

Para la realización del ejercicio de componentes principales (PCA) vamos a usar dos bases de datos. La primera es la que se usó en clases bajo el nombre de INEGI. La segunda corresponde a la variable “delta” (base de datos conseguida para este ejercicio), que incluye información (número de asientos totales, de primera y clase turista, velocidad crucero del avión, rango máximo de la unidad, etcétera) de todos los aviones de la compañía DELTA. El objetivo principal es hacer un ejercicio de componentes principales con la paquetería FactoMineR.

#Se cargan ambas bases de datos desde las respectivas direcciones en línea en las variables INEGI y delta.

The screenshot displays the RStudio interface. On the left, the 'Environment' pane shows the 'delta' dataset loaded as a 'data.frame' with 34 columns and 15.8 KB size. Below it, the 'Console' pane shows the execution of R code to load the 'delta' dataset from a GitHub URL. On the right, the 'R Documentation' pane shows the documentation for 'summary.lm', including its description, usage, and arguments.

Aircraft	Seat.Width.Club	Seat.Pitch.Club	Seat.Club	Seat.Width.First.Class	Seat.Pitch.First.Class	Seats
1 Airbus A319	0.0	0	0	21.0	36.0	
2 Airbus A319 VIP	19.4	44	12	19.4	40.0	
3 Airbus A320	0.0	0	0	21.0	36.0	
4 Airbus A320 32-R	0.0	0	0	21.0	36.0	
5 Airbus A330-200	0.0	0	0	0.0	0.0	
6 Airbus A330-200 (3L2)	0.0	0	0	0.0	0.0	
7 Airbus A330-200 (3L3)	0.0	0	0	0.0	0.0	
8 Airbus A330-300	0.0	0	0	0.0	0.0	
9 Boeing 717	0.0	0	0	19.6	37.0	
10 Boeing 737-700 (73W)	0.0	0	0	21.0	37.0	
11 Boeing 737-800 (738)	0.0	0	0	21.0	37.0	
12 Boeing 737-800 (73H)	0.0	0	0	21.0	38.0	
13 Boeing 737-900ER (739)	0.0	0	0	21.0	37.0	
14 Boeing 747-400 (745)	0.0	0	0	0.0	0.0	
15 Boeing 757-200 (75A)	0.0	0	0	21.0	40.0	

```

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

Durante la inicializaci'on - Warning messages:
1: Setting LC_CTYPE failed, using "C"
2: Setting LC_COLLATE failed, using "C"
3: Setting LC_TIME failed, using "C"
4: Setting LC_MESSAGES failed, using "C"
5: Setting LC_MONETARY failed, using "C"
[Workspace loaded from ~/.RData]

> INEGI <- read.csv(url("https://raw.githubusercontent.com/mauriciogtec/PropedeuticoDataScience2017/master/Datos/DatosINEGI.csv"))
> delta <- read.csv(url("https://raw.githubusercontent.com/mylesmharrison/delta_PCA_kmeans/master/delta.csv"))

```

Summary.lm {stats}

Summarizing Linear Model Fits

Description

summary method for class "lm".

Usage

```
## S3 method for class 'lm'
summary(object, correlation = FALSE, symbolic.cor = FALSE, ...)
```

Arguments

- object**: an object of class "lm", usually, a result of a call to `lm`.
- x**: an object of class "summary.lm", usually, a result of a call to `summary.lm`.
- correlation**: logical; if TRUE, the correlation matrix of the estimated parameters is returned and

Se va a hacer el tratamiento hecho durante clases para hacer ciertas variables per cápita y sea más fácil su manipulación. Además se creará una matriz con el nombre de “X” que será la utilizada en el análisis PCA.

The screenshot shows the RStudio interface. The main editor displays a data frame named 'delta' with 34 variables. The Environment pane on the right shows the global environment with variables like 'col', 'delta', 'INEGI', 'regression', 'rss', and 'X'. The Console pane at the bottom shows the output of a summary function for linear model fits, displaying coefficients and statistics for various models.

omy.	Seats_Economy	Accommodation	Cruising.Speed.mph	Range.miles	Engines	Wingspan.ft	TailHeight.ft	Length.ft	Wing
30.5	96	126	517	2399	2	111.830	38.5830	111.000	
0.0	0	54	517	3119	2	111.830	38.5830	111.000	
31.5	120	150	517	2420	2	111.830	38.5830	123.250	
31.5	120	150	517	2420	2	111.830	38.5830	123.250	
30.5	181	243	531	6536	2	197.830	59.8300	188.670	
30.5	168	243	531	6536	2	197.830	59.8300	188.670	
30.5	227	293	531	5343	2	197.830	56.3300	208.830	
30.5	232	298	531	5343	2	197.830	56.3300	208.830	
31.0	83	110	504	1510	2	93.330	29.0830	120.000	
30.5	94	124	517	2925	2	117.416	41.1670	110.330	
30.5	126	124	517	2925	2	117.416	41.1670	110.330	
30.5	126	160	517	2850	2	117.416	41.1670	129.500	
30.5	139	180	517	2870	2	117.416	41.1670	138.167	
30.5	286	376	564	7365	4	213.000	62.5416	231.830	
31.0	132	174	517	4344	2	134.750	44.5000	155.250	

Showing 1 to 16 of 44 entries

```

Campeche      0.004819799  0.02230312  0.001383686  0.006404107
Coahuila de Zaragoza 0.005530509  0.02181858  0.001334599  0.005597457
Colima        0.005713583  0.02063161  0.001011444  0.005207861

PorcentajePartosHospitales PorcentajeAguaPotable PorcentajeAguaEntubada
Aguascalientes      97.1      98.0      98.9
Baja California      65.7      93.3      95.9
Baja California Sur  95.2      86.7      92.4
Campeche             87.0      89.5      90.3
Coahuila de Zaragoza  90.3      97.9      98.2
Colima               98.5      97.9      98.5

PorcentajeElectricidad PorcentajeParedesSolidas PorcentajePisoTierra
Aguascalientes      99.2      92.3      1.7
Baja California      98.5      77.0      3.3
Baja California Sur  96.7      90.3      5.8
Campeche             96.8      80.7      4.7
Coahuila de Zaragoza  99.1      84.8      1.6
Colima               99.0      94.7      4.5

```

Ahora se instalará la paquetería FactoMineR:

```

> install.packages("FactoMineR")
also installing the dependencies 'ellipse', 'flashClust', 'leaps', 'scatterplot3d'

probando la URL 'https://cran.rstudio.com/bin/macosx/mavericks/contrib/3.3/ellipse_0.3-8.tgz'
Content type 'application/x-gzip' length 47328 bytes (46 KB)
=====
downloaded 46 KB

probando la URL 'https://cran.rstudio.com/bin/macosx/mavericks/contrib/3.3/flashClust_1.01-2.tgz'
Content type 'application/x-gzip' length 22115 bytes (21 KB)
=====
downloaded 21 KB

probando la URL 'https://cran.rstudio.com/bin/macosx/mavericks/contrib/3.3/leaps_3.0.tgz'
Content type 'application/x-gzip' length 66541 bytes (64 KB)
=====
downloaded 64 KB

```

Para realizar el análisis PCA se requiere llamar la paquetería instalada:

```

Console ~/
probando la URL 'https://cran.rstudio.com/bin/macosx/mavericks/contrib/3.3/FactoMineR_1.36.tgz'
Content type 'application/x-gzip' length 3545409 bytes (3.4 MB)

=====
downloaded 3.4 MB

tar: Failed to set default locale
tar: Failed to set default locale
tar: Failed to set default locale
tar: Failed to set default locale
tar: Failed to set default locale

The downloaded binary packages are in
  /var/folders/Zs/_4ty8y0j347gxzj0r3q1yrr80000gn/T//RtmpCNJG5p/downloaded_packages
> library(FactoMineR)
Warning message:
package 'FactoMineR' was built under R version 3.3.2
> |

```

Con la paquetería operando se procede a aplicar el análisis PCA a la variable “X”:

omv.	Seats-Economy	Accommodation	Cruising.Speed.mph.	Range.miles	Engines	Wingspan.ft.	TailHeight.ft.	Length.ft.	Wif
30.5	96	126	517	2399	2	111.830	38.5830	111.000	
0.0	0	54	517	3119	2	111.830	38.5830	111.000	
31.5	120	150	517	2420	2	111.830	38.5830	123.250	
31.5	120	150	517	2420	2	111.830	38.5830	123.250	
30.5	181	243	531	6536	2	197.830	59.8300	188.670	
30.5	168	243	531	6536	2	197.830	59.8300	188.670	
30.5	227	293	531	5343	2	197.830	56.3300	208.830	
30.5	232	298	531	5343	2	197.830	56.3300	208.830	
31.0	83	110	504	1510	2	93.330	29.0830	120.000	
30.5	94	124	517	2925	2	117.416	41.1670	110.330	
30.5	126	124	517	2925	2	117.416	41.1670	110.330	
30.5	126	160	517	2850	2	117.416	41.1670	129.500	
30.5	139	180	517	2870	2	117.416	41.1670	138.167	
30.5	286	376	564	7365	4	213.000	62.5416	231.830	
31.0	132	174	517	4344	2	134.750	44.5000	155.250	

Showing 1 to 16 of 44 entries

Environment History

Name	Type	Length	Size	Value
col	character	1	104 B	"Matrimonios"
delta	data.frame	34	15.8 KB	44 obs. of 34 variables
INEGI	data.frame	16	8.5 KB	32 obs. of 16 variables
model	PCA	5	40 KB	List of 5
regresion	lm	12	15.9 MB	Large lm (12 elements, 15.9 Mb)
rss	numeric	1	48 B	120856978491.536
X	data.frame	14	7.7 KB	32 obs. of 14 variables

Files Plots Packages Help Viewer

Zoom Export Publish

Variables factor map (PCA)

Dim 2 (12.98%)

Dim 1 (49.50%)

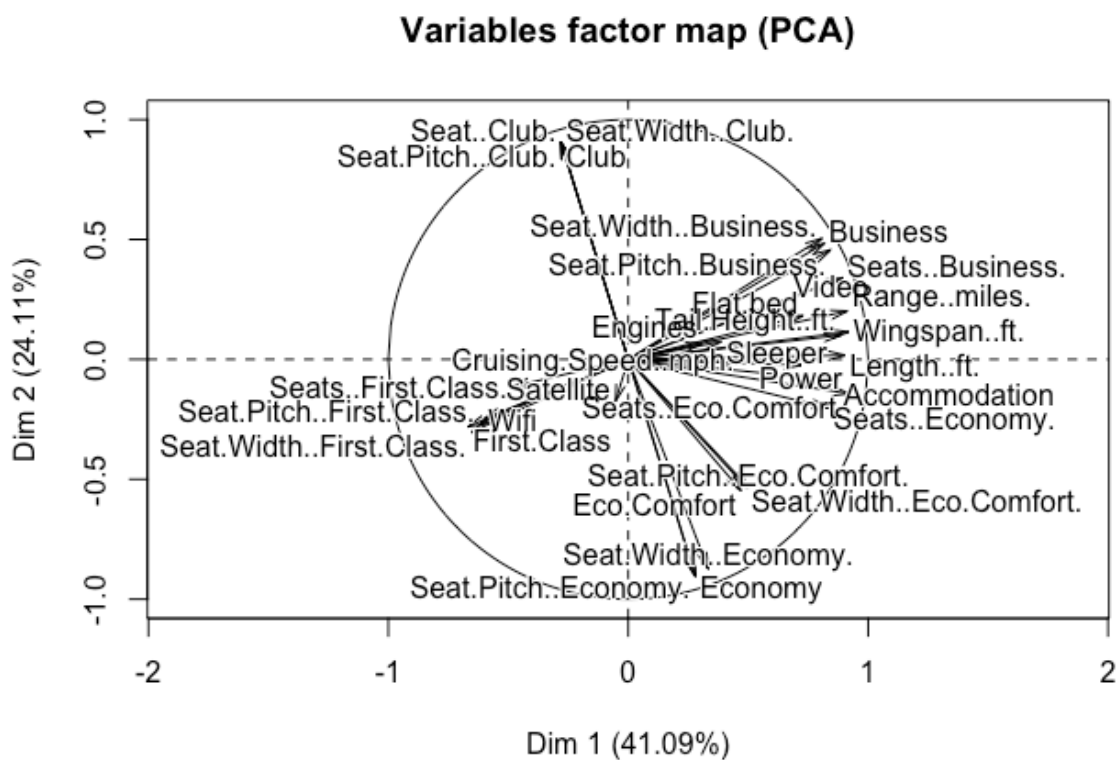
En la parte inferior derecha podemos observar el resultado del análisis. La dimensión 1 (horizontal) captura el 49.5% de la variación total de todas las variables juntas, mientras que la dimensión 2 (vertical) explica el 12.98% de la variación total. A continuación llamaremos el mismo análisis PCA, pero en esta ocasión con la base de datos propia que llamamos “delta”. Dada la estructura de la base de datos “delta”, podemos notar que la primer columna se trata de todos los tipos de aviones que tiene la compañía. En ese sentido, se debe excluir la primer columna para hacer el análisis PCA. En el siguiente código se puede ver como se llama la función PCA sobre la base de datos “delta” excluyendo la primer columna.

```

116
117
118 #PCA
119 library(FactoMineR)
120 model <- PCA(X)
121
122 #PCA DELTA
123 summary(delta)
124
125 model2 <- PCA(delta[-1])
126 model2
126:7 (Untitled) R Script

```

El resultado es el siguiente:



La dimensión 1 (horizontal) captura el 41.09% de la variación total de todas las variables juntas, mientras que la dimensión 2 (vertical) explica el 24.11% de la variación total. Con eso concluye el ejercicio de PCA con la base de datos de INEGI y la propia de "delta".