component of batsmen is calculated by

L1 = 0.965\*Runs + 0.226\*Average – 0.091\*Strike-Rate – 0.047\*Fours + 0.079\*Sixes + 0.002\*HF

**Figure 9. Histogram of Variances of batting variables (6 Components)**



**Figure 10. Screen Plot of Batting Variables**

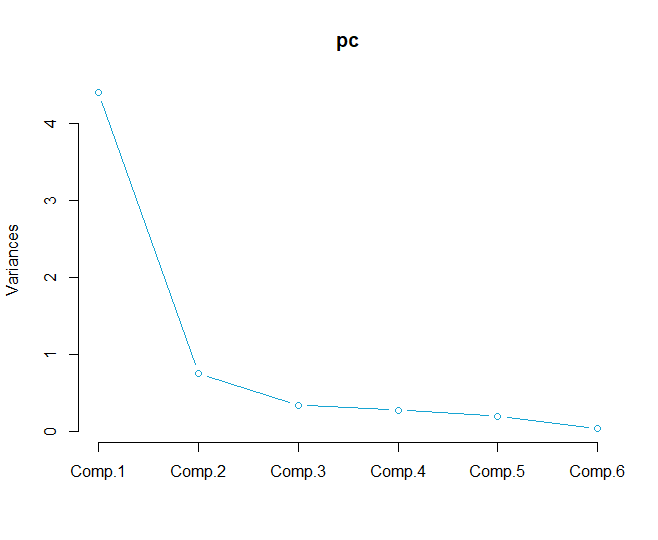


Figure 10 explains that only Comp.1 has variance > 1 and explains more than 73% of total variability

among other batting components. we refer to our first principal component as the general-batting-

performance-index, which is a type of weighted average of all six variables used. Here, the

coefficients of the first principal component are both positive and negative, so larger values of L1

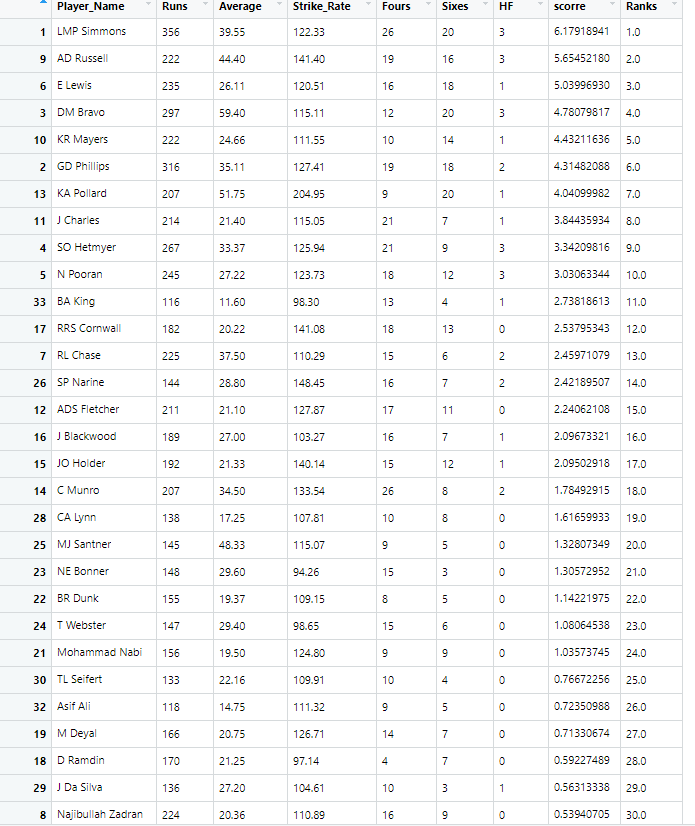
indicate better player performance. This justifies that we should rank (largest to smallest) the

players based on the first principal component.

Figure 11-a ,11-b, 11-c shows the 88 batsmen in CPL 2020 where they were assigned ranks using

Score calculated by First Principal Component discussed earlier.

**Figure 11-a. Batsmen and their Ranks using FPC**



We can see that KA Pollard who was 13th highest run getter was ranked 7th using First Principal

Component also DM Bravo who was 3rd highest run getter was ranked 4th using First Principal

Component and whereas GD Philips 2nd highest run getter was ranked 6th using First Principal

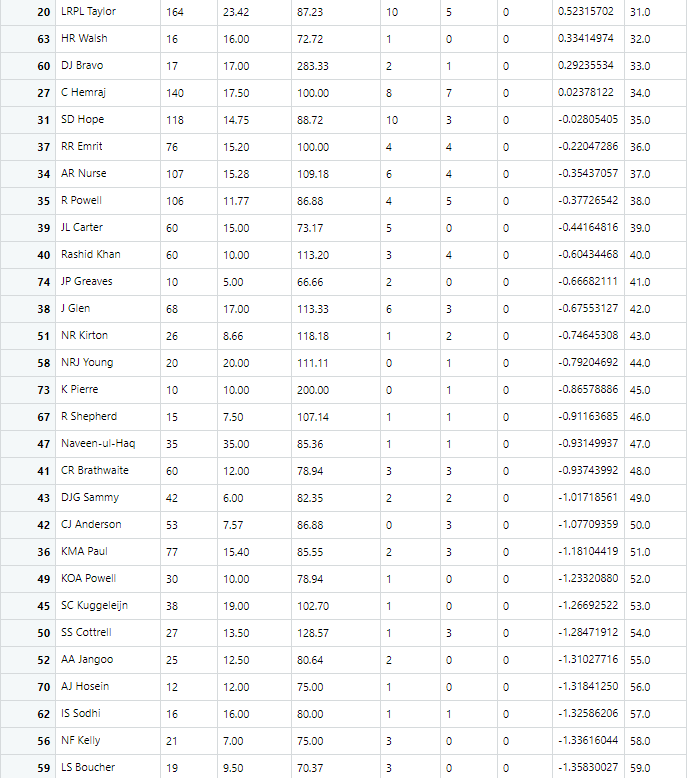
Component and moreover First Principal component explain the variability among all other batting

Variables and some other methods like Ramakrishnan method used by ESPN doesn’t explain the

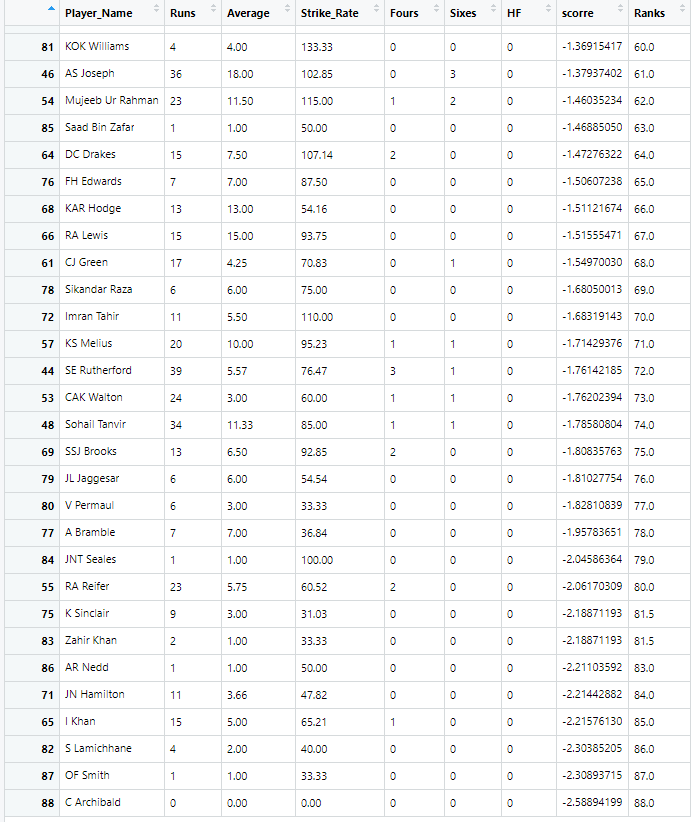
Variability like First Principal Component. Also, First Principal Component explain many other which

methods like Ramakrishnan used by ESPN doesn’t.

**Figure 11-b. Batsmen and their Ranks using FPC**



**Figure 11-c. Batsmen and their Ranks using FPC**



**Figure a. Bar plot of Individual FPC Scores of Batsmen in CPL 2020**

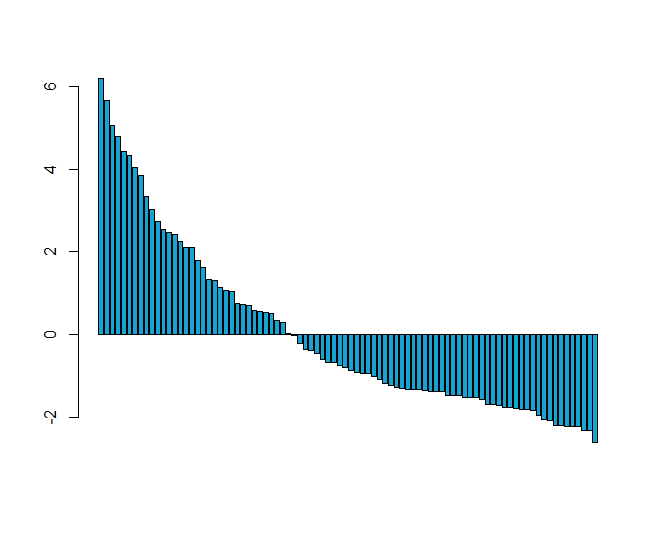


Figure-a shows the bar plot of score of 88 Individual Batsmen calculated using First Principal

Component. There are batsmen with positive and negative score. Higher the score higher the impact

they had in CPL 2020.

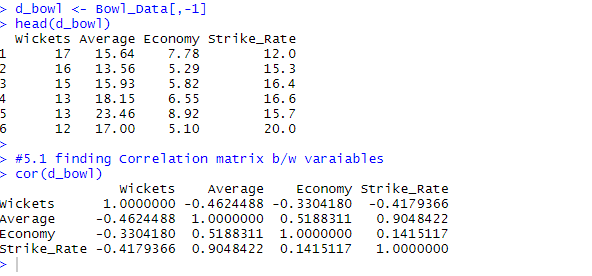
***Ranking Bowlers in CPL 2020:***

This analysis includes the bowling variables Wickets, Average (Bowling), Economy, and

Strike-Rate (Bowling) for all the 56 bowlers with least being ‘1’ wicket taken in CPL 2020. Figure 12

shows the correlation matrix for batting variables of 56 bowlers in CPL 2020.

**Figure-12 correlation matrix for 56 Bowlers**



Values for each of these variables were collected together into a (4 x 1) t column vector of the

form (Wickets, Average, Economy, Strike-Rate) t for each of 56 bowlers. These we call the bowling

vectors. Once data have been obtained, the (4 x 4) sample correlation matrix associated with the

sample bowling vectors may be examined for the correlation structure inherent in these cricket

variables.

Figure 13 shows the standard deviation of 4 components (Bowling Variables) and cumulative

variability or Proportion of 4 components (Bowling variables).

**Figure 13 Ordered component with respected S.D. and Cumulative Proportion**

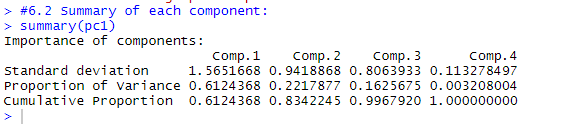


Figure 14. show the eigen values of corresponding batting variables (4 components) like Wickets,

Average (Bowling), Economy and Strike-Rate (Bowling) in CPL 2020.

**Figure 14. Eigen values of corresponding bowling variables (4 Components)**

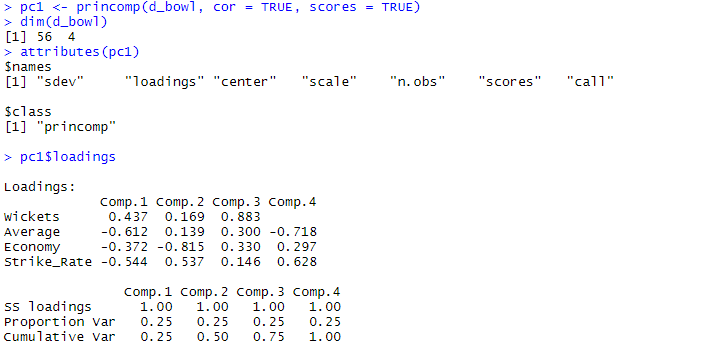
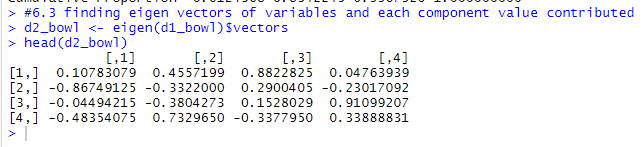


Figure 15. shows the Corresponding Eigen vector pairs for sample correlation matrix i.e., bowling

variables (4 Components) like Wickets, Average (Bowling), Economy and Strike-Rate (Bowling)

in CPL 2020.

**Figure 15. Eigen vector pairs of batting variables (6 Components)**



From figure 13 and figure 16, We can clearly see that only Component [1] has Standard Deviation

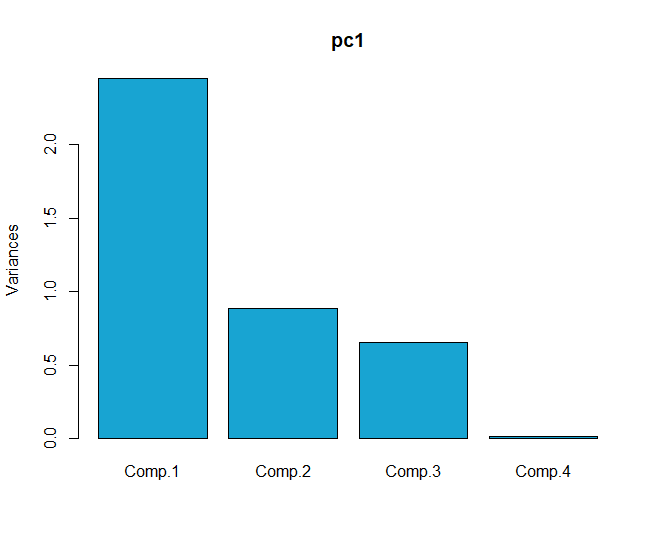
(1.5651) > 1 and variance (2.46) > 1 and also, it explains variability of 61% of total variability.

Hence the corresponding eigen value of Component [1] are column [1] are only consider and the

first principal component of bowler is calculated by

L1 = 0.1087\*Wickets - 0.867\*Average (Bowl.) – 0.0449\*Economy – 0.483\*Strike-Rate (Bowl.)

**Figure 16. Histogram of Variances of bowling variables (4 Components)**



**Figure 17. Screen Plot of Batting Variables**

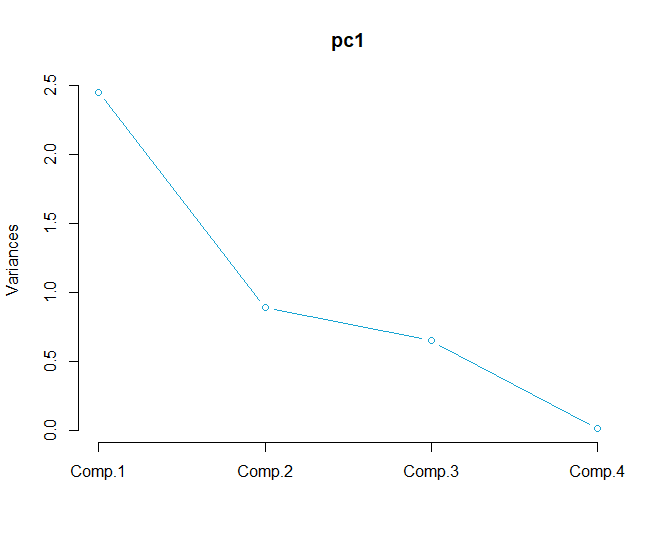


Figure 17 explains that only Comp.1 has variance > 1 and explains more than 61% of total variability

among other bowling components. we refer to our first principal component as the general-bowling-

performance-index, which is a type of weighted average of all four variables used. Here, the

coefficients of the first principal component are both positive and negative, so larger values of L1

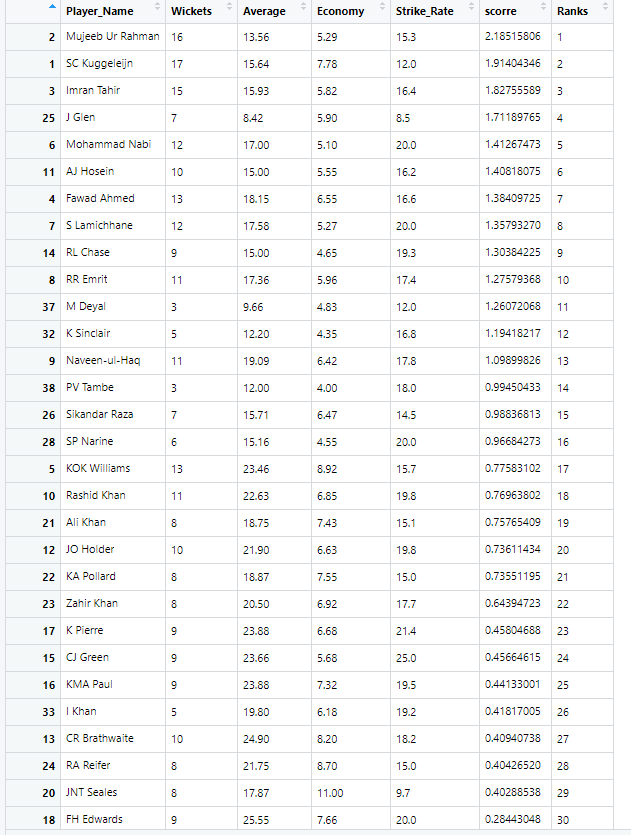
indicate better player performance. This justifies that we should rank (largest to smallest) the

players based on the first principal component.

Figure 18-a ,18-b shows the 56 bowlers in CPL 2020 where they were assigned ranks using

Score calculated by First Principal Component discussed earlier.

**Figure 18-a. Bowlers with ranks and score calculated using FPC**



We can see that Mujeeb who was 2nd highest wicket taker was ranked 1st using First Principal

Component also Imran Tahir who was 3rd highest wicket taker was ranked 3rd using First Principal

Component and whereas J Glen 25th highest wicket taker was ranked 4th using First Principal

Component and moreover First Principal component explain the variability among all other batting

Variables and some other methods like Ramakrishnan method used by ESPN doesn’t explain the

Variability like First Principal Component. Also, First Principal Component explain many other which

methods like Ramakrishnan used by ESPN doesn’t.

**Figure 18-b. Bowlers with their score and rank calculated using FPC**

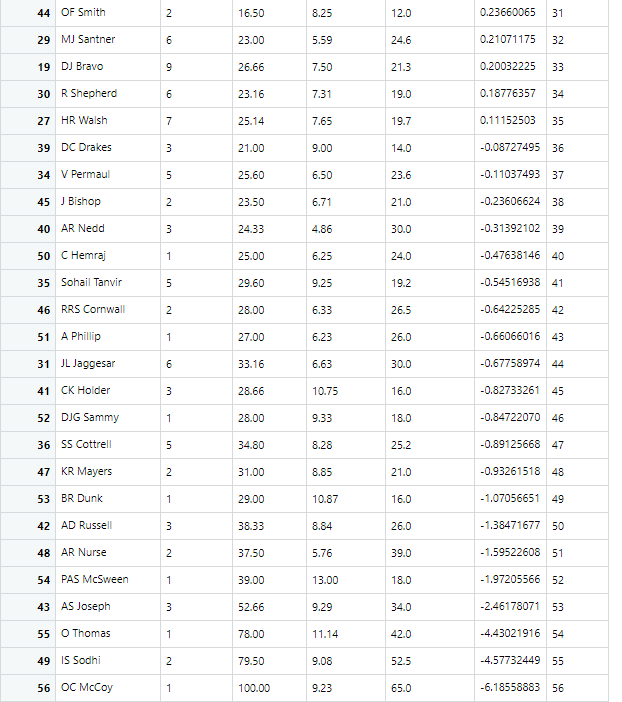
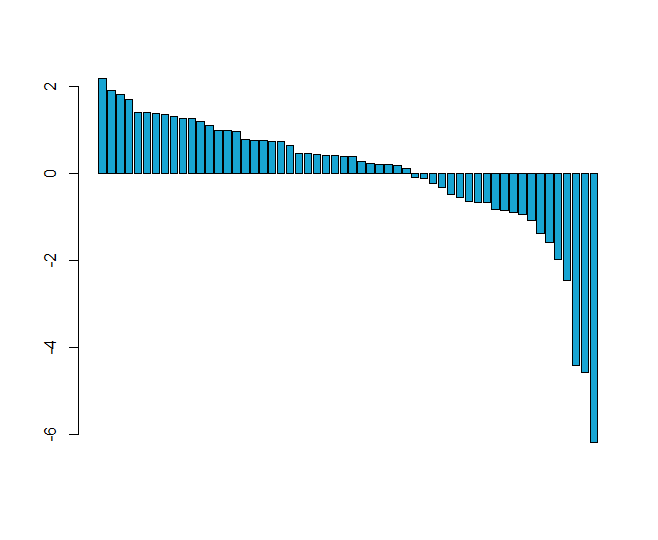


Figure b. Bar plot of scores of 56 Bowlers in CPL 2020 calculated using FPC



From Figure b we can see the scores of 56 bowlers in CPL 2020 calculated using FPC (First Principal

Component). There are bowlers with positive and negative score. Higher the score higher the impact

they had in CPL 2020.

**Conclusion & Importance:**

With the help of Multivariate analysis for both Batsmen and Bowler franchises can sign the correct

player for their team in the auction using previous year data in the tournament (CPL or IPL) and

players recent stats in other leagues or tournaments. They get to know the most impact player in

their team in the tournament and retain them and let off players with least impact in next year

auction and build strong team and strong playing xi which helps franchise to win the Trophy.