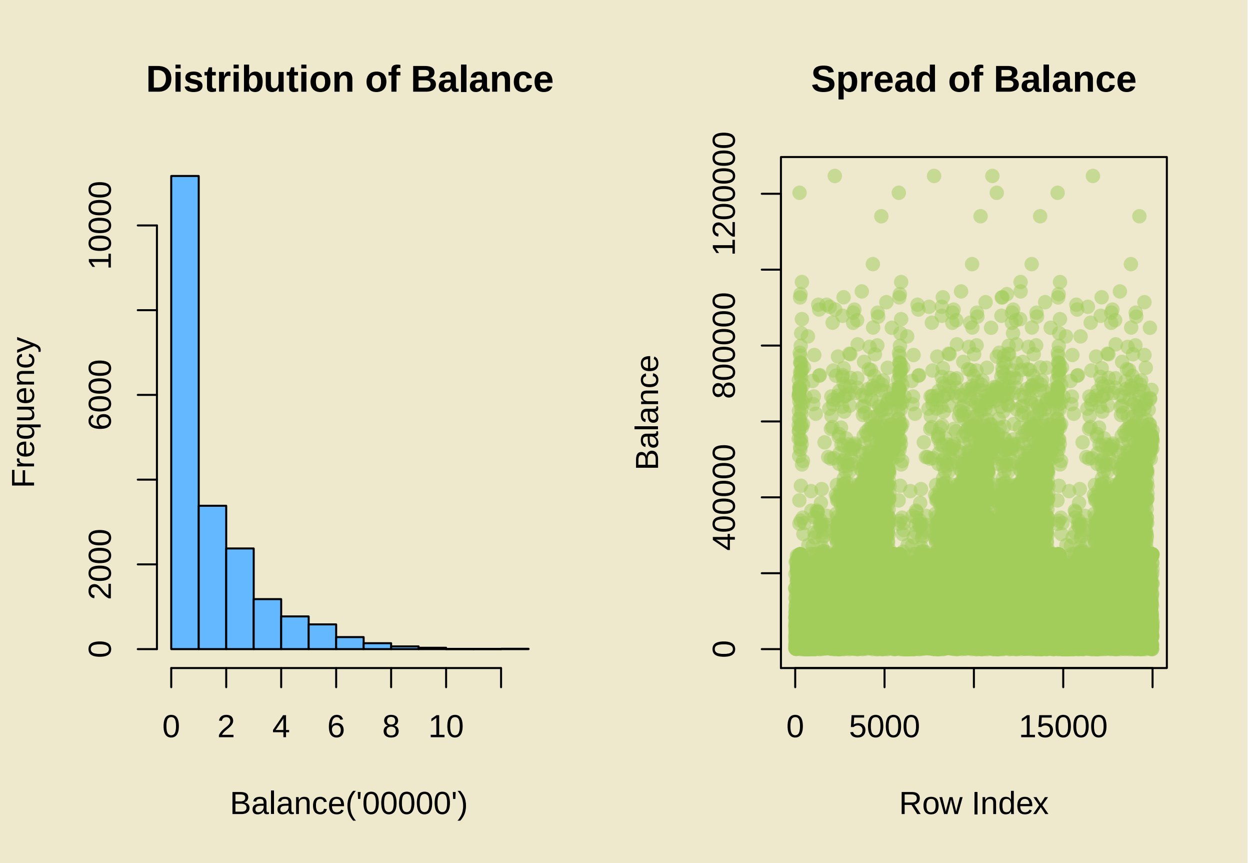
Advanced_statistics_Project

James Peter

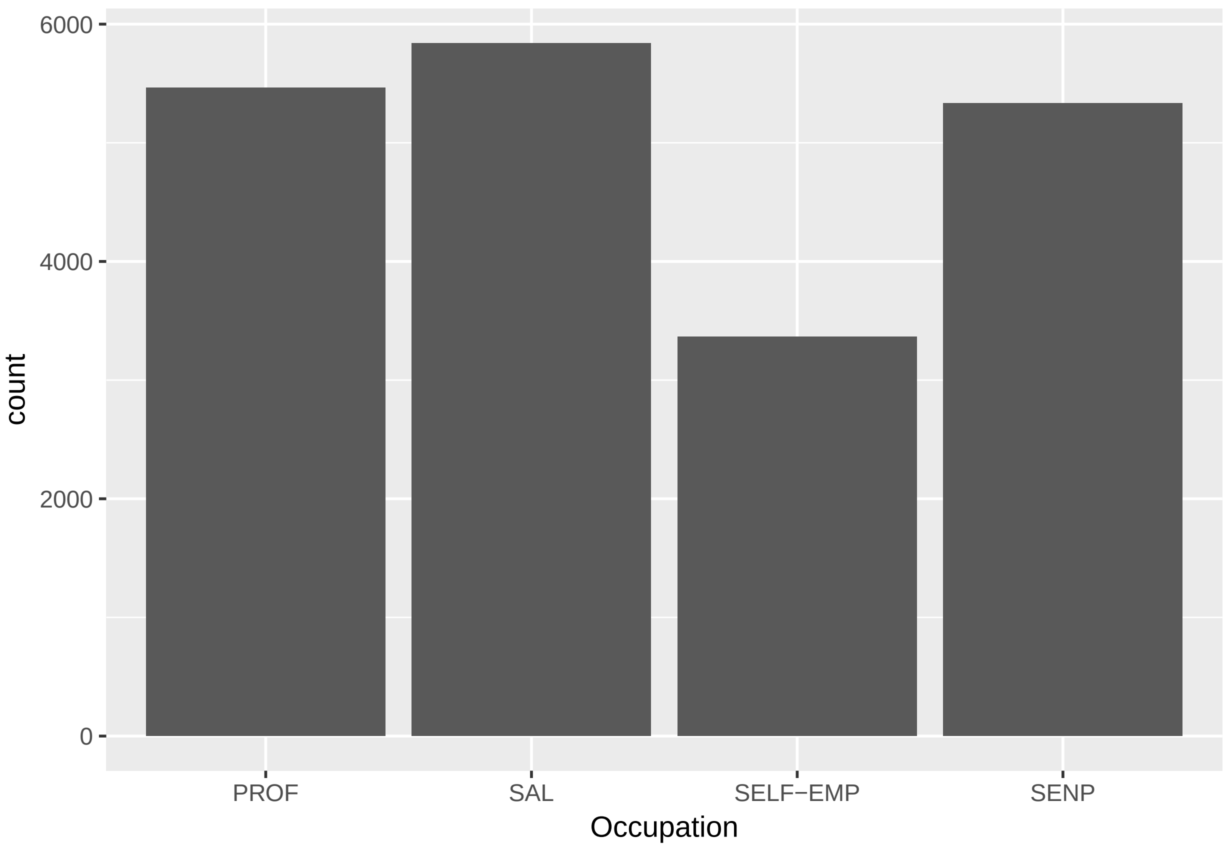
May 7, 2018

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



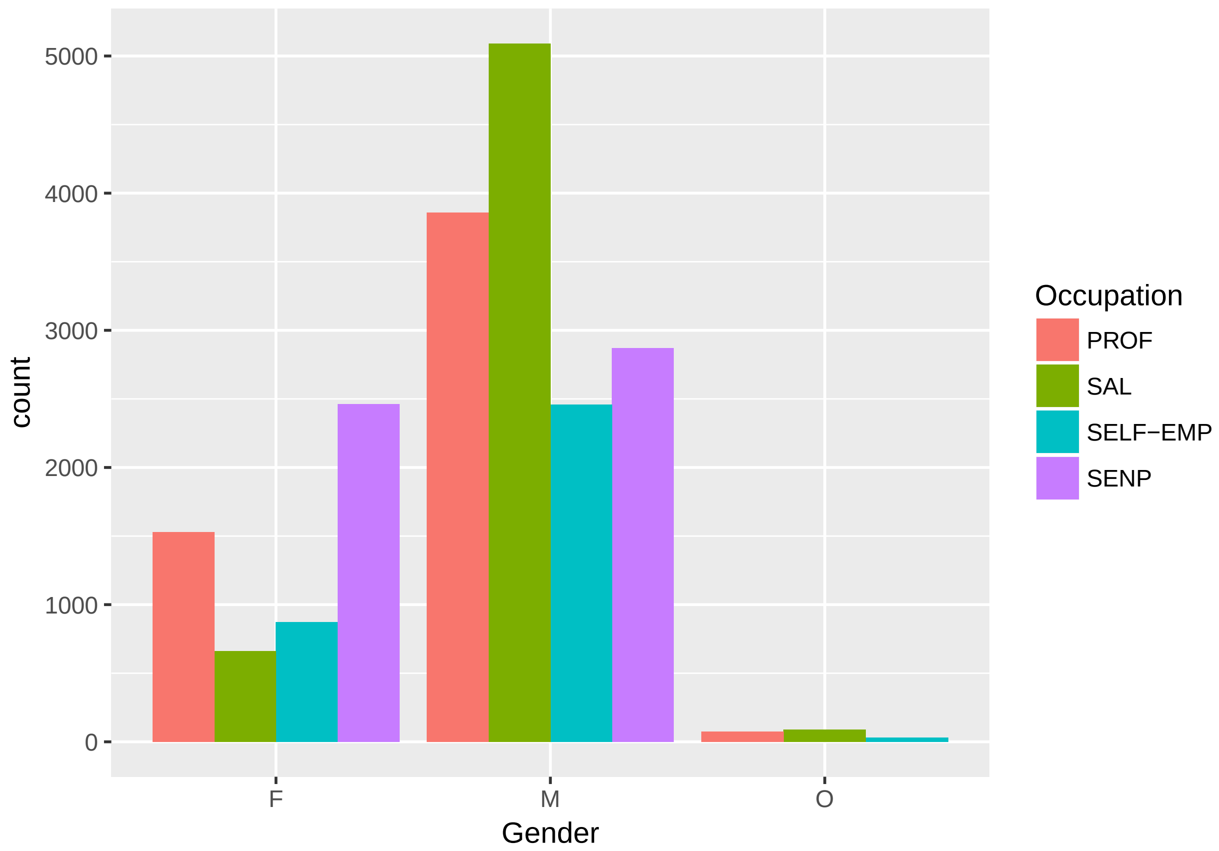
Observation

- The distribution of average quarterly balance of the deposit account holder is right skewed and the difference between the mean and medium is significant which is at 66,425.00, the higher balance amount has huge impact as expected on the total balance.
- The scatter plot shows 12 points which are way above 100,000.00 so it could be considered as outliers, so the effect of this 12 points is higher on mean; if the mean of balance is used as a critical measure for drawing any conclusion. And it will be good measure to check the effects by including and excluding those 12 points while conducting any kind of analysis if required.



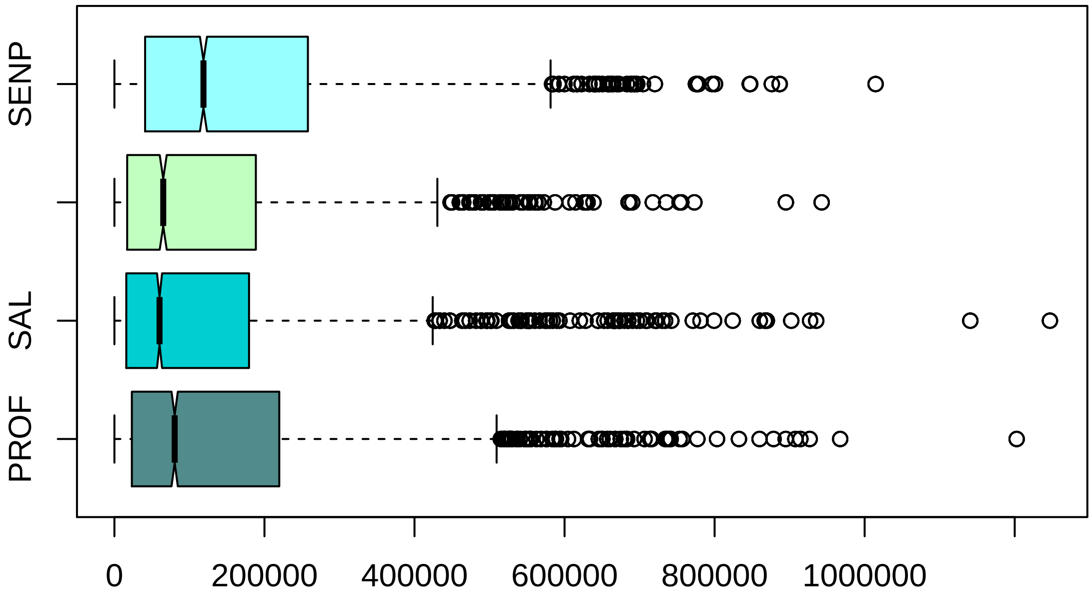
Observation

- The are good total number of observations for all the Occupation types, and the Occupation factor can be considered as balanced factor for conducting the one-way ANOVA.



Observation

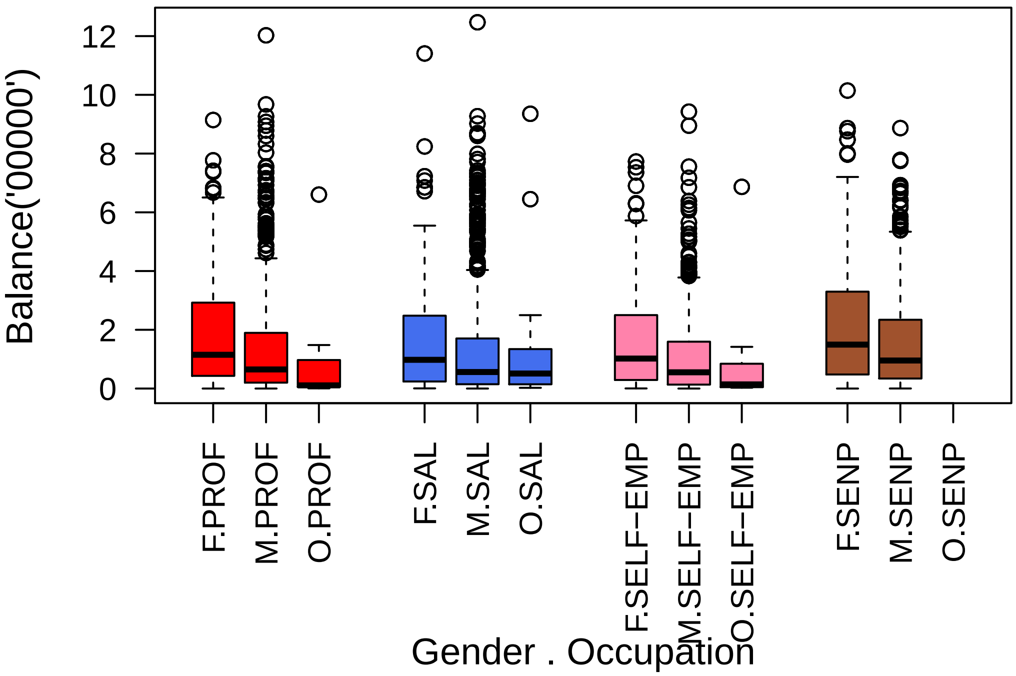
- Withing gender levels theire are good amount of observations for each occupation type. But the "O" Gender type which represents company has every less total number of observations as compared to other two genders; which makes the Gender factor an unbalanced factor for ANOVA analysis, which would require to conduct robust analysis of variance to confirm the significance of variance.



Observation

- The boxplot shows several points which are above the maximum considered by the boxplot and considers those points above the maximum as outliers. All the employments have types have considerable amount of outliers in the data. For the Self-employment and salary employment the values above 400,000.00 are considered outliers and for professional employment type values above approximately 500,000.00 are considered outliers. And for self-employed non-professionals values close to 600,000.00 and above are considered outliers.

Visualizing the Gender:Occupation for two-way ANOVA



Observation

- The interaction boxplot for each factors shows are are considerable amount of outliers and no observation for "O.SENP" which would be treated as 'NA' when two-way ANOVA is conducted.

How to deal with outliers while conducting ANOVA

- Means, standard deviations are highly sensitive to outliers. And since the assumptions of regression and ANOVA, are also based on these statistics. ANOVA and Regression models rely heavily on the normality assumption. So the presence of outliers can severely distort the analysis. To deal with this problem when the assumptions of ANOVA are violated robust method analysis needs to be conducted to verify the ANOVA results.

ANOVA test hypotheses:

```
Null hypothesis: the means of the different groups are the same
Alternative hypothesis: At least one sample mean is not equal to the others.
```

Assumption of ANOVA test: * The observations are obtained independently and randomly from the population defined by the factors levels. * The data of each factor level are normally distributed * The factor levels have a common variance

Normality test

```
## SAL : 3.7e-24 | SELF-EMP : 3.7e-24 | PROF : 3.7e-24 | SENP : 3.7e-24 |
```

Interpretation of Anderson-Darling normality test which is alternative for Shapiro normality test:

- The p-values for all the three factor levels under the Occupation factor is less than 0.05 at 95% significance, which tells that there are not normally distributed, which violates the normality assumption of ANOVA test.

Homogeneity of variance test

```
## 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
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  household$Balance by household$Occupation
## Bartlett's K-squared = 96.401, df = 3, p-value < 2.2e-16
```

Interpretation of LeveneTest and Bartlett test of homogeniety of variances:

- Both the test confirm that the variance of all the factor levels are not homogeneous which means atleast one the pairs in the factor levels in the Occupation factor have different variance.
- Among the two test, the LeveneTest provides accurate results which is immune to the violations of the normality of factor levels.
- A p-value of < 2.2e-16 (significant) tells that the variance observed between factor level pairs are significant, which also the violates the homogeniety of variances assumption of ANOVA.

One-way ANOVA test

```
##                         Df    Sum Sq   Mean Sq F value Pr(>F)
## household$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
```

Interpreting the results of one-way ANOVA tests

- The p-value is less than the significance level 0.05 which is <2e-16 (significant) and tells that there are significant differences in variances between the factor levels of the Occupation factor. The variance of atleat one pair is different from the other factor levels pairs in the Occupation factor.

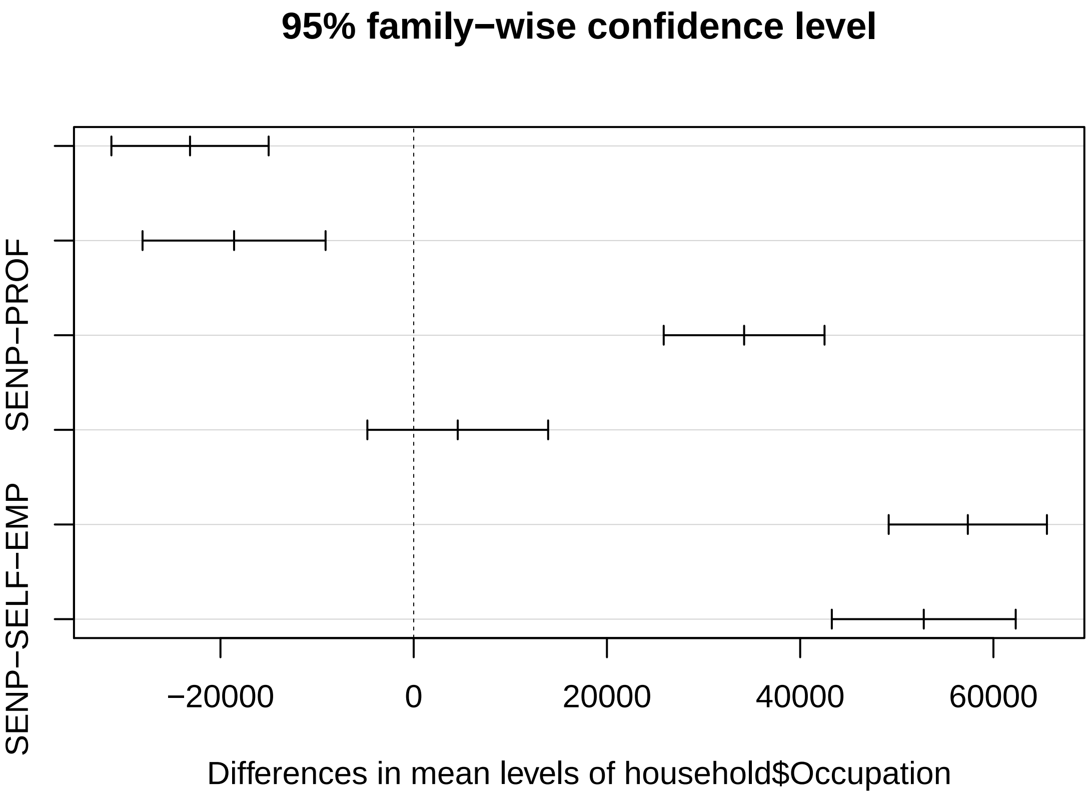
Multiple pairwise comparison between the means of groups

- In one-way ANOVA test, a significant p-value indicates that some of the group means are different, but can tell among the groups which paris of groups are different.
- As the ANOVA test is significant, TukeyHSD (Tukey Honest Significant Differences) function performs multiple pairwise-comparison between the groups and makes it possible to analyze the pairs of groups which different variance.

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

Interpretation of TukeyHSD results:

- diff: difference between means of the two groups
- lwr, upr: the lower and the upper end point of the confidence interval at 95%
- p adj: p-value after adjustment for the multiple comparisons.
- The result tells that there is significance difference in variance between all the pairs except the Self-employed and Salary pair.



Interpretation of TukeyHSD 95% family-wise confidence level plot:

- The 95% confidence interval of the occupation type pairs are far away from the zero variance, except the Self-employed and Salary pair. If all the other pairs had similar interval then the null hypothesis that there is no differece in varaince would have been accepted, and all the interval would have crossed the zero variance as part of the 95% confidence level. But in this case that does not happen.
- Hence the null hypothesis can be rejected and accept the alternate hypothesis that there exist difference in variance of Occupation type and have different means.
- So, while implementing any strategy Occupation type is an important factor to be considered as it has a impact on the average quarterly balance of deposit account cutomers.

Robust methods one-way ANOVA: As Assumptions are violated

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

```
##
## Call:
## lm(formula = Balance ~ Occupation, data = household)
##
## Coefficients:
```

```
##       (Intercept)       OccupationSAL  OccupationSELF-EMP
##            146952              -23151              -18592
##    OccupationSENP
##             34200
```

```
## 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 of Robust methods:

- One-way analysis of means which assumes that factor levels do not have equal variance, provides significant results. The p-value < 2.2e-16 (significant) tells that atleast one pair in factor levels have different variance.
- The results obtained by Anova Table (Type II tests) assumes the Occupation factor has unbalanced observations for the factor levels, But the results provided by the Type II test confirms the result obtained by the previous two ANOVA methods that the difference in variance is significant inspite of taking into considerations the Occupation factor is unbalanced, White.adjust parameter when set to TRUE use a heteroscedasticity-corrected coefficient covariance matrix. Basic forms of models make use of the assumption that the errors have the same variance across all observation points. When this is not the case, the errors are said to be heteroscedastic, Heteroscedasticity-consistent standard errors are used to allow the fitting of a model that does contain heteroscedastic residuals.
- Since, the results obtained by the robust methods also provides significant results. We reject the null hypothesis and accept the alternate hypothesis that exist a difference in the variance for the average quarterly balance and the means of the occupation types are different from one another.

Two-way ANOVA

Two-way ANOVA test hypotheses for Gender and Occupation

1. There is no difference in means of factor Gender
2. There is no difference in means of factor Occupation
3. There is no interaction between factors Gender and Occupation

The alternative hypothesis for cases 1 and 2 is the means are not equal The alternative hypothesis for case 3 is there is an interaction between Gender and Occupation

Assumptions of two-way ANOVA test

Two-way ANOVA test assumes that the observations within each cell are normally distributed and have equal variances.

```
##
##     PROF  SAL SELF-EMP SENP
##   F 1530  660      874 2461
##   M 3857 5091     2460 2871
##   O   76   88       32    0
##         PROF      SAL SELF-EMP     SENP
## F 194154.44 174860.4 184217.4 210156.0
## M 129761.61 116906.7 108727.4 156289.2
## O  69081.36 139669.8 111960.2       NA
```

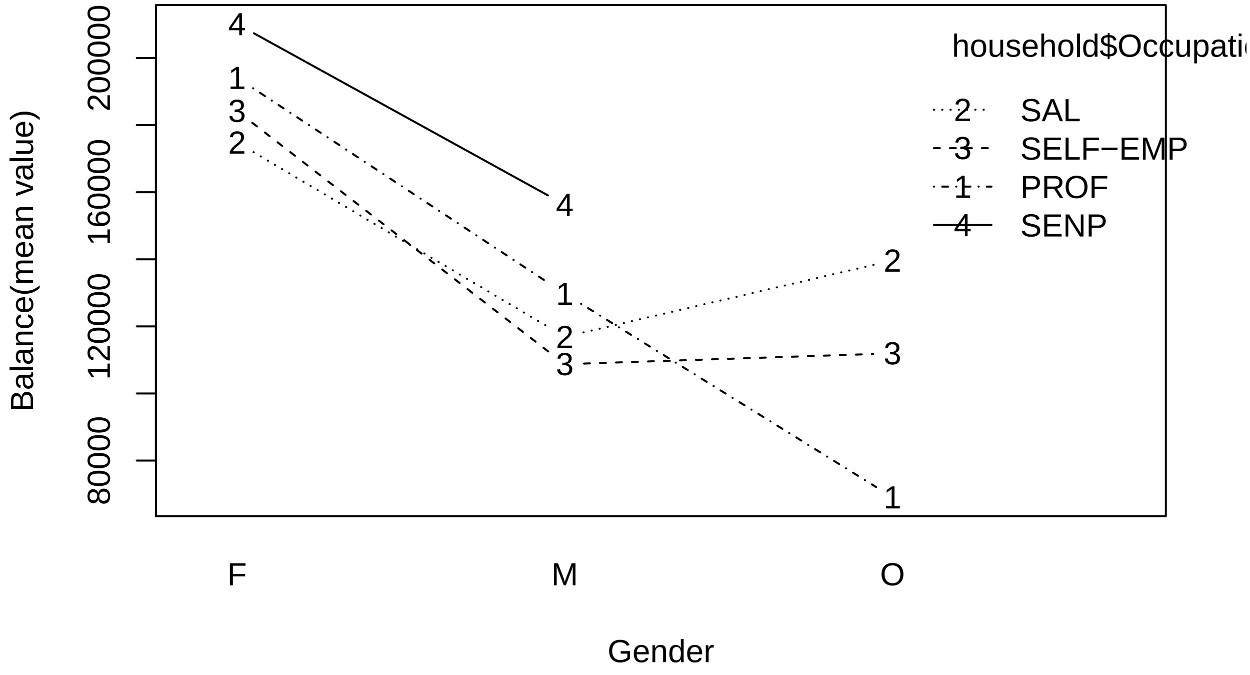
```
## [1] ""
```

```
## [1] "----------------------------------------"
```

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

Observation

- The table value provides an insight the data is unbalanced especially the "O" factor level which has less observations under different employment levels and even has zero observation for SENP (Self-Employed Non-Professionals). The zero observation is treated as NA(Not Available).



Observation

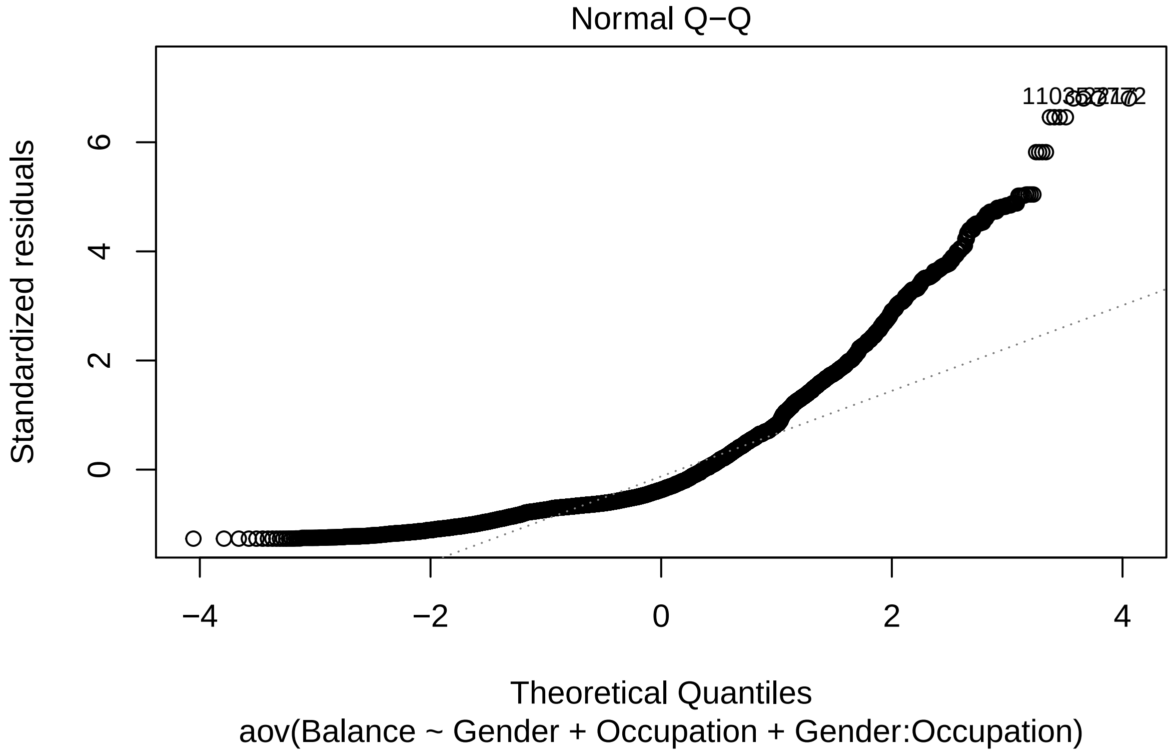
- The Interaction plot shows difference in mean among the Gender, Occupation and interactions.

```
##                      Df    Sum Sq   Mean Sq F value  Pr(>F)
## Gender                2 2.010e+13 1.005e+13 364.176 < 2e-16 ***
## Occupation            3 4.425e+12 1.475e+12  53.449 < 2e-16 ***
## Gender:Occupation     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
```

Results Interpretion:

- the p-value of Gender is < 2e-16 (significant), which indicates that the Genders are associated with significant different Balance amount.
- the p-value of Occupation is < 2e-16 (significant), which indicates that the Occupation type are associated with significant different Balance amount.
- the p-value for the interaction between Gender:Occupation is 0.0022 (significant), which indicates that the relationships between Gender and Occupation depends on the Gender.

Normality test



Anderson-Darling normality test

```
##
##  Anderson-Darling normality test
##
## data:  bal2_residuals
## A = 1041.7, p-value < 2.2e-16
```

Homogeneity of variance test

```
## 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
```

Interpretation of Normality tests and Homogeneity test:

- The Normality plot of the residuals, the normality probability plot of the residuals should approximately follow a straight line. As large number of residuals points do not follow the reference line, we cannot assume normality and the residuals are not normally distributed. And the also the it has some points at the top of the curve which are treated as outliers.
- The Anderson-Darling normality test also provides p-value < 2.2e-16 which confirms the data is not normally distributed and hence the normality assumption is violated.
- The Levene's Test restults reveals that the homogeneity of variance assumption is also violated.

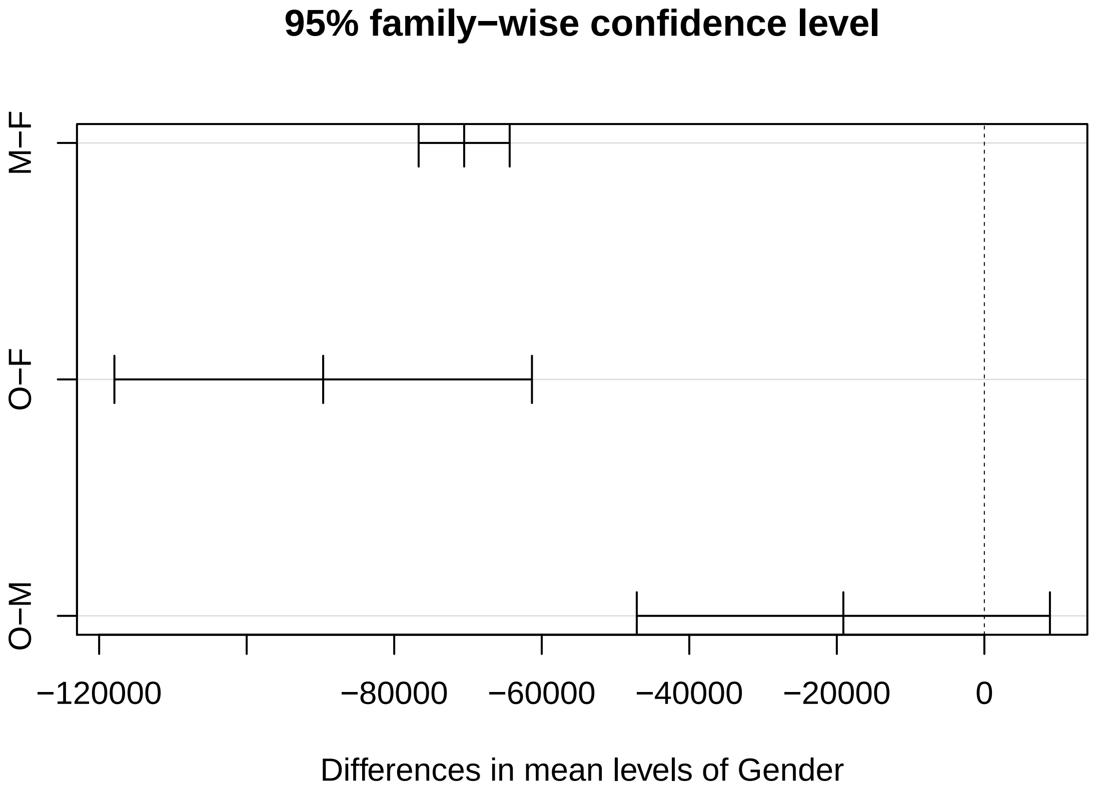
Multiple pairwise comparison between the means of groups

- In ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know which pairs of groups are different. Multiple pariwise comparison can be performed to determine if the mean difference between specific pairs of group are statisfically different

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Balance ~ Gender + Occupation + Gender:Occupation, data = household)
##
## $Occupation
##                    diff       lwr       upr     p adj
## SAL-PROF      -11350.428 -19384.53 -3316.322 0.0016192
## SELF-EMP-PROF -17237.085 -26589.59 -7884.582 0.0000131
## SENP-PROF      21136.106  12919.45 29352.767 0.0000000
## SELF-EMP-SAL   -5886.657 -15123.64  3350.330 0.3575609
## SENP-SAL       32486.534  24401.60 40571.467 0.0000000
## SENP-SELF-EMP  38373.191  28976.99 47769.392 0.0000000
##
## $`Gender:Occupation`
##                              diff         lwr        upr     p adj
## M:PROF-F:PROF          -64392.833  -80797.831 -47987.835 0.0000000
## O:PROF-F:PROF         -125073.083 -188883.781 -61262.386 0.0000000
## F:SAL-F:PROF           -19294.052  -44579.905   5991.801 0.3436234
## M:SAL-F:PROF           -77247.743  -93077.993 -61417.493 0.0000000
## O:SAL-F:PROF           -54484.648 -114006.364   5037.067 0.1108627
## F:SELF-EMP-F:PROF       -9937.073  -32958.860  13084.714 0.9617493
## M:SELF-EMP-F:PROF      -85427.058 -103105.599 -67748.518 0.0000000
## O:SELF-EMP-F:PROF      -82194.280 -179176.705  14788.146 0.1929317
## F:SENP-F:PROF           16001.539   -1675.625  33678.702 0.1210509
## M:SENP-F:PROF          -37865.200  -55051.667 -20678.733 0.0000000
## O:SENP-F:PROF                  NA          NA         NA        NA
## O:PROF-M:PROF          -60680.250 -123573.435   2212.935 0.0705829
## F:SAL-M:PROF            45098.782   22226.921  67970.642 0.0000000
## M:SAL-M:PROF           -12854.910  -24445.615  -1264.205 0.0152781
## O:SAL-M:PROF             9908.185  -48628.831  68445.201 0.9999931
## F:SELF-EMP-M:PROF       54455.761   34114.907  74796.614 0.0000000
## M:SELF-EMP-M:PROF      -21034.225  -35044.166  -7024.284 0.0000596
## O:SELF-EMP-M:PROF      -17801.446 -114182.661  78579.768 0.9999831
## F:SENP-M:PROF           80394.372   66386.169  94402.575 0.0000000
## M:SENP-M:PROF           26527.633   13143.976  39911.291 0.0000000
## O:SENP-M:PROF                  NA          NA         NA        NA
## F:SAL-O:PROF           105779.031   40008.193 171549.870 0.0000096
## M:SAL-O:PROF            47825.340  -14920.381 110571.062 0.3453523
## O:SAL-O:PROF            70588.435  -14436.640 155613.510 0.2194253
## F:SELF-EMP-O:PROF      115136.011   50201.956 180070.065 0.0000004
## M:SELF-EMP-O:PROF       39646.025  -23591.302 102883.352 0.6593423
## O:SELF-EMP-O:PROF       42878.804  -71541.560 157299.168 0.9871046
## F:SENP-O:PROF          141074.622   77837.680 204311.564 0.0000000
## M:SENP-O:PROF           87207.883   24106.351 150309.416 0.0003919
## O:SENP-O:PROF                  NA          NA         NA        NA
## M:SAL-F:SAL            -57953.691  -80416.880 -35490.503 0.0000000
## O:SAL-F:SAL            -35190.596  -96809.040  26427.848 0.7801076
## F:SELF-EMP-F:SAL         9356.979  -18643.034  37356.992 0.9950594
## M:SELF-EMP-F:SAL       -66133.007  -89934.873 -42331.140 0.0000000
## O:SELF-EMP-F:SAL       -62900.228 -161183.436  35382.980 0.6288024
## F:SENP-F:SAL            35295.590   11494.747  59096.434 0.0000805
## M:SENP-F:SAL           -18571.148  -42009.849   4867.552 0.2854352
## O:SENP-F:SAL                   NA          NA         NA        NA
## O:SAL-M:SAL              22763.095  -35615.455  81141.645 0.9822955
## F:SELF-EMP-M:SAL        67310.670   47430.450  87190.890 0.0000000
## M:SELF-EMP-M:SAL        -8179.315  -21511.655   5153.024 0.6901015
## O:SELF-EMP-M:SAL        -4946.537 -101231.589  91338.516 1.0000000
## F:SENP-M:SAL            93249.282   79918.768 106579.795 0.0000000
## M:SENP-M:SAL            39382.543   26709.930  52055.156 0.0000000
## O:SENP-M:SAL                   NA          NA         NA        NA
## F:SELF-EMP-O:SAL        44547.575  -16176.892 105272.043 0.4074842
## M:SELF-EMP-O:SAL       -30942.410  -89849.023  27964.203 0.8608972
## O:SELF-EMP-O:SAL       -27709.631 -139794.640  84375.377 0.9996908
## F:SENP-O:SAL            70486.187   11579.987 129392.387 0.0052347
## M:SENP-O:SAL            16619.448  -42141.363  75380.260 0.9988935
## O:SENP-O:SAL                   NA          NA         NA        NA
## M:SELF-EMP-F:SELF-EMP  -75489.986  -96871.219 -54108.752 0.0000000
## O:SELF-EMP-F:SELF-EMP  -72257.207 -169982.420  25468.006 0.3945523
## F:SENP-F:SELF-EMP       25938.611    4558.517  47318.706 0.0042222
## M:SENP-F:SELF-EMP      -27928.127  -48904.328  -6951.926 0.0008288
## O:SENP-F:SELF-EMP              NA          NA         NA        NA
## O:SELF-EMP-M:SELF-EMP    3232.779  -93373.357  99838.914 1.0000000
## F:SENP-M:SELF-EMP      101428.597   85948.397 116908.797 0.0000000
## M:SENP-M:SELF-EMP       47561.858   32644.448  62479.268 0.0000000
## O:SENP-M:SELF-EMP              NA          NA         NA        NA
## F:SENP-O:SELF-EMP       98195.818    1589.935 194801.702 0.0422783
## M:SENP-O:SELF-EMP       44329.080  -52188.220 140846.380 0.9407042
## O:SENP-O:SELF-EMP              NA          NA         NA        NA
## M:SENP-F:SENP          -53866.739  -68782.516 -38950.961 0.0000000
## O:SENP-F:SENP                  NA          NA         NA        NA
## O:SENP-M:SENP                  NA          NA         NA        NA
```

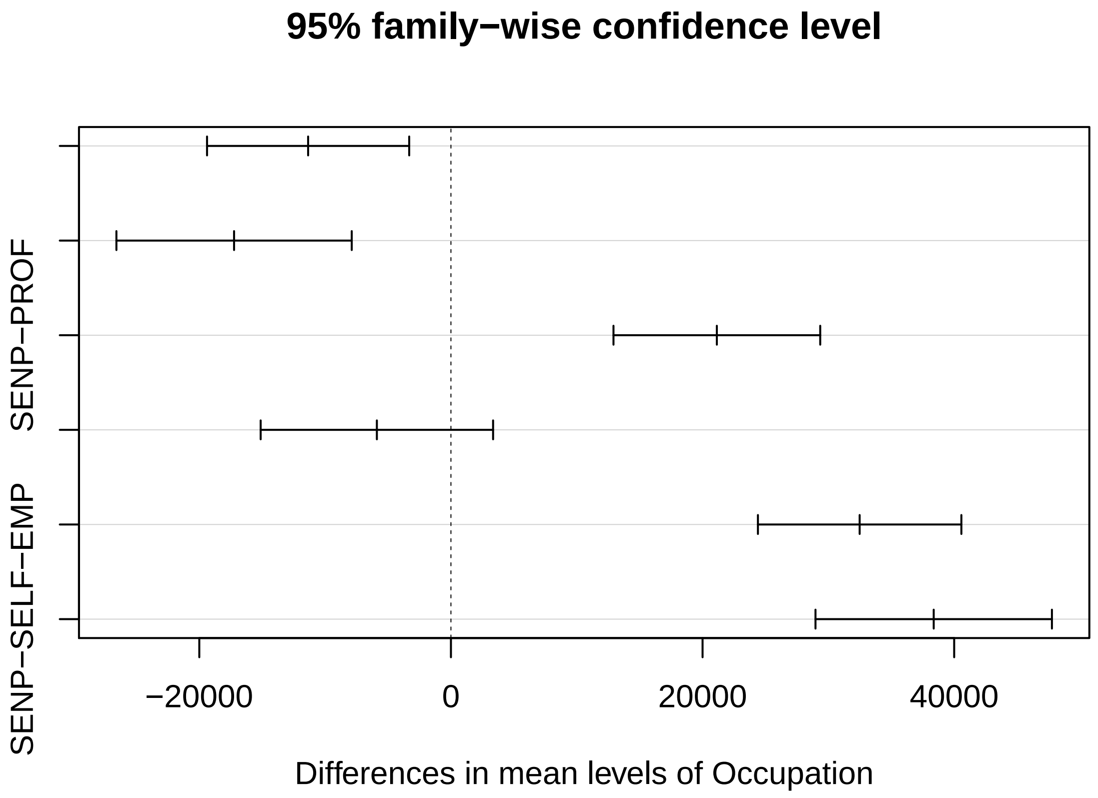
Plot for Gender given by TukeyHSD of two-way ANOVA of Gender and Occupation

```
plot(TukeyHSD(bal2, which = c("Gender")))
```



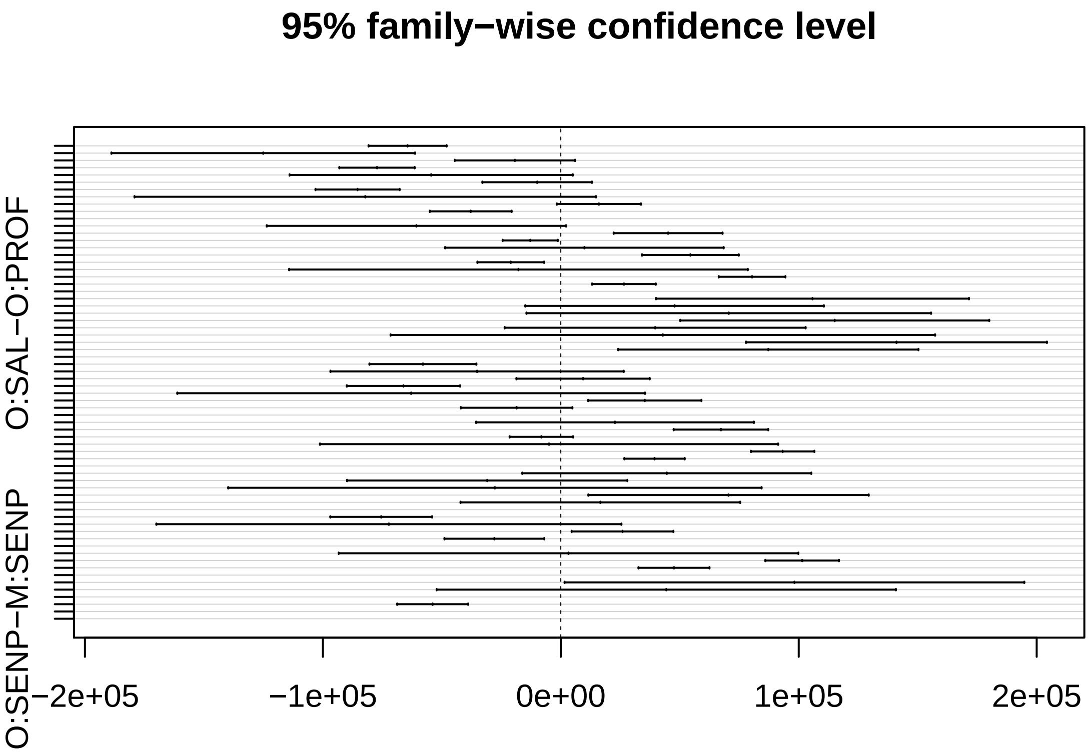
Plot for Occupation given by TukeyHSD of two-way ANOVA of Gender and Occupation

```
plot(TukeyHSD(bal2, which = c("Occupation")))
```



Plot for Gender:Occupation interaction effect given by TukeyHSD of two-way ANOVA of Gender and Occupation

```
plot(TukeyHSD(bal2, which = c("Gender:Occupation")))
```



Interpretation of TukeyHSD results:

- diff: difference between means of two groups
- lwr, upr: the lower and upper end point of the confidence interval at 95%
- p adj: p-value after adjusting for multiple comparisons
- The Self Employed and Salaried factor levels in the Occupation factor has equal variance and all other factor levels have different variances.
- The plots generated using TukeyHSD give significant results all the pairs in Gender and Occupation except the “M-O” pair in Gender and SELF-EMP-SAL pair in Occupation which shows that the variance between those pairs are insignificant and have equal mean confirmed by the 95% confidence level which has 0 mean value as part of the confidence level.
- But whereas for the interaction between Gender and Occupation where large amount of pairs which have insignificant variance among them as most the 95% confidence levels have 0 mean values in the intervals. So, taking call on the importance of interaction between Gender and Occupation will be a subjective call, and moreover results from robust methods can looked to take a scientific conclusion.
- Most of the Interaction effect are not signficant for example O:SAL-M:PROF with p-value close to 1 and NA is for the SENP factor level of employment factor which has zero observation. Largely as there is no significant interaction type II approach can be considered for robust analysis

Robust Two-way ANOVA methods

Type-II robust method

```
## Anova Table (Type II tests)
##
## Response: Balance
##                       Sum Sq    Df  F value    Pr(>F)
## Gender            1.4009e+13     2 253.7972 < 2.2e-16 ***
## Occupation        4.4253e+12     3  53.4493 < 2.2e-16 ***
## Gender:Occupation 5.1450e+11     5   3.7285  0.002247 **
## Residuals         5.5166e+14 19989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type-III robust method

```
## Anova Table (Type III tests)
##
## Response: Balance
##                       Sum Sq    Df  F value    Pr(>F)
## (Intercept)       6.3884e+12     1 231.4806 < 2.2e-16 ***
## Gender            1.2865e+13     2 233.0781 < 2.2e-16 ***
## Occupation        1.3857e+12     3  16.7369 7.427e-11 ***
## Gender:Occupation 5.1450e+11     5   3.7285  0.002247 **
## Residuals         5.5166e+14 19989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of Robust methods:

- When data is unbalanced, there are different ways to calculate the sums of squares for ANOVA. There are at least 3 approaches, commonly called Type I, II and III sums of squares. The aov offers SS type I and is for balanced factor, and Anova-Type II tests when the data is unbalanced. So, when the data is unbalanced the results provided by each method will be slightly different which can be observed comparing the F-statistic given by different methods even though p-value provided may be close, and if yet both the methods provide significant results then at confidence levels at which both the tests were conducted, it can be concluded that there is a signficant difference in the variance among factor levels for the tested factor.
- As in this case the Gender factor was tested for equivalence of variance and results from all the methods tested so far provide significant results, with varying F-statistic because of the violation of assumptions and unbalanced data, it can be concluded at 95% confidence level beyond any doubt that there is a difference in variance between factor levels in the Gender factor, Occupation and Interaction between Gender and Occupation.
- For type II approach that no significant interaction is assumed, due to the way in which the SS are calculated when incorporating the interaction effect, for type III you must specify the contrasts option to obtain sensible results and singula.ok is changed default False to True. Type III approach is this type tests for the presence of a main effect after the other main effect and interaction. This approach is therefore valid in the presence of significant interactions. However, it is often not interesting to interpret a main effect if interactions are present.

- Type-II approach calculates according to the principle of marginality, testing each term after all others, except ignoring the term's higher-order relatives; so-called type-III tests violate marginality, testing each term in the model after all of the others.
- The results given by Type II and Type III provides statistical significant and confirms that there exists difference in variance between the Gender factor and Occupation factor and interaction effect is also significant with p-value 0.00225 at 95% confidence level. Only difference being for the Occupation's main effect from Type III get a p-value of 7.4e-11 as against < 2e-16 given by Type II method.
- Hence, based upon the results obtained from all the three methods, it can be concluded at 95% confidence level beyond any doubt that there is a difference in variance between factor levels in the Gender factor, Occupation and Interaction between Gender and Occupation. And the means are different, to reject the null hypothesis and accept the alternate hypothesis. So, while forming strategies the Gender, Occupation and interaction between Gender and Occupation should be considered as an important factors.

Principal Component Analysis

```
## [1] "Batting Data Summary"
```

```
##       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
```

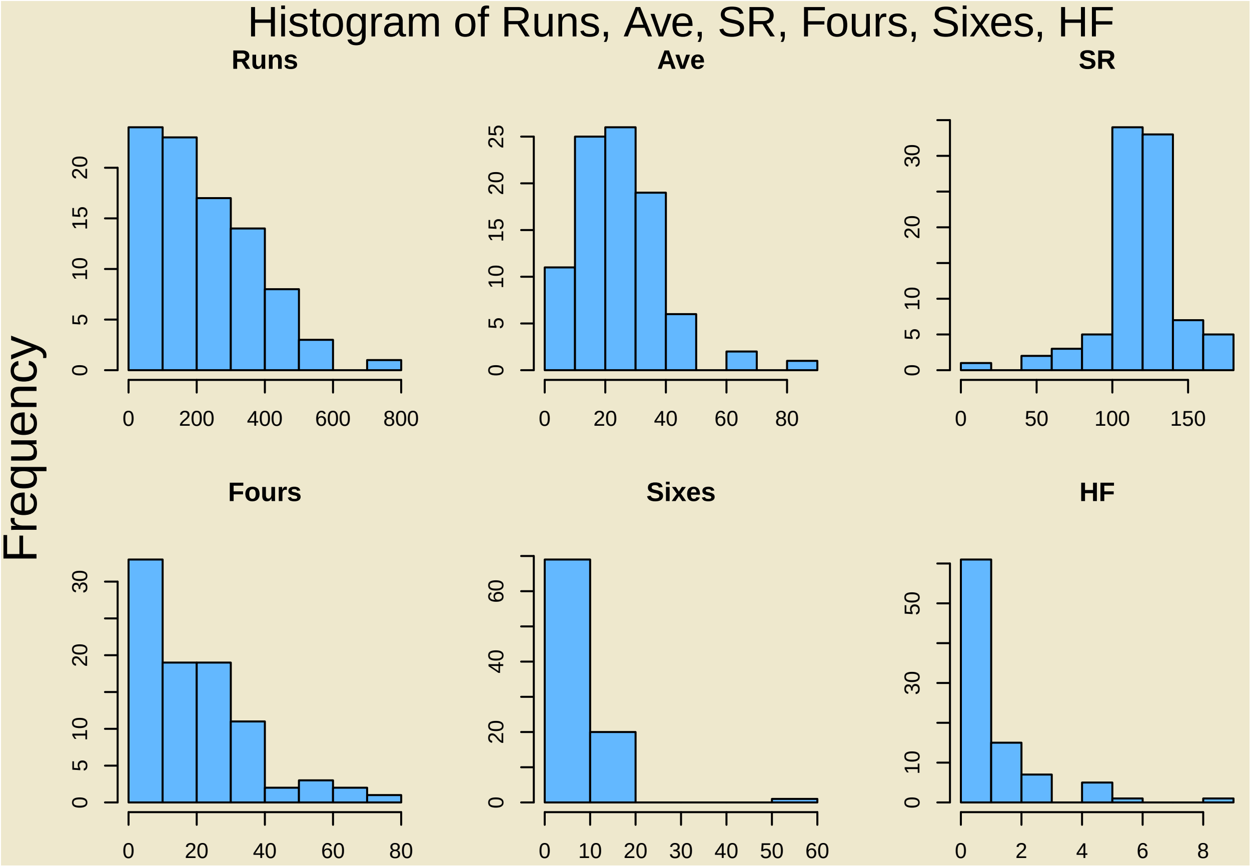
```
## [1] "Bowling Data Summary"
```

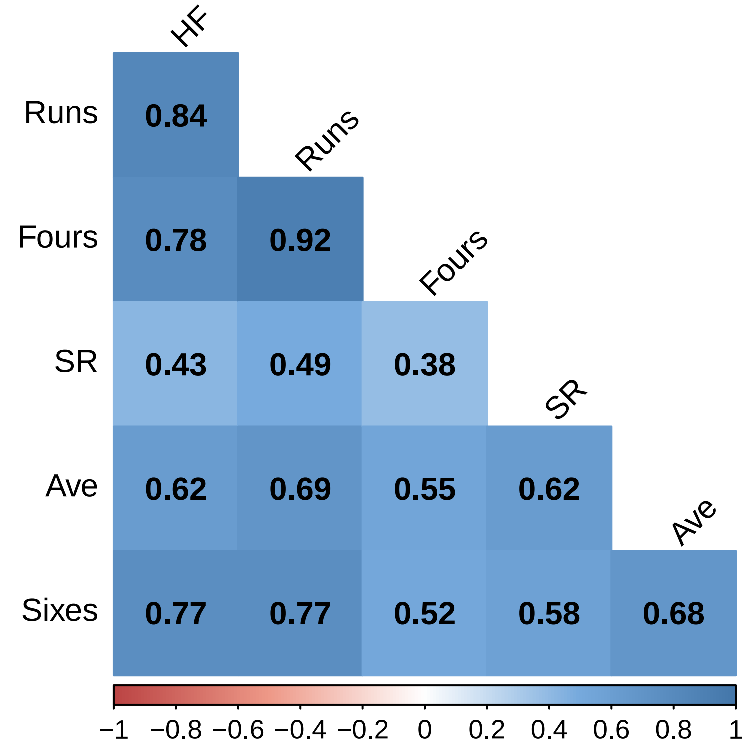
```
##       Wkts            Ave              Econ             SR
##  Min.   : 1.00   Min.   : 12.20   Min.   : 5.400   Min.   :12.00
##  1st Qu.: 5.00   1st Qu.: 22.32   1st Qu.: 6.950   1st Qu.:17.25
##  Median : 8.00   Median : 29.00   Median : 7.530   Median :21.60
##  Mean   : 8.88   Mean   : 34.51   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
```

Observation

- From the summary each variable in batting and bowling is normally distributed, except the half centuries, sixes, fours and ballers' strike rate variables which is slightly right skewed.

Visualization of Batting variables

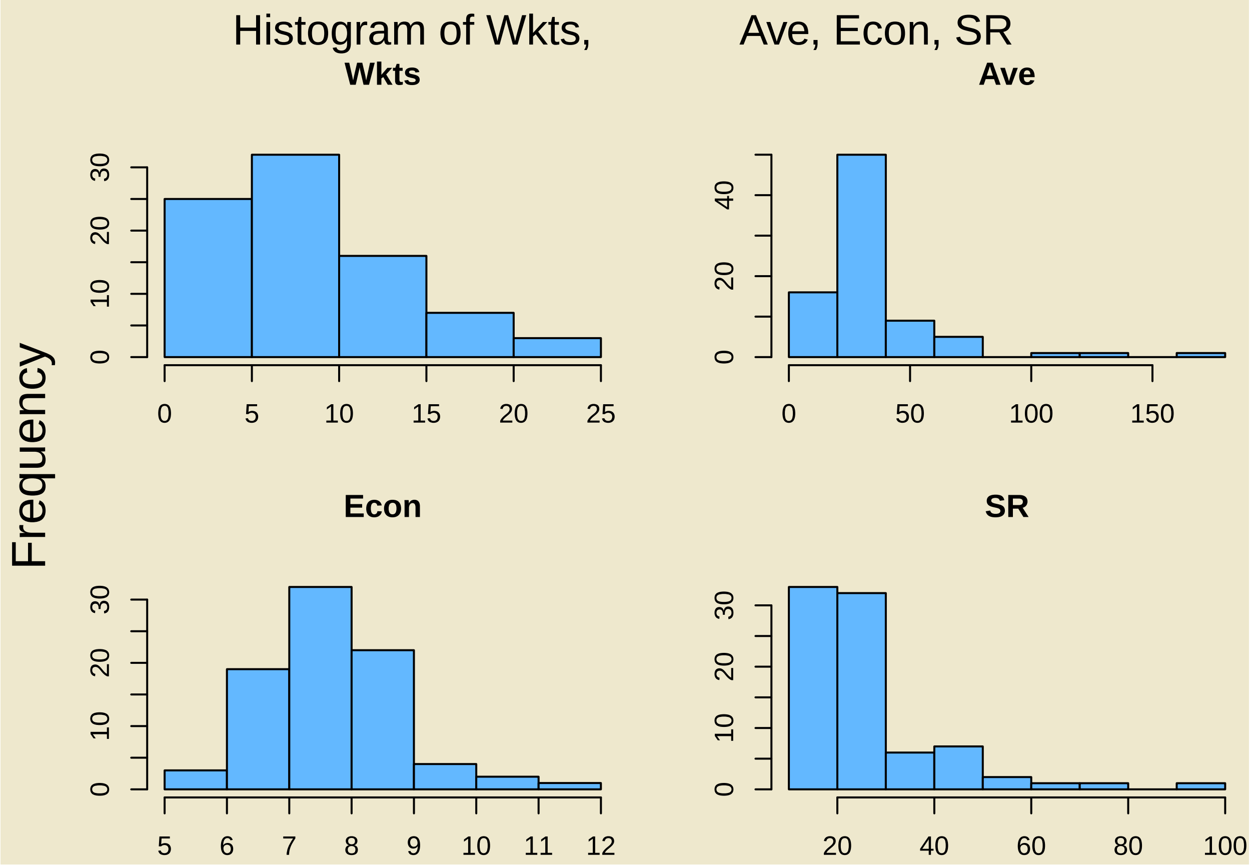


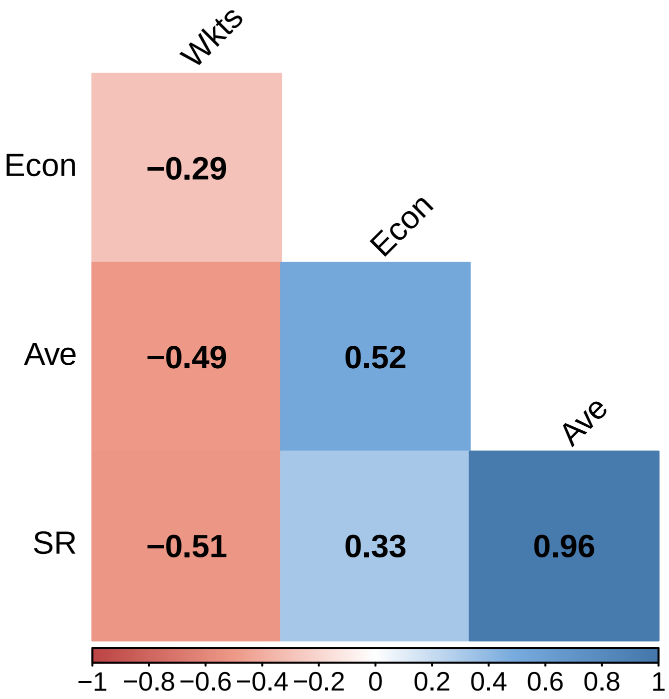


Observation from batting variables plots:

- The first plot shows individual histograms for our “batting variables” based on the performance of the ninety batsmen who played in the 2012 IPL season (as per the dataset).
- Second plot depicts a matrix plot, and we see some significant correlations between these variables. For example, the plot shows that Runs and Fours, Runs and Sixes, and Ave and Runs are considerably correlated, as might be anticipated. This shows the necessity of using a technique which is capable of handling correlated data in any reasonable attempt to study batting performance. High values for each of these variables indicates better batting performance in a univariate sense, and each one measures a different quality of a batsman, but their joint contribution to batting performance in a multivariate sense needs to measured.
- Constructing an overall measure of batting performance by collapsing these correlated variables is possible by using Principal Component Analysis.

Visualization of Balling variables





Observations from balling variables plots

- First plot shows individual histograms for our "bowling variables" based on the performance of the eighty-three bowlers who bowled in the 2012 IPL season (as per the dataset).
- The correlation lower matrix of the four "bowling variables", the 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.
- An overall measure of performance by using these correlated variables can be achieved using Principal Component Analysis technique.

Performing PCA to measure performance

- Principal Component Analysis (PCA) is a nonparametric variable reduction technique well-suited for correlated data that can be effectively used in this 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.
- PCA is particularly useful when data on a number of useful variables has been gathered, and it is probable that there is some redundancy in those variables. Here, redundancy is taken to mean that our cricket performance variables are correlated with one another because, in some unknown sense, they might be measuring similar player-performance attributes. PCA aims to reduce the observed variables down to a smaller number of principal components which account for most of the variation

occurring in the originally observed variables. These can be utilized to provide summarized measures of performance.

```
## [1] "Batting Data-Bartlett Test:"

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

## [1] "Bowling Data-Bartlett Test:"

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

Interpretation of Bartlett test

- The null hypothesis is that the data dimension reduction is not possible. If p-value is less than 0.05, dimension reduction is possible.
- As the p-value is less than 0.05 for both the batting and bowling variables which tells that the variables are correlated with each other and PCA can be used to perform dimension reduction.

Ranking Batsmen using the Principal Components

```
## [1] "Correlation Matrix for 90 Batsmen"

##            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 of correlation matrix

- Correlation matrix associated with the batting vectors may be examined for the correlation structure inherent in these variables.
- Variables are measured on very different scales, they must be standardized before PCA analysis. However, the process of finding the principal components by using the standardized variables is equivalent to finding principal components by using the correlation matrix.

Labeling PC's eingenvectors

```
##         PC1    PC2     PC3     PC4     PC5     PC6
## [1,] -0.458  0.266  0.1098  0.0052 -0.4584  0.7048
## [2,] -0.398 -0.331 -0.0055 -0.8474  0.1012 -0.0606
## [3,] -0.325 -0.698  0.4501  0.4328  0.1189  0.0562
## [4,] -0.406  0.474  0.5082  0.0325 -0.0968 -0.5851
## [5,] -0.417 -0.179 -0.6694  0.2488 -0.3946 -0.3579
## [6,] -0.432  0.276 -0.2808  0.1781  0.7749  0.1610
```

```
## [1] 4.2547 0.8271 0.4120 0.3255 0.1638 0.0169
```

The values given above are the eigenvalues and eigenvectors of batting variables

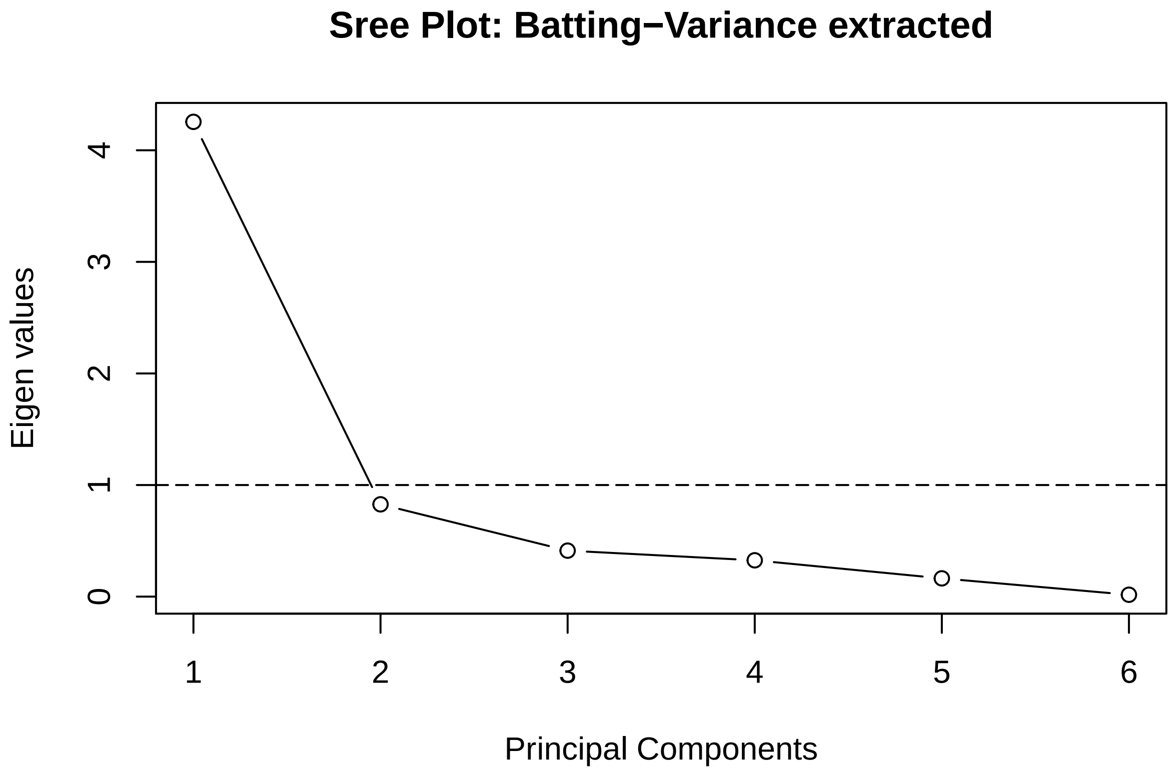
- Eigenvalues measure the magnitude of the vectors and eigenvectors represent the direction of the Principal Component
- So the PC1 will point in the direction, eigenvector of PC1, of highest variance to the extent given by the eigenvalue which is largest of all at 4.255

```
##         PC1    PC2      PC3      PC4     PC5      PC6
## [1,] -0.945  0.242  0.07047  0.00297 -0.1855  0.09156
## [2,] -0.821 -0.301 -0.00353 -0.48342  0.0410 -0.00788
## [3,] -0.671 -0.635  0.28894  0.24688  0.0481  0.00731
## [4,] -0.837  0.431  0.32623  0.01855 -0.0392 -0.07601
## [5,] -0.861 -0.163 -0.42970  0.14193 -0.1597 -0.04649
## [6,] -0.892  0.251 -0.18026  0.10162  0.3136  0.02091
```

Intrepretation of PCA loadings

- The loadings are correlations between the principal components and the individual variables, the correlation between the PC1 and v1(Runs) is highest which tells that runs variable is given highest importance in PC1.

```
## Warning in plot.xy(xy, type, ...): plot type 'both' will be truncated to
## first character
```



```
## [1] 70.912 13.785  6.867  5.424  2.731  0.281
```

```
## [1]  70.9  84.7  91.6  97.0  99.7 100.0
```

Interpretation of Sree plot

- Almost 71% of the Total Variability can be explained by this first principal component. Moreover, its corresponding eigenvalue 4.255 is the only one which is greater than 1.
- Only PC which has eigenvalue greater than 1 is PC1 and so retaining only PC1 to rank the batsmen as the number of useful principal components is then taken to be the abscissa of the point beyond which all remaining eigenvalues add relatively small contributions to the total variability.
- Scree plot is suggesting that since the elbow is at abscissa two, it is reasonable to use only the first principal component which explains 71% of the total variability.

```
##        PC1    PC2      PC3      PC4     PC5      PC6
## [1,] 0.894 0.0587 4.97e-03 8.81e-06 0.03443 8.38e-03
## [2,] 0.674 0.0907 1.25e-05 2.34e-01 0.00168 6.20e-05
## [3,] 0.450 0.4027 8.35e-02 6.10e-02 0.00232 5.34e-05
## [4,] 0.700 0.1855 1.06e-01 3.44e-04 0.00153 5.78e-03
## [5,] 0.741 0.0265 1.85e-01 2.01e-02 0.02551 2.16e-03
## [6,] 0.795 0.0630 3.25e-02 1.03e-02 0.09837 4.37e-04
```

Interpretatio of communalit table

- The total variance extracted by all PC's for six variables given by the rows, and PC1 extracts maximum variance from all the variables and also PC1 has eigenvalue greater than one. So it is best to retain first

Principal Component.
- The first principal component can be referred 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 all positive, so larger values of principal component scores obtained indicate better player performance. This justifies that we should rank (largest to smallest) the players based on the first principal component.

Obtaining PCA scores

```
##     PC1 PC2 PC3 PC4 PC5 PC6
## PC1   1   0   0   0   0   0
## PC2   0   1   0   0   0   0
## PC3   0   0   1   0   0   0
## PC4   0   0   0   1   0   0
## PC5   0   0   0   0   1   0
## PC6   0   0   0   0   0   1
```

```
##    PC1     PC2    PC3     PC4    PC5     PC6
## 1 8.47  0.3526  3.652 -0.9907  0.132  0.2391
## 2 4.59 -1.4685 -0.501 -0.5912 -0.609 -0.1330
## 3 4.12 -0.4688 -0.554 -1.0950 -0.432  0.1713
## 4 3.88 -0.0416  0.261 -0.2268 -0.544 -0.2083
## 5 4.10 -1.3932 -0.116 -0.0146 -0.156 -0.0624
## 6 4.00 -2.0138 -1.239  0.1700 -0.482  0.1534
```

Interpreation

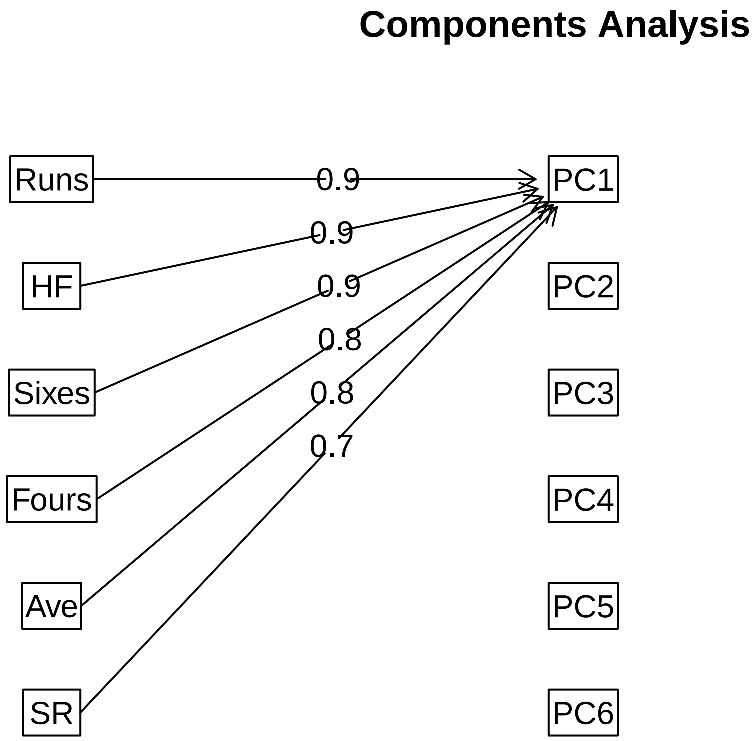
- The correlation between each of the PC's is zero for scores extracted as expected given that PC's are orthogonal to 1st PC and describe maximum variance that remains after the variance explained by the first PC is removed.

```
##         PC1    PC2      PC3      PC4     PC5      PC6
## Runs  0.945 -0.242 -0.07047 -0.00297  0.1855 -0.09156
## Ave   0.821  0.301  0.00353  0.48342 -0.0410  0.00788
## SR    0.671  0.635 -0.28894 -0.24688 -0.0481 -0.00731
## Fours 0.837 -0.431 -0.32623 -0.01855  0.0392  0.07601
## Sixes 0.861  0.163  0.42970 -0.14193  0.1597  0.04649
## HF    0.892 -0.251  0.18026 -0.10162 -0.3136 -0.02091
```

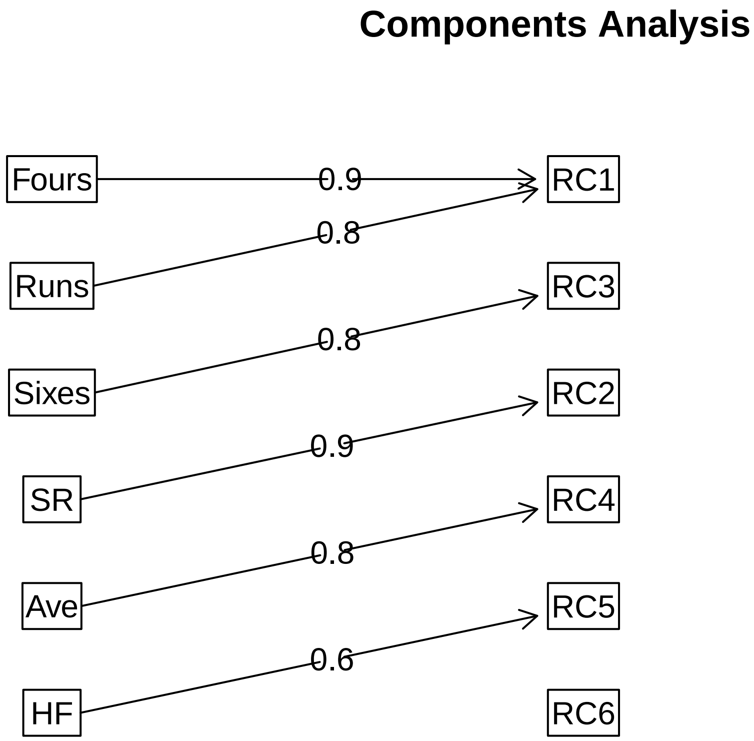
Interpretation

- The given values are the correlation between the batting data variables and Pc's scores which is same as the loadings. And loadings are correlations between the principal components and the individual variables

checking if Rotation is required



Loadings after varimax rotation



Conclusion for requirement of rotation

- Before performing any rotation the latent factors all load into a single PC1 which is also confirmed by results obtained by using Scree plot. After rotation factor loading has increased into five factors, as the purpose for performing PCA is to reduce the dimension the rotation doesn't help in achieving that purpose.
- Hence, after looking at the loadings before and after rotation it is concluded that roation is not required.

Ranks of batsmen

```
##              Name Runs  Ave  SR Fours Sixes HF PC1 Score
## 1        CH Gayle  733 61.1 161    46    59  9      8.47
## 2       G Gambhir  590 36.9 144    64    17  6      4.59
## 3        V Sehwag  495 33.0 161    57    19  5      4.12
## 5        S Dhawan  569 40.6 130    58    18  5      4.10
## 6       AM Rahane  560 40.0 129    73    10  5      4.00
## 4        CL White  479 43.5 150    41    20  5      3.88
## 8       RG Sharma  433 30.9 127    39    18  5      2.90
## 7    KP Pietersen  305 61.0 147    22    20  3      2.86
## 9  AB de Villiers  319 39.9 161    26    15  3      2.31
## 13   F du Plessis  398 33.2 131    29    17  3      2.11
```

Converted to two columns-list for reporting list of ranked Bowlers

```
##      Batsman Name Rank PC1 Score  Batsman Name Rank PC1 Score
## 1        CH Gayle    1      8.47  JEC Franklin   46    -0.540
## 2       G Gambhir    2      4.59   LRPL Taylor   47    -0.540
## 3        V Sehwag    3      4.12     MK Pandey   48    -0.558
## 4        S Dhawan    4      4.10    SC Ganguly   49    -0.577
## 5       AM Rahane    5      4.00    KD Karthik   50    -0.580
## 6        CL White    6      3.88 KC Sangakkara   51    -0.599
## 7       RG Sharma    7      2.90     RA Jadeja   52    -0.640
## 8    KP Pietersen    8      2.86    AL Menaria   53    -0.705
## 9  AB de Villiers    9      2.31  DT Christian   54    -0.722
## 10   F du Plessis   10      2.11     SS Tiwary   55    -0.753
```

Ranking Bowlers using the Principal Components

```
## [1] "Correlation Matrix for 83 Bowlers"
```

```
##        Wkts    Ave   Econ     SR
## Wkts  1.000 -0.491 -0.292 -0.512
## Ave  -0.491  1.000  0.523  0.963
## Econ -0.292  0.523  1.000  0.328
## SR   -0.512  0.963  0.328  1.000
```

Intrepretation of correlation matrix

- Correlation matrix associated with the bowling vectors may be examined for the correlation structure inherent in these variables.
- Variables are measured on very different scales, they must be standardized before PCA analysis. However, the process of finding the principal components by using the standardized variables is equivalent to finding principal components by using the correlation matrix.

Labeling PC's vectors

```
##         PC1     PC2    PC3     PC4
## [1,]  0.428 -0.3349  0.838  0.0382
## [2,] -0.591  0.0476  0.354 -0.7232
## [3,] -0.383 -0.8916 -0.168  0.1724
## [4,] -0.566  0.3010  0.379  0.6677
```

```
## [1] 2.6161 0.7516 0.6202 0.0121
```

The values given above are the eigenvalues and eigenvectors of batting variables

- Eigenvalues measure the magnitude of the vectors and eigenvectors represent the direction of the Principal Component
- So the PC1 will point in the direction, eigenvector of PC1, of highest variance to the extent given by the eigenvalue which is largest of all at 2.616
- The vector values of PC1 except for first vector point which presents wickets is negative, because the bowler performance measured as good performance for higher numbers of wickets and lower values for Average(Ave), Economy(Econ) and Strike Rate(SR)

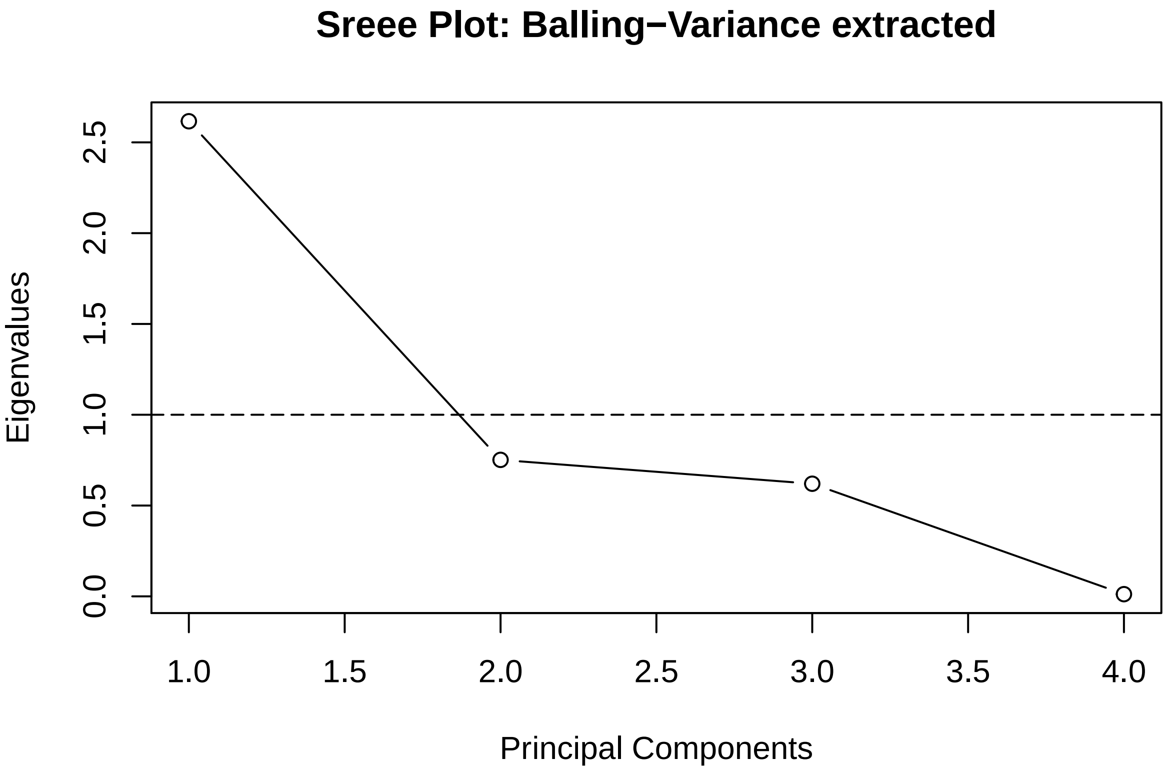
Obtaining PCA loadings

```
##         PC1     PC2    PC3      PC4
## [1,]  0.693 -0.2903  0.660  0.00421
## [2,] -0.956  0.0413  0.279 -0.07971
## [3,] -0.620 -0.7730 -0.132  0.01900
## [4,] -0.915  0.2609  0.298  0.07359
```

Intrepretaion of PCA loadings

- The loadings are correlations between the principal components and the individual variables, the correlation between the PC1 and v1(Runs) is highest which tells that runs variable is given highest importance in PC1.

```
## Warning in plot.xy(xy, type, ...): plot type 'both' will be truncated to
## first character
```



```
## [1] 65.402 18.790 15.505  0.304
```

```
## [1]  65.4  84.2  99.7 100.0
```

Interpretation of Sree plot

- Almost 65.4% of the Total Variability can be explained by this first principal component. Moreover, its corresponding eigenvalue 2.616 is the only one which is greater than 1.

- Only PC which has eigenvalue greater than 1 is PC1 and so retaining only PC1 to rank the batsmen as the number of useful principal components is then taken to be the abscissa of the point beyond which all remaining eigenvalues add relatively small contributions to the total variability.
- Scree plot is suggesting that since the elbow is at abscissa two, it is reasonable to use only the first principal component which explains 65.4% of the total variability.

obtaining commulities

```
##        PC1     PC2    PC3      PC4
## [1,] 0.480 0.08429 0.4360 1.77e-05
## [2,] 0.914 0.00171 0.0777 6.35e-03
## [3,] 0.385 0.59752 0.0175 3.61e-04
## [4,] 0.838 0.06809 0.0890 5.42e-03
```

Interpretation of communality table

- The total variance extracted by all PC's for four variables given by the rows, and PC1 extracts maximum variance from all the variables, expect variable 3 for which maximum variance is extracted by PC2 and also PC1 has eigenvalue greater than one. So it is best to retain only first Principal Component.

```
##     PC1 PC2 PC3 PC4
## PC1   1   0   0   0
## PC2   0   1   0   0
## PC3   0   0   1   0
## PC4   0   0   0   1
```

```
##      PC1    PC2   PC3    PC4
## 1  0.806  0.620 0.945 0.0618
## 2 -0.678  0.973 0.753 0.1862
## 3  2.268 -0.875 1.992 0.0200
## 4  1.417 -0.597 1.234 0.0293
## 5  1.072 -0.535 1.001 0.0381
## 6 -1.423  0.414 0.694 0.1515
```

Interpreation

- The correlation between each of the PC's is zero for scores extracted as expected given that PC's are orthogonal to 1st PC and describe maximum variance that remains after the variance explained by the first PC is removed.
- Since PC1 score is a representation of maximum variance of the four variables, it is reasonable to use the first principle component for ranking. In this context, PC1 score can reasonably be described as the bowling performance index.

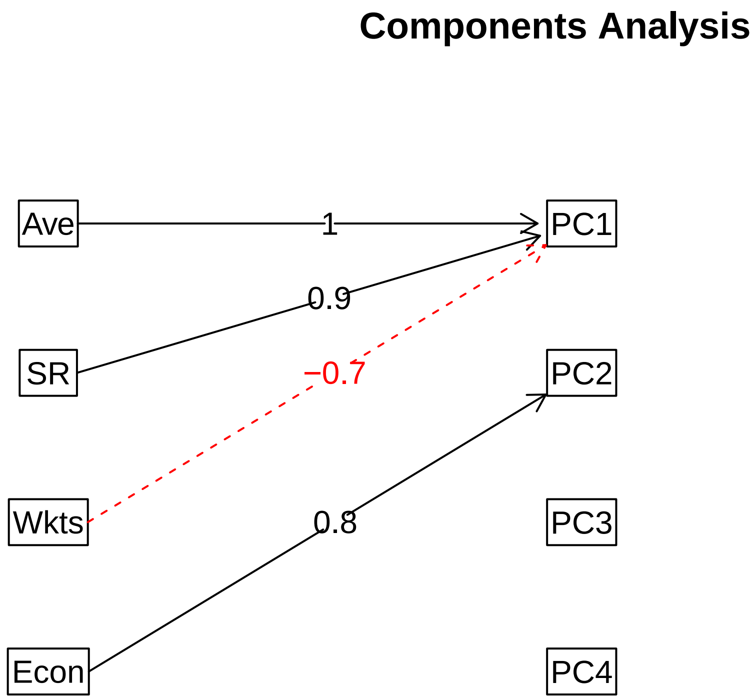
```
##         PC1     PC2    PC3      PC4
## Wkts  0.693 -0.2903  0.660  0.00421
## Ave  -0.956  0.0413  0.279 -0.07971
## Econ -0.620 -0.7730 -0.132  0.01900
## SR   -0.915  0.2609  0.298  0.07359
```

Interpretation

- The given values are the correlation between the balling data variables and Pc's scores which is same the loadings. And loadings are correlations between the principal components and the individual variables

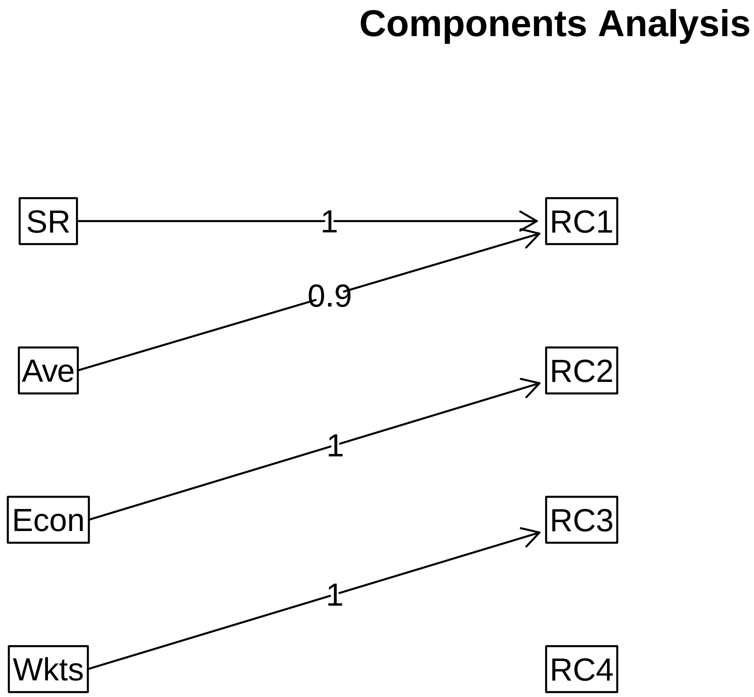
checking if Rotation is required

```
ball_rotation <- principal(pca_ballvar, nfactors = 4, rotate = "none")
fa.diagram(ball_rotation)
```



Loadings after varimax rotation

```
ball_rotated <- principal(pca_ballvar, nfactors = 4, rotate = "varimax")
fa.diagram(ball_rotated)
```



Conclusion for requirement of rotation

- Before performing any rotation the three factors load into PC1 and one into PC2 which is also confirmed by results obtained by using Scree plot. But after performing rotation loadings has increased into three factors with perfect loading of 1 which tells that it is overloading into those factors, as the purpose for performing PCA is to reduce the dimension the rotation doesn't help in achieving that purpose.
- Hence, after looking at the loadings before and after rotation it is concluded that roation is not required.

Ranks of Bowlers

With attached PC1 scores

```
##        Name Wkts  Ave Econ   SR PC1 Score
## 1  R Ashwin   14 30.8 6.54 28.2     0.806
## 2   P Kumar    9 48.2 6.88 42.0    -0.678
## 3  M Morkel   25 18.1 7.19 15.1     2.268
## 4  UT Yadav   19 23.8 7.42 19.2     1.417
## 5    Z Khan   17 26.6 7.55 21.1     1.072
## 6 IK Pathan    8 58.1 7.75 45.0    -1.423
```

Sorted as per PC1 Score

```
## [1] "Sorted list as per PC1 Scores"
##              Name Wkts  Ave Econ   SR PC1 Score
## 7       SP Narine   24 13.5 5.47 14.7      2.92
## 11     SL Malinga   22 15.9 6.30 15.1      2.40
## 3        M Morkel   25 18.1 7.19 15.1      2.27
## 21       DW Steyn   18 15.8 6.10 15.5      2.14
## 41       L Balaji   11 14.7 5.40 16.3      1.84
## 29 M Muralitharan   15 17.3 6.50 16.0      1.71
```

Converted to two columns-list for reporting list of ranked Bowlers

```
bowlers_rank_list <- cbind(as.data.frame(sorted_ball$Name), seq(1,nrow(sorted_ball)), sorted_ball$`PC1 S
#sorted_ball$Name
colnames(bowlers_rank_list) <- c("Bowler Name","Rank","PC1 Score")
bowlers_rank_list <- rbind(bowlers_rank_list, c("James Peter", 84, 7.35))
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = "James Peter"): invalid factor
## level, NA generated
```

```
two_col_bowler_rank_list <- cbind(bowlers_rank_list[1:42,], bowlers_rank_list[43:84,])
head(two_col_bowler_rank_list, 10)
```

```
##       Bowler Name Rank        PC1 Score      Bowler Name Rank
## 1       SP Narine    1 2.91849762426418       AD Mathews   43
## 2      SL Malinga    2 2.39882470947817     Pankaj Singh   44
## 3        M Morkel    3 2.26839707030713         RP Singh   45
## 4        DW Steyn    4 2.14239431559737          J Botha   46
## 5        L Balaji    5 1.83586328336081   V Pratap Singh   47
## 6  M Muralitharan    6 1.71255079295754       KP Appanna   48
## 7   BW Hilfenhaus    7 1.58915781159866    Harmeet Singh   49
## 8  Shakib Al Hasan    8 1.54481523454478 RE van der Merwe   50
## 9        UT Yadav    9  1.4170371768567     DT Christian   51
## 10     AB McDonald   10 1.35663188963888          B Kumar   52
##              PC1 Score
## 1    0.288119764063714
## 2    0.284836933077009
## 3    0.190107009573332
## 4    0.153273524577846
## 5   0.0800473565398499
## 6   0.0248637545833153
## 7  0.00710796863841973
## 8  0.00683443401493736
## 9  -0.0283071956069862
## 10 -0.0382889221438454
```

```
#write.csv(two_col_bowler_rank_list, file = "Bowlers ranked list.csv", sep = ",", row.names = FALSE)
```

Conclusion of Principal Component Analysis

- IPL involves buying players based on the previous performance. There are several indicators to measure the player's performance which make a player a better performer than the rest. However, as this indicators are generally highly correlated with one another, makes it difficult to judge overall player performance.

- So, using principal component analysis the reduced indicators can be used for ranking the players; which can be applied to correlated indicators for measuring performance. Thus, allowing to select players with reduced uncertainty about their potential due to inability to measure players based on the correlated indicators.

Factor Analysis

Test of Assumptions:

- Atleast sizable Inter-item correlation
- Proportion of variance among variables that might be common variance

Inter-item correlations (correlation matrix)

- Are there atleast several sizable correlations, e.g. > 0.5

```
##             exciting dependable luxurious outdoorsy powerful stylish
## exciting        1.00       0.03      0.45      0.18     0.63    0.75
## dependable      0.03       1.00      0.40     -0.07     0.16    0.17
## luxurious       0.45       0.40      1.00     -0.21     0.43    0.63
## outdoorsy       0.18      -0.07     -0.21      1.00     0.32    0.00
## powerful        0.63       0.16      0.43      0.32     1.00    0.59
## stylish         0.75       0.17      0.63      0.00     0.59    1.00
## comfortable     0.04       0.46      0.47      0.02     0.22    0.23
## rugged          0.17       0.00     -0.16      0.79     0.30    0.04
## fun             0.83       0.09      0.46      0.15     0.63    0.74
## safe           -0.15       0.54      0.26      0.02     0.09   -0.02
## performance     0.61       0.23      0.56     -0.14     0.58    0.67
## family         -0.57       0.21     -0.27      0.16    -0.31   -0.52
## versatile      -0.14       0.13     -0.13      0.44     0.07   -0.14
## sports          0.66      -0.09      0.21      0.33     0.54    0.56
## status          0.64       0.22      0.67     -0.08     0.54    0.78
## practical      -0.34       0.29     -0.15      0.17    -0.15   -0.28
## discipline     -0.09       0.08     -0.06      0.05    -0.07   -0.07
##             comfortable rugged   fun
## exciting           0.04   0.17  0.83
## dependable         0.46   0.00  0.09
## luxurious          0.47  -0.16  0.46
## outdoorsy          0.02   0.79  0.15
## powerful           0.22   0.30  0.63
## stylish            0.23   0.04  0.74
## comfortable        1.00   0.05  0.13
## rugged             0.05   1.00  0.16
## fun                0.13   0.16  1.00
## safe               0.58   0.09 -0.08
## performance        0.19  -0.12  0.62
## family             0.24   0.17 -0.54
## versatile          0.25   0.46 -0.10
## sports            -0.12   0.38  0.68
## status             0.27  -0.05  0.66
## practical          0.30   0.18 -0.32
## discipline         0.02   0.02 -0.05
```

```
##               fun  safe performance family versatile sports status
## exciting     0.83 -0.15        0.61  -0.57     -0.14   0.66   0.64
## dependable   0.09  0.54        0.23   0.21      0.13  -0.09   0.22
## luxurious    0.46  0.26        0.56  -0.27     -0.13   0.21   0.67
## outdoorsy    0.15  0.02       -0.14   0.16      0.44   0.33  -0.08
## powerful     0.63  0.09        0.58  -0.31      0.07   0.54   0.54
## stylish      0.74 -0.02        0.67  -0.52     -0.14   0.56   0.78
## comfortable  0.13  0.58        0.19   0.24      0.25  -0.12   0.27
## rugged       0.16  0.09       -0.12   0.17      0.46   0.38  -0.05
## fun          1.00 -0.08        0.62  -0.54     -0.10   0.68   0.66
## safe        -0.08  1.00        0.10   0.42      0.30  -0.22   0.09
## performance  0.62  0.10        1.00  -0.44     -0.20   0.47   0.73
## family      -0.54  0.42       -0.44   1.00      0.54  -0.49  -0.48
## versatile   -0.10  0.30       -0.20   0.54      1.00   0.00  -0.17
## sports       0.68 -0.22        0.47  -0.49      0.00   1.00   0.46
## status       0.66  0.09        0.73  -0.48     -0.17   0.46   1.00
## practical   -0.32  0.41       -0.28   0.69      0.56  -0.29  -0.27
## discipline  -0.05  0.06       -0.04   0.11      0.14  -0.11  -0.01
##             practical discipline
## exciting        -0.34      -0.09
## dependable       0.29       0.08
## luxurious       -0.15      -0.06
## outdoorsy        0.17       0.05
## powerful        -0.15      -0.07
## stylish         -0.28      -0.07
## comfortable      0.30       0.02
## rugged           0.18       0.02
## fun             -0.32      -0.05
## safe             0.41       0.06
## performance     -0.28      -0.04
## family           0.69       0.11
## versatile        0.56       0.14
## sports          -0.29      -0.11
## status          -0.27      -0.01
## practical        1.00       0.08
## discipline       0.08       1.00
```

Interpretaion of correlation matrix:

- The matrix results gives there are sizable variables which have correlated and could be grouped under factors

Assessing the Factorability(structure detection) of the Data

- Evalutation to check whether the "factorability" of the data. Are there meaningful latent factors to be found within the data?
- We can check two things:

(1) Bartlett's test of sphericity
(2) the Kaiser-Meyer-Olkin measure of sampling adequacy.

Bartlett's Test of Sphericity

```
## $chisq
## [1] 3413
##
## $p.value
## [1] 0
##
## $df
## [1] 136
```

Intrepreation of Bartlett's Test of Sphericity:

- Bartlett's Test of Sphericity evaluates whether or not the variables intercorrelate at all, by evaluating the observed correlation matrix against an "identity matrix" (a matrix with ones along the principal diagonal, and zeroes everywhere else). If this test is not statistically significant, we should not employ a factor analysis.
- Bartlett's test was statistically significant, suggesting that the observed correlation matrix among the items is not an identity matrix. This really isn't a particularly powerful indication that we have a factorable dataset, though - all it really tells us that at least some of the variables are correlated with each other.

Kaiser-Meyer-Olkin(KMO)

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mcvar_cor)
## Overall MSA =  0.88
## MSA for each item =
##    exciting  dependable   luxurious   outdoorsy    powerful     stylish
##        0.91        0.83        0.89        0.70        0.93        0.93
## comfortable      rugged         fun        safe performance      family
##        0.81        0.70        0.92        0.81        0.90        0.89
##   versatile      sports      status   practical  discipline
##        0.85        0.89        0.91        0.85        0.54
```

Intrepreation of KMO:

- The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a better measure of factorability. The KMO tests to see if the partial correlations within your data are close enough to zero to suggest that there is at least one latent factor underlying the variables. The minimum acceptable value is 0.50, if the overall MSA is too low, we could look at the item MSA's and drop items that are too low.
- The overall KMO for our data is 0.88 which is excellent - this suggests that we can go ahead with our planned factor analysis.

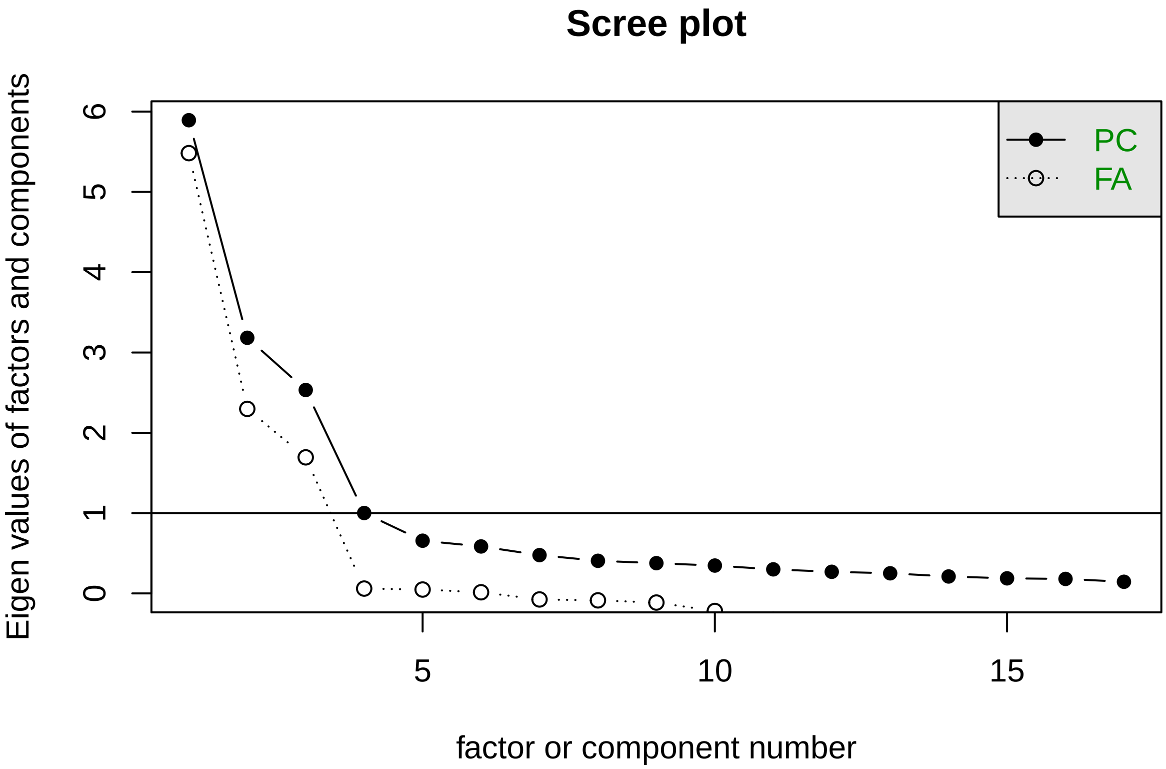
Determining the Number of Factors to Extract

- The decision as to the number of factors that we will need to extract, in order to achieve the most parsimonious (but still interpretatable) factor structure. Parsimonious models are simple models with great explanatory predictive power. They explain data with a minimum number of parameters, or predictor variables. Parsimonious models have optimal parsimony, or just the right amount of predictors

needed to explain the model well.There is generally a tradeoff between goodness of fit and parsimony: low parsimony models (i.e. models with many parameters) tend to have a better fit than high parsimony models. This is not usually a good thing; adding more parameters usually results in a good model fit for the data at hand, but that same model will likely be useless for predicting new data sets.
- The two most commonly employed methods are the scree plot, and parallel analysis which helps to determine number of factors to extract. For this data using scree plot to determine the factors.

Sree Plot

- Eigenvalues are a measure of the amount of variance accounted for by a factor, and so they can be useful in determining the number of factors that we need to extract. In a scree plot, we simply plot the eigenvalues for all of our factors, and then look to see where they drop off sharply.



Interpretation of Scree plot:

- The scree plot technique involves drawing a straight line through the plotted eigenvalues, starting with the largest one. The last point to fall on this line represents the last factor that we extract, with the idea being that beyond this, the amount of additional variance explained is non-meaningful. Regardless of whether we are using a principal components or a principal axis factor extraction, however, there is a very large first factor in this data.
- If we were to draw our straight line starting at this point, would probably conclude that there are only four factors in the dataset.

Conducting the Factor Analysis using 4 factors:

- Having a good idea as to how many factors (4) that we should extract in our analysis of the MBA car data. A decision whether to use "Principal Axis Factoring", or "principal components" analysis. In a very broad sense, "principal axis factoring" is used when we want to identify the latent variables that are underlying a set of variables, while "principal components" analysis is used to reduce a set of variables to a smaller set of factors (i.e., the "principal components" of the data). In other words, Principal Axis Factoring is used when we want to evaluate a theoretical model with a set of variables, and principal components analysis is used for data reduction.
- The primary difference between the way that principal axis factoring and principal component analysis are conducted, is that the correlation matrix on which the factor analysis is based has ones along the principal diagonal in principal components analysis, and the communalities along the principal diagonal in principal axis factor analysis.

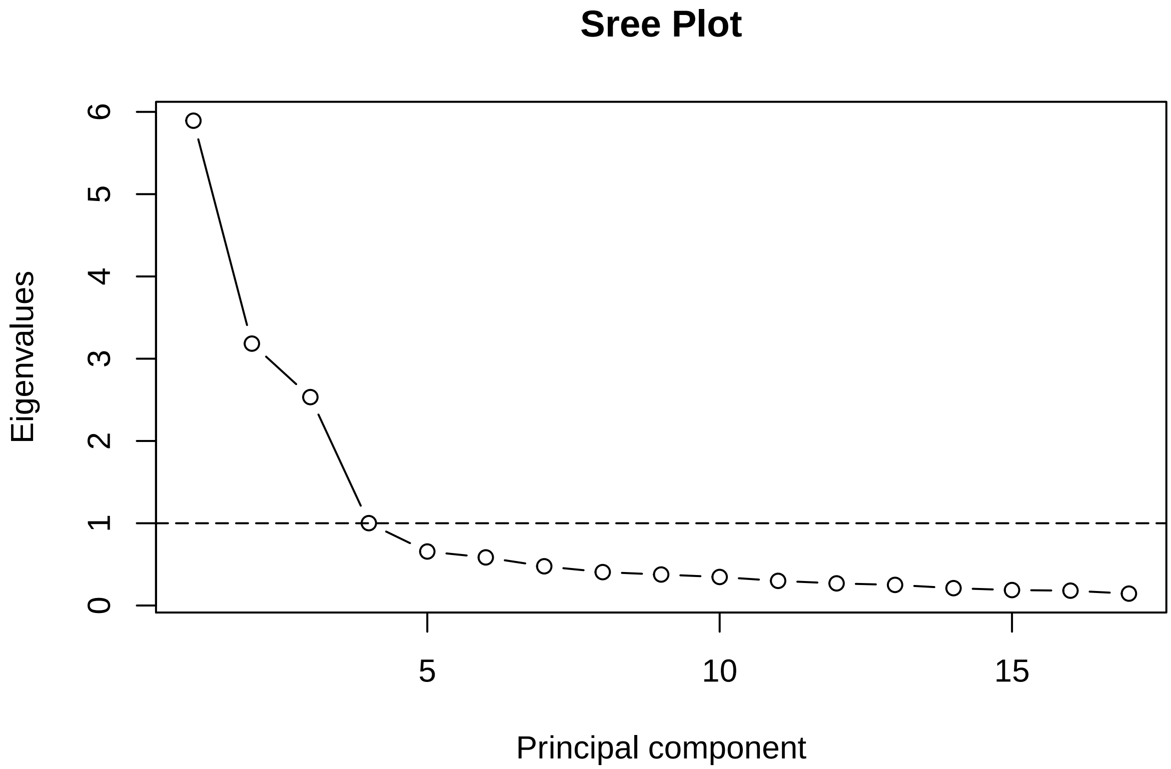
Factor analysis using Principal Components Analysis

Factor loadings using PCA method

```
## Principal Components Analysis
## Call: principal(r = mcvar, nfactors = 4, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##               PC1   PC2   PC3   PC4   h2    u2 com
## exciting     0.87  0.00  0.21  0.01 0.80 0.202 1.1
## dependable   0.13  0.61 -0.43  0.03 0.58 0.424 1.9
## luxurious    0.66  0.28 -0.46 -0.05 0.73 0.271 2.2
## outdoorsy    0.04  0.39  0.81  0.00 0.80 0.195 1.4
## powerful     0.72  0.30  0.23 -0.05 0.67 0.332 1.6
## stylish      0.88  0.11 -0.06  0.01 0.79 0.208 1.0
## comfortable  0.15  0.70 -0.37 -0.10 0.66 0.341 1.7
## rugged       0.05  0.43  0.78 -0.05 0.80 0.203 1.6
## fun          0.87  0.06  0.16  0.04 0.79 0.213 1.1
## safe        -0.07  0.74 -0.35 -0.07 0.69 0.310 1.5
## performance  0.80  0.09 -0.23  0.06 0.70 0.295 1.2
## family      -0.68  0.56 -0.01 -0.03 0.79 0.212 1.9
## versatile   -0.25  0.67  0.38  0.07 0.66 0.336 1.9
## sports       0.71 -0.03  0.49 -0.03 0.75 0.254 1.8
## status       0.84  0.13 -0.22  0.09 0.78 0.223 1.2
## practical   -0.46  0.66  0.01 -0.05 0.64 0.357 1.8
## discipline  -0.11  0.14 -0.01  0.98 0.99 0.013 1.1
##
##                        PC1  PC2  PC3  PC4
## SS loadings           5.89 3.18 2.53 1.00
## Proportion Var        0.35 0.19 0.15 0.06
## Cumulative Var        0.35 0.53 0.68 0.74
## Proportion Explained  0.47 0.25 0.20 0.08
## Cumulative Proportion 0.47 0.72 0.92 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.04
##  with the empirical chi square  156  with prob <  9.7e-08
##
```

```
## Fit based upon off diagonal values = 0.99
```

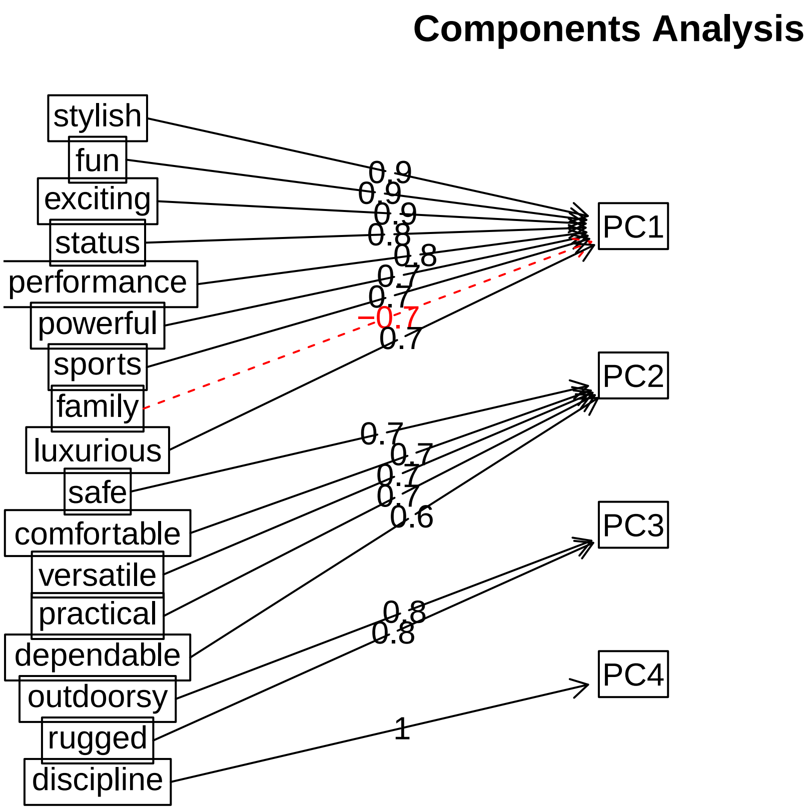
```
## Warning in plot.xy(xy, type, ...): plot type 'both' will be truncated to
## first character
```



### The Sree Plot given using the eigen value usin principal method confirms the earlier results given by using Scree function on the data.

Look of the factor structure for this solution

```
fa.diagram(pc4.out)
```



Interpretation of 4 factor solution

As it seen factor structure using the four factor solution results, remaining factors doesn't have any loadings, fourth factor that has major loadings from only one item.

Variance extracted by factors

```
print("Printed only first five")
```

```
## [1] "Printed only first five"
```

```
print("-----Principal factors Eigne values-----")
```

```
## [1] "-----Principal factors Eigne values-----"
```

```
print(pc4.out$values[1:5])
```

```
## [1] 5.893 3.183 2.533 1.001 0.656
```

```
print("-----Total Variance explained by factors-----")
```

```
## [1] "-----Total Variance explained by factors-----"
```

```
print(cumsum(100*(pc4.out$values[1:5]/length(pc4.out$values))))
```

```
## [1] 34.7 53.4 68.3 74.2 78.0
```

Observation

- 1) Output of this analysis show us that only 4 components have eigenvalues greater than 1, suggesting that we extract 4 components.

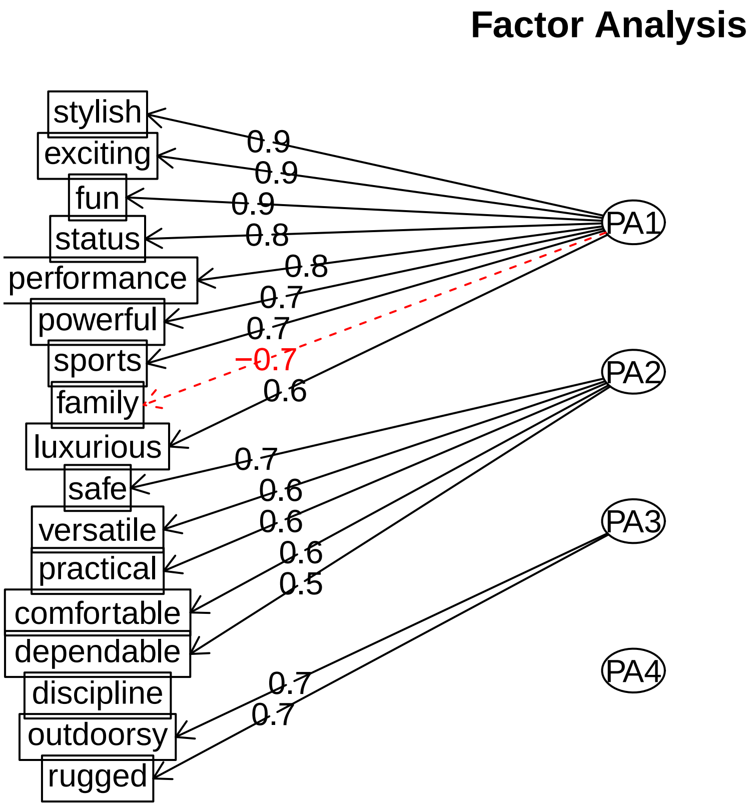
2) The above output also suggests that extracting 4 components explains 74.2% of the total variance.

Communalities

```
##    exciting  dependable   luxurious   outdoorsy    powerful     stylish 
##       0.798       0.576       0.729       0.805       0.668       0.792 
## comfortable      rugged         fun        safe performance      family 
##       0.659       0.797       0.787       0.690       0.705       0.788 
##   versatile      sports      status   practical  discipline 
##       0.664       0.746       0.777       0.643       0.987
```

These are the percentage of variance that can be explained by the retained factors for each variable.

Factor Analysis using the Principal Axis Factoring



### observation * The difference between PCA factor analysis and Pricipal Factors axis is the variable for PC2 has changed from two variables i.e. comfortable and practical from 0.7 to 0.6 respectively and for PC4 it has change from maximum loading of 1 to 0. The fourth factor can be treated as over loading has it has only one variable loading on it which was given by PCA method, this probably represents an overextraction.

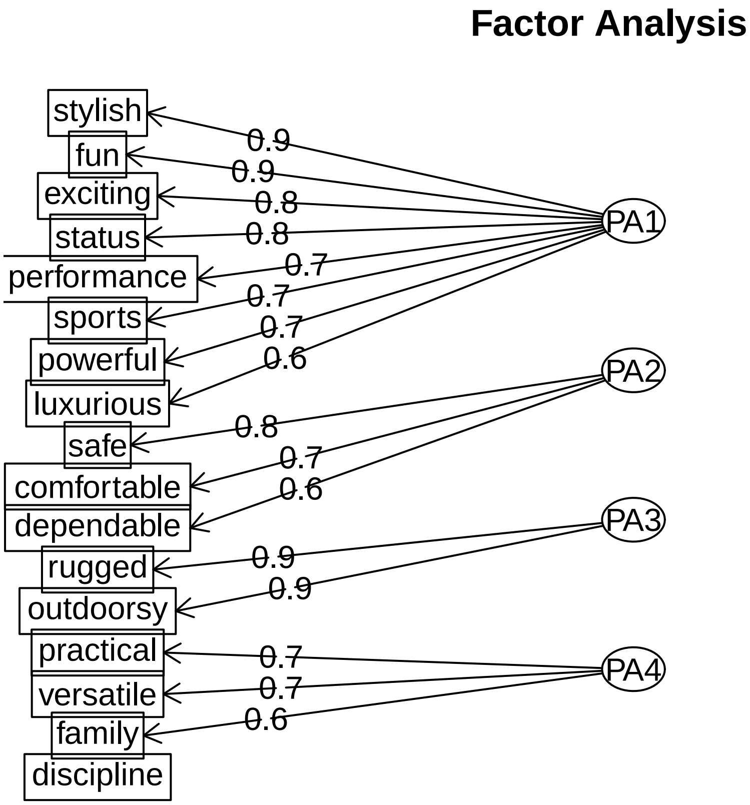
Factor rotation

- Rotation is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved.
- The most common orthogonal rotation method varimax

```r
fa4.out <- fa(mcvar, nfactors = 4, fm="pa", max.iter = 200, rotate = "varimax")
print(fa4.out)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = mcvar, nfactors = 4, rotate = "varimax", max.iter = 200, 
##     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
## 
## Mean item complexity =  1.5
## Test of the hypothesis that 4 factors are sufficient.
## 
## The degrees of freedom for the null model are  136  and the objective function was  11.6 with Chi Squ
## The degrees of freedom for the model are 74  and the objective function was  0.56 
## 
## The root mean square of the residuals (RMSR) is  0.02 
## The df corrected root mean square of the residuals is  0.03 
## 
## The harmonic number of observations is  303 with the empirical chi square  38.4  with prob <  1 
## The total number of observations was  303  with Likelihood Chi Square =  165  with prob <  7.4e-09 
## 
## Tucker Lewis Index of factoring reliability =  0.949
## RMSEA index =  0.065  and the 90 % confidence intervals are  0.051 0.077
## BIC =  -258
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                     PA1  PA2  PA3  PA4
## Correlation of (regression) 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
```

```
fa.diagram(fa4.out)
```



### Observation and Labeling of factors * The rotated solution is more interpretable - in fact, it seems to replicate the expected factor structure nicely. And agrees with number of factors given by the scree plot. * Factor 1 - Youth * The variables like fun, exciting, etc. are more applicable to younger people who are more attracted to a car which is given PA1 loadings, hence labeling the factor as youth describes well the underlying latent factors. * Factor 2 - Professional * The variables like dependable, comfortable, etc. are more applicable to professional people who prefer a car which can serve them well and keep up the strict business schedules which is given PA2 loadings, hence labeling the factor as professional describes well the underlying latent factors. * Factor 3 - Travelling * The variables like rugged, outdoorsy are more applicable to people who like to travel and prefer a car which can perform well on uneven road and rough conditions is given by PA3 loadings, hence labeling the factor as Travelling describes well the underlying latent factors. * Factor 4 - Family * The variables like practical, family, etc. are more applicable for family who a car which is can be used on daily basis and is practical given by PA4 loadings, hence labeling the factor as family describes well the underlying latent factors.

Communalities

The communality for each variable is the percentage of variance that can be explained by the retained factors given by the rotated solution. It's best if the retained factors explain more of the variance in each variable.

```
##    exciting  dependable   luxurious   outdoorsy    powerful     stylish
##      0.7840      0.4438      0.6769      0.7779      0.5951      0.7828
## comfortable      rugged         fun        safe performance      family
##      0.5645      0.8045      0.7695      0.6422      0.6540      0.7832
##   versatile      sports      status   practical  discipline
##      0.6142      0.6905      0.7411      0.6799      0.0207
```

Eigenvalues

```
## [1] 5.893 3.183 2.533 1.001 0.656
```

```
## [1] 5.615 2.827 2.241 0.343 0.178
```

The first eigenvalues derived in the extracted factor solution are stored within e.values. These are the eigenvalues

that were plotted in the scree plots that plotted at the beginning of this factor analysis. The second line is the eigenvalues from the rotated solution

Percentage of Variance accounted for

```
## [1] "-----Plotted in the scree plot at beginning-----"
```

```
## [1] 34.66 18.72 14.90  5.89  3.86
```

```
## [1] "-----By the rotated solution"
```

```
## [1] 33.03 16.63 13.18  2.02  1.05
```

```
## [1] "-----Plotted in the scree plot at beginning-----"
```

```
## [1] 34.7 53.4 68.3 74.2 78.0
```

```
## [1] "-----By the rotated solution"
```

```
## [1] 33.0 49.7 62.8 64.9 65.9
```

Factor loadings

- After viewing the highest-loading items for each factor using fa.diagram, but this only tells us the largest loading for each item. Each item will, however, load on each of the factors to a greater or lesser degree - and we will eventually want to look at the full factor loading matrix. The factor loading matrix shows us the factor loadings for each variable, after they have been rotated to "simple structure." Essentially, we are taking advantage of the fact that there are a number of factor solutions that are equally acceptable to the "optimal" solution that was found within our initial extraction (i.e., that are mathematically equivalent), and rotating the factors so that they are more easily interpreted.

```
## [1] "Full factor loading matrix"
```

```
##
## Loadings:
##             PA1    PA2    PA3    PA4
## exciting     0.849 -0.073  0.150 -0.189
## dependable   0.132  0.626 -0.066  0.172
## luxurious    0.575  0.526 -0.212 -0.160
## outdoorsy    0.082 -0.072  0.857  0.178
## powerful     0.683  0.175  0.304 -0.072
## stylish      0.858  0.147 -0.035 -0.152
## comfortable  0.157  0.706  0.009  0.202
## rugged       0.089  0.002  0.878  0.161
## fun          0.854 -0.006  0.131 -0.153
## safe        -0.060  0.753  0.055  0.262
## performance  0.742  0.216 -0.156 -0.178
## family      -0.528  0.288  0.125  0.637
## versatile   -0.056  0.140  0.405  0.654
## sports       0.700 -0.245  0.353 -0.126
## status       0.788  0.265 -0.122 -0.185
## practical   -0.250  0.273  0.096  0.731
## discipline  -0.060  0.043  0.010  0.123
##
##                  PA1   PA2   PA3   PA4
## SS loadings    5.055 2.153 2.044 1.774
## Proportion Var 0.297 0.127 0.120 0.104
## Cumulative Var 0.297 0.424 0.544 0.649
```

Observation

- The varimax rotation was rather successful in finding a rotation that simplified the complexity of the variables. As each of the variables now load highly on the single factor while the remaining three factors have less loading for a particular variable in a row. Like for the exiting variable first factor has highest loading of 85% and the remaining factors have significant less loadings for that variable. Except for the luxurious variable which loads evenly close loads into factor one and factor two and the discipline which has every less percentage of loadings to all the factor, which can treated as not a strong latent factor.

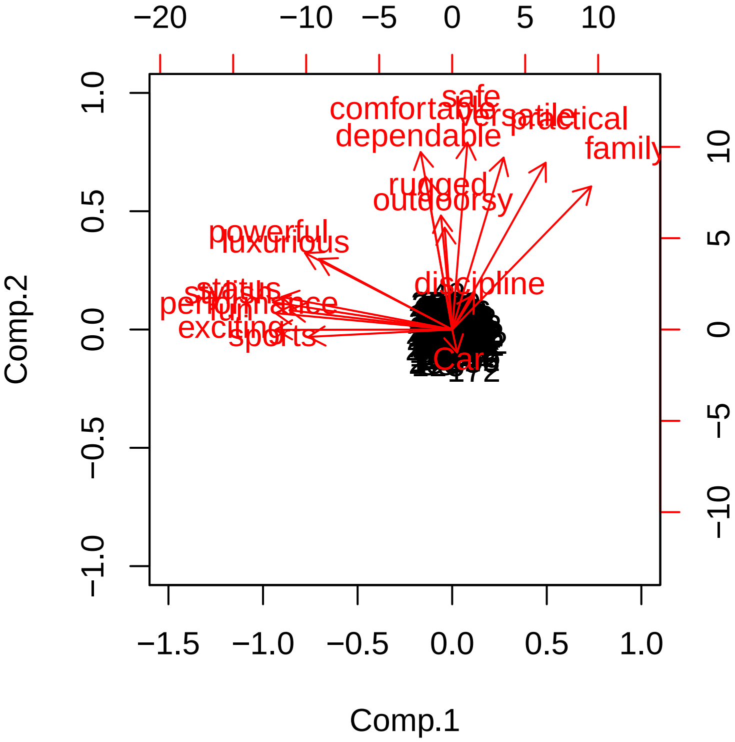
Summary of Factor scores of 4 factors

```
##       PA1              PA2              PA3              PA4
##  Min.   :-2.477   Min.   :-2.494   Min.   :-1.509   Min.   :-2.185
##  1st Qu.:-0.590   1st Qu.:-0.567   1st Qu.:-0.789   1st Qu.:-0.607
##  Median : 0.093   Median : 0.068   Median :-0.168   Median : 0.069
##  Mean   : 0.000   Mean   : 0.000   Mean   : 0.000   Mean   : 0.000
##  3rd Qu.: 0.744   3rd Qu.: 0.681   3rd Qu.: 0.574   3rd Qu.: 0.651
##  Max.   : 1.612   Max.   : 1.826   Max.   : 2.061   Max.   : 1.839
```

Observation

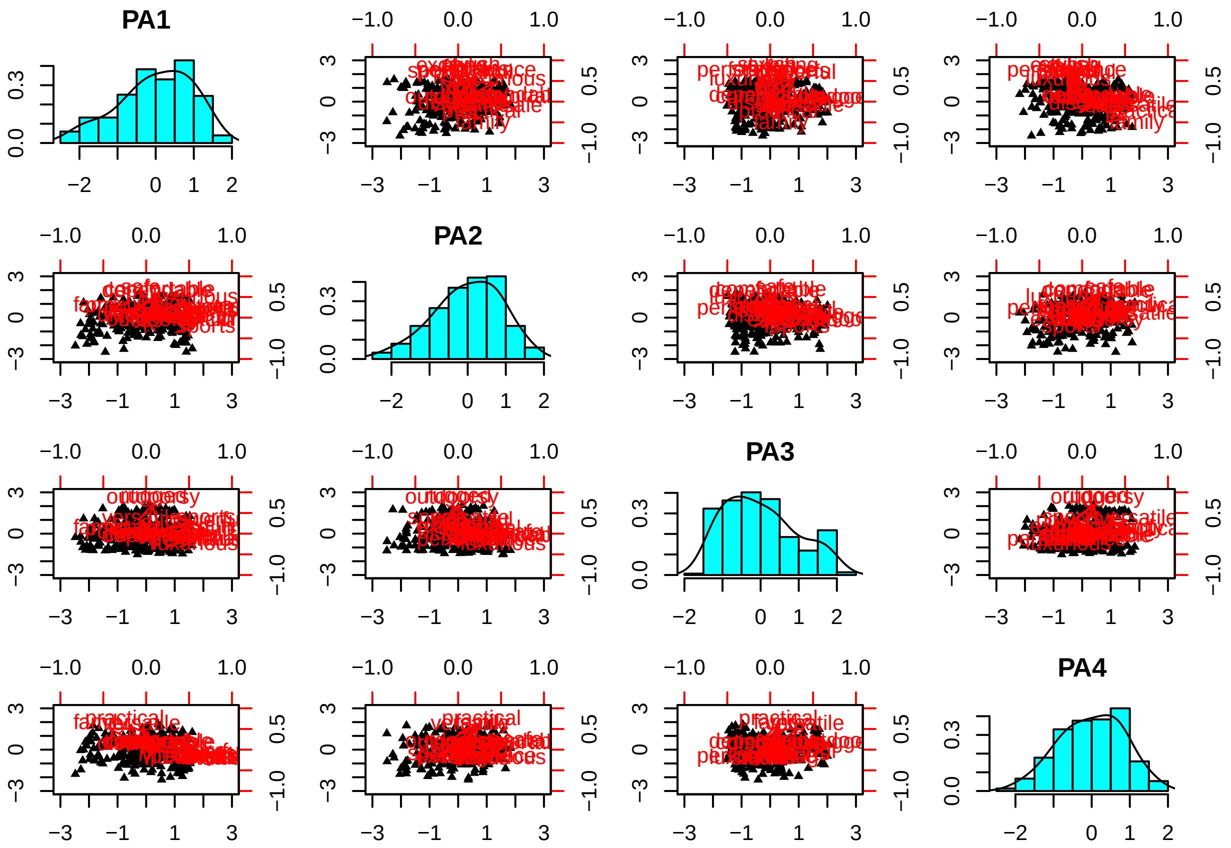
- The factor scores are the weights of each observed variable in producing a score representing the factor, whereas the factor loadings are the weight of each factor on the observed variables.
- The factor loadings are used to interpret what the factor is by judging the relative sizes of the loadings; high loadings suggest stronger factor contributions to those variables. And factor scores on the other hand are composites of the variables that are used to make the latent factor into an observed variable.
- Factor socres are used when the factor was of interest to use as a predictor or outcome in a regression analysis and were not planning to use latent variable modeling methods.
- The factor score computation given by fa() method result in factor scores with mean = 0 and SD = 1.

```
pc.cr <- princomp(mbacar[,2:19], cor=TRUE)
biplot(pc.cr, expand=11, xlim=c(-1.5, 1), ylim=c(-1, 1))
```



```r
#biplot(pc.cr, expand=17, xlim=c(-2, 1.5), ylim=c(-1, 2))
car_scores <- cbind(as.data.frame(car),fa4.out$scores)
```

```r
biplot(fa4.out)
```



```
biplot
```

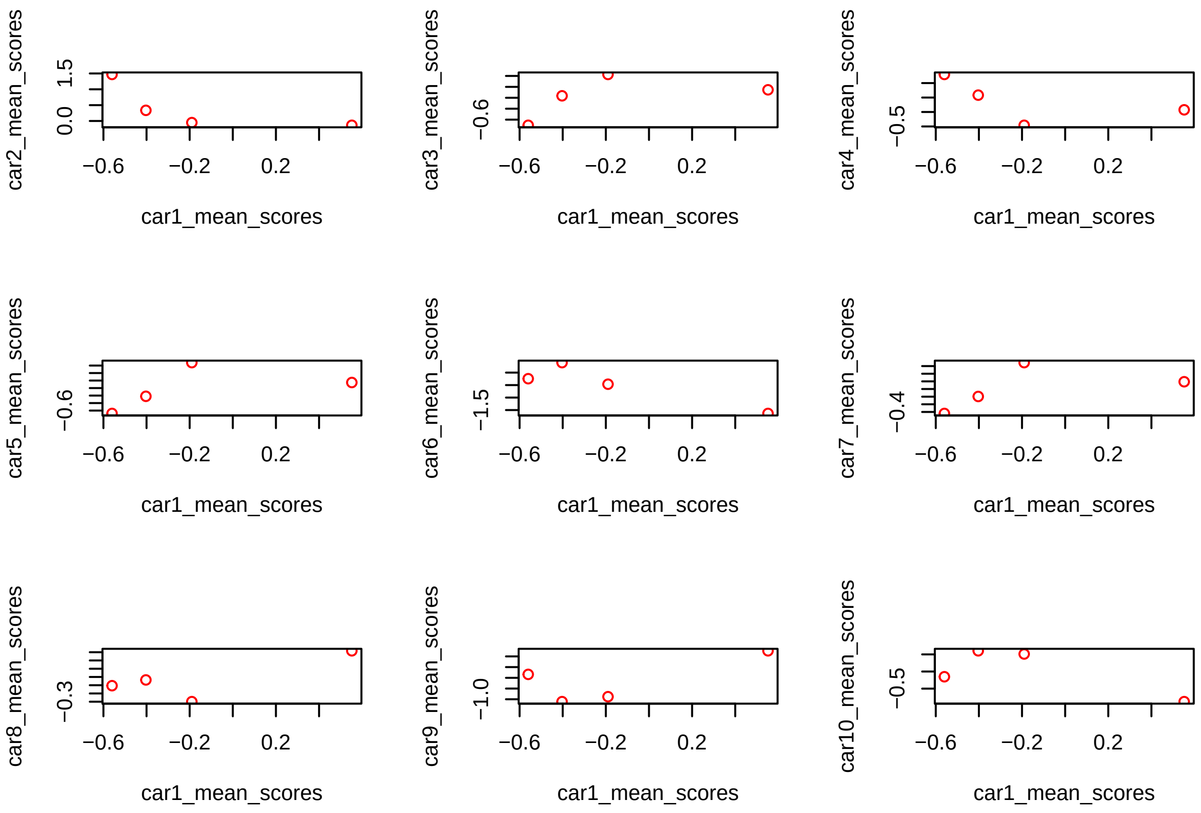
```
## function (x, ...)
## UseMethod("biplot")
## <bytecode: 0x00000000177d5378>
## <environment: namespace:stats>
```

```
car1 <- car_scores[car_scores["car"]==1,c("PA1","PA2","PA3","PA4")]
car2 <- car_scores[car_scores["car"]==2,c("PA1","PA2","PA3","PA4")]
car3 <- car_scores[car_scores["car"]==3,c("PA1","PA2","PA3","PA4")]
car4 <- car_scores[car_scores["car"]==4,c("PA1","PA2","PA3","PA4")]
car5 <- car_scores[car_scores["car"]==5,c("PA1","PA2","PA3","PA4")]
car6 <- car_scores[car_scores["car"]==6,c("PA1","PA2","PA3","PA4")]
car7 <- car_scores[car_scores["car"]==7,c("PA1","PA2","PA3","PA4")]
car8 <- car_scores[car_scores["car"]==8,c("PA1","PA2","PA3","PA4")]
car9 <- car_scores[car_scores["car"]==9,c("PA1","PA2","PA3","PA4")]
car10 <- car_scores[car_scores["car"]==10,c("PA1","PA2","PA3","PA4")]
car1_mean_scores <- c(0.552, -0.190, -0.560, -0.403)
car2_mean_scores <- c(-0.135, -0.051, 1.468, 0.336)
car3_mean_scores <- c(-0.054, 0.227, -0.703, -0.164)
car4_mean_scores <- c(0.075, -0.459, 1.298, 0.583)
car5_mean_scores <- c(0.148, 0.678, -0.677, -0.220)
car6_mean_scores <- c(-1.635, -0.465, -0.245, 0.398)
car7_mean_scores <- c(0.390, 0.890, -0.439, 0.005)
car8_mean_scores <- c(0.317, -0.299, -0.106, -0.036)
car9_mean_scores <- c(1.26, -0.880, 0.161, -1.110)
car10_mean_scores <- c(-0.880, 0.512, -0.153, 0.605)
```

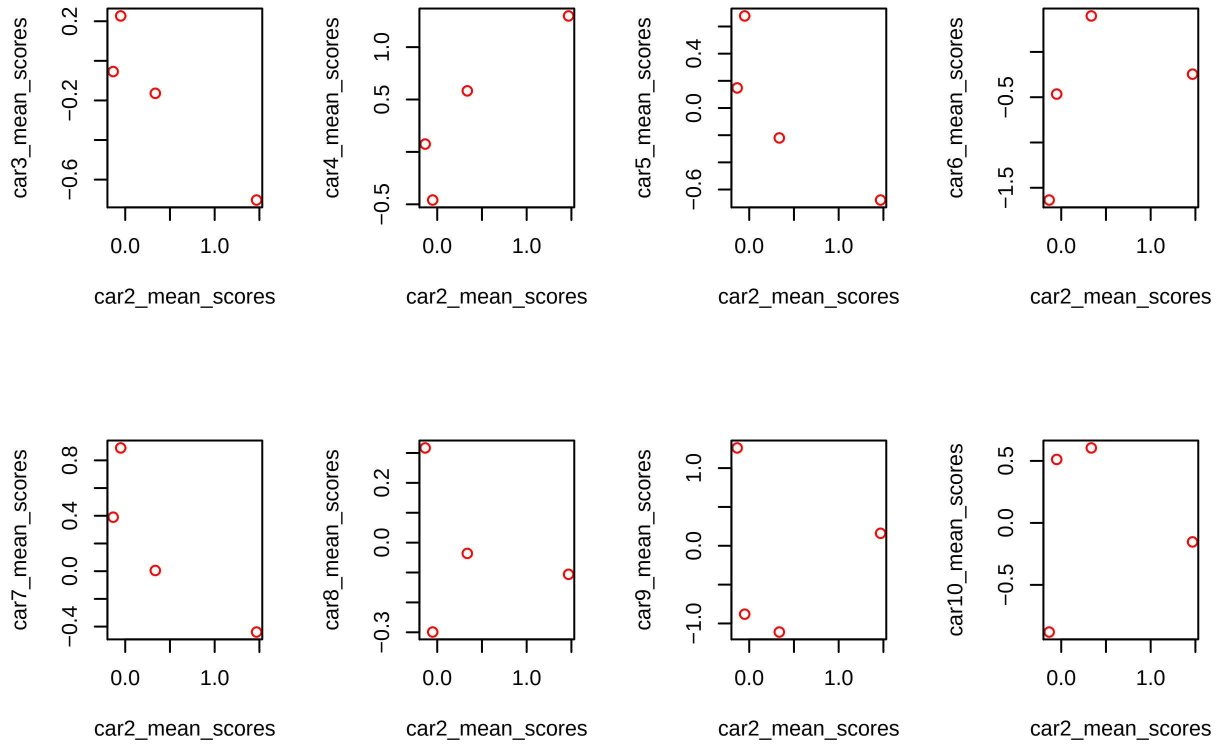
```
summary(car10)
```

```
##       PA1               PA2               PA3               PA4
##  Min.   :-2.283   Min.   :-2.494   Min.   :-1.230   Min.   :-1.448
##  1st Qu.:-1.251   1st Qu.: 0.505   1st Qu.:-0.618   1st Qu.: 0.383
##  Median :-0.958   Median : 0.727   Median :-0.099   Median : 0.562
##  Mean   :-0.880   Mean   : 0.512   Mean   :-0.153   Mean   : 0.605
##  3rd Qu.:-0.441   3rd Qu.: 0.928   3rd Qu.: 0.295   3rd Qu.: 0.920
##  Max.   : 0.419   Max.   : 1.453   Max.   : 1.169   Max.   : 1.839
```

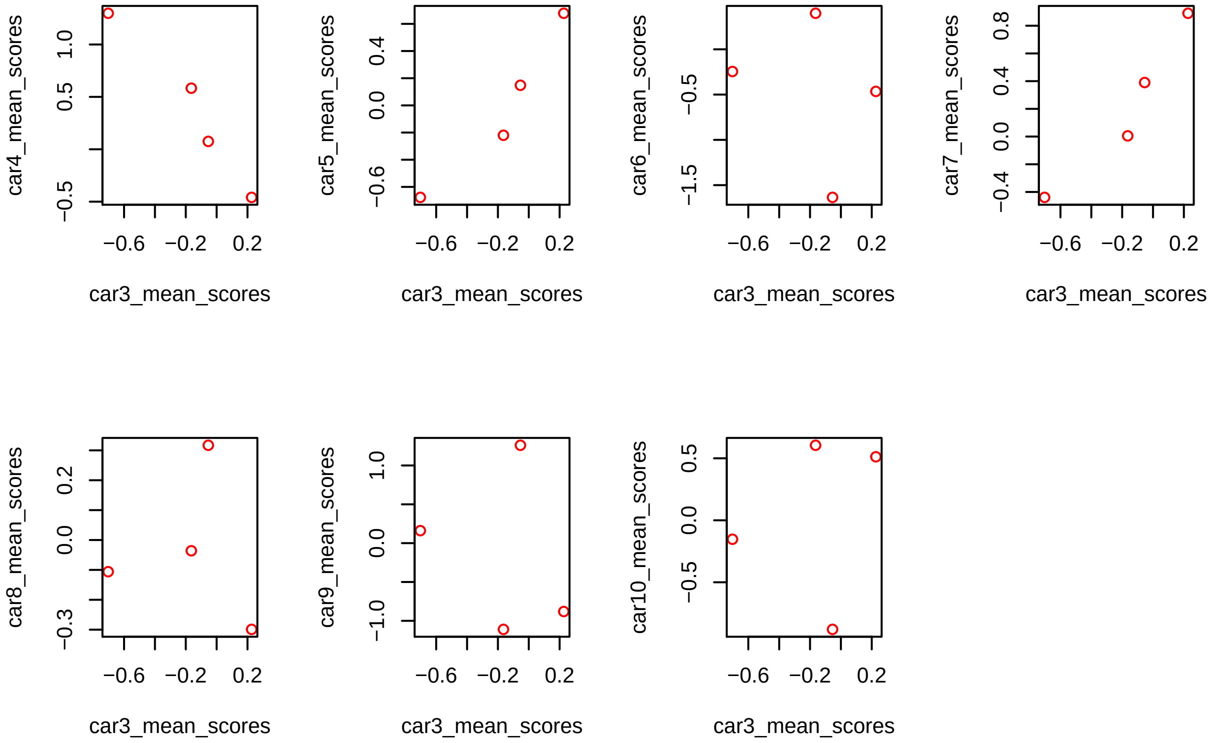
```
par1 <- par(mfrow=c(3,3))
plot(car1_mean_scores, car2_mean_scores, col=c("red"))
plot(car1_mean_scores, car3_mean_scores, col=c("red"))
plot(car1_mean_scores, car4_mean_scores, col=c("red"))
plot(car1_mean_scores, car5_mean_scores, col=c("red"))
plot(car1_mean_scores, car6_mean_scores, col=c("red"))
plot(car1_mean_scores, car7_mean_scores, col=c("red"))
plot(car1_mean_scores, car8_mean_scores, col=c("red"))
plot(car1_mean_scores, car9_mean_scores, col=c("red"))
plot(car1_mean_scores, car10_mean_scores, col=c("red"))
```



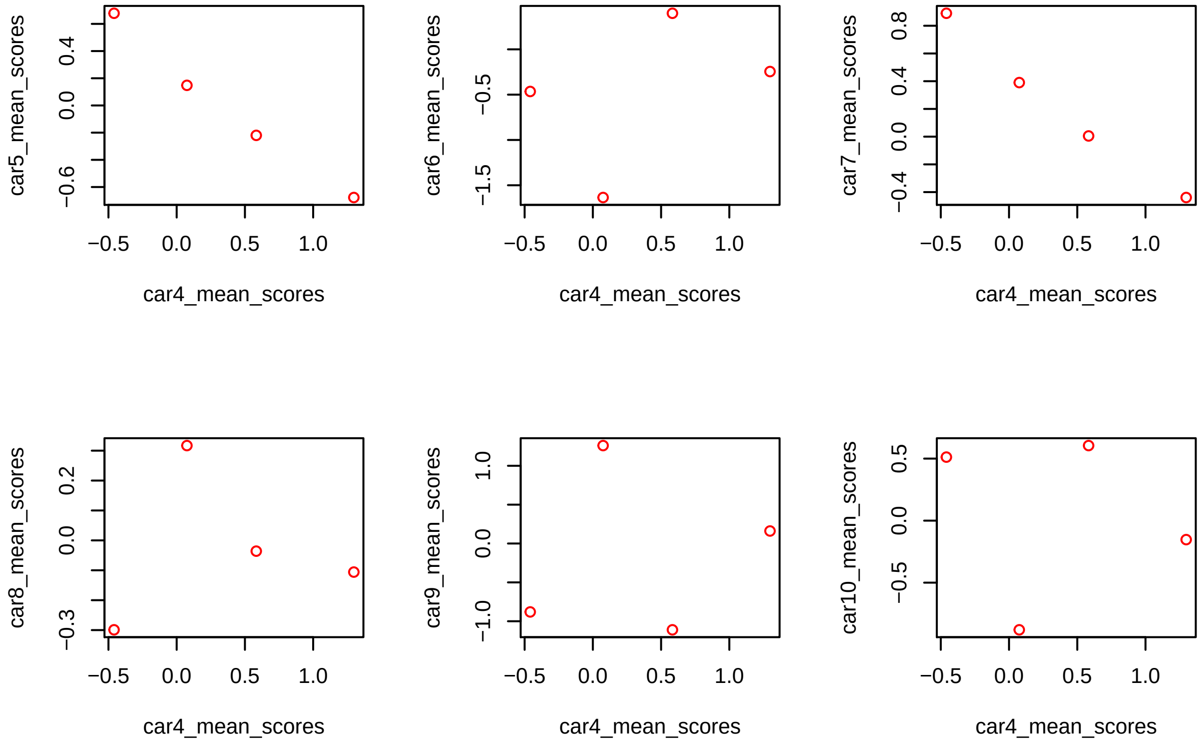
```
par2 <- par(mfrow=c(2,4))
plot(car2_mean_scores, car3_mean_scores, col=c("red"))
plot(car2_mean_scores, car4_mean_scores, col=c("red"))
plot(car2_mean_scores, car5_mean_scores, col=c("red"))
plot(car2_mean_scores, car6_mean_scores, col=c("red"))
plot(car2_mean_scores, car7_mean_scores, col=c("red"))
plot(car2_mean_scores, car8_mean_scores, col=c("red"))
plot(car2_mean_scores, car9_mean_scores, col=c("red"))
plot(car2_mean_scores, car10_mean_scores, col=c("red"))
```



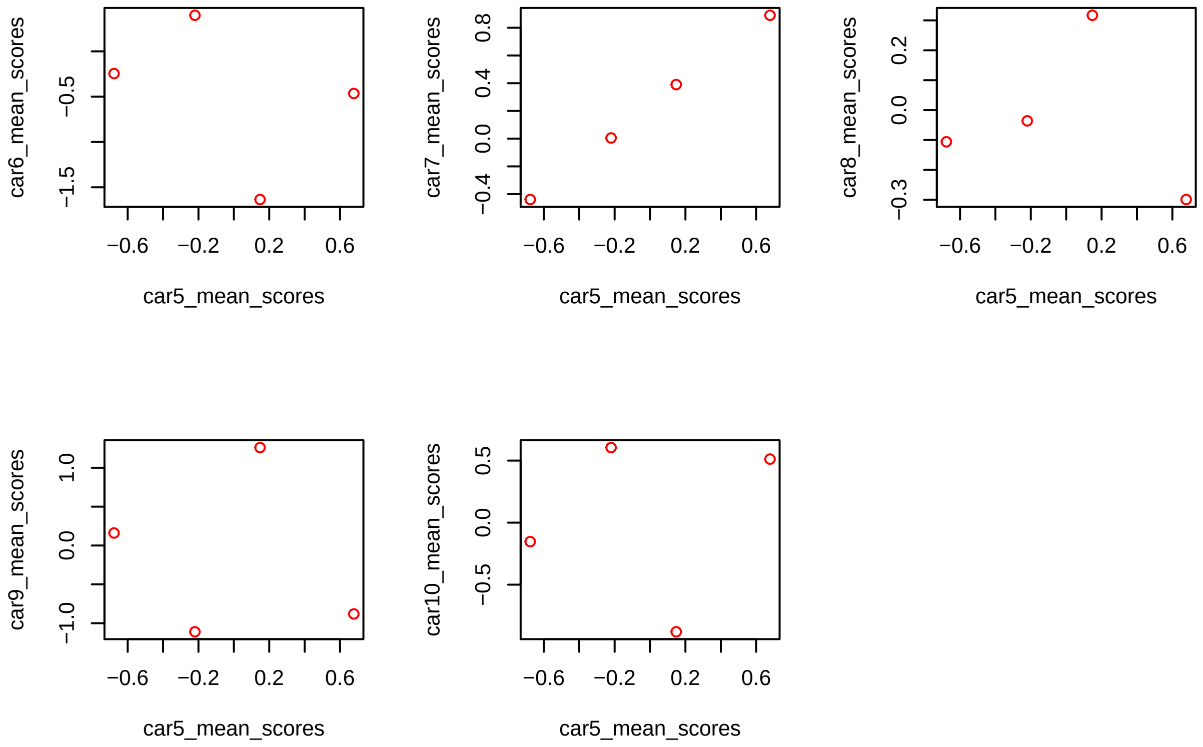
```
par3 <- par(mfrow=c(2,4))
plot(car3_mean_scores, car4_mean_scores, col=c("red"))
plot(car3_mean_scores, car5_mean_scores, col=c("red"))
plot(car3_mean_scores, car6_mean_scores, col=c("red"))
plot(car3_mean_scores, car7_mean_scores, col=c("red"))
plot(car3_mean_scores, car8_mean_scores, col=c("red"))
plot(car3_mean_scores, car9_mean_scores, col=c("red"))
plot(car3_mean_scores, car10_mean_scores, col=c("red"))
```



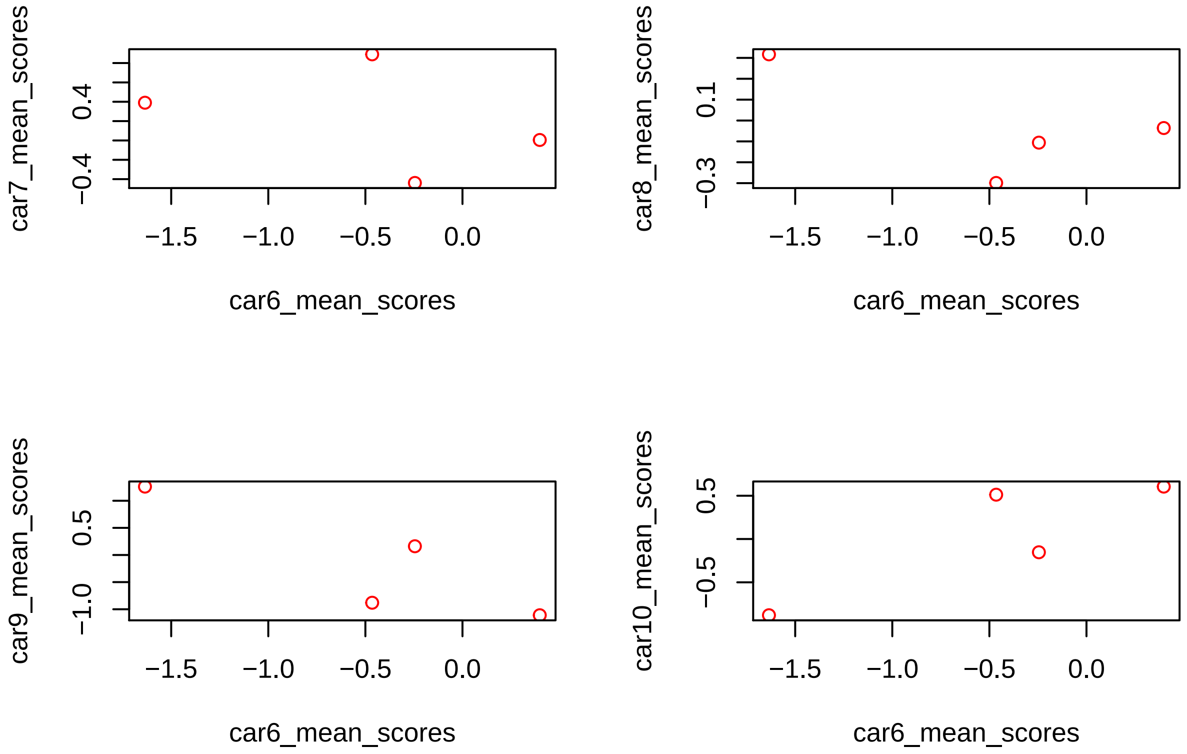
```
par3 <- par(mfrow=c(2,3))
plot(car4_mean_scores, car5_mean_scores, col=c("red"))
plot(car4_mean_scores, car6_mean_scores, col=c("red"))
plot(car4_mean_scores, car7_mean_scores, col=c("red"))
plot(car4_mean_scores, car8_mean_scores, col=c("red"))
plot(car4_mean_scores, car9_mean_scores, col=c("red"))
plot(car4_mean_scores, car10_mean_scores, col=c("red"))
```



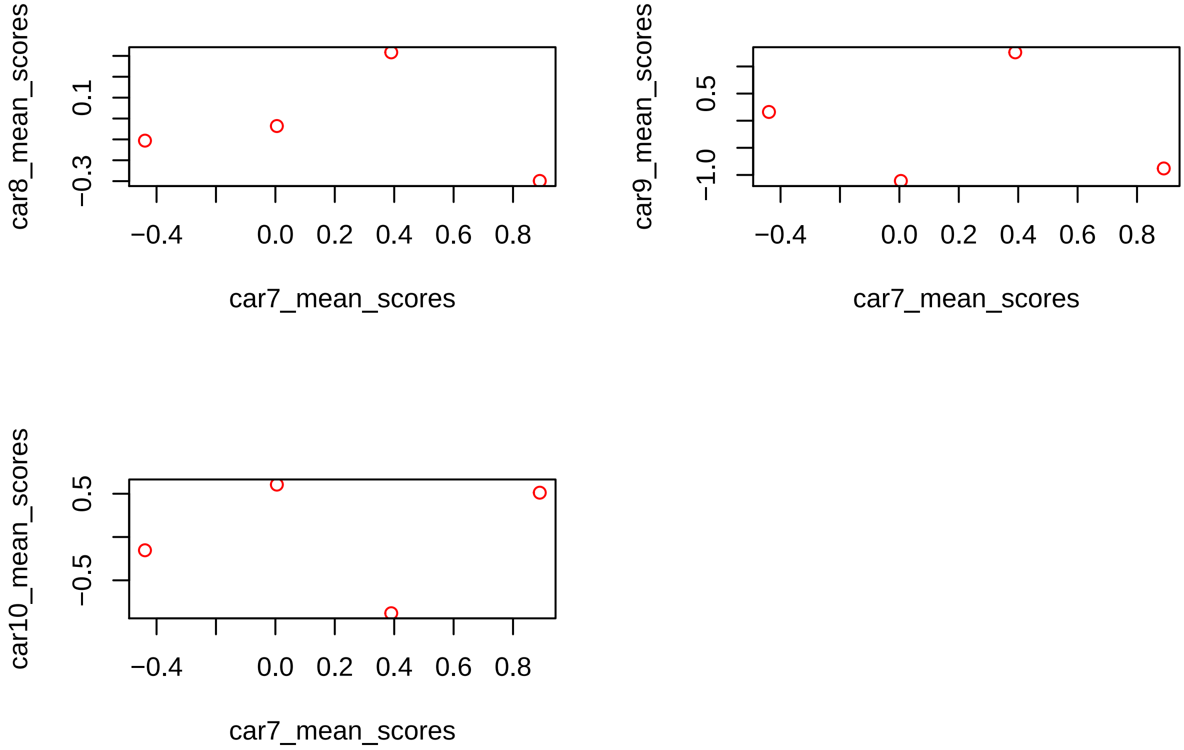
```
par3 <- par(mfrow=c(2,3))
plot(car5_mean_scores, car6_mean_scores, col=c("red"))
plot(car5_mean_scores, car7_mean_scores, col=c("red"))
plot(car5_mean_scores, car8_mean_scores, col=c("red"))
plot(car5_mean_scores, car9_mean_scores, col=c("red"))
plot(car5_mean_scores, car10_mean_scores, col=c("red"))
```



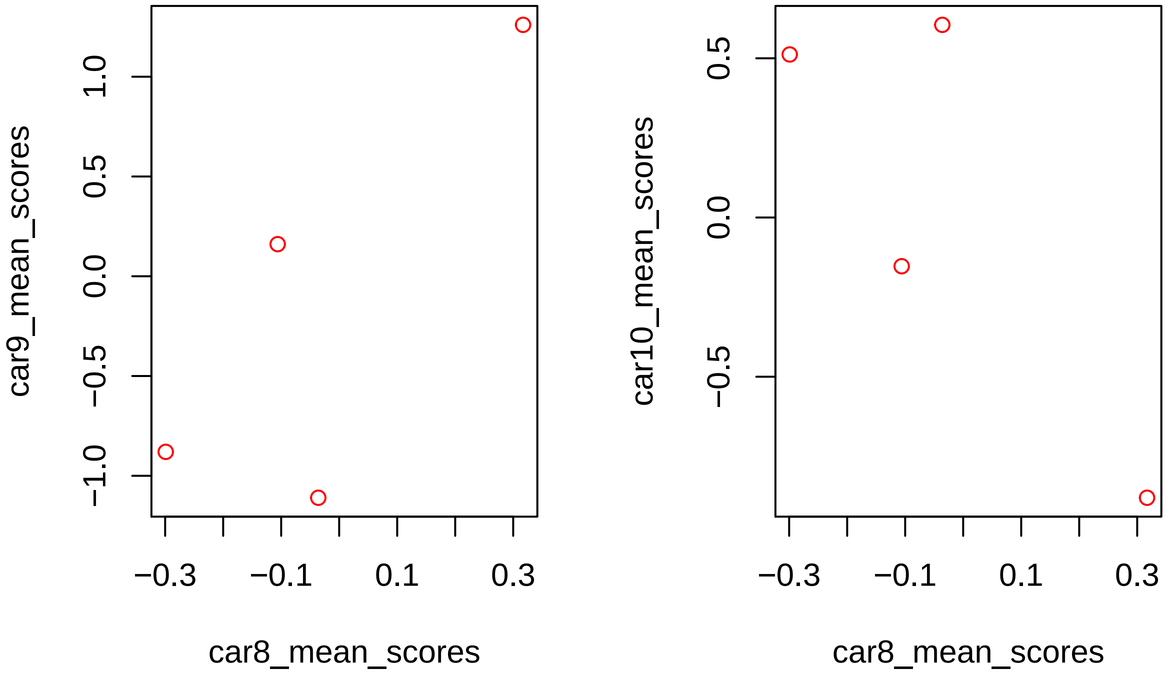
```r
par3 <- par(mfrow=c(2,2))
plot(car6_mean_scores, car7_mean_scores, col=c("red"))
plot(car6_mean_scores, car8_mean_scores, col=c("red"))
plot(car6_mean_scores, car9_mean_scores, col=c("red"))
plot(car6_mean_scores, car10_mean_scores, col=c("red"))
```



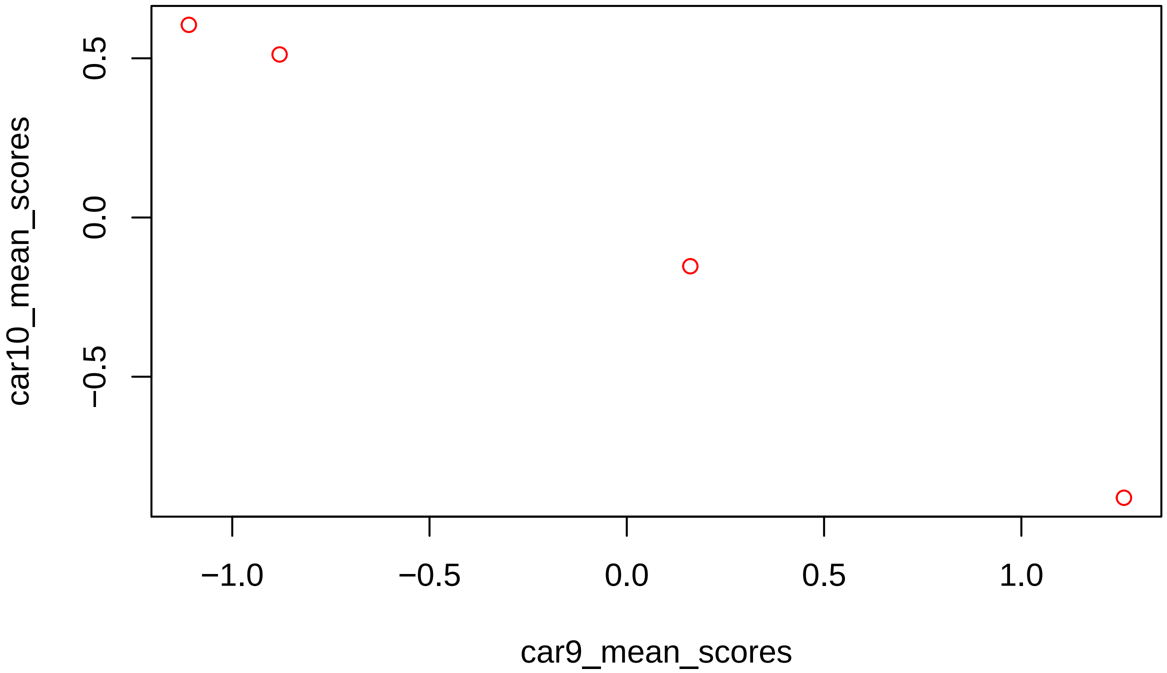
```r
par3 <- par(mfrow=c(2,2))
plot(car7_mean_scores, car8_mean_scores, col=c("red"))
plot(car7_mean_scores, car9_mean_scores, col=c("red"))
plot(car7_mean_scores, car10_mean_scores, col=c("red"))
```



```r
par3 <- par(mfrow=c(1,2))
plot(car8_mean_scores, car9_mean_scores, col=c("red"))
plot(car8_mean_scores, car10_mean_scores, col=c("red"))
```



```r
plot(car9_mean_scores, car10_mean_scores, col=c("red"))
```


0.5
0.0
−0.5
car10_mean_scores
−1.0
−0.5
0.0
0.5
1.0
car9_mean_scores