# Técnicas de Análisis Multivariado - Trabajo 1

*Justo Manrique Urbina - 20091107*

*12/14/2019*

El presente trabajo tiene como objetivo ilustrar el uso de técnicas de análisis multivariado revisadas en la Maestría de Estadística PUCP. Para ello, se hizo uso de distintas bases de datos orientadas a la aplicación de dichas técnicas, las cuales se presentarán en las secciones correspondientes. Cada tipo de análisis tiene como base un problema de negocio o investigación, así como una base de datos la cual es útil para brindar solución al problema. Posteriormente, se analizan los resultados de dichas técnicas y se concluye sobre la misma.

Las técnicas multivariadas utilizadas en el presente informe son:

- Análisis de Componentes Principales.
- Análisis Discriminante.
- Análisis Factorial.

Ver a continuación el uso de cada técnica.

## Análisis de Componentes Principales

### Introducción y Datos

La base de datos sobre la cual aplicaremos el análisis de componentes principales proviene dentro de la instalación base del programa R. La base de datos es conocida como "mtcars" y consiste en aspectos de diseño y rendimiento para 32 automóviles. Dicha base de datos proviene de un estudio del año 1981 de Henderson y Velleman. Dicho estudio fue publicado en el journal *Biometrics*. Mayor información al respecto se puede encontrar en https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html.

Los datos provienen de la revista *Motor Trend*, edición 1974. Dicha base de datos tiene las siguientes variables:

- mpg: Millas por galón.
- cyl: Número de cilindros.
- disp: Desplazamiento (en pies cúbicos).
- hp: Caballos de fuerza.
- drat: Relación del eje trasero.
- wt: Peso (definido en miles de libras).
- qsec: Tiempo para llegar a recorrer un cuarto de libra.
- am: Transmisión (automático o manual).
- vs: Tipo de motor (en forma de V o recto).
- gear: Cantidad de marchas hacia adelante.
- carb: Número de carburadores.

Los objetivos del presente estudio son:

- Conocer si existen grupos de autos con perfiles de rendimiento similares e identificar si existen autos de distinta clase.
- Identificar aquellos autos que lideran cada clase.

Con el propósito de ejecutar el análisis de los datos, realizamos la carga de librerías e importamos los datos. Posteriormente, realizamos un preprocesamiento de los datos para convertir los valores binarios en variables cualitativas.

```
## Carga de datos y librerías ##
library(FactoMineR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
library(reshape2)
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(biotools)
```

```
## Loading required package: rpanel

## Loading required package: tcltk

## Package `rpanel', version 1.1-4: type help(rpanel) for summary information

##
## Attaching package: 'rpanel'

## The following object is masked from 'package:tidyr':
##
##     population

## Loading required package: tkrplot

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: lattice

## Loading required package: SpatialEpi

## Loading required package: sp

## ---
## biotools version 3.1
```

```
##
```

```
library(MASS)
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(klaR)

data("mtcars")

## Preprocesamiento de datos ##

mtcars$vs <- factor(mtcars$vs)
mtcars$vs <- recode_factor(mtcars$vs, `0`="V-shaped",`1`="Straight")
mtcars$am <- factor(mtcars$am)
mtcars$am <- recode_factor(mtcars$am, `0`="Automatic",`1`="Manual")
```

Posteriormente, se observa los primeros valores de la base de datos para entender su estructura.

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec       vs        am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46 V-shaped    Manual    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02 V-shaped    Manual    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61 Straight    Manual    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44 Straight Automatic    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02 V-shaped Automatic    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22 Straight Automatic    3
##                   carb
## Mazda RX4            4
## Mazda RX4 Wag        4
## Datsun 710           1
## Hornet 4 Drive       1
## Hornet Sportabout    2
## Valiant              1
```

Se observa que cada línea corresponde a un modelo de auto específico. Asimismo, se observa que todas las variables son cuantitavias, excepto por las variables 'am' y 'vs'.

## Resultados

Posteriormente, se utilizó la matriz de correlación para entender las relaciones lineales que tiene cada variable respecto a otra. Se utilizaron solo las variables cuantitativas para este análisis:

```
corcars <- cor(mtcars[c(1:7,10:11)])
corcars
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
```

```
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec       gear       carb
## mpg   0.41868403  0.4802848 -0.5509251
## cyl  -0.59124207 -0.4926866  0.5269883
## disp -0.43369788 -0.5555692  0.3949769
## hp   -0.70822339 -0.1257043  0.7498125
## drat  0.09120476  0.6996101 -0.0907898
## wt   -0.17471588 -0.5832870  0.4276059
## qsec  1.00000000 -0.2126822 -0.6562492
## gear -0.21268223  1.0000000  0.2740728
## carb -0.65624923  0.2740728  1.0000000
```

Se observa lo siguiente:

- Se observan correlaciones negativas fuertes en los siguientes pares de variables: (mpg) Millas por galón y (cyl) Números de cilindros; (mpg) Millas por galón y (hp) Caballos de fuerza; (mpg) Millas por galón y (disp) Desplazamiento (en pies cúbicos); (mpg) Millas por galón y (wt) Peso (definido en miles de libras).
- Se observan correlaciones positivas fuertes en los siguientes pares de variables: (cyl) Número de cilintros y (disp) Desplazamiento (en pies cúbicos); (cyl) Número de cilindros y (hp) Caballos de fuerza; (disp) Desplazamiento (en pies cúbicos) y (wt) Peso (definido en miles de libras).

En base a este análisis, podemos intuir que aquellas variables que tengan una correlación positiva fuerte formarán parte de un componente, mientras que aquellos con correlación negativa fuerte formarán parte de distintos componentes.

Posteriormente, utilizamos el test de esfericidad de Bartlett:

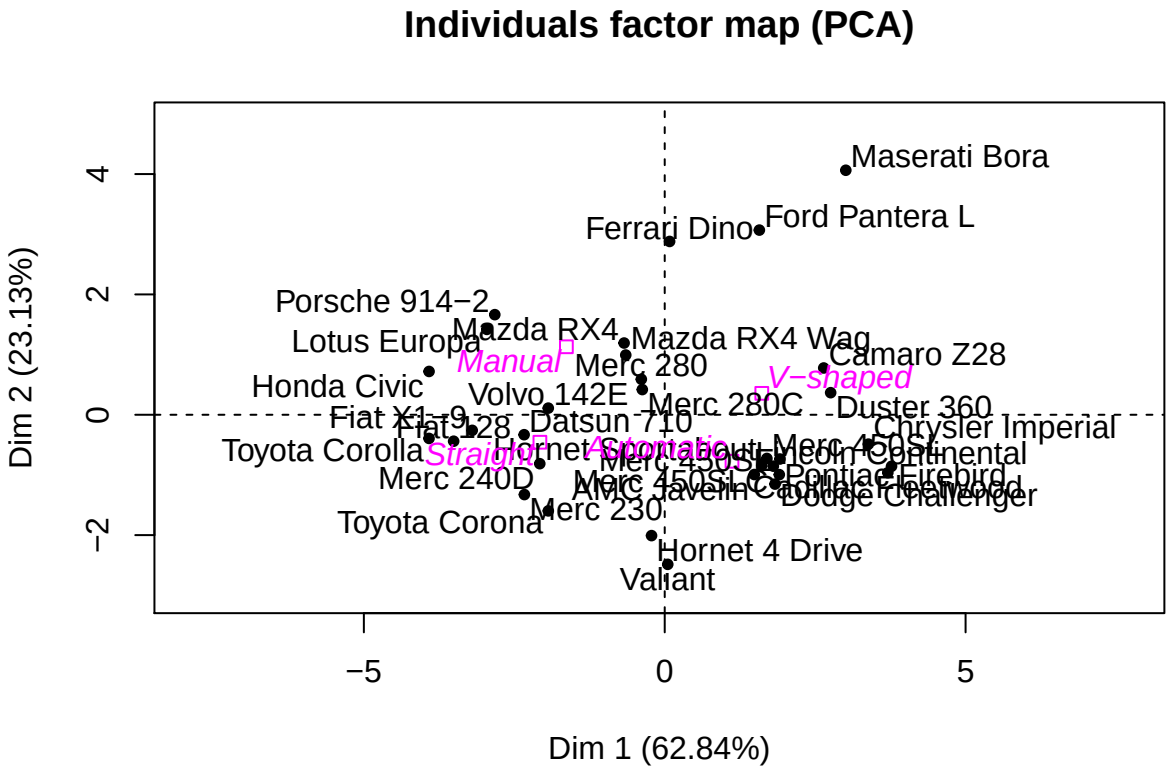
```
cortest.bartlett(corcars,n = 32)
```

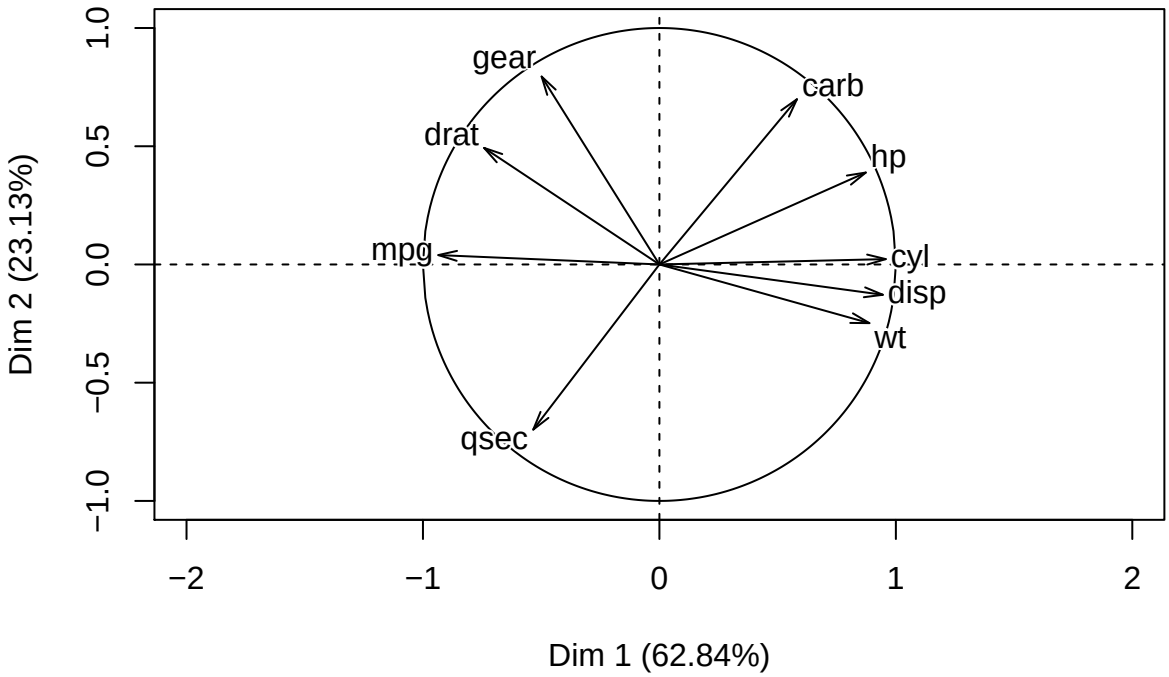
```
## $chisq
## [1] 332.328
##
## $p.value
## [1] 1.203652e-49
##
## $df
## [1] 36
```

De acuerdo a la prueba de esfericidad de Bartlett, observamos que el p-valor es muy pequeño por lo que la hipótesis nula se rechaza. En base a ello podemos concluir que la técnica de componentes principales será de aplicabilidad a la base de datos presentada.

Posteriormente, realizaremos el análisis de componentes principales mediante el siguiente código. Ver a continuación el código y sus salidas:

```
mt_pca <- PCA(mtcars,quali.sup = c(8,9),graph = TRUE,scale.unit = TRUE)
```



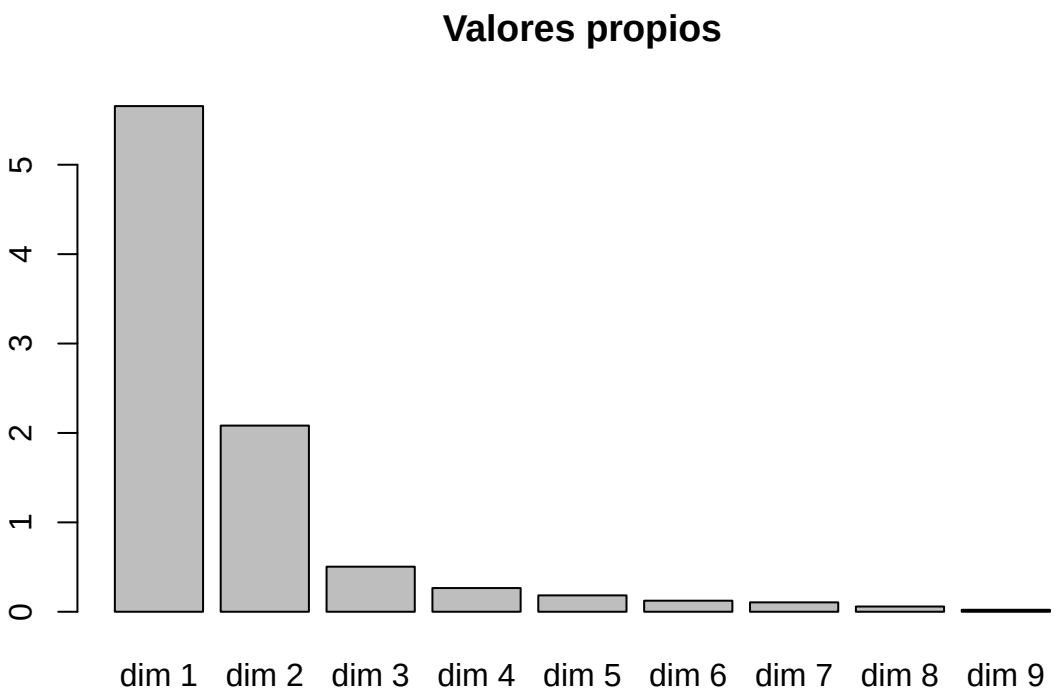

En base a los gráficos mostrados, se observa lo siguiente:

- El primer eje (Dim 1) expresa el 62.84% de la variabilidad de los datos y el segundo eje (Dim 2) el 23.13% de los mismos. En total, los dos primeros ejes expresan juntos el 85.97% de la variabilidad de los datos.
  - Variables factor map (PCA)

    - Se observa que, si un auto se encuentra en el primer cuadrante (esquina superior izquierda), este tendrá mayor millaje por galón, relación del eje trasero y mayor cantidad de marchas hacia adelante. Esto en contraposición del segundo y cuarto cuadrante (parte derecha del gráfico), el cual está asociado a mayor número de cilindros, desplazamiento, peso y caballos de fuerza.
- Individuals factor map (PCA)
    - Se observan que existen autos con perfiles de rendimiento similar. Por ejemplo, en el primer cuadrante se encuentran los autos Porsche 914-2 y Lotus Eurpa, mientras que en el cuarto cuadrante se encuentra el Cadillac Fleetwood y Chrysler Imperial.
    - El perfil de rendimiento similar está asociado a la ubicación del auto en el plano de las dos dimensiones. Para ello, utilizamos el variables factor map para entender qué características tienen en común determinados autos.

Posteriormente a este análisis, realizamos uno más detallado para identificar si existe otro componente (u otros componentes) que puedan incluirse para efectos del estudio. Para ello, generamos un gráfico que nos permita revisar la importancia de todos los componentes.

```
barplot(mt_pca$eig[,1], main="Valores propios", names.arg=paste("dim",1:nrow(mt_pca$eig)))
```




Se observa que existen en total 9 componentes, de los cuales los dos primeros componentes (los cuales explican el 85.97% de los datos) son los que explican en mayor proporción la variabilidad. Los componentes del 3 al 9 tienen un bajo autovalor, por lo que no serán utilizados para el análisis.

Asimismo, identificaremos si existen individuos (en este caso, autos) o variables que contribuyen mucho a los componentes elegidos. Ver código a continuación:

```
round(mt_pca$ind$contrib[,1:2],2)
```

```
##                   Dim.1 Dim.2
## Mazda RX4          0.25  2.13
## Mazda RX4 Wag      0.23  1.48
## Datsun 710         3.02  0.17
## Hornet 4 Drive     0.03  6.05
## Hornet Sportabout  1.44  1.06
```

```
## Valiant              0.00  9.27
## Duster 360           4.20  0.20
## Merc 240D            2.38  0.99
## Merc 230             3.00  2.64
## Merc 280             0.08  0.52
## Merc 280C            0.08  0.26
## Merc 450SE           2.03  0.81
## Merc 450SL           1.59  0.79
## Merc 450SLC          1.80  1.10
## Cadillac Fleetwood   7.60  1.39
## Lincoln Continental  7.85  1.10
## Chrysler Imperial    6.33  0.36
## Fiat 128             6.80  0.29
## Honda Civic          8.47  0.78
## Toyota Corolla       8.48  0.23
## Toyota Corona        2.07  3.83
## Dodge Challenger     1.86  1.99
## AMC Javelin          1.22  1.48
## Camaro Z28           3.86  0.91
## Pontiac Firebird     2.00  1.49
## Fiat X1-9            5.65  0.10
## Porsche 914-2        4.41  4.15
## Lotus Europa         4.83  3.02
## Ford Pantera L       1.37 14.14
## Ferrari Dino         0.00 12.45
## Maserati Bora        5.01 24.78
## Volvo 142E           2.07  0.02
```

```
round(mt_pca$var$contrib[,1:2],2)
```

```
##      Dim.1 Dim.2
## mpg  15.46  0.08
## cyl  16.20  0.02
## disp 15.79  0.79
## hp   13.47  7.26
## drat  9.72 11.67
## wt   13.95  2.96
## qsec  5.03 23.43
## gear  4.39 30.34
## carb  5.98 23.46
```

Se observa lo siguiente:

- Respecto a la importancia de variables en la construcción de componentes:
  - Se observa que, para el segundo componente, la cantidad de carburadores, de marchas hacia adelante y el tiempo para recorrer un cuarto de milla son las variables más importantes.
  - Se observa que, para el primer componente, la cantidad de millas por galón, el número de cilindros, los caballos de fuerza, el peso y el desplazamiento (en pies cúbicos) son las variables más importantes.
  - La relación del eje trasero contribuye a ambos componentes de forma similar.
- Respecto a la importancia de los individuos en la construcción de componentes:
  - Los autos de marca Ford Pantera L, Ferrari Dino y Maserati Bora son los que contribuyen en gran manera a la construcción del segundo componente. Se observa que, en relación a los demás individuos, el aporte de estos individuos es mucho mayor.
  - El aporte de los autos al componente 1 es más equilibrado que el componente 2. Se observa que los autos con mayor aporte son el Toyota Corolla, Honda Civic y Lincoln Continental, sin embargo

en relación a los demás individuos el aporte es regular.

Finalmente, realizamos la descripción de los ejes a través de la correlación de las variables de cada componente:

```
dimdesc(mt_pca,axes = c(1,2))
```

```
## $Dim.1
## $Dim.1$quanti
##      correlation      p.value
## cyl    0.9573620 9.987998e-18
## disp   0.9449932 4.195104e-16
## wt     0.8882114 1.186867e-11
## hp     0.8730011 7.238862e-11
## carb   0.5816671 4.798744e-04
## gear  -0.4981777 3.711660e-03
## qsec  -0.5335561 1.662320e-03
## drat  -0.7415688 1.197792e-06
## mpg   -0.9349924 4.804756e-15
##
## $Dim.1$quali
##           R2      p.value
## vs 0.5916018 2.698147e-07
## am 0.3231649 6.875337e-04
##
## $Dim.1$category
##                Estimate      p.value
## vs=V-shaped   1.843685 2.698147e-07
## am=Automatic  1.376372 6.875337e-04
## am=Manual    -1.376372 6.875337e-04
## vs=Straight  -1.843685 2.698147e-07
##
##
## $Dim.2
## $Dim.2$quanti
##      correlation      p.value
## gear   0.7947510 5.574498e-08
## carb   0.6988388 8.637452e-06
## drat   0.4929872 4.146805e-03
## hp     0.3887501 2.788365e-02
## qsec  -0.6984510 8.780023e-06
##
## $Dim.2$quali
##           R2      p.value
## am 0.4193267 6.169841e-05
##
## $Dim.2$category
##                 Estimate      p.value
## am=Manual      0.9512586 6.169841e-05
## am=Automatic -0.9512586 6.169841e-05
```

En base a lo identificado, la primera dimensión está asociada a las características relacionadas a la potencia del motor (número de cilindros, caballos de fuerza, carburadores y cilindrada) mientras que la segunda dimensión está asociada a la maniobrabilidad del auto (cantidad de marchas, tipo de transmisión, entre otros).

### Discusión y conclusiones

En base al análisis presentado, y en relación a los objetos de estudio, se concluye lo siguiente:

- Existen dos perfiles de autos: aquellos orientados a ser autos potentes (tienen mayor índice en la dimensión 1) y aquellos que son maniobrables (tienen mayor índice en la dimensión 2)
- Utilizando las variables cualitativas, se observa que aquellos autos orientados a ser autos potentes tienen el motor en forma de V, mientras que aquellos que no tienen el motor de forma recta.
- De igual forma, se observa que aquellos autos orientados a ser autos maniobrables tendrían mayor propensión a tener transmisión manual.
- Existen autos que serían tanto maniobrables como potentes: los casos específicos serían el Maserati Bora y Ford Pantera.

## Análisis Discriminante

### Introducción y Datos

La base de datos sobre la cual aplicaremos análisis discriminante proviene de Edgar Anderson, quien publicó en el *Bulletin of the American Iris Society* una base de datos sobre las distintas especies de la planta isis. Esta base de datos ha sido muy estudiada dentro de la comunidad de *machine learning* e inteligencia artificial. Esta base de datos ha sido estudiada por Ronald Fisher en el año 1936, quien publicó su estudio en la revista *Annals of Eugenics*.

La base de datos viene por defecto en la instalación base del programa R. Dicha base de datos tiene las siguientes variables:

- Sepal.Length: El largo del sépalo de una hoja.
- Sepal.Width: El ancho del sépalo de una hoja.
- Petal.Length: El largo del pétalo de una hoja.
- Petal.Width: El ancho del pétalo de una hoja.
- Species: La especie de la planta.

Los objetivos del presente estudio son:

- Crear un clasificador que permita discriminar las especies de plantas a través de sus atributos.
- Identificar si dicho clasificador es lo suficientemente bueno para ser evaluado a posterior con nuevos datos.

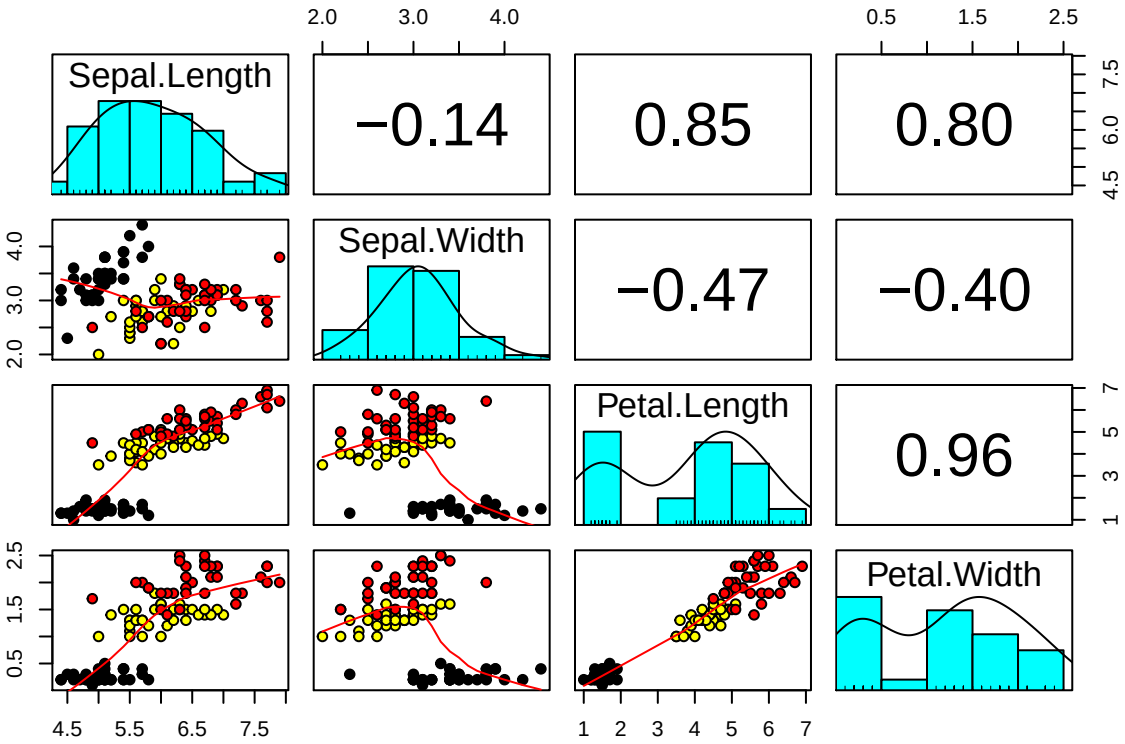
Con el propósito de ejecutar el análisis de los datos, realizamos la carga de librerías e importamos los datos. Posteriormente, realizamos un preprocesamiento de los datos para continuar con el estudio. Este preprocesamiento consiste en particionar los datos en datos de entrenamiento y prueba. Ver a continuación:

```
set.seed(4830201)
data("iris")

index_set <- createDataPartition(iris$Species,times=1,p=0.7,list =F)
train_iris <- iris[index_set,]
test_iris <- iris[-index_set,]
```

Con el próposito de entender la distribución de los datos y la relación entre ellos (así como identificar posibles "reglas de separación" entre clases), realizamos un análisis exploratorio a través del siguiente código:

```
pairs.panels(train_iris[1:4],bg = c("black","yellow","red")[train_iris$Species],pch=21,ellipses = F)
```




Se observa lo siguiente:

- Se observa que la clase "setosa" está asociada a menor largo y anchura del pétalo.
- Se observa que la clase "versicolor" y "virginica" están asociadas a un mayor largo del pétalo y sépalo. Entre ellos dos, la clase "virginica" está asociada a valores más altos.
- Se observa una separación entre la clase "setosa" y las otras dos clases a lo largo de los gráficos de dispersión.

## Resultados

Con el propósito de utilizar el análisis discriminante, primero se validarán los siguientes supuestos:

- La matriz de covarianza es igual en todas las clases y de forma general.
- Las variables asociadas a cada clase se distribuyen de forma normal univariada y multivariada.

Para ello, utilizamos los tests de Shapiro-Wilks, Box y de normalidad multivariante. Ver resultados a continuación y posteriormente un análisis del mismo:

```
train_iris_m <- melt(train_iris)
```

```
## Using Species as id variables
```

```
train_iris_m %>% group_by(Species,variable) %>% summarise(pv_shapiro = round(shapiro.test(value)$p.valu
```

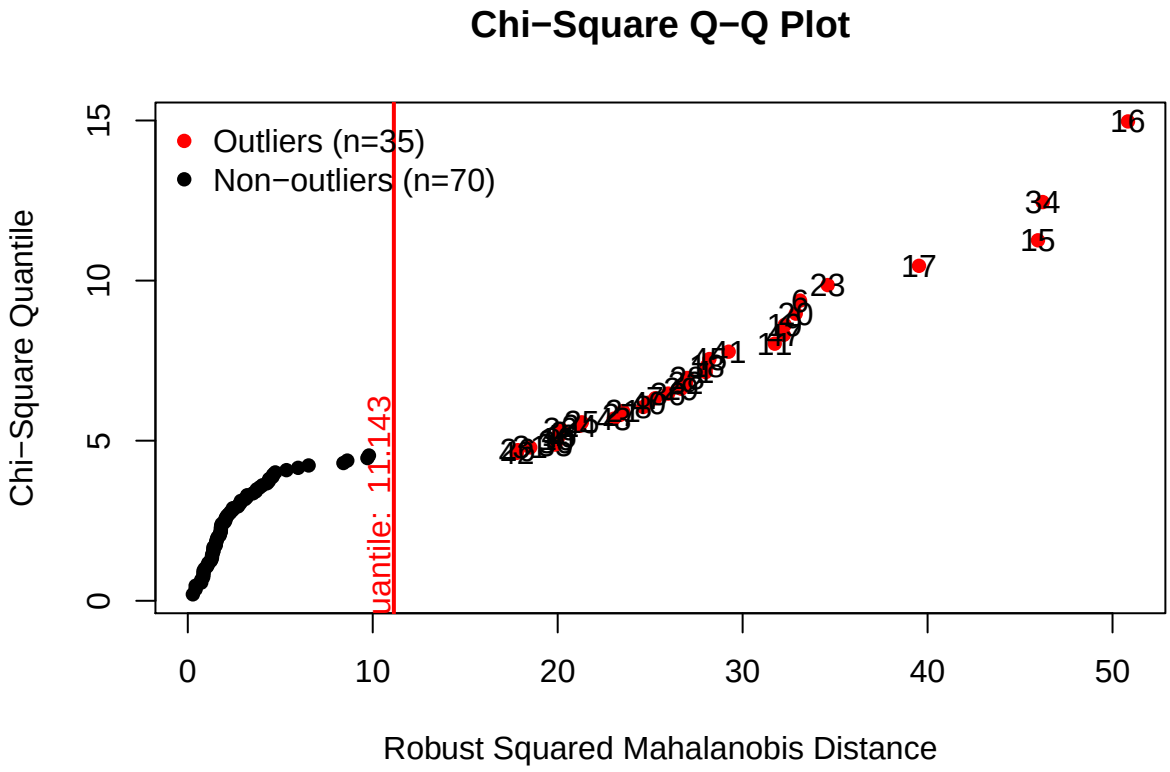
```
## # A tibble: 12 x 3
## # Groups:   Species [3]
##    Species    variable     pv_shapiro
##    <fct>      <fct>             <dbl>
##  1 setosa     Sepal.Length    0.447
##  2 setosa     Sepal.Width     0.250
##  3 setosa     Petal.Length    0.289
##  4 setosa     Petal.Width     0.00001
##  5 versicolor Sepal.Length    0.265
```

```
##  6 versicolor Sepal.Width     0.523
##  7 versicolor Petal.Length    0.378
##  8 versicolor Petal.Width     0.0536
##  9 virginica  Sepal.Length    0.606
## 10 virginica  Sepal.Width     0.515
## 11 virginica  Petal.Length    0.281
## 12 virginica  Petal.Width     0.180
```

```
MVN::mvn(data = train_iris[,1:4],multivariateOutlierMethod = "quan")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## sROC 0.1-2 loaded
```




```
## $multivariateNormality
##              Test          Statistic              p value Result
## 1 Mardia Skewness   54.8276787160279 4.35680187151099e-05     NO
## 2 Mardia Kurtosis -0.354153850016746    0.723223587363986    YES
## 3             MVN               <NA>                 <NA>     NO
##
## $univariateNormality
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk Sepal.Length    0.9738  0.0354      NO
## 2 Shapiro-Wilk Sepal.Width     0.9840  0.2381      YES
## 3 Shapiro-Wilk Petal.Length    0.8760  <0.001      NO
## 4 Shapiro-Wilk Petal.Width     0.8960  <0.001      NO
##
## $Descriptives
##                n     Mean   Std.Dev Median Min Max 25th 75th       Skew
```

```
## Sepal.Length 105 5.876190 0.8217622    5.8 4.4 7.9  5.2  6.4  0.3637159
## Sepal.Width  105 3.061905 0.4466501    3.0 2.0 4.4  2.8  3.3  0.2895417
## Petal.Length 105 3.780000 1.7607035    4.4 1.0 6.9  1.6  5.1 -0.2875537
## Petal.Width  105 1.208571 0.7538509    1.4 0.1 2.5  0.3  1.8 -0.1170858
##                Kurtosis
## Sepal.Length -0.5701195
## Sepal.Width   0.2326692
## Petal.Length -1.3979516
## Petal.Width  -1.3569098
```

```
test_mvn_st <- MVN::mvn(data = train_iris[,1:4],mvnTest="hz")
test_mvn_st$multivariateNormality
```

```
##            Test       HZ      p value MVN
## 1 Henze-Zirkler 1.843637 1.247347e-11  NO
```

```
test_box <- boxM(train_iris[,1:4], grouping = train_iris$Species)
test_box
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  train_iris[, 1:4]
## Chi-Sq (approx.) = 110.68, df = 20, p-value = 1.473e-14
```

Sobre lo anterior, se observa lo siguiente:

- La variable "Petal.Width" no se distribuye de forma normal univariada en las clases "setosa" y "versicolor", pues tiene un p-valor menor a 0.05.
- Se identificaron 9 puntos anómalos que podrían influenciar en el contraste de normalidad multivariante (ver gŕafico)
- Se observa que, de acuerdo al test Henze-Zirkler, el conjunto de datos no sigue una distribución normal univariante. Ello podría deberse a la variable "Petal.Width" por lo anteriormente explicado.
- Se observa que, de acuerdo al test de Box, cada clase tiene una matriz de covarianza distinta a las otras (pues se rechaza la hipótesis nula).

Dado que el conjunto de datos no tiene una distribución normal multivariante, el análisis discriminante perderá precisión en el resultado. No obstante, aún puede ser un buen clasificador. Para ello, analizaremos el rendimiento del clasificador una vez realizado.

El clasificador a utilizar será el análisis discriminante cuadrático, pues de acuerdo al text de Box cada clase tiene una matriz de covarianza distinta. Ver código a continuación:

```
qda_train <- qda(formula = Species ~., data = train_iris)
qda_train
```

```
## Call:
## qda(Species ~ ., data = train_iris)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         5.060000    3.445714         1.48    0.260000
## versicolor     5.980000    2.794286         4.32    1.357143
## virginica      6.588571    2.945714         5.54    2.008571
```

Se observa que las probabilidades a priori del clasificador es 0.33 para cada una de las clases.

Posteriormente se evaluará el rendimiento del clasificador utilizando la partición de prueba del conjunto de datos. Ver a continuación el código:

```
test_predict <- predict(qda_train,test_iris)
table(test_iris$Species,test_predict$class,dnn = c("Clase real","Clase predicha"))
```

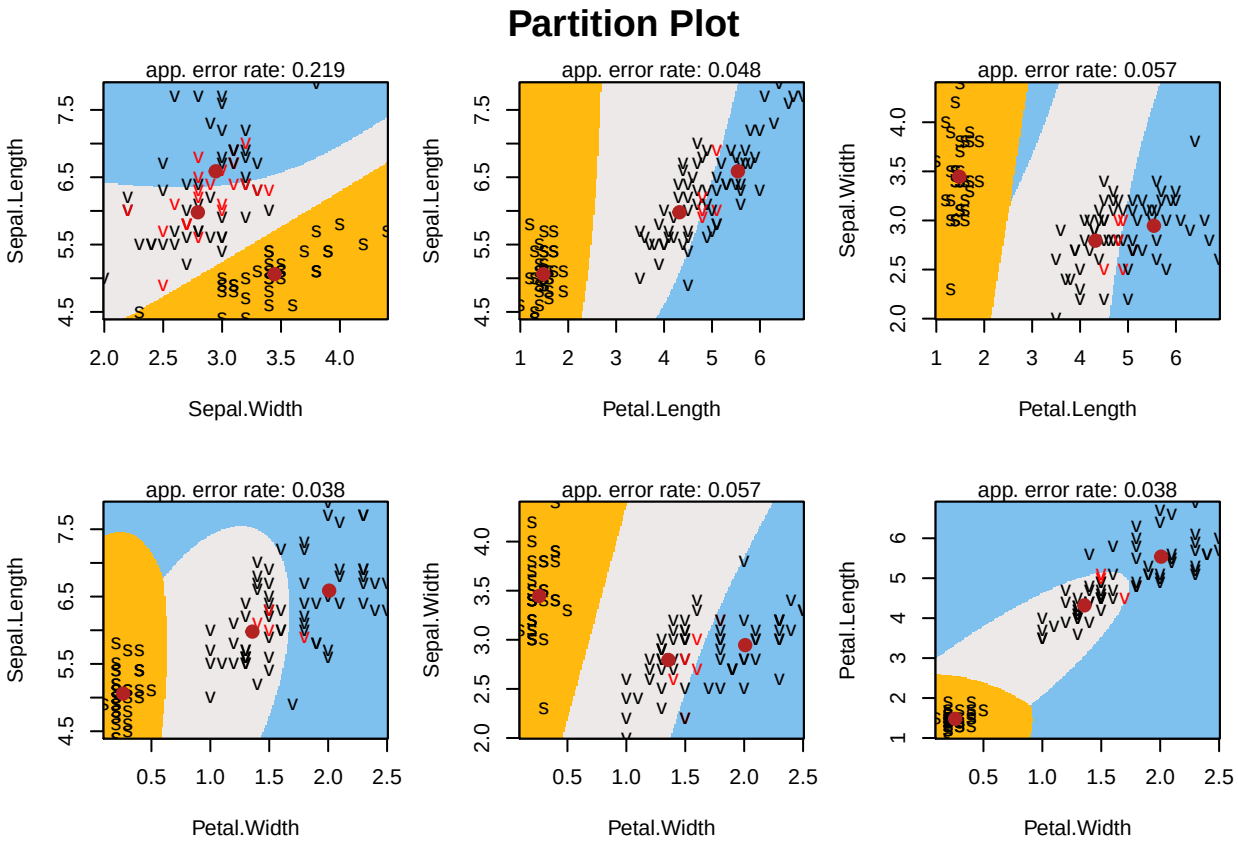
```
##              Clase predicha
## Clase real    setosa versicolor virginica
##   setosa          15          0         0
##   versicolor       0         15         0
##   virginica        0          0        15
```

```
error <- mean(test_iris$Species != test_predict$class) * 100
paste("Error del clasificador: ",error,"%")
```

```
## [1] "Error del clasificador:  0 %"
```

Finalmente, realizaremos un análisis gráfico de las particiones por clase, para observar las fonteras de decisión.

```
partimat(Species~.,data = train_iris,method="qda",prec=200,nplots.hor=3,image.colors = c("darkgoldenrod
```




Sobre ello, se observa lo siguiente:

- Nuevas observaciones serán clasificadas como “setosa” si tienen un bajo valor en el ancho y largo de sus pétalos. Asimismo, serán clasificadas como tal si tienen una baja medición en el ancho de los pétalos y alto valor en el ancho del sépalo.
- Las especies “virginica” y “versicolor” tienen valores muy cercanos en su frontera de decisión, por lo que es más probable que el error de clasificación provenga de dichas clases.

### Discusión y conclusiones

En base al análisis presentado, y en relación a los objetos de estudio, se concluye lo siguiente:

- La clase setosa tiene el menor ancho y largo de pétalo que las otras dos clases.
- Las clases restantes tienen menor diferencia entre sus promedios por clase por lo que podría existir mayor dificultad diferenciándolas a futuro (observado a través del rendimiento del clasificador).
- El análisis discriminante cuadrático es un buen clasificador para las 3 clases, pues su error de clasificación es del 0%.

## Análisis Factorial

### Introducción y Datos

Los datos corresponden a un test de personalidad proveniente de la *International Personality Item Pool*. Los datos corresponden a 2800 sujetos, los cuales realizaron el test. La base de datos viene por defecto en la instalación base del programa R.

En dicho test de personalidad, se indicaron 25 afirmaciones, las cuales cada persona indicó sentirse identificada o no (en una escala del 1 al 10). Estas 25 afirmaciones estaban agrupadas en conjuntos de 5, y cada conjunto intenta medir lo siguiente:

- Grupo 1: Amabilidad.
- Grupo 2: Escruposibilidad.
- Grupo 3: extroversión.
- Grupo 4: Neuroticismo.
- Grupo 5: Apertura hacia otros.

Los objetivos del presente estudio son:

- Identificar si las preguntas de un test de personalidad, las cuales han sido determinadas por el investigador para medir diversos rasgos de personalidad en específico, están correlacionadas entre ellas (por cada rasgo de personalidad).
- Identificar si los factores encontrados en los datos están asociados a cada rasgo de personalidad que el investigador desea encontrar.

Con el propósito de ejecutar el análisis de los datos, realizamos la carga de librerías e importamos los datos. Posteriormente, realizamos un preprocesamiento de los datos para continuar con el estudio. Ver código a continuación:

```
data("bfi")

## Pre-procesamiento de datos ##

bfi=bfi[complete.cases(bfi),]
bfi$education = as.factor(bfi$education)
bfi$gender = as.factor(bfi$gender)

cor_bfi <- cor(bfi[,-c(26,27)])
```

### Resultados

Con el objetivo de aplicar análisis factorial, evaluaremos los supuestos de normalidad multivariante. Este supuesto define el método a través del cual extraeremos los factores.

```
test_mvn_af <- MVN::mvn(data = bfi[,-c(26,27)],mvnTest="hz")
test_mvn_af$multivariateNormality
```

```
##            Test      HZ p value MVN
## 1 Henze-Zirkler 1.03709       0  NO
```

Se observa que, de acuerdo al test Henze-Zirkler, el conjunto de datos no se distribuye de forma normal multivariada. Por lo tanto, utilizaremos el método de máxima verosimilitud para la extracción de factores.

Por otro lado, verificaremos a través del contraste de esfericidad de Bartlett, si hace sentido aplicar el análisis factorial. La hipótesis nula de dicho test es que los coeficientes de correlación para cada tipo de variable son nulos. Ver test a continuación:

```
test_bart_af <- cortest.bartlett(cor_bfi,n=dim(bfi)[1])
test_bart_af
```

```
## $chisq
## [1] 16698.54
##
## $p.value
## [1] 0
##
## $df
## [1] 325
```

Se aprecia que la hipótesis nula puede rechazarse, por lo que es posible seguir trabajando en el análisis factorial.

Posteriormente, se utiliza la medida de adecuación muestral de Kaiser-Meyer-Olkin con el propósito de entender si los datos son adecuados a un modelo de análisis factorial. Ver código a continuación:

```
KMO(cor_bfi)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor_bfi)
## Overall MSA =  0.84
## MSA for each item =
##   A1   A2   A3   A4   A5   C1   C2   C3   C4   C5   E1   E2   E3   E4   E5   N1
## 0.74 0.84 0.87 0.86 0.90 0.83 0.78 0.85 0.83 0.86 0.84 0.88 0.89 0.87 0.89 0.78
##   N2   N3   N4   N5   O1   O2   O3   O4   O5  age
## 0.78 0.86 0.88 0.86 0.86 0.78 0.83 0.78 0.76 0.63
```

Se observa que el índice KMO es de 0.84, lo cual es "Meritorio" de acuerdo a la tabla mostrada en clase.

Finalmente, realizaremos el análisis factorial en base a 5 factores, pues los grupos de pregundas denotados en los datos son 5. Dado que deseamos entender si los factores están específicamente relacionados con las preguntas del investigador por grupo, haremos la rotación varimax .Ver código a continuación:

```
fit.ml.rot <- fa(cor_bfi,nfactors = 5,rotate = "varimax",fm="ml",n.obs = dim(bfi)[1])

fit.ml.rot
```

```
## Factor Analysis using method =  ml
## Call: fa(r = cor_bfi, nfactors = 5, n.obs = dim(bfi)[1], rotate = "varimax",
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       ML1   ML2   ML3   ML5   ML4    h2   u2 com
## A1   0.10  0.04  0.01 -0.38 -0.07 0.162 0.84 1.2
## A2   0.04  0.21  0.15  0.58  0.04 0.401 0.60 1.4
## A3   0.02  0.29  0.11  0.64  0.05 0.515 0.48 1.5
## A4  -0.05  0.18  0.23  0.45 -0.13 0.307 0.69 2.1
## A5  -0.12  0.35  0.08  0.57  0.07 0.476 0.52 1.8
```

```
## C1   0.01  0.05  0.53  0.06  0.21 0.330 0.67 1.4
## C2   0.08  0.00  0.61  0.13  0.12 0.414 0.59 1.2
## C3  -0.03  0.02  0.56  0.12  0.00 0.324 0.68 1.1
## C4   0.22 -0.09 -0.65 -0.01 -0.08 0.487 0.51 1.3
## C5   0.27 -0.20 -0.57 -0.04  0.04 0.439 0.56 1.7
## E1   0.02 -0.58  0.03 -0.13 -0.07 0.359 0.64 1.1
## E2   0.23 -0.68 -0.10 -0.14 -0.06 0.544 0.46 1.4
## E3   0.02  0.51  0.08  0.31  0.31 0.457 0.54 2.4
## E4  -0.12  0.61  0.09  0.37 -0.05 0.540 0.46 1.8
## E5   0.05  0.50  0.32  0.12  0.22 0.416 0.58 2.3
## N1   0.81  0.09 -0.04 -0.20 -0.07 0.719 0.28 1.2
## N2   0.78  0.05 -0.02 -0.20 -0.01 0.658 0.34 1.1
## N3   0.72 -0.08 -0.07  0.00  0.01 0.527 0.47 1.0
## N4   0.56 -0.38 -0.19  0.01  0.08 0.498 0.50 2.1
## N5   0.52 -0.18 -0.06  0.11 -0.15 0.343 0.66 1.6
## O1  -0.02  0.18  0.11  0.07  0.52 0.323 0.68 1.4
## O2   0.18 -0.01 -0.12  0.12 -0.47 0.276 0.72 1.6
## O3   0.02  0.28  0.06  0.14  0.62 0.482 0.52 1.5
## O4   0.21 -0.22 -0.05  0.14  0.36 0.242 0.76 2.7
## O5   0.08 -0.01 -0.08  0.01 -0.53 0.288 0.71 1.1
## age -0.10  0.01  0.09  0.08  0.05 0.027 0.97 3.5
##
##                        ML1  ML2  ML3  ML5  ML4
## SS loadings           2.70 2.37 2.02 1.91 1.57
## Proportion Var        0.10 0.09 0.08 0.07 0.06
## Cumulative Var        0.10 0.19 0.27 0.35 0.41
## Proportion Explained  0.26 0.22 0.19 0.18 0.15
## Cumulative Proportion 0.26 0.48 0.67 0.85 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  325  and the objective function was  7.5 with Chi Squa
## The degrees of freedom for the model are 205  and the objective function was  0.69
##
## 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  2236 with the empirical chi square  1300.1  with prob <  2e-1
## The total number of observations was  2236  with Likelihood Chi Square =  1522.62  with prob <  8.5e-
##
## Tucker Lewis Index of factoring reliability =  0.872
## RMSEA index =  0.054  and the 90 % confidence intervals are  0.051 0.056
## BIC =  -58.43
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                 ML1  ML2  ML3  ML5  ML4
## Correlation of (regression) scores with factors   0.93 0.87 0.86 0.84 0.83
## Multiple R square of scores with factors          0.86 0.76 0.74 0.71 0.69
## Minimum correlation of possible factor scores     0.73 0.52 0.48 0.43 0.39
```

Se observa lo siguiente:

- La media cuadrática de los residuos es de 0.03. Esto es aceptable, pues mientras más cerca a 0, mejor. Asimismo, se observa que el valor corregido es de 0.04, lo cual está debajo del umbral de 0.05 denotado

en clase.
- El índice de Lewis Index es de 0.872, valor el cual está ligeramente por debajo del umbral de 0.9. Por lo tanto, podemos considerar que el valor es aceptable, sin embargo abre la posibilidad de investigar si existen factores adicionales a considerar.

En relación a los factores, se observa que:

- Se observa que las preguntas relacionadas al rasgo de personalidad amabilidad están fuertemente asociadas con el factor 5.
- Se observa que las preguntas relacionadas al rasgo de personalidad escruposibilidad están fuertemente asociadas con el factor 3.
- Se observa que las preguntas relacionadas a la extroversión están fuertemente asociadas con el factor 2.
- Se observa que las preguntas relacionadas al neuroticismo están fuertemente asociadas con el factor 1.
- Se observa que las preguntas relacionadas a la apertura están fuertemente asociadas con el factor 4.

## Discusión y conclusiones

En base a los resultados mostrados, se observa que cada pregunta está relacionada a un factor específico. Por lo tanto, podríamos asegurar que las preguntas determinadas por el investigador están asociadas fuertemente con el rasgo de personalidad que desean investigar.