Mini Project: Advanced Statistics

Model Report

ANOVA, Regression Analysis, PCA, Factor Analysis

GREAT LAKES INSTITUTE OF MANAGEMENT

Table of Contents

1 Project Objective

This project has four parts. Please refer to the ppt for data description of each file and ensure to upload only one document for all three parts and label them correctly in the doc and answer of each part should be in sequence 1, 2,3 (file with different sequence will not be evaluated)

1. ANOVA (Dataset to be used Metadata and PL_X_SELL). For more details refer to the ppt.
    - Conduct a one-way ANOVA analysis to study whether occupation of the account holder affects quarterly average balance in the account
    - Conduct two-way ANOVA analysis on gender and occupation on quarterly average balance.
2. Household expenditure data – Regression.
    - The data set has information on monthly expenditure, annual income of the household, monthly income, household size (number of members in the household), and monthly EMI.
    - Set up a regression model to explain expenditure using the other variables in the data.
3. PCA and Regression(DATASET NAME: Batting_Data.csv , Bowling_Data.csv)
    - Run Principal Component Analysis and
        - Interpret loadings
        - Interpret Communality
        - Number of components to be retained
        - Total variance extracted
        - Check whether rotation is necessary
        - Label the components
        - Use the PC Scores to rank the players
4. Factor Analysis (MBA Car Datafile)
    - The raw data are available in the file labeled mbacar.
    - Conduct a common factor analysis on the data set. How many factors you would retain? How do you interpret them?
    - Save the factor scores and plot the average factor scores against each other for each of the 10 cars evaluated by the students. What do the plots tell you about the similarities of the 10 car models?

2 ANOVA Analysis

2.1 Objective

- Conduct a one-way ANOVA analysis to study whether occupation of the account holder affects quarterly average balance in the account
- Conduct two-way ANOVA analysis on gender and occupation on quarterly average balance.

2.2 Steps to follow

We shall perform ANOVA Analysis in the following sequence:

1. Loading of the data file
2. Descriptive Statistics
3. Data Visualization
4. Check Interaction between Variables (In case of Two Way ANOVA)
5. Test of Assumptions
   - Normality
   - Homogeneity
6. Analysis of Variance: One Way and Two Way
7. Robust Methods execution: These methods shall be executed if Assumptions are violated.
8. Post Hoc Test (Tukey): To see where exactly the differences have occurred between the groups.
9. Summary

2.3 One Way ANOVA

2.3.1 Descriptive Statistics

The outcome of Basic descriptive statistics on the PL_X_SELL Dataset is as follows:

```
# Find out Names of the Columns (Features)
names(PL_X_SELL)

##  [1] "Cust_ID"         "Target"          "Age"             "Gender"
##  [5] "Balance"         "Occupation"      "No_OF_CR_TXNS"   "AGE_BKT"
##  [9] "SCR"             "Holding_Period"

# Find out Class of each Feature, along with internal structure
str(PL_X_SELL)

## 'data.frame':    20000 obs. of  10 variables:
##  $ Cust_ID        : Factor w/ 20000 levels "C1","C10","C100",..: 1 2 3 4
##  $ Target         : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Age            : int  30 41 49 49 43 30 43 53 45 37 ...
##  $ Gender         : Factor w/ 3 levels "F","M","O": 2 2 1 2 2 2 2 2 2 2
##  $ Balance        : num  160379 84371 60849 10559 97100 ...
```

```
##  $ Occupation    : Factor w/ 4 levels "PROF","SAL","SELF-EMP",..: 2 3 1
##  $ No_OF_CR_TXNS : int  2 14 49 23 3 2 23 45 3 33 ...
##  $ AGE_BKT       : Factor w/ 7 levels "<25",">50","26-30",..: 3 6 7 7 6
##  $ SCR           : int  826 843 328 619 397 781 354 239 339 535 ...
##  $ Holding_Period: int  9 9 26 19 8 11 12 5 13 9 ...
```

```
# Provide Summary of a Dataset.
summary(PL_X_SELL)
```

```
##     Cust_ID          Target               Age          Gender
##  C1      :     1   Min.   :0.00000   Min.   :21.0   F: 5525
##  C10     :     1   1st Qu.:0.00000   1st Qu.:30.0   M:14279
##  C100    :     1   Median :0.00000   Median :38.0   O:  196
##  C1000   :     1   Mean   :0.08665   Mean   :38.4
##  C10000  :     1   3rd Qu.:0.00000   3rd Qu.:47.0
##  C10001  :     1   Max.   :1.00000   Max.   :55.0
##  (Other):19994
##     Balance            Occupation    No_OF_CR_TXNS     AGE_BKT
##  Min.   :       0   PROF     :5463   Min.   : 0.00   <25  :1784
##  1st Qu.:   23737   SAL      :5839   1st Qu.: 7.00   >50  :3020
##  Median :   79756   SELF-EMP:3366   Median :13.00   26-30:3404
##  Mean   : 146181   SENP     :5332   Mean   :16.65   31-35:3488
##  3rd Qu.: 217311                    3rd Qu.:22.00   36-40:2756
##  Max.   :1246967                    Max.   :50.00   41-45:3016
##                                                      46-50:2532
##       SCR          Holding_Period
##  Min.   :100.0   Min.   : 1.00
##  1st Qu.:333.0   1st Qu.: 8.00
##  Median :560.0   Median :16.00
##  Mean   :557.1   Mean   :15.34
##  3rd Qu.:784.0   3rd Qu.:23.00
##  Max.   :999.0   Max.   :31.00
##
```

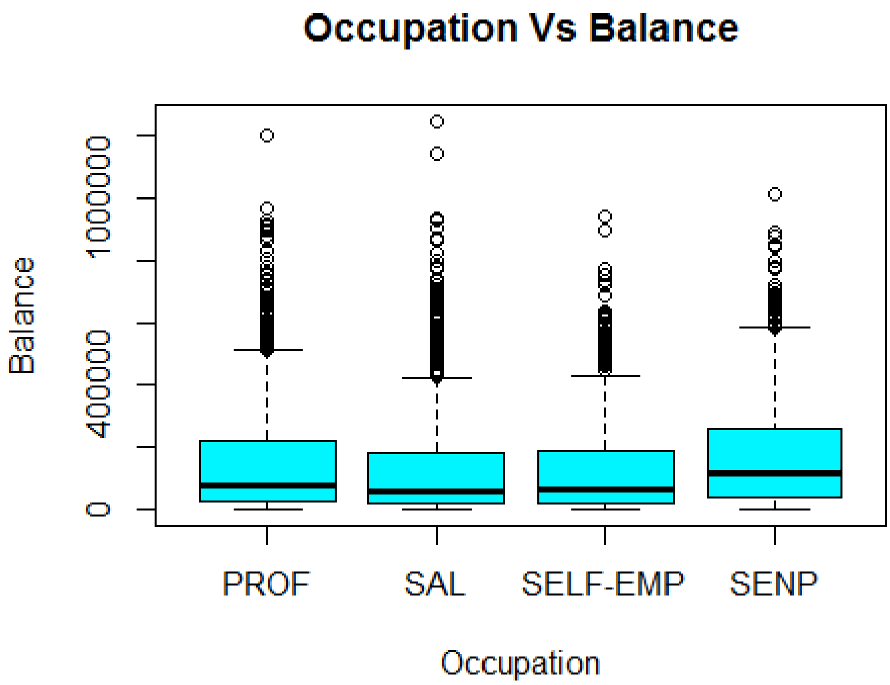
Summary of Data Features:

Sr. No.	Feature Name	Feature Type	Feature Description
1	Cust_ID	Factor	Customer ID, a Unique Identifier
2	Target	Integer	The Labelled Class having two categories. 0 - Representing the Non-Responder Segment and 1 -Representing the Responder Segment.
3	Age	Integer	Age of Customer
4	Gender	Factor	Gender of the customer; Male, Female and O; O represents Companies or Firms.
5	Balance	Numeric	Average Quarterly Balance maintained by the customer in deposit account

Sr. No.	Feature Name	Feature Type	Feature Description
6	Occupation	Factor	Occupation of the Customer
7	No_of_Cr_TXNS	Numeric	No of Credit Transactions recorded on the account in last month
8	AGE_BKT	Factor	Age Slabs
9	SCR	Integer	Generic Marketing Score of the customer
10	Holding_Period	Integer	Ability of the customer to hold money in the account measured in Number of days. Value range is between 0 to 31 days

2.3.2 Data Visualization

Following Boxplot shows the relationship between Occupation Vs Balance:



<u>Interpretation:</u>

All the occupation types have outliers in Balance, however, most of the **occupation type** have some variations in the mean but the difference is not significant. However, it seems that **SENP (Self Employed but not professional) type of Occupation has maximum average balance** and their holding period is also high.

2.3.3 Testing of Assumptions: One Way ANOVA

2.3.3.1 Test of Normality: The Anderson-Darling Test

Shapiro test cannot be performed here as the sample size is too large. Hence, Anderson- Darling normality test was performed which is from 'nortest' package.

```
#Anderson Darling Test for Normality

for(i in
    unique(factor(Occupation)))
  {cat(ad.test(PL_X_SELL[PL_X_SELL$Occupation==i,]$Balance)$p.value,"")}

## 3.7e-24 3.7e-24 3.7e-24 3.7e-24
```

Interpretation:

P value is less than 0.05 for all occupation types, showing the **occupations are not normally distributed.**

2.3.3.2 Homogeneity in Variance Test

We performed Levene's Test and Barlett's Test to check Homogeneity in Variance.

```
# Homogeneity in Variance Test using Levenes Test
leveneTest(Balance~Occupation)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     3  54.545 < 2.2e-16 ***
##       19996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

The Levene's test and Bartlett test has rejected the null hypothesis with p value <0.05 .Hence we may conclude that there is a **difference in the variance**. Hence the assumption of **Homogeneous in variance is also violated.**

2.3.4 One Way ANOVA

```
# ANOVA - Analysis of Variance
aov1 <- aov(Balance~Occupation)
summary(aov1)

##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Occupation      3 1.052e+13 3.506e+12   123.8 <2e-16 ***
## Residuals   19996 5.662e+14 2.831e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the normality violated, **Kruskal.test** test was conducted to understand if any difference in average account balance in different occupation. The test result shows a significant p value, concluding that there is significant difference between different occupations on account balance at 0.05 significance level.

```
kruskal.test(Balance~Occupation)

##
##  Kruskal-Wallis rank sum test
##
## data:  Balance by Occupation
## Kruskal-Wallis chi-squared = 582.54, df = 3, p-value < 2.2e-16
```

As the homogeneity of variance also violated, robust method is done to test the significance.

```
# Robust Method for One Way Anova
#
oneway.test(Balance ~ Occupation, var.equal=FALSE)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Balance and Occupation
## F = 120.18, num df = 3, denom df = 10305, p-value < 2.2e-16

model1<- lm(Balance ~ Occupation)

Anova(model1,Type="II", white.adjust=TRUE)

## Analysis of Deviance Table (Type II tests)
##
## Response: Balance
##               Df      F    Pr(>F)
## Occupation     3 120.18 < 2.2e-16 ***
## Residuals  19996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

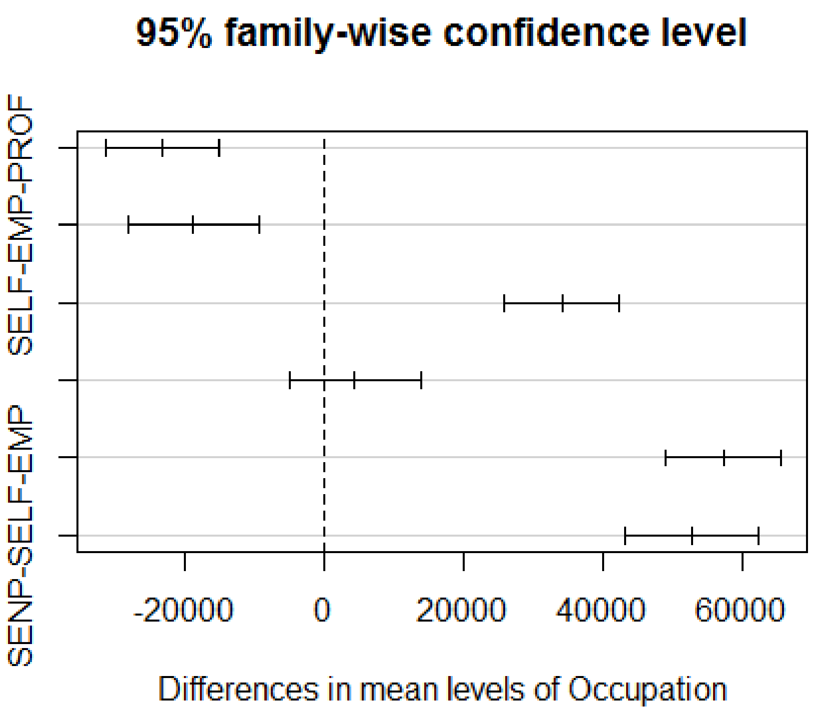
Kruskal Wallis test and Robust method shows the same significance of ordinary ANOVA. Hence there is a significant difference between the account balance between different occupation types.

2.3.5 Post Hoc Test (Tukey)

Further, we conducted Post-Hoc Tukey Test to see where exactly the difference in Variance is:

```
# Post-Hoc Test (Tukey)
TukeyHSD(aov1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Balance ~ Occupation)
```

```
##
## $Occupation
##                      diff        lwr        upr     p adj
## SAL-PROF       -23151.230 -31288.977 -15013.482 0.0000000
## SELF-EMP-PROF -18592.178 -28065.330  -9119.026 0.0000028
## SENP-PROF       34199.915  25877.257  42522.573 0.0000000
## SELF-EMP-SAL     4559.052  -4797.095  13915.198 0.5936835
## SENP-SAL        57351.145  49161.914  65540.376 0.0000000
## SENP-SELF-EMP   52792.093  43274.678  62309.508 0.0000000

plot(TukeyHSD(aov1))
```



Interpretation:

As can be seen from TukeyHSD test, and the plot, there is significant difference in mean levels of Occupation of `SELF-EMP-SAL` type.

2.4 One Way ANOVA Summary:

Since the data had violated both the assumption of normality and homogenous of variance, Kruskal Wallis Test and Robust method was performed which showed the similar result of normal ANOVA test. Hence we can conclude that there is significant difference seen between occupation and average amount of account balance.
Post hoc test was conducted to find that there is significant difference across the occupations except between self-employed and salaried person.

2.5 Two Way ANOVA:

2.5.1 Create Factor Variables:

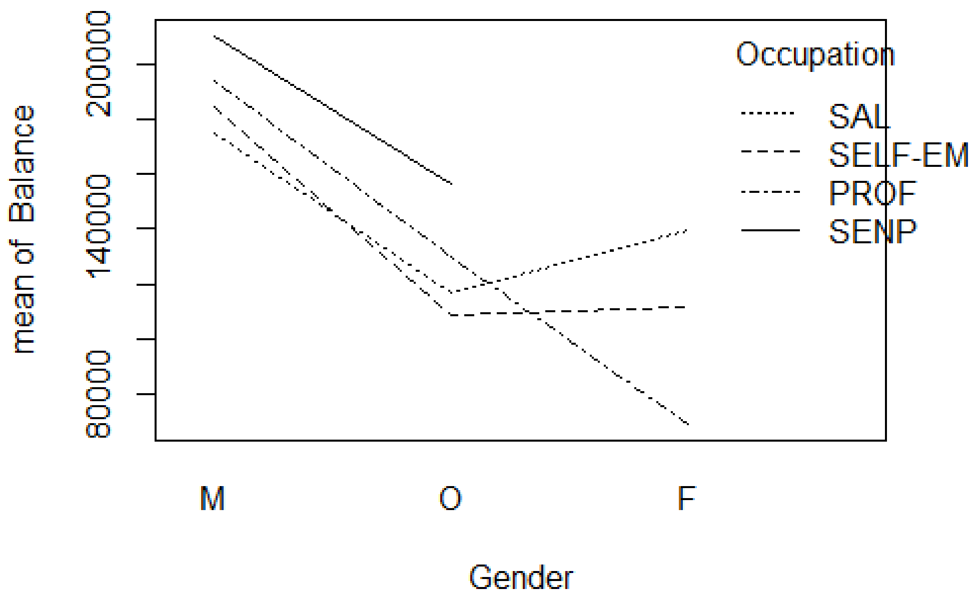
```
# Factors:
Gender<-factor(Gender,labels=c("M","O","F"))
Occupation<-factor(Occupation,labels=c("PROF","SAL","SELF-EMP","SENP"))
```

2.5.2 Data Visualization:

```
# Data Visualization
tapply(Balance,list(Gender,Occupation),mean)

##          PROF      SAL  SELF-EMP     SENP
## M 194154.44 174860.4 184217.4 210156.0
## O 129761.61 116906.7 108727.4 156289.2
## F  69081.36 139669.8 111960.2       NA

tapply(Balance,list(Gender,Occupation),sd)

##        PROF      SAL  SELF-EMP     SENP
## M 186560.2 198516.3 188517.8 193518.9
## O 164560.0 152463.1 138033.2 158528.5
## F 147598.3 224833.8 225062.5       NA
```

2.5.3 Check Interaction between Variables:

```
# Interaction Plot
#
interaction.plot(Gender,Occupation,Balance)
```



2.5.4 Testing for Assumptions:

Normality Test: Anderson Darling Test:

The normality was violated for occupation in the one way ANOVA analysis. Given below is the normality seen for Gender is also violated as p<0.05. Hence normality is violated.

```
# Testing of Assumptions
# Normality Test - Anderson Darling Test
for(i in
    unique(factor(Gender)))
{cat(ad.test(PL_X_SELL[PL_X_SELL$Gender==i,]$Balance)$p.value,"")}

## 3.7e-24 3.7e-24 3.7e-24
```

Homogeneity of Variance Test: LeveneTest

```
# Homogeneity of Variance Test
#
leveneTest(Balance~Occupation*Gender)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group    10  52.553 < 2.2e-16 ***
##       19989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the P Value, (less than 0.05), we conclude that **Homogeneity of Variance is also violated.**

2.5.5 Two Way ANOVA Execution:

```
aov2 <- aov(Balance~Occupation+Gender+Occupation:Gender)
summary(aov2)

##                      Df    Sum Sq   Mean Sq F value   Pr(>F)
## Occupation            3 1.052e+13 3.506e+12 127.035 < 2e-16 ***
## Gender                2 1.401e+13 7.004e+12 253.797 < 2e-16 ***
## Occupation:Gender     5 5.145e+11 1.029e+11   3.729 0.00225 **
## Residuals         19989 5.517e+14 2.760e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

The Two Way ANOVA Test confirms that there is significant difference in the Variance.

2.5.6 Post Hoc Test (Tukey)

```
TukeyHSD(aov2)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

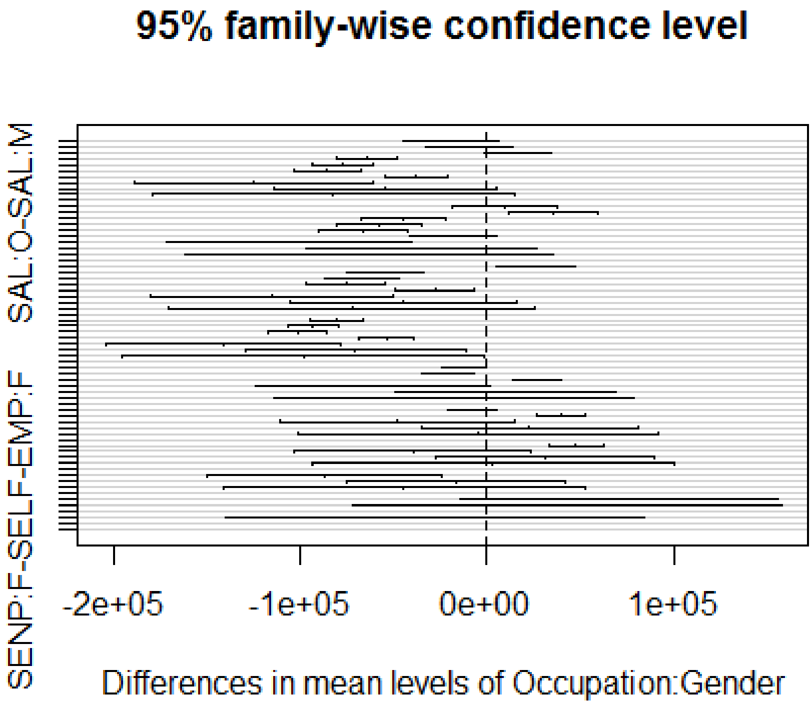
```
##
## Fit: aov(formula = Balance ~ Occupation + Gender + Occupation:Gender)
##
## $Occupation
##                   diff        lwr        upr     p adj
## SAL-PROF      -23151.230 -31185.335 -15117.124 0.0000000
## SELF-EMP-PROF -18592.178 -27944.680  -9239.676 0.0000020
## SENP-PROF      34199.915  25983.254  42416.576 0.0000000
## SELF-EMP-SAL    4559.052  -4677.935  13796.039 0.5833428
## SENP-SAL       57351.145  49266.212  65436.078 0.0000000
## SENP-SELF-EMP  52792.093  43395.892  62188.294 0.0000000
##
## $Gender
##          diff       lwr       upr     p adj
## O-M -56408.97 -62578.25 -50239.69 0.0000000
## F-M -66673.63 -94975.49 -38371.76 0.0000001
## F-O -10264.66 -38267.73  17738.41 0.6660915
##
## $`Occupation:Gender`
##                                diff         lwr        upr     p adj
## SAL:M-PROF:M             -19294.052  -44579.905   5991.801 0.3436234
## SELF-EMP:M-PROF:M         -9937.073  -32958.860  13084.714 0.9617493
## SENP:M-PROF:M             16001.539   -1675.625  33678.702 0.1210509
## PROF:O-PROF:M            -64392.833  -80797.831 -47987.835 0.0000000
## SAL:O-PROF:M             -77247.743  -93077.993 -61417.493 0.0000000
## SELF-EMP:O-PROF:M        -85427.058 -103105.599 -67748.518 0.0000000
## SENP:O-PROF:M            -37865.200  -55051.667 -20678.733 0.0000000
## PROF:F-PROF:M           -125073.083 -188883.781 -61262.386 0.0000000
## SAL:F-PROF:M             -54484.648 -114006.364   5037.067 0.1108627
## SELF-EMP:F-PROF:M        -82194.280 -179176.705  14788.146 0.1929317
## SENP:F-PROF:M                    NA          NA         NA        NA
## SELF-EMP:M-SAL:M           9356.979  -18643.034  37356.992 0.9950594
## SENP:M-SAL:M              35295.590   11494.747  59096.434 0.0000805
## PROF:O-SAL:M             -45098.782  -67970.642 -22226.921 0.0000000
## SAL:O-SAL:M              -57953.691  -80416.880 -35490.503 0.0000000
## SELF-EMP:O-SAL:M         -66133.007  -89934.873 -42331.140 0.0000000
## SENP:O-SAL:M             -18571.148  -42009.849   4867.552 0.2854352
## PROF:F-SAL:M            -105779.031 -171549.870 -40008.193 0.0000096
## SAL:F-SAL:M              -35190.596  -96809.040  26427.848 0.7801076
## SELF-EMP:F-SAL:M         -62900.228 -161183.436  35382.980 0.6288024
## SENP:F-SAL:M                     NA          NA         NA        NA
## SENP:M-SELF-EMP:M         25938.611    4558.517  47318.706 0.0042222
## PROF:O-SELF-EMP:M        -54455.761  -74796.614 -34114.907 0.0000000
## SAL:O-SELF-EMP:M         -67310.670  -87190.890 -47430.450 0.0000000
## SELF-EMP:O-SELF-EMP:M    -75489.986  -96871.219 -54108.752 0.0000000
## SENP:O-SELF-EMP:M        -27928.127  -48904.328  -6951.926 0.0008288
## PROF:F-SELF-EMP:M       -115136.011 -180070.065 -50201.956 0.0000004
## SAL:F-SELF-EMP:M         -44547.575 -105272.043  16176.892 0.4074842
## SELF-EMP:F-SELF-EMP:M    -72257.207 -169982.420  25468.006 0.3945523
## SENP:F-SELF-EMP:M                NA          NA         NA        NA
## PROF:O-SENP:M            -80394.372  -94402.575 -66386.169 0.0000000
## SAL:O-SENP:M             -93249.282 -106579.795 -79918.768 0.0000000
## SELF-EMP:O-SENP:M       -101428.597 -116908.797 -85948.397 0.0000000
## SENP:O-SENP:M            -53866.739  -68782.516 -38950.961 0.0000000
```

```
## PROF:F-SENP:M          -141074.622 -204311.564 -77837.680 0.0000000
## SAL:F-SENP:M            -70486.187 -129392.387 -11579.987 0.0052347
## SELF-EMP:F-SENP:M       -98195.818 -194801.702  -1589.935 0.0422783
## SENP:F-SENP:M                   NA          NA         NA        NA
## SAL:O-PROF:O            -12854.910  -24445.615  -1264.205 0.0152781
## SELF-EMP:O-PROF:O       -21034.225  -35044.166  -7024.284 0.0000596
## SENP:O-PROF:O            26527.633   13143.976  39911.291 0.0000000
## PROF:F-PROF:O           -60680.250 -123573.435   2212.935 0.0705829
## SAL:F-PROF:O              9908.185  -48628.831  68445.201 0.9999931
## SELF-EMP:F-PROF:O       -17801.446 -114182.661  78579.768 0.9999831
## SENP:F-PROF:O                   NA          NA         NA        NA
## SELF-EMP:O-SAL:O         -8179.315  -21511.655   5153.024 0.6901015
## SENP:O-SAL:O             39382.543   26709.930  52055.156 0.0000000
## PROF:F-SAL:O            -47825.340 -110571.062  14920.381 0.3453523
## SAL:F-SAL:O              22763.095  -35615.455  81141.645 0.9822955
## SELF-EMP:F-SAL:O         -4946.537 -101231.589  91338.516 1.0000000
## SENP:F-SAL:O                    NA          NA         NA        NA
## SENP:O-SELF-EMP:O        47561.858   32644.448  62479.268 0.0000000
## PROF:F-SELF-EMP:O       -39646.025 -102883.352  23591.302 0.6593423
## SAL:F-SELF-EMP:O         30942.410  -27964.203  89849.023 0.8608972
## SELF-EMP:F-SELF-EMP:O     3232.779  -93373.357  99838.914 1.0000000
## SENP:F-SELF-EMP:O               NA          NA         NA        NA
## PROF:F-SENP:O           -87207.883 -150309.416 -24106.351 0.0003919
## SAL:F-SENP:O            -16619.448  -75380.260  42141.363 0.9988935
## SELF-EMP:F-SENP:O       -44329.080 -140846.380  52188.220 0.9407042
## SENP:F-SENP:O                   NA          NA         NA        NA
## SAL:F-PROF:F             70588.435  -14436.640 155613.510 0.2194253
## SELF-EMP:F-PROF:F        42878.804  -71541.560 157299.168 0.9871046
## SENP:F-PROF:F                   NA          NA         NA        NA
## SELF-EMP:F-SAL:F        -27709.631 -139794.640  84375.377 0.9996908
## SENP:F-SAL:F                    NA          NA         NA        NA
## SENP:F-SELF-EMP:F               NA          NA         NA        NA
```

Interpretation:

From the above Tukey test the difference between and within is clearly seen. All the variables with p value <0.05 shows a significant difference between and within group. As identified in one way ANOVA test, the Self-employed and salaries person does not have significant difference. Also, the Firms are not significant with Males. When the two groups combined, it is observed that Salaried female has no significant difference with Professional Females and self-employed female. Similarly, Self-employed and professional females also has no difference and SENP and Professional females also have no difference. Salaried Firm and self-employed firm also has no significant difference with Professional females. The same has been plotted in graph below.

```
plot(TukeyHSD(aov2))
```



2.6 Two Way ANOVA Summary:

The two way ANOVA test of gender and occupation with average account balance shows significant difference between them. But the normality assumption and homogenous in variance assumption were violated. In which, Kurskal wallis test was conducted which also showed the significant difference among Gender and average account balance. Post hoc test was conducted to identify the between group and within group difference.

3 Regression Analysis

3.1 Objective

The House Hold Expenditure data set has information on monthly expenditure, annual income of the household, monthly income, household size (number of members in the household), and monthly EMI.

Set up a regression model to explain expenditure using the other variables in the data.

3.2 Steps to Follow

We shall perform Regression Analysis in the following sequence:

1. Data Loading and Descriptive Statistics
2. Visualization
3. Running the Regression and Interpretation
4. Testing of Assumptions
    - Mean of the Residuals is zero
    - Homoscedasticity of Residuals
    - Correlation
5. Data Transformation (Optional)
6. Regression using Logarithmic Terms (Optional)
7. Multivariate Regression
8. Testing of Assumptions
    - Mean of the Residuals is zero
    - Homoscedasticity of Residuals
    - Correlation
9. Robust Regression (Optional)
10. Parsimony

3.3 Descriptive Statistics

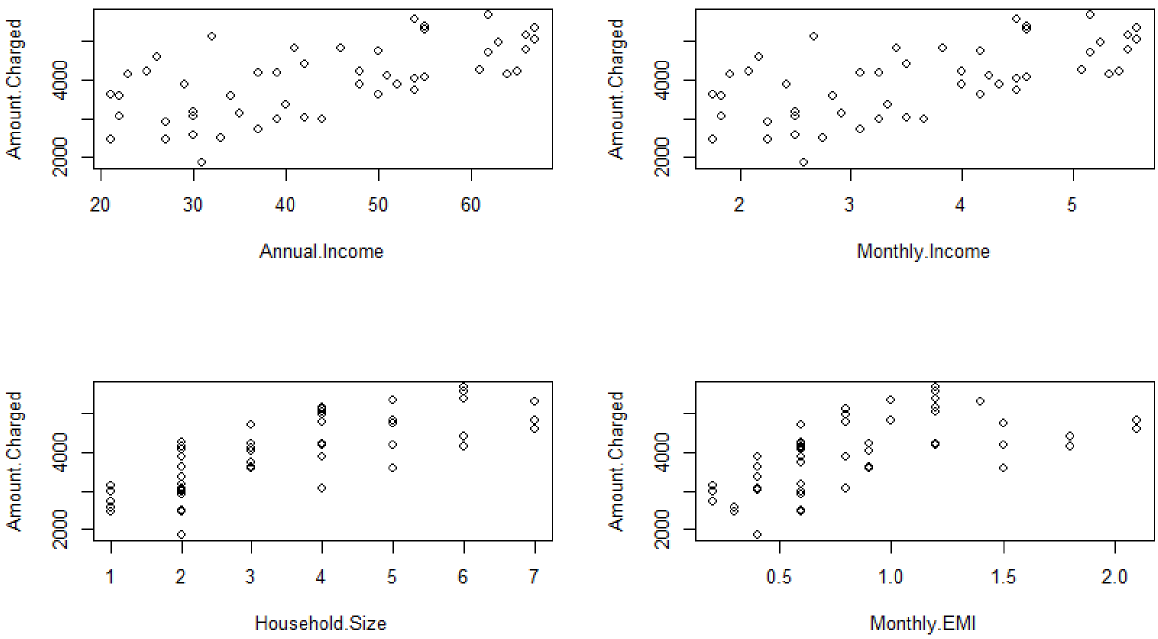
The outcome of Basic descriptive statistics on Household expenditure Dataset is as follows:

```
# Find out Names of the Columns (Features)
names(Household_Data)

## [1] "Annual.Income"  "Monthly.Income" "Household.Size" "Amount.Charged"
## [5] "Monthly.EMI"

# Find out Class of each Feature, along with internal structure
str(Household_Data)

## 'data.frame':    50 obs. of  5 variables:
##  $ Annual.Income : int  54 30 32 50 31 55 37 40 66 51 ...
##  $ Monthly.Income: num  4.5 2.5 2.67 4.17 2.58 ...
##  $ Household.Size: int  3 2 4 5 2 2 1 2 4 3 ...
```

```
##  $ Amount.Charged: int  4016 3159 5100 4742 1864 4070 2731 3348 4764
##  $ Monthly.EMI   : num  0.9 0.6 0.8 1.5 0.4 0.6 0.2 0.4 0.8 0.6 ...

# Provide Summary of a Dataset.
summary(Household_Data)

##  Annual.Income   Monthly.Income   Household.Size Amount.Charged
##  Min.   :21.00   Min.   :1.750   Min.   :1.00   Min.   :1864
##  1st Qu.:30.25   1st Qu.:2.521   1st Qu.:2.00   1st Qu.:3130
##  Median :42.00   Median :3.500   Median :3.00   Median :4090
##  Mean   :43.48   Mean   :3.623   Mean   :3.42   Mean   :3964
##  3rd Qu.:54.75   3rd Qu.:4.562   3rd Qu.:4.75   3rd Qu.:4733
##  Max.   :67.00   Max.   :5.583   Max.   :7.00   Max.   :5678
##   Monthly.EMI
##  Min.   :0.200
##  1st Qu.:0.600
##  Median :0.800
##  Mean   :0.862
##  3rd Qu.:1.200
##  Max.   :2.100
```

3.4 Data Visualization

The correlation of each variable against expenditure (Amount Charged) were plotted and is shown in the following graph:



Interpretation:

A linearity is somewhat seen between Expenditure (Amount Charged) and other variables.

3.5 Running the Regression Analysis

We perform Regression Analysis on one variable at a time, and then keep on adding variables.

```
# Building Regression equation for one variable at a time.
#=========================================================================
# Annual Income
resultAI<-lm(formula = Amount.Charged~Annual.Income)
# Monthly Income
resultMI<-lm(formula = Amount.Charged~Monthly.Income)
# Household Size
resultHS<-lm(formula = Amount.Charged~Household.Size)
# Monthly EMI
resultME<-lm(formula = Amount.Charged~Monthly.EMI)
# Building Regression equation for Multiple variables.
#=========================================================================
# Annual Income + Monthly Income
resultAIMI<-lm(formula =
Amount.Charged~Annual.Income+Monthly.Income,data=Household_Data)
# Annual Income + Household Size
resultAIHS<-lm(formula = Amount.Charged~Annual.Income+Household.Size)
# Annual Income + Monthly EMI
resultAIME<-lm(formula = Amount.Charged~Annual.Income+Monthly.EMI)
# Monthly Income + Household Size
resultMIHS<-lm(formula = Amount.Charged~Monthly.Income+Household.Size)
# Monthly Income + Monthly EMI
resultMIME<-lm(formula = Amount.Charged~Monthly.Income+Monthly.EMI)
# Household Size + Monthly EMI
resultHSME<-lm(formula = Amount.Charged~Household.Size+Monthly.EMI)
# Annual Income + Household Size + Monthly EMI
resultAIHSME<-lm(formula =
Amount.Charged~Annual.Income+Household.Size+Monthly.EMI)
# Annual Income + Monthly Income + Household Size + Monthly EMI
resultAIMIHSME<-lm(formula =
Amount.Charged~Annual.Income+Monthly.Income+Household.Size+Monthly.EMI)
```

Summary of the Regression outcome for various variables is as follows:

Independent Variable	Intercept	Estimated Standard	P Value	F Stats	R Squared	Adj. R
Annual Income	2204	40.48	9.01e-07 ***	9.01E-07	0.3981	0.3856
Monthly Income	2204	485.76	9.01e-07 ***	9.01E-07	0.3981	0.3856
Household Size	2581.9	404.1	2.86e-10 ***	2.87E-10	0.5668	0.5577
Monthly EMI	2949.5	1177	3.11e-06 ***	3.11E-06	0.3672	0.354
Annual Income + Monthly Inc.	2204	40.480 + NA	9.01e-07 *** NA	9.01E-07	0.3981	0.3856

Independent Variable	Intercept	Estimated Standard	P Value	F Stats	R Squared	Adj. R
Annual Income + H Hold Size.	1304.91	33.133 356.296	7.68e-11 *** 3.12e-14 ***	2.20E-16	0.8256	0.8181
Annual Income + Monthly EMI	1259.04	39.495 1145.897	7.17e-11 *** 2.37e-10 ***	1.04E-14	0.7459	0.7351
Monthly Inc + H Hold Size.	1304.91	397.6 356.3	7.68e-11 *** 3.12e-14 ***	2.20E-16	0.8256	0.8181
Monthly Inc. + Monthly EMI	1259.04	473.94 1145.9	7.17e-11 *** 2.37e-10 ***	1.04E-14	0.7459	0.7351
H Hold Size + Monthly EMI	2578.5	657.8 -1002.4	2.11e-06 *** 0.0274 *	2.48E-10	0.6098	0.5932
Annual Income + H Hold Size + Monthly EMI	1339.84	32.208 409.73 -205.876	7.17e-11 *** 2.37e-10 ***	1.04E-14	0.7459	0.7351
Annual Inc. + Monthly Inc + H Hold Size + Monthly EMI	1339.84	32.208 NA 409.73 -205.876	1.13e-09 *** NA 2.82e-05 *** 0.516	2.20E-16	0.8272	0.8159

<u>Interpretation:</u>

Single Variable Regression Analysis:

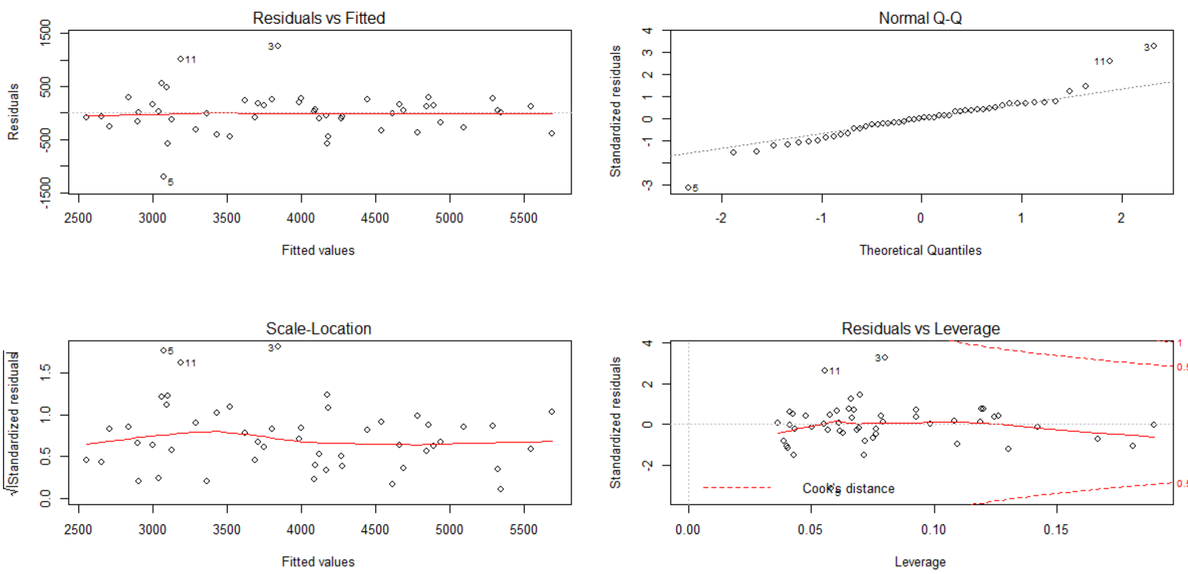
- All Variables are significant and positively influencing the Expenditure.
- However, R Square suggests that Household Size is having maximum impact, and is able to explain 55.77 % Variance in the Expenditure.

Multi Variate Regression Analysis:

- Annual Income + Household Size combined explains close to 82% variance in the Expenditure.
- Monthly Income + Household Size combined explains close to 82% variance in the Expenditure.

3.6 Testing of Assumptions

```
# Testing of Assumptions (Mean of Residuals)
#==========================================================================
mean(resultAIMIHSME$residuals)

## [1] 2.628453e-15

par(mfrow=c(2,2))
# Testing of homoscedasticity
plot(resultAIMIHSME)
```




```
# Testing the Correlation between Errors and Explanatory Variables
cor.test(Annual.Income,resultAIMIHSME$residuals)

##
##  Pearson's product-moment correlation
##
## data:  Annual.Income and resultAIMIHSME$residuals
## t = -1.8627e-16, df = 48, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2783477  0.2783477
## sample estimates:
##           cor
## -2.688527e-17
```

Interpretation:

- The assumption of **mean of Residuals is zero holds True**.
- The Residuals plot confirms **normal linear regression** model.
- The Q-Q Plot confirms that our dependent variable is **normally distributed**.
- The Scale Location graph confirms **homoscedasticity** in the data.

- The Residuals Vs Leverage graphs shows some extreme Cooks distance lines, which suggests **impactful outliers present** in the data.
- This **recommends** to go for **Robust Regression** Model.
- Correlation between Errors and Explanatory Variables:
    - P Value is less than 0.05, which suggests that Errors and **Explanatory Variables are correlated**.

3.7 Robust Regression

```
# Robust Regression
#==========================================================================
result2 <- rlm(Amount.Charged~Annual.Income+Monthly.EMI+Household.Size)
aov(result2)

## Call:
##    aov(formula = result2)
##
## Terms:
##                  Annual.Income Monthly.EMI Household.Size Residuals
## Sum of Squares        16999745    14851119        3468789   7379496
## Deg. of Freedom              1           1              1        46
##
## Residual standard error: 400.5294
## Estimated effects may be unbalanced

#
summary(result2)

##
## Call: rlm(formula = Amount.Charged ~ Annual.Income + Monthly.EMI +
##     Household.Size)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1178.41  -184.38    28.73   206.42  1315.42
##
## Coefficients:
##                 Value     Std. Error t value
## (Intercept)    1267.3989  159.3779     7.9522
## Annual.Income    34.4282    3.2778    10.5034
## Monthly.EMI    -146.2721  243.1861    -0.6015
## Household.Size  383.1238   68.2088     5.6169
##
## Residual standard error: 302.6 on 46 degrees of freedom
```

Interpretation:

We see that there is not much change in the equation while doing robust estimation. So our original model remains sound. (Monthly Income + Household Size).

3.8 Parsimony Model:

After performing the Multi Variate Regression Analysis and confirming that Monthly Income + Household Size are the variables which are having maximum impact on Expenditure, we perform Parsimony Model, which would help us in confirming a simplified approach for our Regression Model.

The Parsimony Model is performed first with Forward Selection, then Backward Selection and then using both directions.

```
# Parsimony
#
library(leaps)

## Warning: package 'leaps' was built under R version 3.3.3

Null<-lm(Amount.Charged~1)
Full<-
lm(Amount.Charged~Annual.Income+Monthly.Income+Monthly.EMI+Household.Size)
step(Null,scope = list(lower=Null, upper=Full),direction = "forward")

## Step:  AIC=601.57
## Amount.Charged ~ Household.Size + Monthly.Income
##
##                  Df Sum of Sq     RSS    AIC
## <none>                        7448393 601.57
## + Monthly.EMI  1     68897 7379496 603.11

##
## Call:
## lm(formula = Amount.Charged ~ Household.Size + Monthly.Income)
##
## Coefficients:
##    (Intercept)  Household.Size  Monthly.Income
##         1304.9           356.3           397.6

# Parsimony Backward Selection
step(Full, direction = "backward")

##
## Step:  AIC=601.57
## Amount.Charged ~ Annual.Income + Household.Size
##
##                     Df Sum of Sq      RSS    AIC
## <none>                            7448393 601.57
## - Annual.Income   1  11050038 18498431 645.06
## - Household.Size  1  18251011 25699404 661.50

##
## Call:
## lm(formula = Amount.Charged ~ Annual.Income + Household.Size)
##
## Coefficients:
##    (Intercept)   Annual.Income  Household.Size
##        1304.90           33.13          356.30

# Parsimony Stepwise both forward and Backward Selection
step(Null,scope = list(upper=Full),data=Household_Data,direction="both")
```

```
##
## Step:  AIC=601.57
## Amount.Charged ~ Household.Size + Monthly.Income
##
##                  Df Sum of Sq      RSS    AIC
## <none>                         7448393 601.57
## + Monthly.EMI     1     68897  7379496 603.11
## - Monthly.Income  1  11050038 18498431 645.06
## - Household.Size  1  18251011 25699404 661.50

##
## Call:
## lm(formula = Amount.Charged ~ Household.Size + Monthly.Income)
##
## Coefficients:
##    (Intercept)  Household.Size  Monthly.Income
##         1304.9           356.3           397.6
```

Interpretation:

The parsimony analysis in forward, backward and both model confirm to retain **household size** and **monthly (or annual) income** are the regressors which can be included in the model.

3.9 Conclusion:

Following Regression Model can be set up to explain Expenditure using Household Size and Monthly Income.

Expenditure = 1304.9(Intercept) + 356.3*Household Size + 397.6* Monthly Income.

OR

Expenditure = 1304.9(Intercept) + 356.3*Household Size + 33.13* Annual Income.

4 Principal Component Analysis

4.1 Objective

Run Principal Component Analysis on the given Batting and Bowling data sets and perform following activities:

- Interpret loadings
- Interpret Communality
- Number of components to be retained
- Total variance extracted
- Check whether rotation is necessary
- Label the components
- Use the PC Scores to rank the players

4.2 Steps to follow

We shall perform Principal Component Analysis on Batting and Bowling Datasets in the following manner:

- Descriptive Statistics
- Generate Correlation Matrix
- Check if Dimensionality Reduction is possible or not.
- Perform Principal Component Analysis using Singular Value Decomposition.
- Interpret the Results

4.3 PCA on Batting Dataset

4.3.1 Descriptive Statistics

```
# Read Input file
batting<-read.csv("batting_bowling_ipl_bat.csv", header=TRUE)

str(batting)

## 'data.frame':    180 obs. of  7 variables:
##  $ Name : Factor w/ 91 levels "","A Ashish Reddy",..: 1 14 1 27 1 86 1 
15 1 72 ...
##  $ Runs : int  NA 733 NA 590 NA 495 NA 479 NA 569 ...
##  $ Ave  : num  NA 61.1 NA 36.9 NA ...
##  $ SR   : num  NA 161 NA 144 NA ...
##  $ Fours: int  NA 46 NA 64 NA 57 NA 41 NA 58 ...
##  $ Sixes: int  NA 59 NA 17 NA 19 NA 20 NA 18 ...
##  $ HF   : int  NA 9 NA 6 NA 5 NA 5 NA 5 ...

pcabat <- na.omit(batting[2:7]) # Omit NA records
summary(pcabat)

##       Runs             Ave               SR              Fours      
##  Min.   :  2.0   Min.   : 0.50   Min.   : 18.18   Min.   : 0.00  
##  1st Qu.: 98.0   1st Qu.:14.66   1st Qu.:108.75   1st Qu.: 6.25  
##  Median :196.5   Median :24.44   Median :120.14   Median :16.00  
##  Mean   :219.9   Mean   :24.73   Mean   :119.16   Mean   :19.79  
```

```
##   3rd Qu.:330.8   3rd Qu.:32.20   3rd Qu.:132.00   3rd Qu.:28.00
##   Max.   :733.0   Max.   :81.33   Max.   :164.10   Max.   :73.00
##       Sixes              HF
##   Min.   : 0.000   Min.   :0.000
##   1st Qu.: 3.000   1st Qu.:0.000
##   Median : 6.000   Median :0.500
##   Mean   : 7.578   Mean   :1.189
##   3rd Qu.:10.000   3rd Qu.:2.000
##   Max.   :59.000   Max.   :9.000
```

4.3.2 Generate Correlation Matrix

```
# Understanding Correlation
batcorr <- cor(pcabat)
batcorr

##            Runs       Ave        SR     Fours     Sixes        HF
## Runs  1.0000000 0.6929845 0.4934887 0.9188086 0.7697776 0.8351477
## Ave   0.6929845 1.0000000 0.6236059 0.5462114 0.6824143 0.6207537
## SR    0.4934887 0.6236059 1.0000000 0.3848104 0.5839428 0.4275835
## Fours 0.9188086 0.5462114 0.3848104 1.0000000 0.5225736 0.7836888
## Sixes 0.7697776 0.6824143 0.5839428 0.5225736 1.0000000 0.7676964
## HF    0.8351477 0.6207537 0.4275835 0.7836888 0.7676964 1.0000000
```

Interpretation:

The above Correlation matrix shows strong correlation between the variables having correlation coefficient > 0.5. (Highlighted in Yellow)

4.3.3 Dimensionality Reduction Check

We shall perform Barlett Sphericity Test to check if Dimension Reduction is possible or not.

```
# Barlett Sphericity Test for checking the possibility
# of data dimension reduction
#
print(cortest.bartlett(batcorr,nrow(pcabat)))

## $chisq
## [1] 572.3093
##
## $p.value
## [1] 2.693573e-112
##
## $df
## [1] 15
```

Interpretation:

P Value is less than 0.05, hence we reject the Null Hypothesis and confirm that Dimensionality Reduction is possible.

4.3.4 Principal Component Analysis

```
# Finding out the Eigen Values and Eigen Vectors

A<-eigen(batcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

## [1] 4.25471977 0.82707395 0.41202798 0.32546749 0.16383742 0.01687338
```

Inference:

- PC1 has extracted 4.25 units of variances from the 6 variables.
- i.e. PC explains 4.25/6 = 70% of variance
- PC2 has extracted 0.82 units of variance from the 6 variables = 13% of variance. Rest All can be ignored.
- We will consider only PC1 as it is > 1 unit

```
eigenvectors

##            [,1]       [,2]         [,3]         [,4]        [,5]
## [1,] -0.4582608  0.2664321  0.109779419  0.005201415 -0.45840889
## [2,] -0.3979731 -0.3311176 -0.005504861 -0.847363074  0.10122837
## [3,] -0.3253838 -0.6978033  0.450134482  0.432750288  0.11890348
## [4,] -0.4057417  0.4735580  0.508235378  0.032523046 -0.09676885
## [5,] -0.4173346 -0.1790246 -0.669425885  0.248781566 -0.39458014
## [6,] -0.4323718  0.2759323 -0.280825406  0.178117767  0.77486668
##             [,6]
## [1,]  0.70483594
## [2,] -0.06063730
## [3,]  0.05624934
## [4,] -0.58514214
## [5,] -0.35786211
## [6,]  0.16096217
```

Inference:

- Since we are considering only PC1, the eigen vectors for PC1 are highlighted above. These values shall be used while scoring the players.

```
#
# Getting the Loadings and Cummunality

pc1<-principal(pcabat,nfactors = length(pcabat),rotate="none")
pc1

## Principal Components Analysis
## Call: principal(r = pcabat, nfactors = length(pcabat), rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##        PC1   PC2   PC3   PC4   PC5   PC6 h2       u2 com
## Runs  0.95 -0.24  0.07  0.00 -0.19 -0.09  1 -2.2e-16 1.2
## Ave   0.82  0.30  0.00 -0.48  0.04  0.01  1  6.7e-16 1.9
## SR    0.67  0.63  0.29  0.25  0.05 -0.01  1  1.0e-15 2.7
## Fours 0.84 -0.43  0.33  0.02 -0.04  0.08  1  1.0e-15 1.9
```

```
## Sixes 0.86  0.16 -0.43  0.14 -0.16  0.05  1  7.8e-16 1.7
## HF   0.89 -0.25 -0.18  0.10  0.31 -0.02  1  1.1e-15 1.5
##
##                        PC1  PC2  PC3  PC4  PC5  PC6
## SS loadings           4.25 0.83 0.41 0.33 0.16 0.02
## Proportion Var        0.71 0.14 0.07 0.05 0.03 0.00
## Cumulative Var        0.71 0.85 0.92 0.97 1.00 1.00
## Proportion Explained  0.71 0.14 0.07 0.05 0.03 0.00
## Cumulative Proportion 0.71 0.85 0.92 0.97 1.00 1.00
##
```

Inference:

- SS Loadings for each Principal Component can be calculated as
- The sum of (square of each parameter).
- PC1 Loading = 0.95^2+0.82^2+0.67^2+0.84^2+0.86^2+0.89^2 = **4.25**

Communality Interpretation:

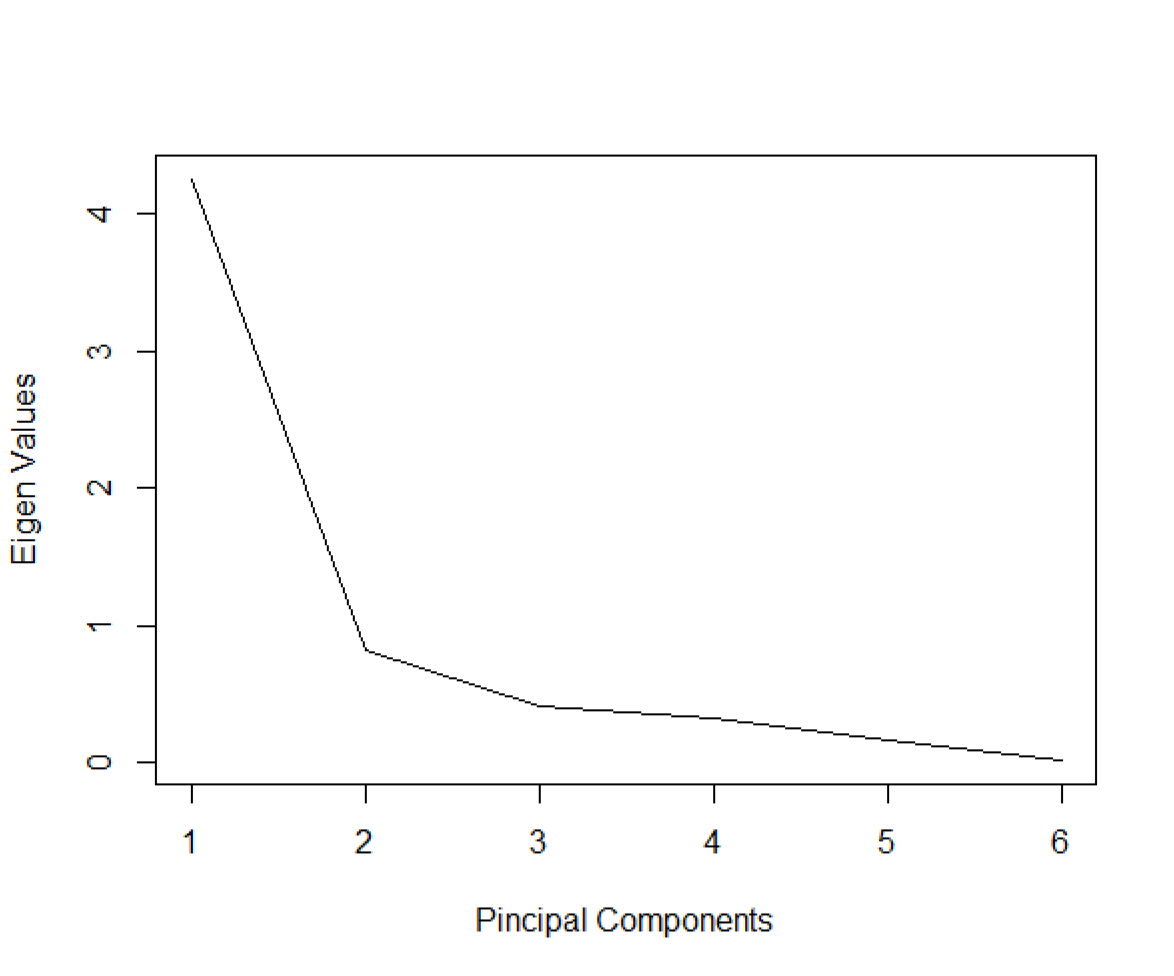
Attribute	PC1	PC1 Variance (Sq of PC1)	Remark
Runs	0.95	0.9	0.90 <= Communality between PC1 and Runs
Average	0.82	0.67	0.67 <= Communality between PC1 and Average
Strike Rate	0.67	0.45	0.45 <= Communality between PC1 and St. Rate
Fours	0.84	0.71	0.71 <= Communality between PC1 and Fours
Sixes	0.86	0.74	0.74 <= Communality between PC1 and Sixes
Half Century	0.89	0.79	0.79 <= Communality between PC1 and H Century
Total Variance:		**4.26**	

```
# Interpreting the variance
#
part.pca<-eigenvalues/sum(eigenvalues)*100
part.pca

## [1] 70.9119961 13.7845659  6.8671330  5.4244582  2.7306237  0.2812231
```

Inference:

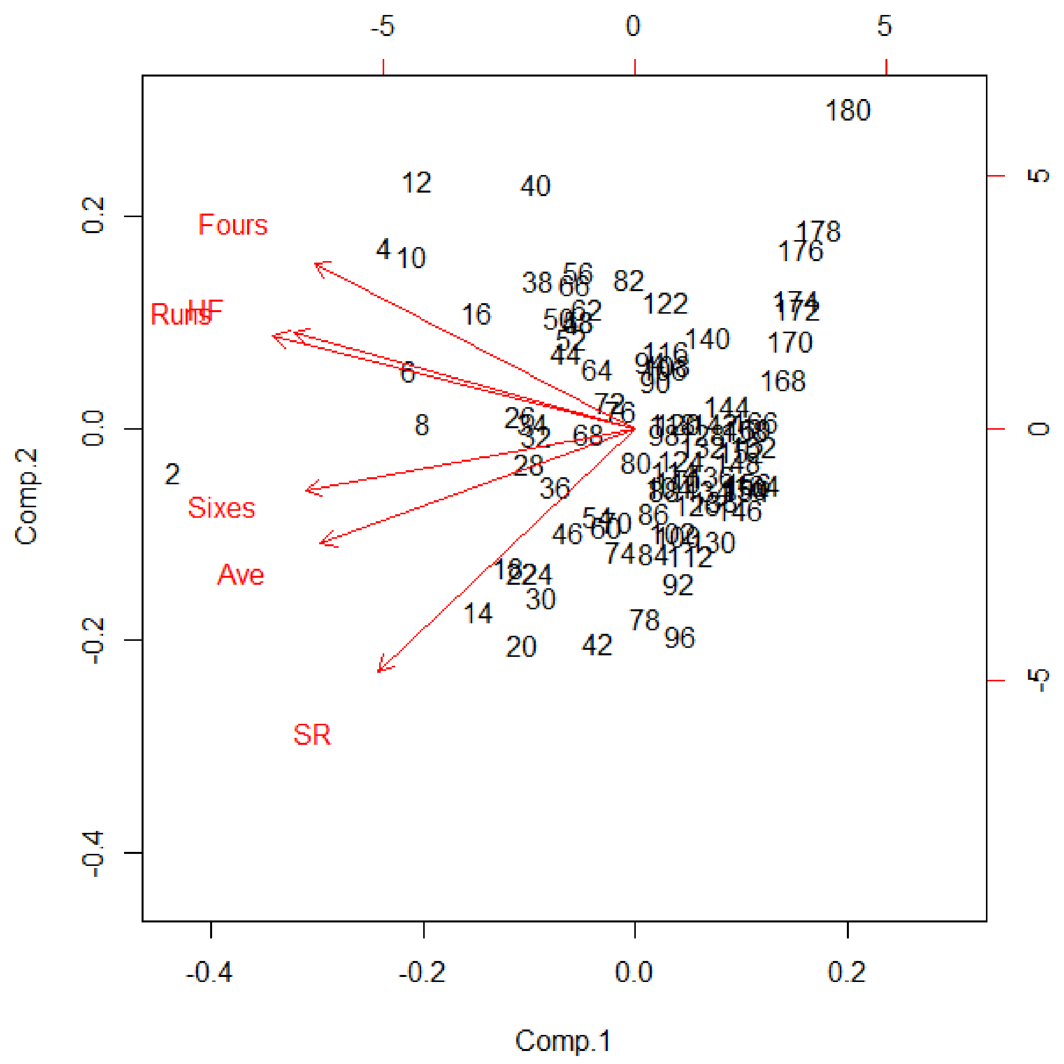
- PC1 is extracting 71% of variance while
- PC2 is extracting 14% of variance and so on….

```
#Plotting Scree Graphs
plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")
```




```
# Principal Components Scoring and Perceptual Map

pcabatsc<-scale(pcabat)
z<-as.matrix(pcabatsc%*%eigenvectors)
z

pc.cr<-princomp(pcabatsc,cor=TRUE)
summary(pc.cr)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     2.062697 0.9094361 0.64189406 0.57049758
0.40476836
## Proportion of Variance 0.709120 0.1378457 0.06867133 0.05424458
0.02730624
## Cumulative Proportion  0.709120 0.8469656 0.91563695 0.96988153
0.99718777
##                             Comp.6
## Standard deviation     0.129897588
```

```
## Proportion of Variance 0.002812231
## Cumulative Proportion  1.000000000

biplot(pc.cr)
```



4.3.5 Top 10 Batsmen Ranking

Using the Eigen vectors we calculated in previous step, and ranking them in descending order, we come up with top 10 Batsmen as follows:

Sr. No.	Name	PC1	Runs	Ave	SR	Fours	Six	HF
1	CH Gayle	8.469	733	61.08	160.7	46	59	9
2	G Gambhir	4.593	590	36.87	143.6	64	17	6
3	V Sehwag	4.119	495	33	161.2	57	19	5
4	S Dhawan	4.097	569	40.64	129.6	58	18	5

Sr. No.	Name	PC1	Runs	Ave	SR	Fours	Six	HF
5	AMRahane	4.002	560	40	129.3	73	10	5
6	CLWhite	3.878	479	43.54	149.7	41	20	5
7	RG Sharma	2.903	433	30.92	126.6	39	18	5
8	KP Pietersen	2.863	305	61	147.3	22	20	3
9	AB de Villiers	2.314	319	39.87	161.1	26	15	3
10	F du Plessis	2.113	398	33.16	130.9	29	17	3

4.4 PCA on Bowling Dataset

4.4.1 Descriptive Statistics

```
# Read Input file
bowling<-read.csv("batting_bowling_ipl_bowl.csv", header=TRUE)
# Descriptive Statistics
str(bowling)

## 'data.frame':    166 obs. of  5 variables:
##  $ Name: Factor w/ 84 levels "","A Ashish Reddy",..: 1 59 1 53 1 46 1
79 1 84 ...
##  $ Wkts: int  NA 14 NA 9 NA 25 NA 19 NA 17 ...
##  $ Ave : num  NA 30.8 NA 48.2 NA ...
##  $ Econ: num  NA 6.54 NA 6.88 NA 7.19 NA 7.42 NA 7.55 ...
##  $ SR  : num  NA 28.2 NA 42 NA 15.1 NA 19.2 NA 21.1 ...

pcabowl <- na.omit(bowling[2:5]) # Omit NA records
summary(pcabowl)

##       Wkts            Ave              Econ              SR
##  Min.   : 1.00   Min.   : 12.20   Min.   : 5.400   Min.   :12.00
##  1st Qu.: 5.00   1st Qu.: 22.32   1st Qu.: 6.950   1st Qu.:17.25
##  Median : 8.00   Median : 29.00   Median : 7.530   Median :21.60
##  Mean   : 8.88   Mean   : 34.51   Mean   : 7.656   Mean   :26.33
##  3rd Qu.:12.50   3rd Qu.: 36.44   3rd Qu.: 8.280   3rd Qu.:28.90
##  Max.   :25.00   Max.   :161.00   Max.   :11.650   Max.   :96.00
```

4.4.2 Generate Correlation Matrix

```
# Understanding Correlation
bowlcorr <- cor(pcabowl)
bowlcorr

##            Wkts        Ave       Econ         SR
## Wkts  1.0000000 -0.4905337 -0.2924540 -0.5123438
## Ave  -0.4905337  1.0000000  0.5226172  0.9630984
## Econ -0.2924540  0.5226172  1.0000000  0.3277374
## SR   -0.5123438  0.9630984  0.3277374  1.0000000
```

Interpretation:

The above Correlation matrix shows strong correlation between the variables having correlation coefficient > 0.5. (Highlighted in Yellow)

4.4.3 Dimensionality Reduction Check

We shall perform Barlett Sphericity Test to check if Dimension Reduction is possible or not.

```
#
# Barlett Sphericity Test for checking the possibility
# of data dimension reduction
#
print(cortest.bartlett(bowlcorr,nrow(pcabowl)))

## $chisq
## [1] 336.2771
##
## $p.value
## [1] 1.36091e-69
##
## $df
## [1] 6
```

Interpretation:

P Value is less than 0.05, hence we reject the Null Hypothesis and confirm that Dimensionality Reduction is possible.

4.4.4 Principal Component Analysis

```
#
# Finding out the Eigen Values and Eigen Vectors

A<-eigen(bowlcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

## [1] 2.61606918 0.75160217 0.62018101 0.01214765
```

Inference:

- PC1 has extracted 2.61 units of variances from the 4 variables.
- i.e. PC explains 2.61/4 = 65% of variance
- PC2 has extracted 0.75 units of variance from the 4 variables = 19% of variance. Rest All can be ignored.
- We will consider only PC1 as it is > 1 unit

```
eigenvectors

##            [,1]        [,2]       [,3]        [,4]
## [1,]  0.4282076 -0.33487615  0.8384720  0.03822333
## [2,] -0.5911683  0.04764188  0.3539052 -0.72318835
```

```
## [3,] -0.3834154 -0.89162604 -0.1681540  0.17239454
## [4,] -0.5658188  0.30098375  0.3787349  0.66769582
```

Inference:

- Since we are considering only PC1, the eigen vectors for PC1 are highlighted above. These values shall be used while scoring the players.

```
#
# Getting the loadings and Cummunality

pc1<-principal(pcabowl,nfactors = length(pcabowl),rotate="none")
pc1

## Principal Components Analysis
## Call: principal(r = pcabowl, nfactors = length(pcabowl), rotate =
"none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##        PC1   PC2   PC3   PC4 h2      u2 com
## Wkts -0.69  0.29  0.66  0.00  1 3.3e-16 2.3
## Ave   0.96 -0.04  0.28 -0.08  1 1.6e-15 1.2
## Econ  0.62  0.77 -0.13  0.02  1 3.3e-16 2.0
## SR    0.92 -0.26  0.30  0.07  1 7.8e-16 1.4
##
##                        PC1  PC2  PC3  PC4
## SS loadings           2.62 0.75 0.62 0.01
## Proportion Var        0.65 0.19 0.16 0.00
## Cumulative Var        0.65 0.84 1.00 1.00
## Proportion Explained  0.65 0.19 0.16 0.00
## Cumulative Proportion 0.65 0.84 1.00 1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <   NA
##
## Fit based upon off diagonal values = 1
```

Inference:

- SS Loadings for each Principal Component can be calculated as
- The sum of (square of each parameter).
- PC1 Loading = -0.69^2+0.96^2+0.62^2+0.92^2= **2.62**
- **Communality Interpretation:**

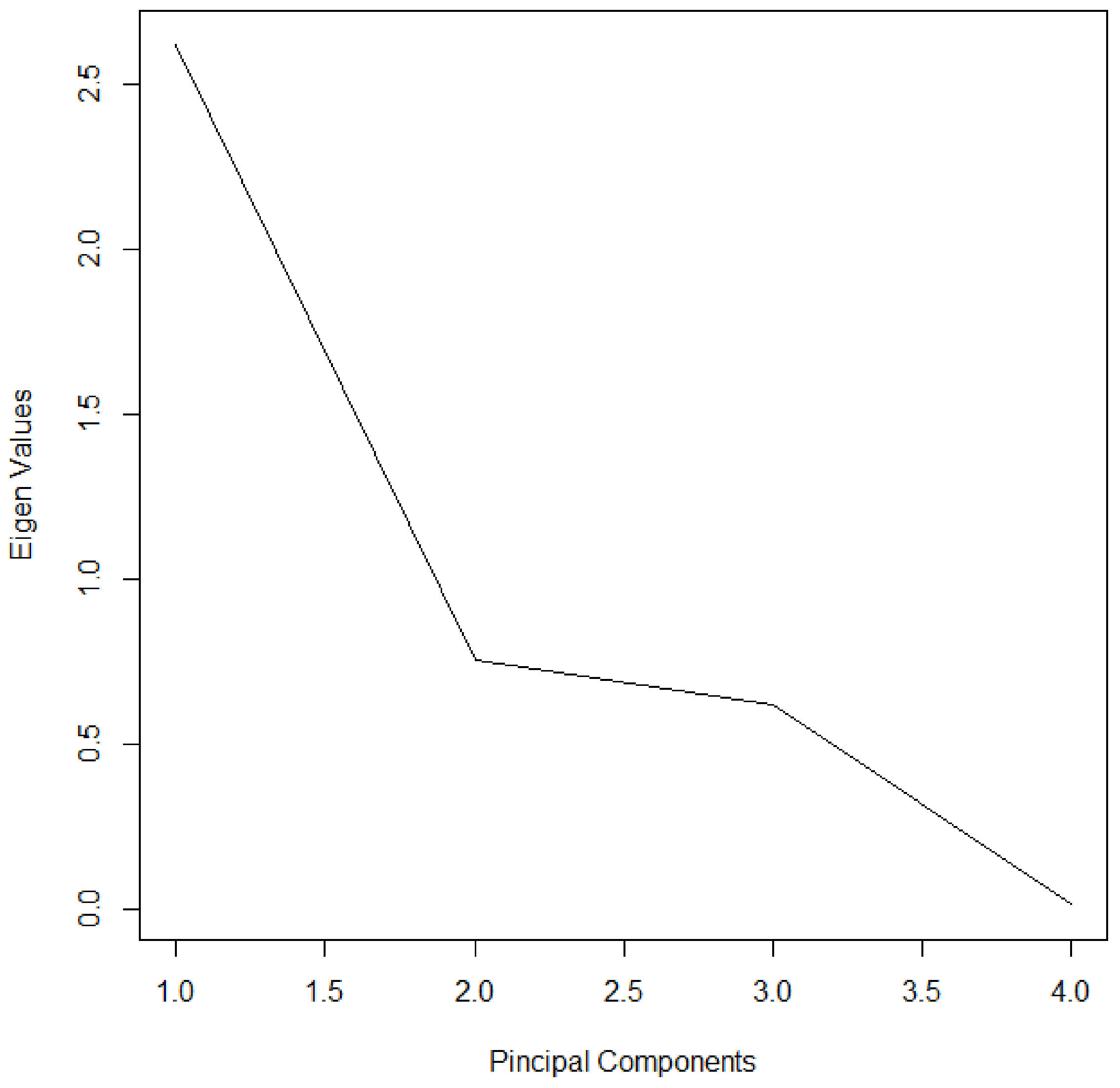
Attribute	PC1	PC1 Variance (Sq of PC1)	Remark
Wickets	-0.69	0.48	0.48 <= Communality between PC1 and wickets
Average	0.96	0.92	0.92 <= Communality between PC1 and Average
Economy	0.63	0.38	0.38 <= Communality between PC1 and Economy
Strike Rate	0.92	0.84	0.84 <= Communality between PC1 and St. Rate
Total Variance:		**2.62**	

```
# Interpreting the variance
#
part.pca<-eigenvalues/sum(eigenvalues)*100
part.pca

## [1] 65.4017296 18.7900541 15.5045251  0.3036911
```

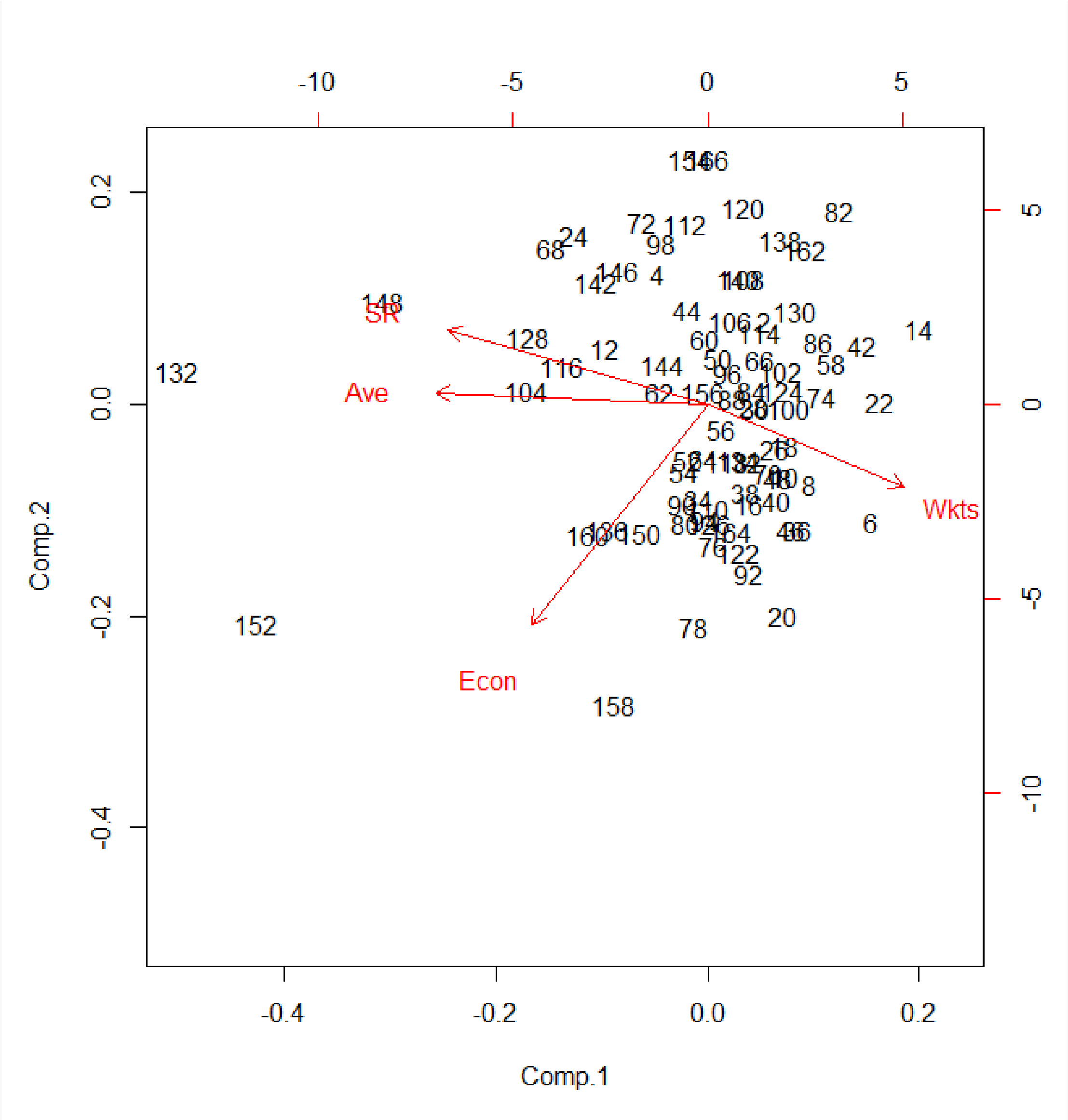
Inference:

- PC1 is extracting 65% of variance while
- PC2 is extracting 19% of variance and so on….

```
#
#Plotting Scree Graphs


plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")
```




```
# Principal Components Scoring and Perceptual Map

pcabowlsc<-scale(pcabowl)
z<-as.matrix(pcabowlsc%*%eigenvectors)
z

pc.cr<-princomp(pcabowlsc,cor=TRUE)
summary(pc.cr)

## Importance of components:
##                           Comp.1    Comp.2    Comp.3      Comp.4
## Standard deviation     1.6174267 0.8669499 0.7875157 0.110216359
## Proportion of Variance 0.6540173 0.1879005 0.1550453 0.003036911
## Cumulative Proportion  0.6540173 0.8419178 0.9969631 1.000000000

biplot(pc.cr)
```



4.4.5 Top 10 Bowler Ranking

Using the Eigen vectors we calculated in previous step, and ranking them in descending order, we get the top 10 Bowlers as follows:

Ranking	Name	PC1	Wkts	Ave	Econ	SR
1	SP Narine	2.918	24	13.5	5.47	14.7
2	SLMalinga	2.399	22	15.9	6.3	15.1
3	MMorkel	2.268	25	18.12	7.19	15.1
4	DWSteyn	2.142	18	15.83	6.1	15.5
5	L Balaji	1.836	11	14.72	5.4	16.3
6	MMuralitharan	1.713	15	17.33	6.5	16
7	BWHilfenhaus	1.589	14	16.64	6.85	14.5
8	Shakib Al Hasan	1.545	12	16.25	6.5	15
9	UT Yadav	1.417	19	23.84	7.42	19.2
10	AB McDonald	1.357	5	12.2	6.1	12

5 Factor Analysis

5.1 Objective

- The raw data are available in the file labeled mbacar.
- Conduct a common factor analysis on the data set. How many factors you would retain? How do you interpret them?
- Save the factor scores and plot the average factor scores against each other for each of the 10 cars evaluated by the students.
- What do the plots tell you about the similarities of the 10 car models?

5.2 Steps to follow

We shall perform Factor Analysis on the given MBACar Database in the following manner:

1. Data Loading and Descriptive Statistics.
2. Verify test assumptions of Factor Analysis such as Factorability.
3. Conduct exploratory Factor Analysis (EFA)
4. Determine the number of Factors
5. Label the factors.
6. Examine correlations among factors if we use oblique rotations. (Optional)
7. Summarize Factor Scores.

5.3 Descriptive Statistics

```
# Load Data
# Read Input file
MBACar=read.csv("MBAcar_Datafile.csv", header = TRUE)
attach(MBACar)
dim(MBACar)

## [1] 303  19

# Find out Names of the Columns (Features)
names(MBACar)

##  [1] "student"     "Car"         "exciting"    "dependable"
"luxurious"
##  [6] "outdoorsy"   "powerful"    "stylish"     "comfortable" "rugged"
## [11] "fun"         "safe"        "performance" "family"
"versatile"
## [16] "sports"      "status"      "practical"   "discipline"

# Find out Class of each Feature, along with internal structure
str(MBACar)

## 'data.frame':    303 obs. of  19 variables:
##  $ student     : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010
##  $ Car         : int  5 6 7 8 9 10 1 2 3 4 ...
##  $ exciting    : int  3 2 3 3 5 2 3 2 2 5 ...
```

```
## $ dependable : int  5 3 5 2 4 4 2 3 5 3 ...
## $ luxurious  : int  5 2 4 3 4 3 3 3 4 2 ...
## $ outdoorsy  : int  1 1 1 1 2 3 1 5 1 5 ...
## $ powerful   : int  3 2 4 2 3 2 3 3 3 4 ...
## $ stylish    : int  4 1 4 3 5 2 4 4 3 4 ...
## $ comfortable: int  5 2 4 3 3 4 2 3 4 3 ...
## $ rugged     : int  1 2 1 1 3 1 2 5 1 5 ...
## $ fun        : int  3 2 3 3 5 2 2 4 2 4 ...
## $ safe       : int  5 3 5 2 4 5 3 3 4 4 ...
## $ performance: int  3 2 3 3 5 2 2 2 2 3 ...
## $ family     : int  5 4 3 2 1 5 2 4 4 4 ...
## $ versatile  : int  3 3 4 2 2 4 2 5 2 5 ...
## $ sports     : int  1 1 2 3 5 1 4 5 2 5 ...
## $ status     : int  4 1 4 4 5 3 1 3 3 4 ...
## $ practical  : int  3 4 3 2 2 4 2 3 4 5 ...
## $ discipline : int  1 1 1 0 0 0 0 1 0 1 ...
```

5.4 Testing of Assumptions:

5.4.1 Inter-item correlation:

See if there are at least several sizeable correlations, e.g. > 0.5?

```
#=========================================================================
# Test of Assumptions
#========================================================================
# Inter-item correlations (correlation matrix) -
# Create a correlation matrix
MBACarMatrix<-cor(MBACar)
round(MBACarMatrix, 2)

##             student   Car exciting dependable luxurious outdoorsy powerful
## student        1.00 -0.01     0.01      -0.02     -0.05      0.11    -0.03
## Car           -0.01  1.00    -0.05       0.08     -0.05     -0.13    -0.02
## exciting       0.01 -0.05     1.00       0.03      0.45      0.18     0.63
## dependable    -0.02  0.08     0.03       1.00      0.40     -0.07     0.16
## luxurious     -0.05 -0.05     0.45       0.40      1.00     -0.21     0.43
## outdoorsy      0.11 -0.13     0.18      -0.07     -0.21      1.00     0.32
## powerful      -0.03 -0.02     0.63       0.16      0.43      0.32     1.00
## stylish       -0.06 -0.09     0.75       0.17      0.63      0.00     0.59
## comfortable    0.06 -0.15     0.04       0.46      0.47      0.02     0.22
## rugged         0.03 -0.14     0.17       0.00     -0.16      0.79     0.30
## fun            0.02 -0.12     0.83       0.09      0.46      0.15     0.63
## safe           0.03  0.09    -0.15       0.54      0.26      0.02     0.09
## performance   -0.01  0.17     0.61       0.23      0.56     -0.14     0.58
## family         0.01  0.04    -0.57       0.21     -0.27      0.16    -0.31
## versatile      0.01 -0.09    -0.14       0.13     -0.13      0.44     0.07
## sports        -0.03  0.01     0.66      -0.09      0.21      0.33     0.54
## status        -0.05  0.03     0.64       0.22      0.67     -0.08     0.54
## practical      0.06 -0.05    -0.34       0.29     -0.15      0.17    -0.15
## discipline     0.05 -0.04    -0.09       0.08     -0.06      0.05    -0.07
##             stylish comfortable rugged   fun  safe performance family
## student       -0.06        0.06   0.03  0.02  0.03       -0.01   0.01
## Car           -0.09       -0.15  -0.14 -0.12  0.09        0.17   0.04
## exciting       0.75        0.04   0.17  0.83 -0.15        0.61  -0.57
## dependable     0.17        0.46   0.00  0.09  0.54        0.23   0.21
```

```
## luxurious      0.63         0.47  -0.16  0.46  0.26         0.56  -0.27
## outdoorsy      0.00         0.02   0.79  0.15  0.02        -0.14   0.16
## powerful       0.59         0.22   0.30  0.63  0.09         0.58  -0.31
## stylish        1.00         0.23   0.04  0.74 -0.02         0.67  -0.52
## comfortable    0.23         1.00   0.05  0.13  0.58         0.19   0.24
## rugged         0.04         0.05   1.00  0.16  0.09        -0.12   0.17
## fun            0.74         0.13   0.16  1.00 -0.08         0.62  -0.54
## safe          -0.02         0.58   0.09 -0.08  1.00         0.10   0.42
## performance    0.67         0.19  -0.12  0.62  0.10         1.00  -0.44
## family        -0.52         0.24   0.17 -0.54  0.42        -0.44   1.00
## versatile     -0.14         0.25   0.46 -0.10  0.30        -0.20   0.54
## sports         0.56        -0.12   0.38  0.68 -0.22         0.47  -0.49
## status         0.78         0.27  -0.05  0.66  0.09         0.73  -0.48
## practical     -0.28         0.30   0.18 -0.32  0.41        -0.28   0.69
## discipline    -0.07         0.02   0.02 -0.05  0.06        -0.04   0.11
##            versatile sports status practical discipline
## student         0.01  -0.03  -0.05      0.06       0.05
## Car            -0.09   0.01   0.03     -0.05      -0.04
## exciting       -0.14   0.66   0.64     -0.34      -0.09
## dependable      0.13  -0.09   0.22      0.29       0.08
## luxurious      -0.13   0.21   0.67     -0.15      -0.06
## outdoorsy       0.44   0.33  -0.08      0.17       0.05
## powerful        0.07   0.54   0.54     -0.15      -0.07
## stylish        -0.14   0.56   0.78     -0.28      -0.07
## comfortable     0.25  -0.12   0.27      0.30       0.02
## rugged          0.46   0.38  -0.05      0.18       0.02
## fun            -0.10   0.68   0.66     -0.32      -0.05
## safe            0.30  -0.22   0.09      0.41       0.06
## performance    -0.20   0.47   0.73     -0.28      -0.04
## family          0.54  -0.49  -0.48      0.69       0.11
## versatile       1.00   0.00  -0.17      0.56       0.14
## sports          0.00   1.00   0.46     -0.29      -0.11
## status         -0.17   0.46   1.00     -0.27      -0.01
## practical       0.56  -0.29  -0.27      1.00       0.08
## discipline      0.14  -0.11  -0.01      0.08       1.00
```

Interpretation:

We could see significant correlation (Coefficient > 0.5) in many variables, as highlighted in yellow.

5.4.2 KMO Test to see if the data is likely to factor or not:

```
# Kaiser-Meyer-Olkin (KMO) Test :
KMO(r=MBACarMatrix)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = MBACarMatrix)
## Overall MSA =  0.87
## MSA for each item =
##     student         Car    exciting  dependable   luxurious   outdoorsy
##        0.32        0.40        0.91        0.82        0.89        0.70
##    powerful     stylish comfortable      rugged         fun        safe
##        0.93        0.93        0.81        0.70        0.91        0.81
## performance      family   versatile      sports      status   practical
```

```
##        0.90        0.89        0.85        0.89        0.91        0.84
##  discipline
##        0.53
```

Interpretation:

The KMO suggests strong MSA (Measure of Sample Adequacy) for individual as well as overall variables. This suggests we may proceed with Factoring.

5.4.3 Barlett Sphericity Test

We shall conduct Barlett Spherecity Test to see if Dimension Reduction is possible or not.

```
print(cortest.bartlett(MBACarcorr,nrow(fcah)))
```

```
## $chisq
## [1] 3413.468
##
## $p.value
## [1] 0
##
## $df
## [1] 136
```

Interpretation:

The Barlett Sphericity Test also suggests that Data Dimension Reduction is possible. (P value is less than 0.05, failing to accept the null hypothesis)

5.5 Deciding Number of Factors:

Let's calculate the eigen values and eigen vectors.

```
A<-eigen(MBACarcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues
```

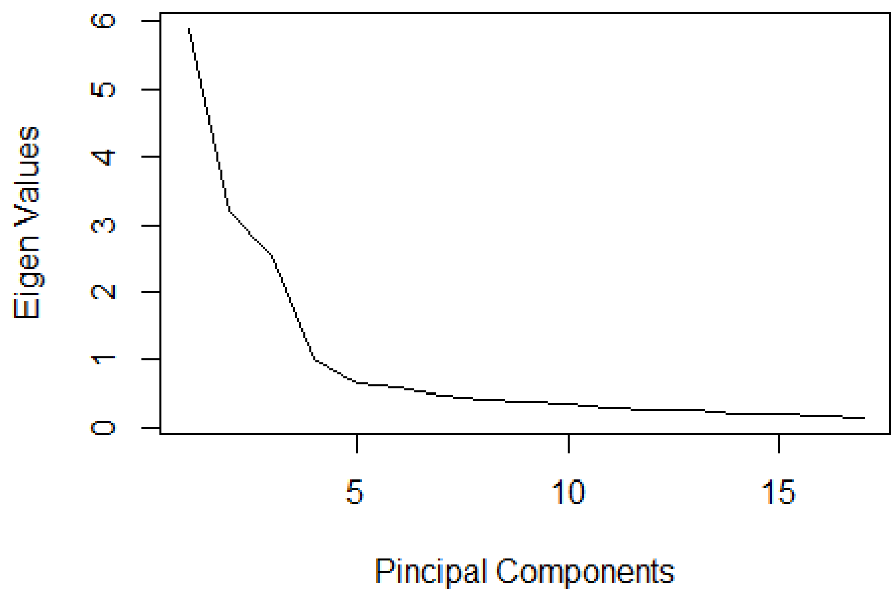
```
##  [1] 5.8927230 3.1832144 2.5331040 1.0011058 0.6562549 0.5848610
0.4767852
##  [8] 0.4058686 0.3765764 0.3467376 0.2995466 0.2690347 0.2505522
0.2108704
## [15] 0.1877133 0.1803949 0.1446570
```

Interpretation:

Based on the fact that there are 4 factors having Eigen Value > 1, we can decide for 4 Factors.

5.5.1 The SREE plot

```
plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")
```



Interpretation:

The SREE plot also suggests that we may go ahead with 4 factors.

5.6 Conduct Factor Analysis

5.6.1 Factor Analysis without rotation

```
# Factor Analysis using Principal Axis Factoring using 4 factors
#
solution<-fa(r=MBACarcorr,nfactors=4,rotate = "none",fm="pa")
solution

## Factor Analysis using method =  pa
## Call: fa(r = MBACarcorr, nfactors = 4, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##               PA1  PA2   PA3   PA4    h2   u2 com
## exciting     0.86 0.02  0.20  0.09 0.784 0.22 1.1
## dependable   0.12 0.51 -0.41 -0.08 0.444 0.56 2.1
## luxurious    0.63 0.24 -0.46 -0.09 0.677 0.32 2.2
## outdoorsy    0.04 0.44  0.75 -0.16 0.778 0.22 1.7
## powerful     0.69 0.30  0.17 -0.04 0.595 0.40 1.5
## stylish      0.87 0.10 -0.08  0.10 0.783 0.22 1.1
## comfortable  0.14 0.61 -0.39 -0.12 0.565 0.44 1.9
## rugged       0.05 0.49  0.72 -0.21 0.804 0.20 2.0
## fun          0.86 0.07  0.15  0.10 0.769 0.23 1.1
## safe        -0.07 0.67 -0.39 -0.18 0.642 0.36 1.8
## performance  0.77 0.07 -0.22  0.07 0.654 0.35 1.2
## family      -0.68 0.55 -0.06  0.13 0.783 0.22 2.0
## versatile   -0.24 0.65  0.28  0.24 0.614 0.39 2.0
## sports       0.69 0.01  0.46  0.09 0.690 0.31 1.8
```

```
## status       0.82 0.12 -0.22  0.04 0.741 0.26 1.2
## practical   -0.45 0.62 -0.05  0.30 0.680 0.32 2.3
## discipline  -0.09 0.10 -0.01  0.05 0.021 0.98 2.5
##
##                        PA1  PA2  PA3  PA4
## SS loadings           5.61 2.83 2.24 0.34
## Proportion Var        0.33 0.17 0.13 0.02
## Cumulative Var        0.33 0.50 0.63 0.65
## Proportion Explained  0.51 0.26 0.20 0.03
## Cumulative Proportion 0.51 0.77 0.97 1.00
##
##                                                   PA1  PA2  PA3  PA4
## Correlation of scores with factors               0.98 0.95 0.94 0.73
## Multiple R square of scores with factors         0.95 0.90 0.89 0.54
## Minimum correlation of possible factor scores    0.91 0.79 0.78 0.07
```

Interpretation:

Although we have three factors having loadings > 1, PA1 is heavily loaded.

Hence we may look for rotation to see if loadings can be balanced.

5.6.2 Factor Analysis with rotation

```
# Factor Analysis using Principal Axis Factoring
# using 4 factors with Rotation

solution1<-fa(r=MBACarcorr,nfactors=4,
              rotate = "varimax",fm="pa")

print(solution1)

## Factor Analysis using method =  pa
## Call: fa(r = MBACarcorr, nfactors = 4, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##               PA1   PA2   PA3   PA4    h2   u2 com
## exciting     0.85 -0.07  0.15 -0.19 0.784 0.22 1.2
## dependable   0.13  0.63 -0.07  0.17 0.444 0.56 1.3
## luxurious    0.57  0.53 -0.21 -0.16 0.677 0.32 2.4
## outdoorsy    0.08 -0.07  0.86  0.18 0.778 0.22 1.1
## powerful     0.68  0.17  0.30 -0.07 0.595 0.40 1.6
## stylish      0.86  0.15 -0.03 -0.15 0.783 0.22 1.1
## comfortable  0.16  0.71  0.01  0.20 0.565 0.44 1.3
## rugged       0.09  0.00  0.88  0.16 0.804 0.20 1.1
## fun          0.85 -0.01  0.13 -0.15 0.769 0.23 1.1
## safe        -0.06  0.75  0.05  0.26 0.642 0.36 1.3
## performance  0.74  0.22 -0.16 -0.18 0.654 0.35 1.4
## family      -0.53  0.29  0.12  0.64 0.783 0.22 2.5
## versatile   -0.06  0.14  0.41  0.65 0.614 0.39 1.8
## sports       0.70 -0.25  0.35 -0.13 0.690 0.31 1.8
## status       0.79  0.27 -0.12 -0.19 0.741 0.26 1.4
## practical   -0.25  0.27  0.10  0.73 0.680 0.32 1.6
## discipline  -0.06  0.04  0.01  0.12 0.021 0.98 1.7
##
##                        PA1  PA2  PA3  PA4
```

```
## SS loadings           5.05 2.15 2.04 1.77
## Proportion Var        0.30 0.13 0.12 0.10
## Cumulative Var        0.30 0.42 0.54 0.65
## Proportion Explained  0.46 0.20 0.19 0.16
## Cumulative Proportion 0.46 0.65 0.84 1.00
##
##                                                   PA1  PA2  PA3  PA4
## Correlation of scores with factors               0.96 0.89 0.93 0.84
## Multiple R square of scores with factors         0.92 0.79 0.86 0.70
## Minimum correlation of possible factor scores    0.83 0.59 0.73 0.41
```

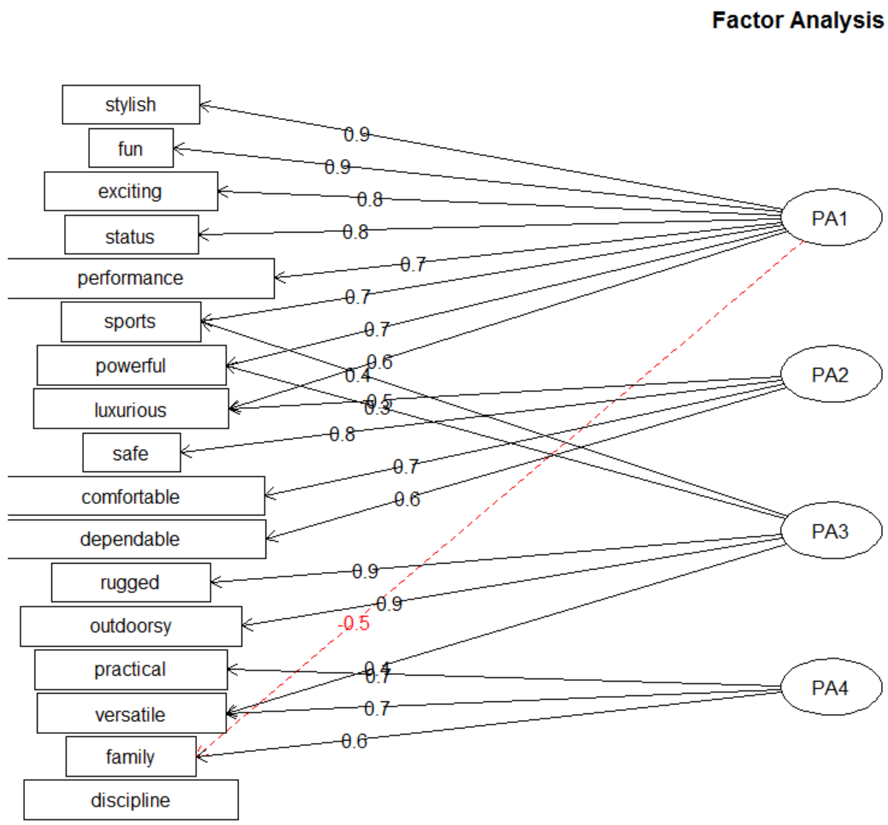
Interpretation:

We could see slight improvement in the Loadings.

5.7 Labelling the Factors

Let's draw the factors diagram and see how each factor describes various characteristics of the cars.

```
fa.diagram(solution1,simple=FALSE)
```



Interpretation:

Factor wise characteristics are as follows:

PA1: Stylish, Fun, Exciting, Status, Performance, Sports, Powerful, Luxurious.

It can be categorised as a **Premium Status Car**.

PA2: Luxurious, Safe, Comfortable, and Dependable.

It can be categorised as a **Sedan Car**.

PA3: Sports, Powerful, Rugged, Outdoorsy, Versatile.

This car can be categorised as a **Sports Utility Vehicle**.

PA4: Practical, Versatile, Family.

It can be categorised as an **Economical Car**.

6 Appendix – Source Code

```
#=======================================================================
#
#   Mini Project 3: Advanced Statistics
#
#   1. ANOVA Analysis
#   2. Regression Analysis
#   3. Principal Component Analysis
#   4. Factor Analysis
#
#=======================================================================
# Environment Set up and Data Import
#=======================================================================
# Install Libraries and Packages
#=======================================================================
#
library(psych)

library(car)

library(foreign)
library(MASS)
library(lattice)

library(nortest) # Anderson Darling
#
#=======================================================================
# Setup Working Directory
#=======================================================================
setwd("D:/Moderator/00 Great Lakes Engagement/20 BACP.Aug17 AS")
getwd()

#
#=======================================================================
# 1. One Way ANOVA Analysis
#=======================================================================
# Read Input File
PL_X_SELL=read.csv("PL_X_SELL.csv")
attach(PL_X_SELL)
#
# Find out Total Number of Rows and Columns
dim(PL_X_SELL)

# Find out Names of the Columns (Features)
names(PL_X_SELL)

# Find out Class of each Feature, along with internal structure
str(PL_X_SELL)

#
# Check top 6 and bottom 6 Rows of the Dataset
head(PL_X_SELL)

tail(PL_X_SELL)
```

```
# head(PL_X_SELL,10) # To obtain desired number of rows, here 10.
#
#Check for Missing Values
colSums(is.na(PL_X_SELL))

#
# Provide Summary of a Dataset.
summary(PL_X_SELL)

#===========================================================================
# Data Visualization using Graphs:
#===========================================================================
# Boxplot of Balance v/s Occupation
boxplot(Balance~Occupation,main='Occupation Vs Balance',xlab =
"Occupation",
        ylab = "Balance",col = "turquoise1",horizontal = FALSE)

#===========================================================================
# Testing of Assumptions
#===========================================================================
#
#Anderson Darling Test for Normality

for(i in
    unique(factor(Occupation)))
  {cat(ad.test(PL_X_SELL[PL_X_SELL$Occupation==i,]$Balance)$p.value,"")}

## 3.7e-24 3.7e-24 3.7e-24 3.7e-24

#
# Homogeneity in Variance Test using Levenes Test
leveneTest(Balance~Occupation)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     3  54.545 < 2.2e-16 ***
##       19996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#
# Barlett's Test
# barlett.test(Balance~Occupation)
#
# ANOVA - Analysis of Variance
aov1 <- aov(Balance~Occupation)
summary(aov1)

kruskal.test(Balance~Occupation)

# Post-Hoc Test (Tukey)
TukeyHSD(aov1)

plot(TukeyHSD(aov1))
```

```
#
# Robust Method for One Way Anova
#
oneway.test(Balance ~ Occupation, var.equal=FALSE)

Anova(model1,Type="II", white.adjust=TRUE)

#
#============================================================================
# 1B. Two Way ANOVA Analysis
#============================================================================
# Same PL_X_SELL file is being used for Two Way ANOVA
#
summary(PL_X_SELL)

# Factors:
Gender<-factor(Gender,labels=c("M","O","F"))
Occupation<-factor(Occupation,labels=c("PROF","SAL","SELF-EMP","SENP"))
# Data Visualization
tapply(Balance,list(Gender,Occupation),mean)

tapply(Balance,list(Gender,Occupation),sd)

#
# Interaction Plot
#
interaction.plot(Gender,Occupation,Balance)

#
# Testing of Assumptions
# Normality Test - Anderson Darling Test
#
for(i in
    unique(factor(Gender)))
{cat(ad.test(PL_X_SELL[PL_X_SELL$Gender==i,]$Balance)$p.value,"")}

## 3.7e-24 3.7e-24 3.7e-24

# Homogeneity of Variance Test
#
leveneTest(Balance~Occupation*Gender)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group    10  52.553 < 2.2e-16 ***
##       19989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#
# Anova Output  and Interpretations

aov2 <- aov(Balance~Occupation+Gender+Occupation:Gender)
summary(aov2)

TukeyHSD(aov2)
```

```
plot(TukeyHSD(aov2))

#
#==========================================================================
# 2. Regression Analysis
#==========================================================================
#
# Read Input File
Household_Data=read.csv("Household_Data.csv")
attach(Household_Data)
#
# Find out Total Number of Rows and Columns
dim(Household_Data)

## [1] 50  5

# Find out Names of the Columns (Features)
names(Household_Data)

## [1] "Annual.Income"  "Monthly.Income" "Household.Size" "Amount.Charged"
## [5] "Monthly.EMI"

# Find out Class of each Feature, along with internal structure
str(Household_Data)

#
# Check top 6 and bottom 6 Rows of the Dataset
head(Household_Data)

tail(Household_Data)

# head(Household_Data,10) # To obtain desired number of rows, here 10.

#
#Check for Missing Values
colSums(is.na(Household_Data))

#
# Provide Summary of a Dataset.
summary(Household_Data)

#
#==========================================================================
# Data Visualization using Graphs:
#==========================================================================
# Check relation of Amount Charged with other Variables
#
par(mfrow=c(2,2))
plot(Amount.Charged~Annual.Income)
plot(Amount.Charged~Monthly.Income)
plot(Amount.Charged~Household.Size)
plot(Amount.Charged~Monthly.EMI)

#
#==========================================================================
# Building Regression equation for one variable at a time.
```

```r
#===========================================================================
# Annual Income
resultAI<-lm(formula = Amount.Charged~Annual.Income)
resultAI

summary(resultAI)

# Monthly Income
resultMI<-lm(formula = Amount.Charged~Monthly.Income)
resultMI

summary(resultMI)

# Household Size
resultHS<-lm(formula = Amount.Charged~Household.Size)
resultHS

summary(resultHS)

# Monthly EMI
resultME<-lm(formula = Amount.Charged~Monthly.EMI)
resultME

summary(resultME)

#===========================================================================
# Building Regression equation for Multiple variables.
#===========================================================================
# Annual Income + Monthly Income
resultAIMI<-lm(formula =
Amount.Charged~Annual.Income+Monthly.Income,data=Household_Data)
resultAIMI

summary(resultAIMI)

# Annual Income + Household Size
resultAIHS<-lm(formula = Amount.Charged~Annual.Income+Household.Size)
resultAIHS

summary(resultAIHS)

# Annual Income + Monthly EMI
resultAIME<-lm(formula = Amount.Charged~Annual.Income+Monthly.EMI)
resultAIME

summary(resultAIME)

# Monthly Income + Household Size
resultMIHS<-lm(formula = Amount.Charged~Monthly.Income+Household.Size)
resultMIHS

summary(resultMIHS)

# Monthly Income + Monthly EMI
resultMIME<-lm(formula = Amount.Charged~Monthly.Income+Monthly.EMI)
resultMIME
```

```
summary(resultMIME)

# Household Size + Monthly EMI
resultHSME<-lm(formula = Amount.Charged~Household.Size+Monthly.EMI)
resultHSME

summary(resultHSME)

# Annual Income + Household Size + Monthly EMI
resultAIHSME<-lm(formula =
Amount.Charged~Annual.Income+Household.Size+Monthly.EMI)
resultAIHSME

summary(resultAIHSME)

# Annual Income + Monthly Income + Household Size + Monthly EMI
resultAIMIHSME<-lm(formula =
Amount.Charged~Annual.Income+Monthly.Income+Household.Size+Monthly.EMI)
resultAIMIHSME

summary(resultAIMIHSME)

#
#==========================================================================
# Testing of Assumptions ( Mean and Variance)
#==========================================================================
mean(resultAIMIHSME$residuals)

## [1] 2.628453e-15

par(mfrow=c(2,2))
# Testing of homoscedisticty
plot(resultAIMIHSME)

# Testing the Correlation between Errors and Explanatory Variables
cor.test(Annual.Income,resultAIMIHSME$residuals)

#==========================================================================
# Robust Regression
#==========================================================================
result2 <- rlm(Amount.Charged~Annual.Income+Monthly.EMI+Household.Size)
aov(result2)

#
summary(result2)

#
# Parsimony
#
library(leaps)

Null<-lm(Amount.Charged~1)
Full<-
lm(Amount.Charged~Annual.Income+Monthly.Income+Monthly.EMI+Household.Size)
step(Null,scope = list(lower=Null, upper=Full),direction = "forward")
```

```
# Parsimony Backward Selection
step(Full, direction = "backward")

# Parsimony Stepwise both forward and Backward Selection
step(Null,scope = list(upper=Full),data=Household_Data,direction="both")

#
#===========================================================================
# 3A. Principal Component Analysis - Batting
#===========================================================================
# Libraries
#install.packages("ggfortify")
#install.packages("nFactors")
#install.packages("Hmisc")
#install.packages("GPArotation")
library(corpcor)

library(GPArotation)
library(psych)
library(ggplot2)

library(ggfortify)

library(nFactors)

library(dplyr)

library(expm)

library(Hmisc)

#
# Read Input file
batting<-read.csv("batting_bowling_ipl_bat.csv", header=TRUE)
batting

attach(batting)
# Descriptive Statistics
str(batting)

summary(batting)

pcabat <- na.omit(batting[2:7]) # Omit NA records
summary(pcabat)

#
# Understanding Correlation
batcorr <- cor(pcabat)
batcorr

#
# Barlett Sphericity Test for checking the possibility
# of data dimension reduction
#
print(cortest.bartlett(batcorr,nrow(pcabat)))
```

```
#
# Finding out the Eigen Values and Eigen Vectors

A<-eigen(batcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

#
# Getting the loadings and Cummunality

pc1<-principal(pcabat,nfactors = length(pcabat),rotate="none")
pc1

# Interpreting the variance
#
part.pca<-eigenvalues/sum(eigenvalues)*100
part.pca

#
#Plotting Scree Graphs
dev.off() # To Reset the earlier partition command.

plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")

# Principal Components Scoring and Perceptual Map

pcabatsc<-scale(pcabat)
z<-as.matrix(pcabatsc%*%eigenvectors)
pc.cr<-princomp(pcabatsc,cor=TRUE)
summary(pc.cr)

biplot(pc.cr)

#
#=========================================================================
# 3B.Principal Component Analysis -Bowling
#=========================================================================
# Read Input file
bowling<-read.csv("batting_bowling_ipl_bowl.csv", header=TRUE)
bowling

attach(bowling)

# Descriptive Statistics
str(bowling)

summary(bowling)

pcabowl <- na.omit(bowling[2:5]) # Omit NA records
summary(pcabowl)

#
# Understanding Correlation
```

```
bowlcorr <- cor(pcabowl)
bowlcorr

#
# Barlett Sphericity Test for checking the possibility
# of data dimension reduction
#
print(cortest.bartlett(bowlcorr,nrow(pcabowl)))

#
# Finding out the Eigen Values and Eigen Vectors

A<-eigen(bowlcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

eigenvectors

#
# Getting the loadings and Cummunality

pc1<-principal(pcabowl,nfactors = length(pcabowl),rotate="none")
pc1

# Interpreting the variance
#
#Plotting Scree Graphs
dev.off() # To Reset the earlier partition command.

plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")

# Principal Components Scoring and Perceptual Map

pcabowlsc<-scale(pcabowl)
z<-as.matrix(pcabowlsc%*%eigenvectors)
z

pc.cr<-princomp(pcabowlsc,cor=TRUE)
summary(pc.cr)

biplot(pc.cr)
#
#==========================================================================
# 4. Factor Analysis
#==========================================================================
# Install necessary Packages
# install.packages("MVN")
# install.packages("psy")

# Load Required Libraries
library(corpcor)
library(GPArotation)
library(psych)
library(ggplot2)
```

```r
library(MASS)
library(MVN)

library(psy)

#
# Load Data
# Read Input file
MBACar=read.csv("MBAcar_Datafile.csv", header = TRUE)
attach(MBACar)
# detach(MBACar)
#
# Find out Total Number of Rows and Columns
dim(MBACar)

# Find out Names of the Columns (Features)
names(MBACar)

# Find out Class of each Feature, along with internal structure
str(MBACar)

#
# Check top 6 and bottom 6 Rows of the Dataset
head(MBACar)

tail(MBACar)

# head(MBACar,10) # To obtain desired number of rows, here 10.
#
#Check for Missing Values
colSums(is.na(MBACar))

#
# Provide Summary of a Dataset.
summary(MBACar)

#==========================================================================
# Test of Assumptions
#==========================================================================
# Testing of assumptions for Factor Analysis
# Inter-item correlations (correlation matrix) -
# are there at least several sizable correlations - e.g. > 0.5?
# Create a correlation matrix
MBACarMatrix<-cor(MBACar)
round(MBACarMatrix, 2)

# Kaiser-Meyer-Olkin (KMO) Test :
KMO(r=MBACarMatrix)

# Bartlett's Test of Sphericity:

fcah<-MBACar[3:19]
summary(fcah)

MBACarcorr<-cor(fcah)
MBACarcorr
```

```r
print(cortest.bartlett(MBACarcorr,nrow(fcah)))

#
A<-eigen(MBACarcorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

#
plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")

# Factor Analysis using Principal Axis Factoring using 4 factors
#
solution<-fa(r=MBACarcorr,nfactors=4,rotate = "none",fm="pa")
solution

# Factor Analysis using Principal Axis Factoring
# using 4 factors with Rotation

solution1<-fa(r=MBACarcorr,nfactors=4,
              rotate = "varimax",fm="pa")

print(solution1)

# With 3 Factors

solution2<-fa(r=MBACarcorr,nfactors=3,rotate = "varimax",fm="pa")
print(solution2)

fa.diagram(solution1,simple=FALSE)

#=======================================================================
#
#                                T H E - E N D
#
#=======================================================================
```