## MBA Car Problem Factor Analysis

### Problem statement

In January 1998, 303 MBA students were asked about their evaluations of and preferences for 10 different automobiles. The automobiles, listed in order of presentation in the survey were: (where 5= **Extremely Descriptive** and 1 = **Not at all descriptive** ):The words are: exciting, dependable, luxurious. Outdoorsy, powerful, stylish, comfortable, and rugged

1. BMW328i
2. Ford Explorer
3. Infiniti J30
4. Jeep Grand Cherokee
5. Lexus ES300
6. Chrysler Town & Country
7. Mercedes C280
8. Saab 9000
9. Porche Boxster
10. Volvo V90

Each student rated all cars. For the purposes of this exercise, one car was selected randomly from each student, resulting in a sample size of 303 evaluations (approx. 30 observations per type of car).

The students rated each car on 16 attributes. The first eight questions asked the students to assess the extent to which each of the following words was descriptive of a particular car .

The next eight questions asked the students to rate their level of agreement with each of the following statements about a particular car (where 5= âStrongly Agreeâ and 1 = âStrongly Disagreeâ). The questions are:

The raw data are available in the file labeled mbacar.

#### a. Conduct a common factor analysis on the data set. How many factors you would retain? How do you interpret

them?

#### b. Save the factor scores and plot the average factor scores for each of the 10 cars evaluated by the students.

What do the plots tell you about the similarities of the 10 car models.

### Approach

Factor Analysis is a method for analyzing the covariation among the observed variables to address the questions:

1. How many latent factors are needed to account for most of the variation among the observed variables?
2. Which variables appear to define each factor; hence what labels should we give to these factors? If the observed covariation can be explained by a small number of factors (e.g., 2-5), this would increase our understanding of the relationships among the traits or variables!

#### Load required Libraries

library(MASS)  
library(corrplot)

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

library(psych)

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

library(GPArotation)  
library(psy)

##   
## Attaching package: 'psy'

## The following object is masked from 'package:psych':  
##   
## wkappa

library(sqldf)

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

## Loading required package: gsubfn

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

## Loading required package: proto

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

## Loading required package: RSQLite

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

#### Load user defined functions

#### Load the data set and see the data structure

### Load data  
setwd('D:/GL/AS/AS Group assignment')  
  
MBA\_car\_df = read.table("MBA Car Problem Data.txt",header=TRUE) # read text file   
str(MBA\_car\_df)

## '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 ...

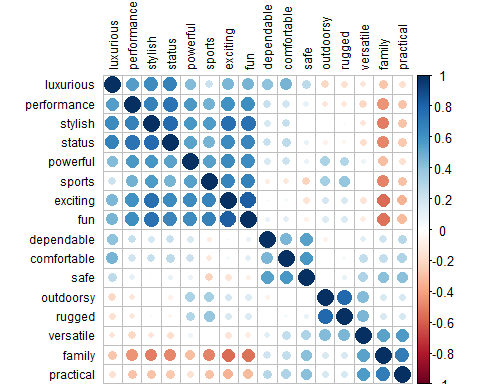
#### Observation

1. There are 303 observations of 19 variables.

For dataframes with many columns, corrplot can be useful to get a sense of the structure in the data (including larger scale organization)

We omit the columns, **student, car and discipline** from the data frame because they do not any value.

d <- MBA\_car\_df[,-c(1,2,19)]  
  
corrplot(cor(d), order = "hclust", tl.col='black', tl.cex=.75)



#### Observation

So there appears to be some structure in the data, with 2 general classes of attributes (e.g. negative and positive), and within each cluster, there are some sub-clusters that hang together (performance, stylish, status are negatively related with family and practical).

### Steps in Factor Analysis

1. Test assumptions of Factor Analysis such as Factorability
2. Exploratory Factor Analysis (EFA)
3. Determine the number of Factors
4. Identify which item belong to which factor
5. Drop items as necessary and repeat steps c and d
6. Name and define factors. For example - For Factor 1 - Pain Reliever Action and Factor 2 - Pain Reliever side effects. You may give good names.
7. Examine correlations among factors

Ref: Revelle, W.(2011),psych:Procedure for Personality and Psychological Research. <http://personality-project.org/r/psych.manual.pdf>)

### A) Test assumptions of Factor Analysis such as Factorability

#### 1) The variables used in factor analysis should be linearly related to each other.

This can be checked by looking at a scatter plots of pairs of variables.

Obviously the variables must be at least moderately correlated with each other, otherwise the number of factors will be almost the same as the number of original variables, which means that carrying out a factor analysis would be pointless.

*Correlation Matrix:* To do the factor analysis we must have variables that correlate fairly well with each other. The correlation matrix is generated in R to check the pattern of relationship between variables.

#create a correlation matrix  
dCar<-cor(d)  
round(dCar, 2)

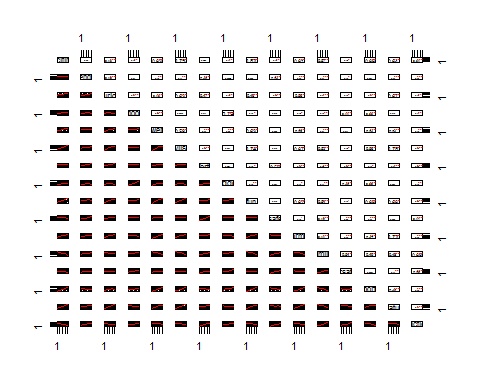
## exciting dependable luxurious outdoorsy powerful stylish  
## exciting 1.00 0.03 0.45 0.18 0.63 0.75  
## dependable 0.03 1.00 0.40 -0.07 0.16 0.17  
## luxurious 0.45 0.40 1.00 -0.21 0.43 0.63  
## outdoorsy 0.18 -0.07 -0.21 1.00 0.32 0.00  
## powerful 0.63 0.16 0.43 0.32 1.00 0.59  
## stylish 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  
## comfortable rugged fun safe performance family versatile  
## exciting 0.04 0.17 0.83 -0.15 0.61 -0.57 -0.14  
## dependable 0.46 0.00 0.09 0.54 0.23 0.21 0.13  
## luxurious 0.47 -0.16 0.46 0.26 0.56 -0.27 -0.13  
## outdoorsy 0.02 0.79 0.15 0.02 -0.14 0.16 0.44  
## powerful 0.22 0.30 0.63 0.09 0.58 -0.31 0.07  
## stylish 0.23 0.04 0.74 -0.02 0.67 -0.52 -0.14  
## comfortable 1.00 0.05 0.13 0.58 0.19 0.24 0.25  
## rugged 0.05 1.00 0.16 0.09 -0.12 0.17 0.46  
## fun 0.13 0.16 1.00 -0.08 0.62 -0.54 -0.10  
## safe 0.58 0.09 -0.08 1.00 0.10 0.42 0.30  
## performance 0.19 -0.12 0.62 0.10 1.00 -0.44 -0.20  
## family 0.24 0.17 -0.54 0.42 -0.44 1.00 0.54  
## versatile 0.25 0.46 -0.10 0.30 -0.20 0.54 1.00  
## sports -0.12 0.38 0.68 -0.22 0.47 -0.49 0.00  
## status 0.27 -0.05 0.66 0.09 0.73 -0.48 -0.17  
## practical 0.30 0.18 -0.32 0.41 -0.28 0.69 0.56  
## sports status practical  
## exciting 0.66 0.64 -0.34  
## dependable -0.09 0.22 0.29  
## luxurious 0.21 0.67 -0.15  
## outdoorsy 0.33 -0.08 0.17  
## powerful 0.54 0.54 -0.15  
## stylish 0.56 0.78 -0.28  
## comfortable -0.12 0.27 0.30  
## rugged 0.38 -0.05 0.18  
## fun 0.68 0.66 -0.32  
## safe -0.22 0.09 0.41  
## performance 0.47 0.73 -0.28  
## family -0.49 -0.48 0.69  
## versatile 0.00 -0.17 0.56  
## sports 1.00 0.46 -0.29  
## status 0.46 1.00 -0.27  
## practical -0.29 -0.27 1.00

#### Observation

From the above correlation matrix we can see that all the variables value of correlation coefficient greater than 0.46 with at least one other variable. Hence we can assume that variables are fairly correlated with each other and we can run Factor Analysis analysis on this data.

##### Look at a scatter plots of pairs of variables

par(mar=c(1,1,1,1)+0.1)  
pairs(d, lower.panel=panel.smooth, upper.panel=panel.cor)



#### Observation

**We observe the variables are moderately correlated with each other the highest being 0.83 and there are several values in the range 0.40 - 0.83**

#### 2) Use KMO to find out whether we have sufficient items for each factor.

Henry Kaiser (1970) introduced an Measure of Sampling Adequacy (MSA) of factor analytic data matrices. Kaiser and Rice (1974) then modified it. This is just a function of the squared elements of the `image' matrix compared to the squares of the original correlations. The overall MSA as well as estimates for each item are found. The index is known as the Kaiser-Meyer-Olkin (KMO) index.

#Apply Kaiser-Meyer-Olkin (KMO) Test, KMO should be > 0.5 :  
  
KMO(r = d)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = d)  
## Overall MSA = 0.88  
## MSA for each item =   
## exciting dependable luxurious outdoorsy powerful stylish   
## 0.91 0.84 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   
## 0.86 0.90 0.91 0.86

#### Observation

Since MSA being 0.88 > 0.5, we can run Factor Analysis on this data.

#### 3) Bartlett's test of sphericity

Bartlettâs test is used to check that the original variables are sufficiently correlated. This test should come out significant (p < 0.05) â if not, factor analysis will not be appropriate.

#Bartlett's test of sphericity (Should be significant)  
cortest.bartlett(d)

## R was not square, finding R from data

## $chisq  
## [1] 3395.747  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 120

#### Observation

The Bartlett's test has come out significant (p = 0); hence the original variables are sufficiently correlated and factor analysis is appropriate.

#### 4) Determinant of the correlation Matrix

det(dCar)

## [1] 1.034942e-05

#### Observation

The value of determinant of the correlation matrix as calculated by R is 0.00001034942. This value is greater than the necessary value of 0.00001, hence we can proceed further with our factor analysis.

#### 5) Anti-image correlation matrix diagonals - they should be > 0.5

X <- cor(d)   
iX <- ginv(X)   
S2 <- diag(diag((iX^-1)))  
AIS <- S2%\*%iX%\*%S2 # anti-image covariance matrix  
IS <- X+AIS-2\*S2 # image covariance matrix  
Dai <- sqrt(diag(diag(AIS)))  
IR <- ginv(Dai)%\*%IS%\*%ginv(Dai) # image correlation matrix  
AIR <- ginv(Dai)%\*%AIS%\*%ginv(Dai) # anti-image correlation matrix  
  
print(diag(AIR), row.names = FALSE)

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

#### Observation

We observe that the diagonals of the Anti - Image Correlation matrix to be 1.

##### Since Anti - Image Correlation matrix diagonals > 0.5, we can run **Factor Analysis** on this data.

#### 6) sample size should be large enough to yield reliable estimates of correlations among the variables:

1. Ideally, there should be a large ratio of N / k (Cases / items)
2. EFA can still be reasonably done with > ~ 5:1

In this case we have N = 303 and k = 16 and the ratio is **303 : 16** = ~ **19 : 1**.

##### **Hence the sample size is large enough to yield reliable estimates of the correlations among the variables.**

### B) Exploratory Factor Analysis (EFA)

**Exploratory Factor Analysis (EFA)** is generally used to discover structure of a measure and to examine its internal reliability. EFA is often recommended when researchers have no hypotheses about the nature of the underlying factor structure of their measure.

Exploratory factor analysis has three basic decision points:

* 1. Decide the number of factors
  2. Choosing an extraction method
  3. Choosing a rotation method

### C) Determine the number of Factors

#### (1) Decide the number of factors

The most common approach to deciding the number of factors is to generate a scree plot.

The scree plot is a two-dimensional graph with factors on the x-axis and *eigenvalues* on the y-axis.

Eigenvalues are produced by a process called **principal component analysis** (PCA) and represent the variance accounted for by each underlying factor. They are represented by scores that total to the number of items.

A 12-item scale will theoretically have 12 possible underlying factors, each factor will have an eigen value that indicates the amount of variation in the items accounted by each factor. If a, the first factor has an eigen value of 3.0, it accounts for 25% of the variance (3/12= .25). The total of all the eigen values will be 12 if there are 12 items, so some factors will have smaller eigenvalues. They are typically arranged in a scree plot in decending order.

#### (2) Choosing an extraction method and extraction

Once the number of factors are decided, you need to decide which mathematical solution to find the loadings. There are five basic extraction methods:

* 1. PCA - which assumes there is no measurement error and is considered not to be true exploratory factor analysis.
  2. Maximum Likelihood (a.k.a canonical factoring)
  3. Alpha Factoring
  4. Image Factoring
  5. Principal axis factoring with iterated communalities (a.k.a least squares)

#### Calculate initial factor loadings:

This can be done in a number of different ways: the two most common methods are described very briefly below:

* Principal Component Analysis (PCA) Method

As the name suggests, this method uses the method used to carry out a principal component analysis. However, the factors obtained will not actually be the principal components (although the loadings for the kth factor will be proportional to the coefficients of the kth principal component.

* Principal Axis Factoring

This is a method which tries to find the lowest number of factors which can account for the variability in the original variables, that is associated with these factors (this is in contrast to the principal components method which looks for a set of factors which can account for the total variability in the original variables).

These two methods will tend to give similar results if the variables are quite highly correlated and / or the number of original variables is quite high. Whichever method is used, the resulting factors at this stage will be uncorrelated.

### Principal axis factoring

We shall use **Principal axis factoring**, (fm="pa") because we are most interested in identifying the underlying constructs in the data.

The extraction method will produce factor loadings for every item in every extracted factor.

Now, We will use **fa()** function from the *psych* package, which received the following primary arguments:

* r: the correlation matrix
* nfactors: number of factors to be extracted (default 1)
* rotate: one of several matrix rotation methods, such as "varimax" or "oblimin" or "none"
* fm: one of several factoring methodsm such as **pa** (principal axis) or **ml** (maximum likelihood)

### We start by standardizing the variables.

d\_std = as.data.frame(scale(d))  
  
  
solution <- fa(r=d\_std, nfactors = 6, rotate="none",fm="pa")

## maximum iteration exceeded

###   
  
print(solution)

## Factor Analysis using method = pa  
## Call: fa(r = d\_std, nfactors = 6, rotate = "none", fm = "pa")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PA1 PA2 PA3 PA4 PA5 PA6 h2 u2 com  
## exciting 0.85 0.01 0.21 0.03 0.12 -0.11 0.80 0.199 1.2  
## dependable 0.13 0.51 -0.39 -0.05 -0.07 -0.11 0.45 0.549 2.2  
## luxurious 0.64 0.25 -0.46 -0.19 0.10 0.15 0.75 0.250 2.6  
## outdoorsy 0.04 0.42 0.76 -0.12 -0.10 0.06 0.78 0.223 1.7  
## powerful 0.69 0.29 0.19 0.01 -0.09 0.03 0.60 0.401 1.6  
## stylish 0.87 0.10 -0.07 -0.01 0.19 0.09 0.81 0.187 1.2  
## comfortable 0.15 0.62 -0.37 -0.17 0.05 0.03 0.58 0.420 2.0  
## rugged 0.05 0.47 0.73 -0.16 -0.12 0.04 0.80 0.199 1.9  
## fun 0.86 0.07 0.16 0.03 0.15 -0.23 0.85 0.146 1.3  
## safe -0.07 0.70 -0.38 -0.11 -0.21 -0.16 0.72 0.278 2.0  
## performance 0.81 0.07 -0.25 0.36 -0.33 0.07 0.98 0.023 2.0  
## family -0.67 0.56 -0.05 0.14 0.04 0.04 0.79 0.214 2.1  
## versatile -0.24 0.63 0.29 0.17 0.14 0.07 0.60 0.400 2.1  
## sports 0.68 -0.01 0.46 0.10 -0.01 -0.03 0.69 0.313 1.8  
## status 0.82 0.11 -0.21 0.01 0.03 0.12 0.75 0.252 1.2  
## practical -0.44 0.63 -0.04 0.27 0.20 -0.03 0.71 0.291 2.5  
##   
## PA1 PA2 PA3 PA4 PA5 PA6  
## SS loadings 5.67 2.84 2.26 0.38 0.34 0.17  
## Proportion Var 0.35 0.18 0.14 0.02 0.02 0.01  
## Cumulative Var 0.35 0.53 0.67 0.70 0.72 0.73  
## Proportion Explained 0.49 0.24 0.19 0.03 0.03 0.01  
## Cumulative Proportion 0.49 0.73 0.92 0.96 0.99 1.00  
##   
## Mean item complexity = 1.8  
## Test of the hypothesis that 6 factors are sufficient.  
##   
## The degrees of freedom for the null model are 120 and the objective function was 11.48 with Chi Square of 3395.75  
## The degrees of freedom for the model are 39 and the objective function was 0.21   
##   
## The root mean square of the residuals (RMSR) is 0.01   
## The df corrected root mean square of the residuals is 0.02   
##   
## The harmonic number of observations is 303 with the empirical chi square 10.19 with prob < 1   
## The total number of observations was 303 with Likelihood Chi Square = 62.09 with prob < 0.011   
##   
## Tucker Lewis Index of factoring reliability = 0.978  
## RMSEA index = 0.046 and the 90 % confidence intervals are 0.022 0.064  
## BIC = -160.75  
## Fit based upon off diagonal values = 1  
## Measures of factor score adequacy   
## PA1 PA2 PA3 PA4 PA5  
## Correlation of scores with factors 0.98 0.95 0.95 0.84 0.84  
## Multiple R square of scores with factors 0.97 0.91 0.90 0.71 0.71  
## Minimum correlation of possible factor scores 0.94 0.81 0.80 0.41 0.42  
## PA6  
## Correlation of scores with factors 0.67  
## Multiple R square of scores with factors 0.45  
## Minimum correlation of possible factor scores -0.11

#### Observation

1. From the SS Loadings entry, we infer that all three eigen values are > 1 and explains 67% of the total variation.
2. Factor loadings are not clear. There are many **double-loading** (Refer dependable, luxurious, outdoorsy, comfortable, outdoorsy, etc.)

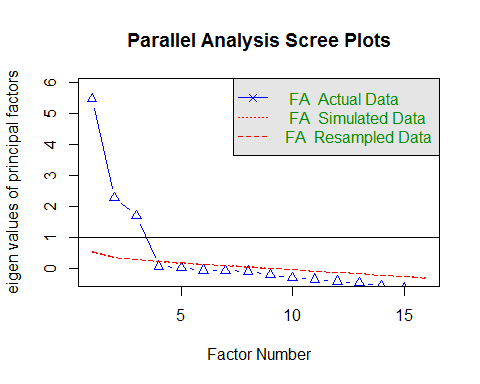
#### Number of Factors

Next we find out the number of factors that we all be selecting for factor analysis. This can be evaluated via method such as Parallel Analysis and eigenvalue, etc.

#### Parallel Analysis

We use Psych packages fa.parallel function to execute parallel analysis. Here we specify the data frame and factor method (minres in our case). Run the following to find acceptable number of factors and generate the scree plot:

parallel <- fa.parallel(d, fm = 'minres', fa = 'fa')



## Parallel analysis suggests that the number of factors = 3 and the number of components = NA

#### Observation

The blue line shows eigenvalues of actual data and the two red lines (placed on top of each other) show simulated and resampled data. Here we look at the large drops in the actual data and spot the point where it levels off to the right. Also we locate the point of inflection  the point where the gap between simulated data and actual data tends to be minimum.

Looking at this plot and parallel analysis, anywhere between 2 to 4 factors factors would be good choice.

**So we shall use 3 as the number of factors for performing Factor Analysis.**

#### (3) Choosing a rotation method

We observe that the **factor loadings** are not clear. Once an initial solution is obtained, the loadings are rotated. **Factor Rotation is used to increase interpretability**.

Rotation is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved.

There are two types of rotation method, **orthogonal** and **oblique** rotation. In orthogonal rotation the rotated factors will remain uncorrelated whereas in oblique rotation the resulting factors will be correlated.

There are a number of different methods of rotation of each type.

The most common orthogonal method is called **varimax** rotation; this is the method that many books recommend.

<http://www.statstutor.ac.uk/resources/uploaded/factoranalysis.pdf>

## D) Identify which item belong to which factor

**Criteria for Practical and Statistical Significance of Factor Loadings:**

Factor loading can be classified based on their magnitude.

|  |  |
| --- | --- |
| Value | Interpretation |
| > .30 | Minimum consideration level |
| > .40 | More important |
| > .50 | Practically significant |

#### Orthogonal Rotation (varimax):

Assuming that there is no correlation between the extracted factors, we will carry out a varimax rotation.

Rotated component matrix obtained after **varimax** rotation is shown below:

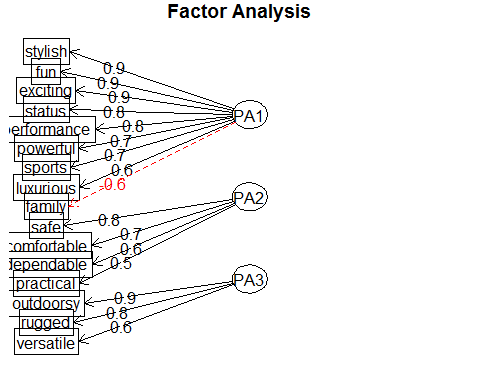
# Orthogonal varimax rotation  
  
solution1 <- fa(r=d, nfactors = 3, rotate="varimax",fm="pa")  
print(solution1)

## Factor Analysis using method = pa  
## Call: fa(r = d, nfactors = 3, rotate = "varimax", fm = "pa")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PA1 PA2 PA3 h2 u2 com  
## exciting 0.86 -0.14 0.12 0.77 0.23 1.1  
## dependable 0.15 0.64 -0.06 0.44 0.56 1.1  
## luxurious 0.64 0.43 -0.28 0.67 0.33 2.2  
## outdoorsy 0.08 -0.05 0.86 0.75 0.25 1.0  
## powerful 0.71 0.12 0.27 0.60 0.40 1.3  
## stylish 0.87 0.09 -0.06 0.77 0.23 1.0  
## comfortable 0.18 0.72 0.01 0.55 0.45 1.1  
## rugged 0.10 0.01 0.85 0.73 0.27 1.0  
## fun 0.86 -0.06 0.11 0.76 0.24 1.0  
## safe -0.04 0.77 0.06 0.60 0.40 1.0  
## performance 0.77 0.14 -0.19 0.65 0.35 1.2  
## family -0.64 0.52 0.30 0.77 0.23 2.4  
## versatile -0.19 0.38 0.59 0.53 0.47 1.9  
## sports 0.69 -0.28 0.35 0.68 0.32 1.8  
## status 0.83 0.18 -0.17 0.74 0.26 1.2  
## practical -0.40 0.55 0.32 0.56 0.44 2.5  
##   
## PA1 PA2 PA3  
## SS loadings 5.58 2.61 2.38  
## Proportion Var 0.35 0.16 0.15  
## Cumulative Var 0.35 0.51 0.66  
## Proportion Explained 0.53 0.25 0.23  
## Cumulative Proportion 0.53 0.77 1.00  
##   
## Mean item complexity = 1.4  
## Test of the hypothesis that 3 factors are sufficient.  
##   
## The degrees of freedom for the null model are 120 and the objective function was 11.48 with Chi Square of 3395.75  
## The degrees of freedom for the model are 75 and the objective function was 0.8   
##   
## The root mean square of the residuals (RMSR) is 0.03   
## The df corrected root mean square of the residuals is 0.04   
##   
## The harmonic number of observations is 303 with the empirical chi square 57.75 with prob < 0.93   
## The total number of observations was 303 with Likelihood Chi Square = 233.86 with prob < 3.2e-18   
##   
## Tucker Lewis Index of factoring reliability = 0.922  
## RMSEA index = 0.085 and the 90 % confidence intervals are 0.072 0.096  
## BIC = -194.67  
## Fit based upon off diagonal values = 0.99  
## Measures of factor score adequacy   
## PA1 PA2 PA3  
## Correlation of scores with factors 0.98 0.93 0.94  
## Multiple R square of scores with factors 0.95 0.87 0.89  
## Minimum correlation of possible factor scores 0.91 0.74 0.78

print(solution1$loadings, cutoff = 0.5, sort=TRUE)

##   
## Loadings:  
## PA1 PA2 PA3   
## exciting 0.860   
## luxurious 0.638   
## powerful 0.713   
## stylish 0.871   
## fun 0.861   
## performance 0.770   
## family -0.636 0.524   
## sports 0.694   
## status 0.826   
## dependable 0.642   
## comfortable 0.722   
## safe 0.771   
## practical 0.545   
## outdoorsy 0.858  
## rugged 0.848  
## versatile 0.594  
##   
## PA1 PA2 PA3  
## SS loadings 5.581 2.607 2.381  
## Proportion Var 0.349 0.163 0.149  
## Cumulative Var 0.349 0.512 0.661

fa.diagram(solution1)



**Observation:**

The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). The curved arrows are the correlations between the factors. If no curved arrow is present, then the correlation between the factors is not great.

From the above output we can see that on applying varimax rotation, the component loadings are very clear.

We see two variables have double-loading for the variable *family*.

We consider the loadings more than 0.5 and not loading on more than one factor. Note that negative values are acceptable here.

There are double loading.

Let us try with four factors and to check whether it eliminates double-loading.

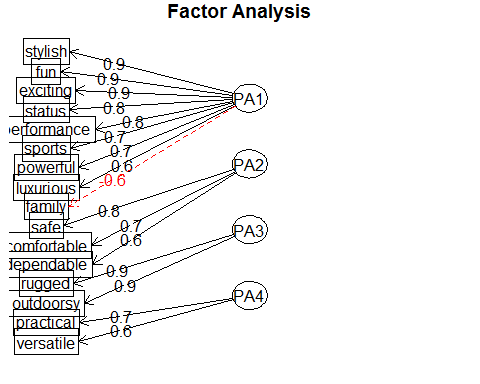
# Orthogonal varimax rotation  
  
solution1 <- fa(r=d, nfactors = 4, rotate="varimax",fm="pa")  
print(solution1)

## Factor Analysis using method = pa  
## Call: fa(r = d, nfactors = 4, rotate = "varimax", fm = "pa")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PA1 PA2 PA3 PA4 h2 u2 com  
## exciting 0.86 -0.09 0.14 -0.13 0.78 0.22 1.1  
## dependable 0.12 0.64 -0.06 0.13 0.44 0.56 1.2  
## luxurious 0.58 0.51 -0.22 -0.16 0.68 0.32 2.4  
## outdoorsy 0.08 -0.06 0.86 0.15 0.78 0.22 1.1  
## powerful 0.69 0.17 0.30 -0.05 0.59 0.41 1.5  
## stylish 0.87 0.14 -0.04 -0.10 0.78 0.22 1.1  
## comfortable 0.15 0.72 0.02 0.15 0.57 0.43 1.2  
## rugged 0.08 0.01 0.88 0.14 0.80 0.20 1.1  
## fun 0.86 -0.02 0.12 -0.10 0.77 0.23 1.1  
## safe -0.08 0.77 0.07 0.19 0.64 0.36 1.2  
## performance 0.75 0.20 -0.16 -0.14 0.65 0.35 1.3  
## family -0.57 0.34 0.15 0.57 0.78 0.22 2.8  
## versatile -0.10 0.20 0.43 0.59 0.59 0.41 2.2  
## sports 0.71 -0.26 0.34 -0.06 0.69 0.31 1.8  
## status 0.80 0.25 -0.13 -0.16 0.74 0.26 1.3  
## practical -0.29 0.33 0.12 0.71 0.71 0.29 1.9  
##   
## PA1 PA2 PA3 PA4  
## SS loadings 5.23 2.28 2.09 1.40  
## Proportion Var 0.33 0.14 0.13 0.09  
## Cumulative Var 0.33 0.47 0.60 0.69  
## Proportion Explained 0.48 0.21 0.19 0.13  
## Cumulative Proportion 0.48 0.68 0.87 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 120 and the objective function was 11.48 with Chi Square of 3395.75  
## The degrees of freedom for the model are 62 and the objective function was 0.5   
##   
## 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 29.12 with prob < 1   
## The total number of observations was 303 with Likelihood Chi Square = 147.72 with prob < 5.9e-09   
##   
## Tucker Lewis Index of factoring reliability = 0.949  
## RMSEA index = 0.069 and the 90 % confidence intervals are 0.054 0.082  
## BIC = -206.53  
## Fit based upon off diagonal values = 1  
## Measures of factor score adequacy   
## PA1 PA2 PA3 PA4  
## Correlation of scores with factors 0.96 0.90 0.93 0.82  
## Multiple R square of scores with factors 0.93 0.82 0.87 0.67  
## Minimum correlation of possible factor scores 0.86 0.64 0.74 0.34

print(solution1$loadings, cutoff = 0.5, sort=TRUE)

##   
## Loadings:  
## PA1 PA2 PA3 PA4   
## exciting 0.862   
## luxurious 0.584 0.512   
## powerful 0.689   
## stylish 0.867   
## fun 0.863   
## performance 0.751   
## family -0.568 0.566  
## sports 0.709   
## status 0.798   
## dependable 0.639   
## comfortable 0.722   
## safe 0.773   
## outdoorsy 0.864   
## rugged 0.882   
## versatile 0.595  
## practical 0.706  
##   
## PA1 PA2 PA3 PA4  
## SS loadings 5.232 2.285 2.092 1.401  
## Proportion Var 0.327 0.143 0.131 0.088  
## Cumulative Var 0.327 0.470 0.601 0.688

fa.diagram(solution1)



#### Observation

We observe that there is double-loading with item, family connected to two factors with values -0.568 and 0.566.

Let us try Ml (maximum likelihood) factoring method and to check whether it eliminates double-loading.

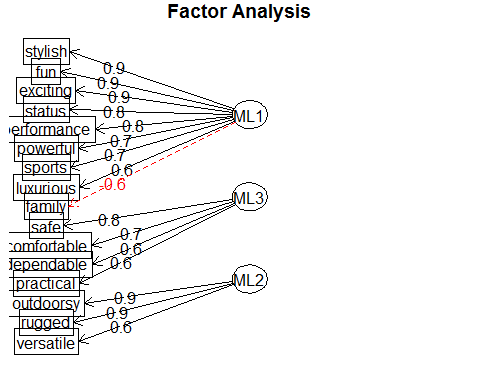
# Orthogonal varimax rotation - Factoring method - ml  
  
solution2 <- fa(r=d, nfactors = 3, rotate="varimax",fm="ml")  
print(solution2)

## Factor Analysis using method = ml  
## Call: fa(r = d, nfactors = 3, rotate = "varimax", fm = "ml")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## ML1 ML3 ML2 h2 u2 com  
## exciting 0.86 -0.15 0.13 0.78 0.22 1.1  
## dependable 0.15 0.64 -0.06 0.43 0.57 1.1  
## luxurious 0.64 0.41 -0.28 0.66 0.34 2.1  
## outdoorsy 0.07 -0.04 0.87 0.77 0.23 1.0  
## powerful 0.71 0.11 0.28 0.59 0.41 1.4  
## stylish 0.88 0.08 -0.06 0.78 0.22 1.0  
## comfortable 0.19 0.72 0.01 0.56 0.44 1.1  
## rugged 0.09 0.02 0.86 0.76 0.24 1.0  
## fun 0.86 -0.08 0.12 0.77 0.23 1.1  
## safe -0.03 0.76 0.06 0.58 0.42 1.0  
## performance 0.77 0.13 -0.20 0.65 0.35 1.2  
## family -0.63 0.53 0.27 0.75 0.25 2.3  
## versatile -0.18 0.40 0.57 0.52 0.48 2.0  
## sports 0.69 -0.28 0.35 0.68 0.32 1.9  
## status 0.83 0.17 -0.16 0.74 0.26 1.2  
## practical -0.39 0.56 0.29 0.55 0.45 2.3  
##   
## ML1 ML3 ML2  
## SS loadings 5.58 2.61 2.38  
## Proportion Var 0.35 0.16 0.15  
## Cumulative Var 0.35 0.51 0.66  
## Proportion Explained 0.53 0.25 0.22  
## Cumulative Proportion 0.53 0.78 1.00  
##   
## Mean item complexity = 1.4  
## Test of the hypothesis that 3 factors are sufficient.  
##   
## The degrees of freedom for the null model are 120 and the objective function was 11.48 with Chi Square of 3395.75  
## The degrees of freedom for the model are 75 and the objective function was 0.79   
##   
## The root mean square of the residuals (RMSR) is 0.03   
## The df corrected root mean square of the residuals is 0.04   
##   
## The harmonic number of observations is 303 with the empirical chi square 60.52 with prob < 0.89   
## The total number of observations was 303 with Likelihood Chi Square = 230.9 with prob < 8.9e-18   
##   
## Tucker Lewis Index of factoring reliability = 0.923  
## RMSEA index = 0.085 and the 90 % confidence intervals are 0.071 0.095  
## BIC = -197.63  
## Fit based upon off diagonal values = 0.99  
## Measures of factor score adequacy   
## ML1 ML3 ML2  
## Correlation of scores with factors 0.98 0.93 0.95  
## Multiple R square of scores with factors 0.95 0.87 0.90  
## Minimum correlation of possible factor scores 0.91 0.73 0.79

print(solution2$loadings, cutoff = 0.5, sort=TRUE)

##   
## Loadings:  
## ML1 ML3 ML2   
## exciting 0.862   
## luxurious 0.644   
## powerful 0.707   
## stylish 0.878   
## fun 0.864   
## performance 0.769   
## family -0.627 0.535   
## sports 0.688   
## status 0.829   
## dependable 0.635   
## comfortable 0.723   
## safe 0.759   
## practical 0.561   
## outdoorsy 0.873  
## rugged 0.865  
## versatile 0.569  
##   
## ML1 ML3 ML2  
## SS loadings 5.576 2.612 2.375  
## Proportion Var 0.348 0.163 0.148  
## Cumulative Var 0.348 0.512 0.660

fa.diagram(solution2)



#### Observation

From the diagram, we infer 1) The variable, family is connected to the factor MA1 by a red line and the loading value is negative. 2) Factor loading of the attributes as:

|  |  |
| --- | --- |
| Factor | Items belonging to factor |
| ML1 | stylish, fun, exciting, status, performance, powerful, sports, luxurious |
| ML2 | outdoorsy, rugged, versatile |
| ML3 | safe, comfortable, dependable, practical |

### E) Drop items as necessary and repeat steps c and d

* Based on the absolute value of item per loading, the diagram is drawn.
* From factor loadings table, We observe that there is double-loading with item, family connected to two factors with values -0.627 and 0.535.
* We have tried different factoring methods - ML and PA and still we find that the item family is connected to with ML1 and ML3. But we observe that **family** is more connected to ML3 than ML1.
* Let us drop item **family** and repeat steps C and D.

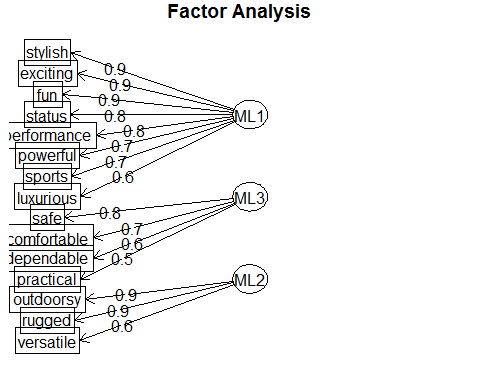
# Orthogonal varimax rotation - Factoring method - ml  
d\_family\_removed <- subset(d,select = -family)  
solution3 <- fa(r=d\_family\_removed, nfactors = 3, rotate="varimax",fm="ml")  
print(solution3)

## Factor Analysis using method = ml  
## Call: fa(r = d\_family\_removed, nfactors = 3, rotate = "varimax", fm = "ml")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## ML1 ML3 ML2 h2 u2 com  
## exciting 0.87 -0.14 0.11 0.79 0.21 1.1  
## dependable 0.14 0.65 -0.04 0.44 0.56 1.1  
## luxurious 0.63 0.44 -0.27 0.66 0.34 2.2  
## outdoorsy 0.08 -0.05 0.88 0.78 0.22 1.0  
## powerful 0.71 0.12 0.27 0.59 0.41 1.4  
## stylish 0.88 0.09 -0.07 0.78 0.22 1.0  
## comfortable 0.17 0.75 0.03 0.59 0.41 1.1  
## rugged 0.10 0.01 0.87 0.78 0.22 1.0  
## fun 0.87 -0.07 0.10 0.77 0.23 1.0  
## safe -0.05 0.78 0.08 0.62 0.38 1.0  
## performance 0.77 0.14 -0.21 0.65 0.35 1.2  
## versatile -0.17 0.34 0.56 0.46 0.54 1.9  
## sports 0.70 -0.28 0.33 0.67 0.33 1.8  
## status 0.82 0.20 -0.17 0.74 0.26 1.2  
## practical -0.37 0.49 0.29 0.46 0.54 2.5  
##   
## ML1 ML3 ML2  
## SS loadings 5.17 2.33 2.29  
## Proportion Var 0.34 0.16 0.15  
## Cumulative Var 0.34 0.50 0.65  
## Proportion Explained 0.53 0.24 0.23  
## Cumulative Proportion 0.53 0.77 1.00  
##   
## Mean item complexity = 1.4  
## Test of the hypothesis that 3 factors are sufficient.  
##   
## The degrees of freedom for the null model are 105 and the objective function was 10.24 with Chi Square of 3032.41  
## The degrees of freedom for the model are 63 and the objective function was 0.65   
##   
## The root mean square of the residuals (RMSR) is 0.03   
## The df corrected root mean square of the residuals is 0.04   
##   
## The harmonic number of observations is 303 with the empirical chi square 54.33 with prob < 0.77   
## The total number of observations was 303 with Likelihood Chi Square = 190.96 with prob < 8.3e-15   
##   
## Tucker Lewis Index of factoring reliability = 0.927  
## RMSEA index = 0.084 and the 90 % confidence intervals are 0.069 0.095  
## BIC = -169.01  
## Fit based upon off diagonal values = 0.99  
## Measures of factor score adequacy   
## ML1 ML3 ML2  
## Correlation of scores with factors 0.97 0.92 0.95  
## Multiple R square of scores with factors 0.95 0.85 0.90  
## Minimum correlation of possible factor scores 0.90 0.70 0.79

print(solution3$loadings,cutoff = 0.5, sort=TRUE)

##   
## Loadings:  
## ML1 ML3 ML2   
## exciting 0.872   
## luxurious 0.628   
## powerful 0.708   
## stylish 0.877   
## fun 0.870   
## performance 0.765   
## sports 0.698   
## status 0.820   
## dependable 0.648   
## comfortable 0.747   
## safe 0.782   
## outdoorsy 0.877  
## rugged 0.875  
## versatile 0.561  
## practical   
##   
## ML1 ML3 ML2  
## SS loadings 5.167 2.330 2.294  
## Proportion Var 0.344 0.155 0.153  
## Cumulative Var 0.344 0.500 0.653

fa.diagram(solution3)



From the diagram, we infer:

1. There is only single loading of items and there is no double-loading.
2. Factor loading of the attributes as:

|  |  |
| --- | --- |
| Factor | Items belonging to factor |
| ML1 | stylish, exciting, fun, status, performance, powerful, sports, luxurious |
| ML2 | outdoorsy, rugged, versatile |
| ML3 | safe, comfortable, dependable, practical |

### F) Name and define factors

|  |  |
| --- | --- |
| Factor | Name |
| ML1 | Delux |
| ML2 | Adventure |
| ML3 | Family |

### G) Examine correlations among factors

solution4 <- fa(r=d\_family\_removed, nfactors = 3, rotate="oblimin",fm="ml")  
print(solution4)

## Factor Analysis using method = ml  
## Call: fa(r = d\_family\_removed, nfactors = 3, rotate = "oblimin", fm = "ml")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## ML1 ML3 ML2 h2 u2 com  
## exciting 0.87 -0.16 0.13 0.79 0.21 1.1  
## dependable 0.14 0.64 -0.04 0.44 0.56 1.1  
## luxurious 0.64 0.41 -0.26 0.66 0.34 2.1  
## outdoorsy 0.05 -0.01 0.88 0.78 0.22 1.0  
## powerful 0.70 0.11 0.29 0.59 0.41 1.4  
## stylish 0.88 0.06 -0.05 0.78 0.22 1.0  
## comfortable 0.17 0.74 0.03 0.59 0.41 1.1  
## rugged 0.06 0.05 0.88 0.78 0.22 1.0  
## fun 0.87 -0.09 0.11 0.77 0.23 1.1  
## safe -0.06 0.79 0.08 0.62 0.38 1.0  
## performance 0.77 0.11 -0.20 0.65 0.35 1.2  
## versatile -0.20 0.38 0.56 0.46 0.54 2.0  
## sports 0.69 -0.28 0.34 0.67 0.33 1.9  
## status 0.83 0.16 -0.15 0.74 0.26 1.2  
## practical -0.39 0.51 0.28 0.46 0.54 2.5  
##   
## ML1 ML3 ML2  
## SS loadings 5.17 2.33 2.29  
## Proportion Var 0.34 0.16 0.15  
## Cumulative Var 0.34 0.50 0.65  
## Proportion Explained 0.53 0.24 0.23  
## Cumulative Proportion 0.53 0.77 1.00  
##   
## With factor correlations of   
## ML1 ML3 ML2  
## ML1 1.00 0.03 0.02  
## ML3 0.03 1.00 -0.05  
## ML2 0.02 -0.05 1.00  
##   
## Mean item complexity = 1.4  
## Test of the hypothesis that 3 factors are sufficient.  
##   
## The degrees of freedom for the null model are 105 and the objective function was 10.24 with Chi Square of 3032.41  
## The degrees of freedom for the model are 63 and the objective function was 0.65   
##   
## The root mean square of the residuals (RMSR) is 0.03   
## The df corrected root mean square of the residuals is 0.04   
##   
## The harmonic number of observations is 303 with the empirical chi square 54.33 with prob < 0.77   
## The total number of observations was 303 with Likelihood Chi Square = 190.96 with prob < 8.3e-15   
##   
## Tucker Lewis Index of factoring reliability = 0.927  
## RMSEA index = 0.084 and the 90 % confidence intervals are 0.069 0.095  
## BIC = -169.01  
## Fit based upon off diagonal values = 0.99  
## Measures of factor score adequacy   
## ML1 ML3 ML2  
## Correlation of scores with factors 0.97 0.92 0.95  
## Multiple R square of scores with factors 0.95 0.85 0.90  
## Minimum correlation of possible factor scores 0.90 0.70 0.79

#### Observation

##### Correlation between factors

|  |  |  |  |
| --- | --- | --- | --- |
| Factor | ML1 | ML2 | ML3 |
| ML1 | 100% | 3% | 2% |
| ML2 | 3% | 100% | -5% |
| ML3 | 2% | -5% | 100% |

From the above table, we observe that: a) Correlation between ML1 and ML2 is 3% b) Correlation between ML1 and ML3 is 2% c) Correlation between ML2 and ML3 is -5%

In the diagram, the curved arrows are the correlations between the factors.

Since no curved arrow is present, then the correlation between the factors is not great.

#### b. Save the factor scores and plot the average factor scores for each of the 10 cars evaluated by the students.

#### What do the plots tell you about the similarities of the 10 car models.

fac\_scores <- solution3$scores  
fac\_scores\_df <- data.frame(Car = MBA\_car\_df$Car, factor\_scores = fac\_scores)  
fac\_scores\_sorted\_df <- fac\_scores\_df[order(fac\_scores\_df[,1]),]  
colnames(fac\_scores\_sorted\_df) <- c("Car","ML1","ML2","ML3")  
avg\_scores <- sqldf('SELECT Car, AVG(ML1) AS ML1\_Score, AVG(ML2) AS ML2\_Score, AVG(ML3) AS ML3\_Score FROM fac\_scores\_sorted\_df GROUP BY Car')  
###  
  
avg\_scores[] <- lapply(avg\_scores, abs)   
  
y1 <- round(avg\_scores$ML1\_Score \* 100,0)  
y2 <- round(avg\_scores$ML2\_Score \* 100,0)  
y3 <- round(avg\_scores$ML3\_Score \* 100,0)  
  
  
print(y1)

## [1] 54 12 3 1 23 171 43 29 138 96

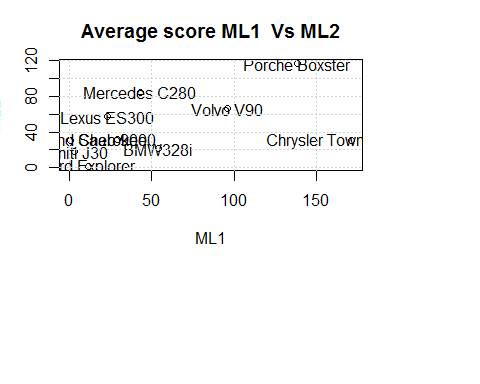
print(y2)

## [1] 21 1 18 31 57 30 85 32 117 66

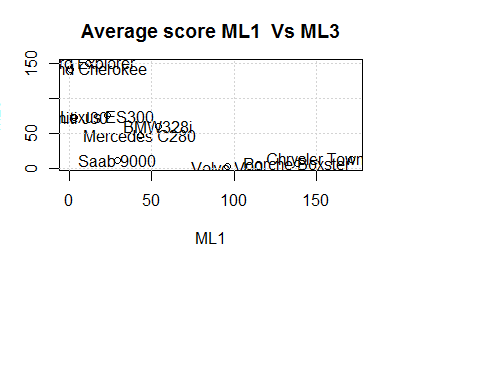
print(y3)

## [1] 60 150 73 142 75 12 47 11 7 2

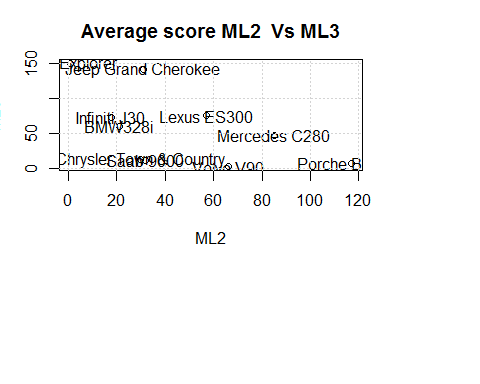
xnames <- c("BMW328i","Ford Explorer","Infiniti J30","Jeep Grand Cherokee","Lexus ES300","Chrysler Town & Country","Mercedes C280","Saab 9000","Porche Boxster","Volvo V90")  
  
par(mar=c(11, 3, 3, 6) + 0.1)  
  
### Factor ML1 Vs ML2  
  
plot(x=y1, y= y2, main="Average score ML1 Vs ML2",xlab = "ML1", ylab = "ML2")  
text (x=y1, y= y2,xnames)  
grid()



### Factor ML1 Vs ML3  
  
plot(x=y1, y= y3, main="Average score ML1 Vs ML3",xlab = "ML1", ylab = "ML3")  
text (x=y1, y= y3,xnames)  
grid()



### Factor ML2 Vs ML3  
  
plot(x=y2, y= y3, main="Average score ML2 Vs ML3",xlab = "ML2", ylab = "ML3")  
text (x=y2, y= y3,xnames)  
grid()



##### Interpretation

ML1 Vs ML2

|  |  |
| --- | --- |
| Cars | Remarks |
| Chrysler Town & Country | Closer to ML1 |
| Lexus ES300, Mercedes C280, Saab 9000 | Closer to ML2 |
| Porche Boxster,Volvo V90 | Closer to both ML1 and ML2 |

**Note**: Cars - BMW328i, Ford Explorer, Infiniti J30,Jeep Grand Cherokee is neither close to ML1 nor ML2

ML1 Vs ML3

|  |  |
| --- | --- |
| Cars | Remarks |
| Ford Explorer,Jeep Grand Cherokee,Jeep Grand Cherokee,Lexus ES300 | Closer to ML3 |
| Volvo V90, Porche Boxster, Chrysler Town & Country | Closer to ML1 |

**Note**: Cars - Saab 9000 is neither close to ML2 nor ML3

ML2 Vs ML3

|  |  |
| --- | --- |
| Cars | Remarks |
| Volvo V90, Porche Boxster | Closer to ML2 |
| Infiniti J30, BMW328i, Ford Explorer, Jeep Grand Cherokee | Closer to ML3 |
| Lexus ES300, Mercedes C280 | Closer to both ML2 and ML3 |

**Note**: Cars - Chrysler Town & Country, Saab 9000 are neither close to ML2 nor ML3