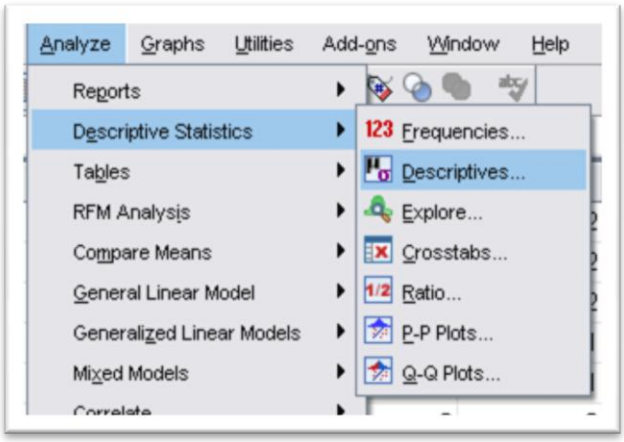
Los métodos vistos se pueden clasificar de la siguiente forma

- **Exploratorios y Test**
- **Clasificatorios**
  - **Los métodos clasificatorios pretenden clasificar clúster, existen 2 tipos**
    - **Supervisados**
      - **Intentan predecir una variable discreta en base a otras, se puede usar para predecir por ejemplo: número de hijos, clúster al que pertenece, grupo socio económico, pero no para determinar un sueldo, por que el sueldo es continuo**
    - **No Supervisados**
      - **Busca grupos de individuos similares, estos son llamados clúster, los individuos son similares cuando tienen similar posición en el espacio de los atributos seleccionados. Sirve para crear segmentaciones en base al comportamiento de clientes, como por ejemplo R,M,F.**
- **Reducción de Variables**
  - **Son para reducir el número de conceptos, en cierta forma son una segmentación de columnas o atributos, al encontrar atributos de comportamiento similar, los agrupa en un eje. Sirve cuando se quiere simplificar una encuesta crear constructos que permitan comprender mejor a los clientes.**

1.- Exploratorios y Test

Análisis Descriptivo



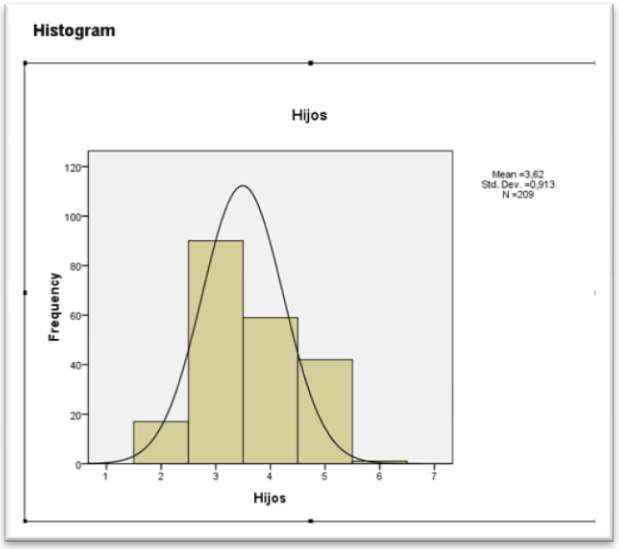
Frecuencies

Muestra el histograma de la variable, sirve para hacerse una idea de la distribución de las variables pidiendo el histograma (se pide en charts)

Frequency Table

Hijos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	17	8,1	8,1	8,1
	3	90	43,1	43,1	51,2
	4	59	28,2	28,2	79,4
	5	42	20,1	20,1	99,5
	6	1	,5	,5	100,0
	Total	209	100,0	100,0	



Descriptives

Muestra información básica como promedios, mínimo, máximo, desviación estándar.

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
R	209	19	0	19	1102	5,27	,354	5,123	26,247
F	209	15	4	19	2247	10,75	,363	5,247	27,534
M	209	$9,500	$500	$10,000	$527,500	$2,523.92	$167.206	$2,417.266	5843174,917
Valid N (listwise)	209								

CrossTabs

Muestra el cruce entre 2 variables y se puede pedir el test chi^2 (se pide en statistics), esto es para variables discretas, en caso de haber muchos casos distintos, no sirve.

Sexo * GSE Crosstabulation

Count

		GSE		
		C1	C3	Total
Sexo	Hombre	22	37	59
	Mujer	67	83	150
Total		89	120	209

Chi-Square Tests

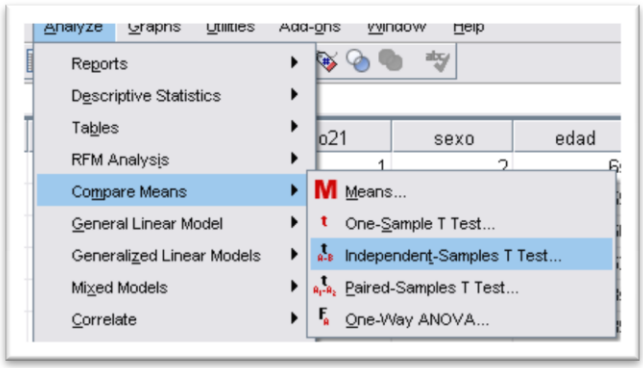
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,943[a]	1	,332		
Continuity Correction[b]	,665	1	,415		
Likelihood Ratio	,951	1	,330		
Fisher's Exact Test				,355	,208
N of Valid Cases	209				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 25.12.

Valor del test chi^2

Recordar que el test chi^2 es un test de dependencia de grupos, la hipótesis nula es “no existe dependencia de grupo”, y se rechaza con un valor, en este caso, a un 5%, no puedo decir que los grupos son distintos.

Teste sobre las Medias



One-Sample T Test

El test es para probar la probabilidad de que dada una muestra el valor real del promedio sea un parámetro

	N	Mean	Std. Deviation	Std. Error Mean
Hijos	209	3,62	,913	,063

One-Sample Test

	Test Value = 2					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Hijos	25,608	208	,000	1,617	1,49	1,74

Por ejemplo, si el promedio es 3,62 puedo evaluar la probabilidad de que sea 2, la probabilidad es 0 y al menos hay que sumarle 1,49 o a los mas 1,74 para quedar dentro del intervalo con 95% de probabilidad.

Independient-Samples T Test

Es como el test chi^2 pero para variables continuas solo en 2 grupos, como por ejemplo, el número de hijos, necesita una variable de agrupación, que en este caso fue GSE.

Group Statistics

	GSE	N	Mean	Std. Deviation	Std. Error Mean
Hijos	C1	89	3,31	1,029	,109
	C3	120	3,84	,745	,068

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hijos	Equal variances assumed	8,896	,003	-4,296	207	,000	-,527	,123	-,769	-,285
	Equal variances not assumed			-4,101	152,607	,000	-,527	,129	-,781	-,273

E este caso al 5% existe dependencia de grupo, los promedios son estadísticamente distintos.

One_Way Anova

Es el equivalente al chi^2 para variables continuas, funciona de forma similar al test anterior, pero comprar varios grupos simultáneamente y solo se puede concluir que existe un grupo distinto.

La hipótesis nula es que no existe dependencia de grupo

ANOVA

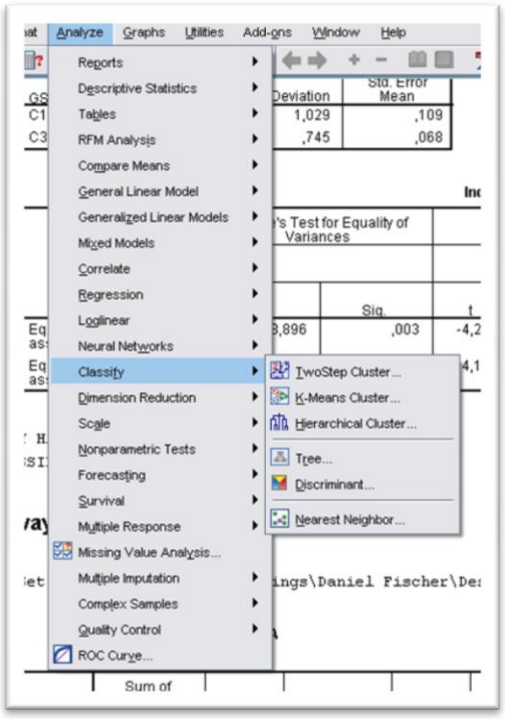
Hijos

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	122,862	3	40,954	166,195	,000
Within Groups	50,516	205	,246		
Total	173,378	208			

El resultado es el mismo que en el test-t, pero con significancia menor.

2.- Clasificatorios

La diferencia entre un método supervisado y uno no supervisado es que el no supervisado busca identificar clúster, los crea y el supervisado, intenta predecir clúster previamente hechos



a.- No Supervisado

K-Means

Se le asigna un numero de clúster y busca una segmentación, entrega los centros de masa de los clústeres y la idea es intentar ponerles nombre a los clústeres, si se encuentran 2 que no son conceptualmente distintos es que ya no vale la pena separarlos

Final Cluster Centers

	Cluster			
	1	2	3	4
R	17	1	18	6
F	5	17	5	7
M	$8,053	$647	$9,214	$2,602

Number of Cases in each Cluster

Cluster	1	19,000
	2	85,000
	3	7,000
	4	98,000
Valid		209,000
Missing		,000

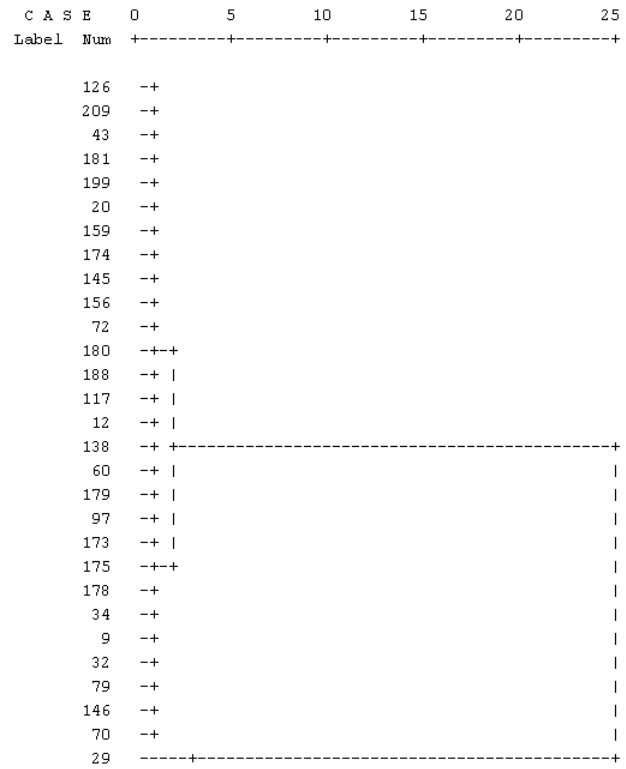
Vemos una clasificación de 4 clústeres, al grupo 1 es el de los muy gastadores que van poco, el 2 el de los que van mucho pero gastan poco, el 3 es casi idéntico al 1 y el 4 es el intermedio, en conclusión, hay que hacerlo con un clúster menos.

Además hay que fijarse que no aparezca un clúster con muy poca gente, como el 3.

- R: Hace cuanto fue la última vez
- M:Monto promedio gastado
- F: Frecuencia promedio

Hierarchical Cluster

Es para lo mismo que k-means, pero más lento por otro lado hace mejores clasificaciones, itera agregando en cada iteración al vecino más cercano a cada clúster y entrega un dendograma que muestra la distancia ente cada clúster que se unió.



Hay un salto muy grande entre 2 a 1 clúster, claramente hay que separar en al menos 2, por orto lado, habría que ver si es mejor 3 clústeres que 2, eso se puede hacer con test a ANOVA o graficando.

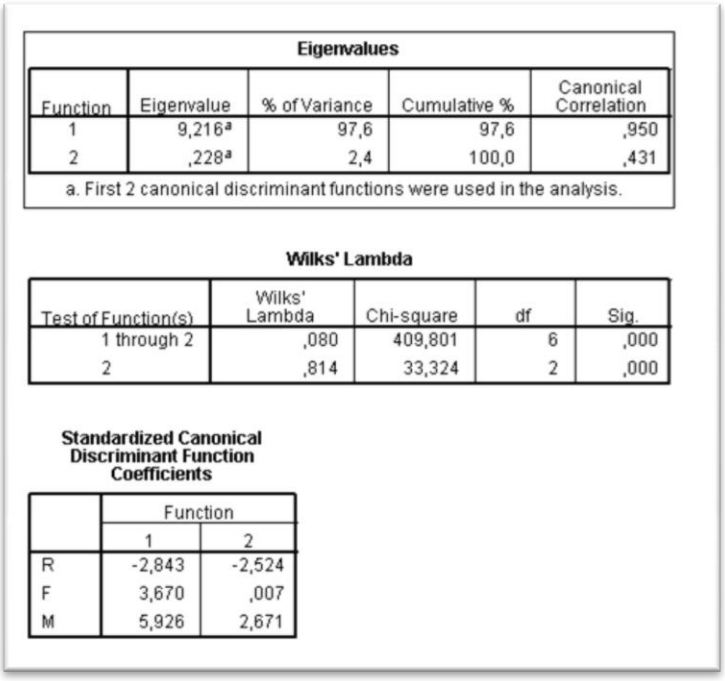
b.- Supervisado

Intenta predecir una segmentación previa con otros parámetros

Discriminant

Crea ejes que maximizan la varianza y luego corta por estos ejes para minimizar la varianza dentro de cada grupo, la calidad de la clasificación se resume con Lambda de Wilkinson

Mientras más chico, mejor, se mueve entre 0 y 1, menor a 0,3 es bueno.



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	9,216[a]	97,6	97,6	,950
2	,228[a]	2,4	100,0	,431

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,080	409,801	6	,000
2	,814	33,324	2	,000

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
R	-2,843	-2,524
F	3,670	,007
M	5,926	2,671

En este caso, intentamos predecir el número de hijos con R,M,F, si usamos solo la segunda función, lambda es muy malo, pro con ambas es muy bueno, por otro lado, la función 1 depende principalmente de M, y la segunda no depende de F (ultimo cuadro).

Classification Results[b,c]

		Hijos	Predicted Group Membership 2	3	4	Total
Original	Count	2	17	0	0	17
		3	0	86	4	90
		4	0	0	59	59
		Ungrouped cases	0	0	43	43
	%	2	100,0	,0	,0	100,0
		3	,0	95,6	4,4	100,0
		4	,0	,0	100,0	100,0
		Ungrouped cases	,0	,0	100,0	100,0
Cross-validated[a]	Count	2	17	0	0	17
		3	0	86	4	90
		4	0	1	58	59
	%	2	100,0	,0	,0	100,0
		3	,0	95,6	4,4	100,0
		4	,0	1,7	98,3	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 97.6% of original grouped cases correctly classified.

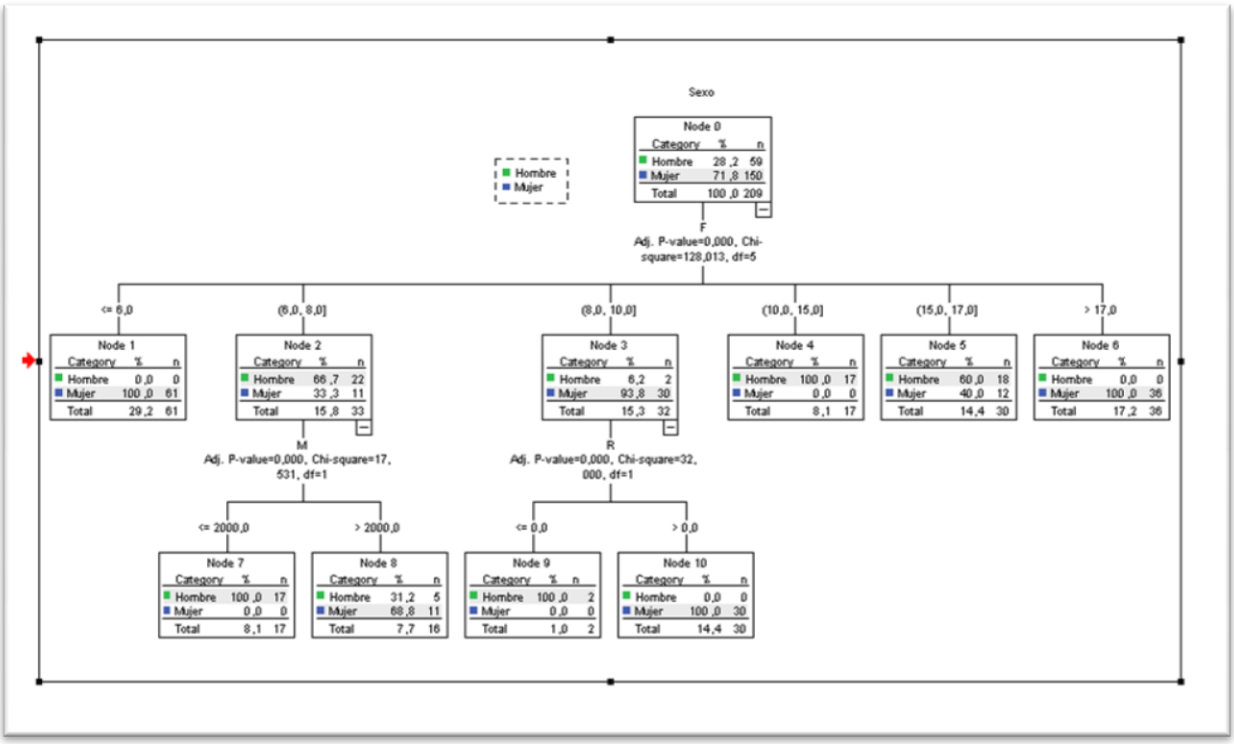
c. 97.0% of cross-validated grouped cases correctly classified.

La tabla de confusión también permite ver la calidad del ajuste

Podemos ver que no hay errores, de los con 2 hijos, los 17 fueron bien clasificados, pero hay casos que no pudieron clasificarse. Esto es porque en el rango de clasificación se puso de 0 a 4 hijos y habían casos con más hijos

Chaid (árbol)

Ahora intentemos predecir el sexo con R,M,F.



La primera variable por la que e clasifico fue F, creando 5 cortes (6 grupos).

El primero es el con F <=6

El segundo entre (6 – 8]

En el segundo conviene volver a cortar con M en 2000

Los que gastan menos de 2000 y tienen F entre (6 – 8] son todos hombres, mientras que los que gastan más de 2000 y tienen F (6 – 8] son 68,8% mujeres, por ende, se asumirá que si alguien cae en esta hoja, es mujer (de ahí el error de predicción)

Y también podemos ver la tabla de confucion:

Classification

Observed	Predicted		
	Hombre	Mujer	Percent Correct
Hombre	54	5	91,5%
Mujer	12	138	92,0%
Overall Percentage	31,6%	68,4%	91,9%

Growing Method: CHAID
Dependent Variable: Sexo

En el 91% de los casos el modelo acerta en el sexo

De los que son mujeres, acerta en el 92%, mientras que en los hombres el 90%

54 hombres fueron bien predichos y 5 fueron predichos como mujer.

3.- Reducción de variables

Cuando hay respuestas muy correlacionadas se utiliza este método para identificarlas e intentar encontrar lo que realmente esta tras esas respuestas.

¿Cuándo es adecuado realizar un AF?

Un **AF** resultará adecuado cuando existan altas correlaciones entre las variables, que es cuando podemos suponer que se explican por factores comunes. El análisis de la matriz de correlaciones será pues el primer paso a dar. Analíticamente, podemos comprobar el grado de correlación con las siguientes pruebas o test:

Test de esfericidad de Bartlett.

Es necesario suponer la normalidad de las variables. Contrasta la $\mathbf{H_0}$ de que la matriz de correlaciones es una matriz identidad (incorrelación lineal entre las variables). Si, como resultado del contraste, no pudiésemos rechazar esta $\mathbf{H_0}$, y el tamaño de la muestra fuese razonablemente grande, deberíamos reconsiderar la realización de un **AF**, ya que las variables no están correlacionadas.

El estadístico de contraste del test de Bartlett es:

$B = - (n - 1 - (2p + 5)/6) \ln | R^* |$
bajo la hipótesis nula resulta $X^2_{(p2 - p)/2}$

donde:

- **p** es el número de variables y
- $| R^* |$ es el determinante de la matriz de correlaciones muestrales.

Indice KMO (Kaiser-Meyer-Olkin) de adecuación de la muestra.

KMO se calcula como:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ji}^2}{\sum\sum_{i \neq j} r_{ji}^2 + \sum\sum_{i \neq j} a_{ji}^2}$$

donde:

- r_{ji} - coeficiente de correlación observada entre las variables **j** e **i**.
- a_{ji} - coeficiente de correlación parcial entre las variables **j** e **i**.

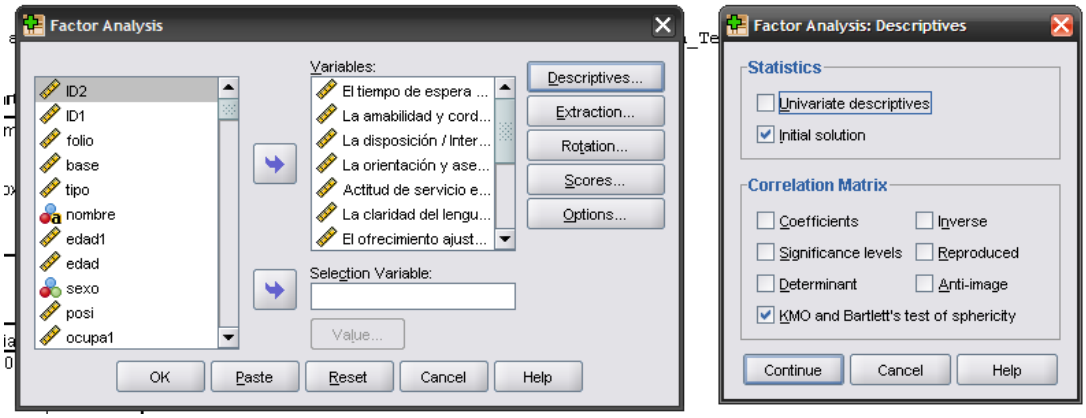
Estos coeficientes miden la correlación existente entre las variables **j** e **i**, una vez eliminada la influencia que las restantes variables ejercen sobre ellas. Estos efectos pueden interpretarse como los efectos correspondientes a los factores comunes, y por tanto, al eliminarlos, a_{ji} - representará la correlación entre los factores únicos de las dos variables, que teóricamente tendría que ser nula. Si hubiese

correlación entre las variables (en cuyo caso resultaría apropiado un **AF**), estos coeficientes deberían estar próximos a 0, lo que arrojaría un **KMO** próximo a 1. Por el contrario, valores del **KMO** próximos a 0 desaconsejarían el **AF**.

Está comúnmente aceptado que:

- Si **KMO** < 0.5 no resultaría aceptable para hacer un **AF**.
- Si 0.5 < **KMO** < 0.6 grado de correlación medio, y habría aceptación media.
- Si **KMO** > 0.7 indica alta correlación y, por tanto, conveniencia de **AF**.

En SPSS se debe pedir en el test mismo en la sección descriptives:



El output es el siguiente:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,947
Bartlett's Test of Sphericity	Approx. Chi-Square	4597,864
	df	91
	Sig.	,000

KMO es muy bueno: 0.947

Y se rechaza la hipótesis nula de Matriz Corr = I con significancia 0,000

Por ende está bien aplicar un método de análisis factorial

Para descibir un espacio de N dimensiones, se necesitan N vectores, pero solo se extraen los que aportan una cantidad considerable de información, estos son los con el valor propio mayor que 1 (Eigen Values)

ACP (Análisis de componentes principales)

Este es uno de los métodos de análisis factorial.

Crea ejes para explicar la máxima varianza, los ejes creados son ortogonales.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7,302	52,160	52,160	7,302	52,160	52,160	4,326	30,899	30,899
2	1,339	9,565	61,725	1,339	9,565	61,725	4,316	30,825	61,725
3	,763	5,449	67,174						
4	,587	4,191	71,365						
5	,572	4,085	75,450						
6	,529	3,780	79,230						
7	,472	3,368	82,598						
8	,439	3,132	85,730						
9	,405	2,890	88,621						
10	,387	2,763	91,384						
11	,377	2,695	94,079						
12	,335	2,393	96,471						
13	,260	1,856	98,327						
14	,234	1,673	100,000						

Extraction Method: Principal Component Analysis.

En este caso solo se extrajeron 2 ejes, explicando el 61% de la varianza

Communalities

	Initial	Extraction
El tiempo de espera para ser atendido por una operadora	1,000	,368
La amabilidad y cordialidad en la atención de la operadora	1,000	,657
La disposición / Interés en escuchar sus consultas o necesidades	1,000	,703
La orientación y asesoría para escoger el plan adecuado	1,000	,647
Actitud de servicio en general	1,000	,735
La claridad del lenguaje utilizado por el ejecutivo	1,000	,516
El ofrecimiento ajustado a sus necesidades	1,000	,555
La claridad de la explicación recibida acerca de los beneficios adicionales a su plan	1,000	,540
La claridad de la información recibida acerca de la duración del servicio y condiciones de término	1,000	,711
La claridad de la explicación recibida acerca de los costos	1,000	,566
La claridad de la explicación recibida de la forma en que se facturan los costos	1,000	,633
Desempeño de las funciones en general del Vendedor	1,000	,643
Y, ¿ Qué nota le pone a la claridad y sencillez de la información del contrato?	1,000	,666
Y en resumen, ¿Cómo califica el Proceso de la	1,000	,701

De la pregunta 1 se extrajo poca información (36%) , mientras que de la 2 se extrajo el 65%

Pero lo más importante es intentar encontrar los conceptos tras los ejes, eso se ve con las proyecciones de las preguntas sobre los ejes.

Component Matrix[a]

	Component	
	1	2
El tiempo de espera para ser atendido por una operadora	,550	,255
La amabilidad y cordialidad en la atención de la operadora	,645	,491
La disposición / Interés en escuchar sus consultas o necesidades	,707	,450
La orientación y asesoría para escoger el plan adecuado	,784	,177
Actitud de servicio en general	,796	,320
La claridad del lenguaje utilizado por el ejecutivo	,711	,106
El ofrecimiento ajustado a sus necesidades	,736	,118
La claridad de la explicación recibida acerca de los beneficios adicionales a su plan	,702	-,217
La claridad de la información recibida acerca de la duración del servicio y condiciones de término	,674	-,507
La claridad de la explicación recibida acerca de los costos	,708	-,253
La claridad de la explicación recibida de la forma en que se facturan los costos	,661	-,444
Desempeño de las funciones en general del Vendedor	,800	-,058
Y, ¿ Qué nota le pone a la claridad y sencillez de la información del contrato?	,757	-,305

Rotated Component Matrix[a]

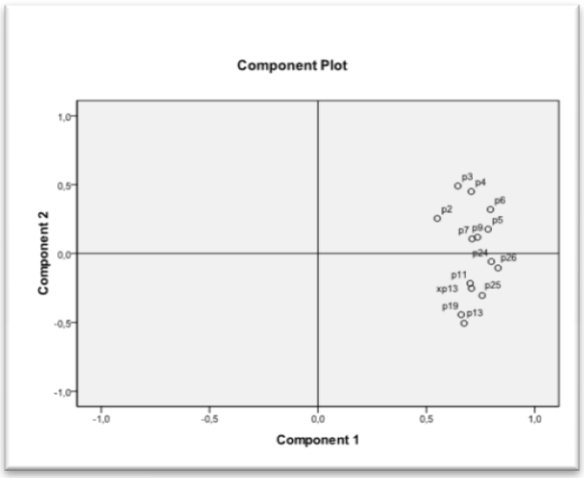
	Component	
	1	2
El tiempo de espera para ser atendido por una operadora	,569	,208
La amabilidad y cordialidad en la atención de la operadora	,803	,109
La disposición / Interés en escuchar sus consultas o necesidades	,819	,181
La orientación y asesoría para escoger el plan adecuado	,680	,429
Actitud de servicio en general	,789	,336
La claridad del lenguaje utilizado por el ejecutivo	,578	,427
El ofrecimiento ajustado a sus necesidades	,604	,436
La claridad de la explicación recibida acerca de los beneficios adicionales a su plan	,344	,649
La claridad de la información recibida acerca de la duración del servicio y condiciones de término	,119	,835
La claridad de la explicación recibida acerca de los costos	,322	,680
La claridad de la explicación recibida de la forma en que se facturan los costos	,154	,781
Desempeño de las funciones en general del Vendedor	,525	,606
Y, ¿ Qué nota le pone a la claridad y sencillez de la información del contrato?	,321	,751

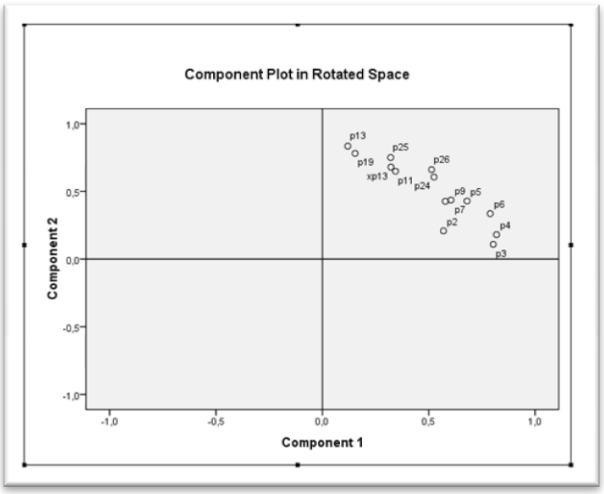
En la componente 1 tenemos mayor proyección sobre las preguntas relacionadas con la actitud del operador, mientras que en la 2, sobre la claridad de la información, se puede hacer una rotación para verlo más claro (podemos ver ambas tablas).

Es importante comprender el concepto tras los ejes, interpretando los ejes rotados, en el primer eje, tenemos mayor proyección sobre las preguntas relacionadas con la amabilidad de la operadora, mientras que en el segundo eje sobre la claridad de las condiciones del contrato.

Entonces s puede concluir que el 62% de la satisfacción está explicada por el trato de la operadora y la claridad del contrato.

Y también se pueden graficar las proyecciones

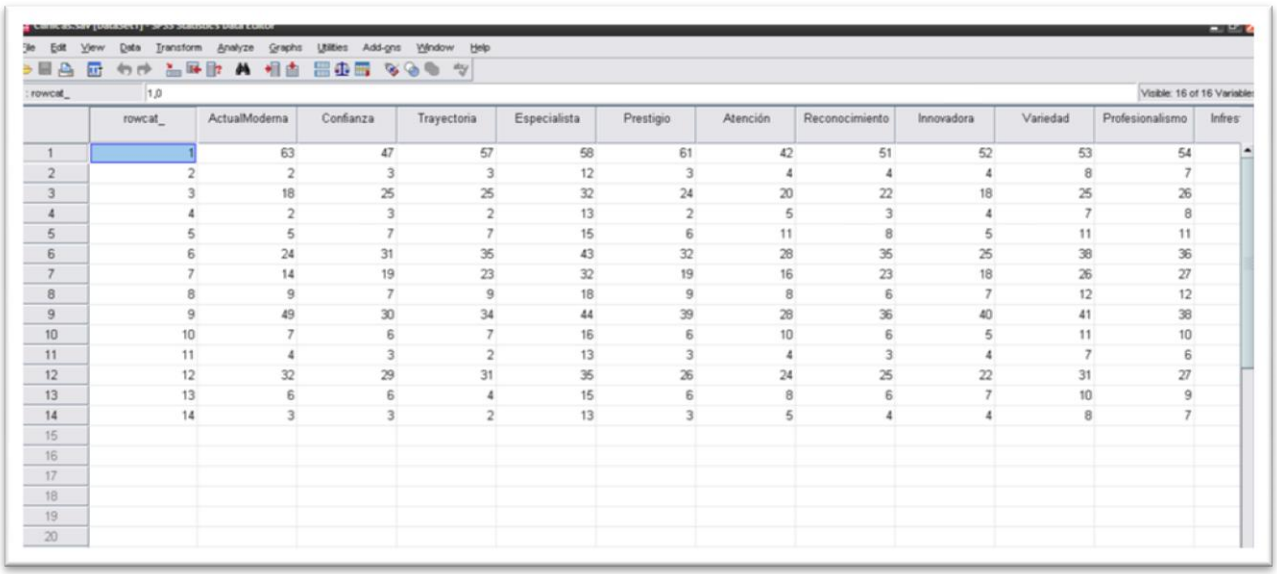




ANACORR

Otro método de reducción de variables, trabaja con las correlaciones de las variables.

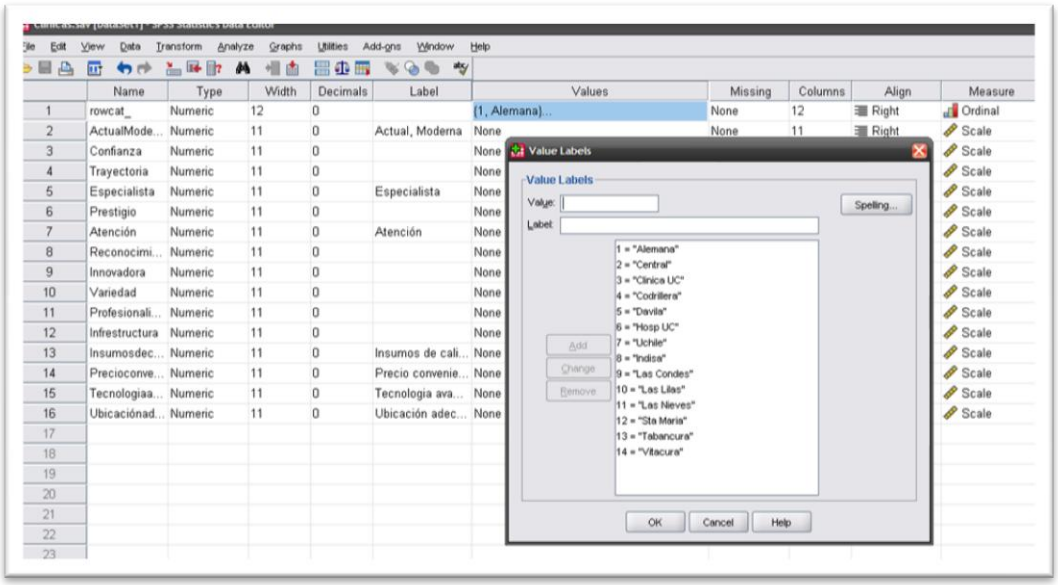
Conviene utilizarlo para mapas de posicionamiento como el siguiente:



File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

: rowcat_ 1,0 Visible: 16 of 16 Variables

	rowcat_	ActualModerna	Confianza	Trayectoria	Especialista	Prestigio	Atención	Reconocimiento	Innovadora	Variedad	Profesionalismo	Infres
1	1	63	47	57	58	61	42	51	52	53	54	
2	2	2	3	3	12	3	4	4	4	8	7	
3	3	18	25	25	32	24	20	22	18	25	26	
4	4	2	3	2	13	2	5	3	4	7	8	
5	5	5	7	7	15	6	11	8	5	11	11	
6	6	24	31	35	43	32	28	35	25	38	36	
7	7	14	19	23	32	19	16	23	18	26	27	
8	8	9	7	9	18	9	8	6	7	12	12	
9	9	49	30	34	44	39	28	36	40	41	38	
10	10	7	6	7	16	6	10	6	5	11	10	
11	11	4	3	2	13	3	4	3	4	7	6	
12	12	32	29	31	35	26	24	25	22	31	27	
13	13	6	6	4	15	6	8	6	7	10	9	
14	14	3	3	2	13	3	5	4	4	8	7	
15												
16												
17												
18												
19												
20												



Se ejecuta con el comando

```
CORRESPONDENCE
  table= ALL(14,15)
  /DIMENSIONS = 2
  /MEASURE = eucl
  /STANDARDIZE = RCMEAN
  /NORMALIZATION = SYMMETRICAL
  /PRINT = TABLE RPOINTS CPOINTS
  /PLOT = NDIM(1,MAX) BIPLOT(20) .
```

Intentara encontrar los atributos que están correlacionados con un output como el de ACP con las proyecciones sobre los ejes, pero proyectará los atributos y las marcas

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
Row	Mass	1	2	Inertia	1	2	1	2	Total
Alemana	,071	-1,098	,178	,014	,527	,026	,971	,014	,985
Central	,071	,218	-,139	,001	,021	,016	,738	,161	,899
Clinica UC	,071	,045	,363	,001	,001	,107	,020	,700	,720
Codrillera	,071	,252	-,240	,001	,028	,047	,638	,309	,947
Davila	,071	,269	,032	,001	,032	,001	,735	,006	,741
Hosp UC	,071	,139	,585	,003	,008	,279	,081	,767	,849
Uchile	,071	,255	,276	,002	,028	,062	,352	,221	,574
Indisa	,071	,178	-,101	,001	,014	,008	,489	,084	,573
Las Condes	,071	-,786	-,445	,009	,270	,162	,820	,141	,961
Las Lilas	,071	,194	-,113	,001	,016	,010	,531	,096	,627
Las Nieves	,071	,163	-,317	,001	,012	,082	,309	,626	,935
Sta Maria	,071	-,183	,375	,003	,015	,115	,148	,333	,482
Tabancura	,071	,148	-,213	,001	,010	,037	,422	,464	,886
Vitacura	,071	,206	-,241	,001	,019	,048	,534	,391	,924
Active Total	1,000			,039	1,000	1,000			

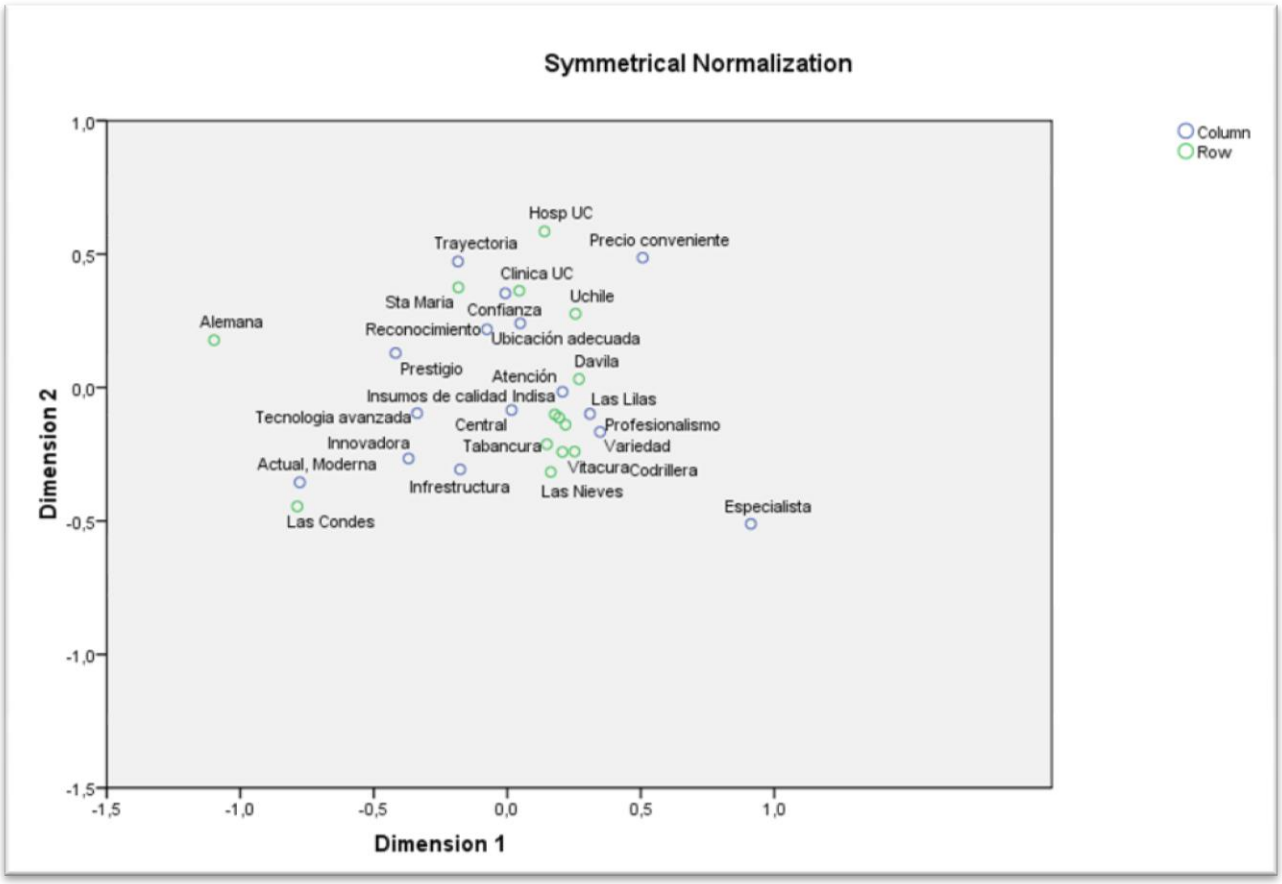
a. Symmetrical normalization

Overview Column Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
Column	Mass	1	2	Inertia	1	2	1	2	Total
Actual, Moderna	,067	-,777	-,355	,008	,246	,096	,839	,094	,933
Confianza	,067	-,007	,353	,001	,000	,095	,001	,781	,782
Trayectoria	,067	-,185	,472	,002	,014	,170	,210	,735	,945
Especialista	,067	,912	-,511	,011	,339	,199	,851	,143	,994
Prestigio	,067	-,418	,129	,002	,071	,013	,807	,041	,848
Atención	,067	,207	-,016	,001	,017	,000	,523	,002	,525
Reconocimiento	,067	-,076	,218	,001	,002	,036	,064	,283	,347
Innovadora	,067	-,369	-,266	,002	,056	,054	,687	,191	,878
Variedad	,067	,347	-,166	,002	,049	,021	,796	,097	,893
Profesionalismo	,067	,310	-,099	,001	,039	,007	,731	,040	,771
Infrestructura	,067	-,177	-,307	,001	,013	,072	,295	,477	,772
Insumos de calidad	,067	,016	-,084	,000	,000	,005	,013	,193	,207
Precio conveniente	,067	,507	,487	,005	,105	,180	,608	,300	,909
Tecnologia avanzada	,067	-,338	-,096	,002	,046	,007	,766	,033	,800
Ubicación adecuada	,067	,048	,241	,001	,001	,044	,026	,345	,371
Active Total	1,000			,039	1,000	1,000			

a. Symmetrical normalization

Se puede ver el resumen de ambas tablas en el siguiente gráfico:



Se interpreta de la misma forma que ACP

Ejercicio: interpretar los ejes.