MAESTRÍA EN MÉTODOS CUANTITATIVOS PARA EL ANÁLISIS Y GESTIÓN DE DATOS EN LAS ORGANIZACIONES

M71V/M72V 06 MÉTODOS DE ANÁLISIS MULTIVARIADO

SEGUNDA ENTREGA TRABAJO FINAL – ANÁLISIS DE COMPONENTES PRINCIPALES

DOCENTES: DEL DUCA, Silvina - VIETRI, Silvia

GRUPO N° 3:

BRAVO, Juan
FILOMENO, Antonella
FUNES, Gustavo
LAMBRECHT, Lea

INTRODUCCIÓN

Continuando con el objetivo de comprender las operaciones de la cadena de supermercados, la finalidad principal de este informe es transmitir los resultados de un proceso de identificación de patrones subyacentes y estructuras latentes mediante el análisis de componentes principales. Para ello se parte de la misma base de datos que ya se posee, y se utilizó para el análisis de conglomerados.

Los resultados obtenidos se presentarán como un complemento a la segmentación previamente realizada mediante técnicas de conglomerado. Estos hallazgos proporcionarán una visión más clara de las relaciones entre variables y contribuirán a las recomendaciones estratégicas que persiguen los análisis realizados.

ANÁLISIS DESCRIPTIVO

Se utilizó la misma base de datos que la utilizada en para el análisis de conglomerados. La base de datos cuenta con 17 variables, las que presentan cierta dificultad de interpretación a la hora de realizar un análisis estratégico **(Anexo II - Descripción de las variables)**. De este grupo se excluyen las variables categóricas. Este trabajo pretende evaluar la interrelación entre las mismas, y analizar la posibilidad de reducir las variables con las que se va continuar trabajando, sin que por ello se pierda la información que ellas aportan. El método de componentes principales logra transformar las variables originales en un nuevo conjunto de variables no correlacionadas entre sí. Aquí se muestra la salida de R para los primeros registros de las primeras variables.

```
> head(data)
# A tibble: 6 × 17
  `N° SUC` `Ubicación SUC` Zona      Año `Unidades Vendidas` `$ Precio Venta Promedio` `$ Costo Promedio` `$ Venta c/IVA`
     <dbl> <chr>           <chr>   <dbl>               <dbl>                    <dbl>               <dbl>           <dbl>
1       55 CHACABUCO       NORTE 1  2022             4163172.                      203.                152.      1132795902.
2       57 CHIVILCOY       NORTE 1  2022             5048535.                      198.                145.      1328470644.
3       58 BRAGADO         NORTE 1  2022             4919319.                      219.                161.      1413269313.
4       85 JUNIN           NORTE 1  2022             8184715.                      226.                160.      2460599681.
5       89 COLON           NORTE 1  2022             3993697.                      183.                133.       950401220.
6       96 SALTO           NORTE 1  2022             3705004.                      203.                151.       989021425.
# ℹ 9 more variables: `$ Venta s/IVA` <dbl>, `$ Margen AM` <dbl>, `Cant. Mermas` <dbl>, `$ Mermas` <dbl>,
#   `$ Utilidad Bruta` <dbl>, `Costo Laboral (Sin seguridad y Limpieza)` <dbl>, `$Bufet` <dbl>,
#   `Cant. Donaciones` <dbl>, `$ Donacion Mercaderia` <dbl>
```

A continuación se muestra información obtenida con la función “summary” sin escalar las variables. Como complemento en el **Anexo II - Tabla 2** se presenta la media y desvío estándar con la función “apply”.

Tabla 1: Resumen estadístico – Función Summary (sin variables escaladas)

summary	Unid Vendidas	$ Venta s/IVA	$ Margen AM	$ Mermas	Costo Laboral	$ Donación Merc
Min.	282970	$ 111.400.000.000	$ 46.060.000.000	-$ 73.641.397	$ 29.621.436	-$ 6.942.885,00
1st Qu.	3293303	$ 750.200.000.000	$ 223.200.000.000	-$ 22.301.418	$ 76.567.271	-$ 1.230.101,00
Median	4862077	$ 1.079.000.000.000	$ 344.300.000.000	-$ 13.602.923	$ 114.879.686	-$ 562.171,00
Mean	5258509	$ 1.307.000.000.000	$ 429.500.000.000	-$ 16.970.719	$ 142.948.403	-$ 855.951,00
3rd Qu.	6742179	$ 1.730.000.000.000	$ 566.700.000.000	-$ 8.074.249	$ 176.780.924	-$ 227.559,00
Max.	14471548	$ 4.006.000.000.000	$ 1.703.000.000.000	-$ 554.307	$ 554.306.426	$ 0,00

SUPUESTOS TEÓRICOS

El supuesto teórico más relevante para la aplicación de PCA es la linealidad entre las variables, es decir que exista una relación lineal entre las mismas. Este supuesto se valida utilizando la matriz de correlación. Para el caso de los datos analizados la matriz de correlación es la siguiente:

Matriz de Correlación

Para evaluar la relación lineal entre dos variables cuantitativas, se utilizó el método de correlación de Pearson, que proporciona un valor que indica la fuerza y la dirección de la relación entre las dos variables.

Los elementos de la diagonal principal de la matriz son iguales a 1, ya que representan la correlación de cada variable consigo misma. Los elementos fuera de la diagonal principal (las correlaciones entre pares de variables distintas) varían entre 1 y -1.

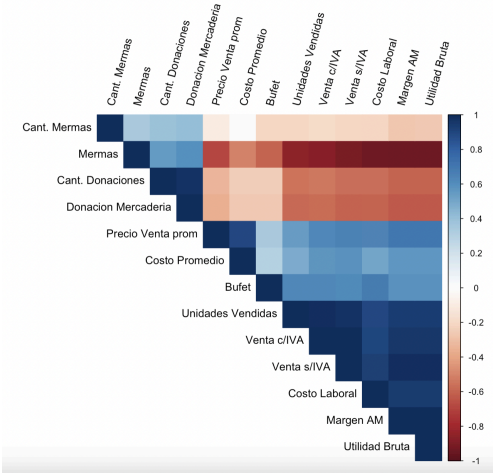
Si se toma como ejemplo la variable “unidades vendidas”, se interpreta que con respecto a sí misma tiene una correlación perfecta. En cuanto a Ventas s/IVA, tiene una fuerte correlación positiva, lo que significa que a medida que una variable aumenta, la otra también lo hace. Con respecto a la variable Merma, tiene una fuerte correlación negativa, y significa que a medida que una variable aumenta, la otra disminuye. En cambio, la correlación con Cant. Mermas, tiene un valor cercano a cero, lo que indica que hay poca correlación lineal.

Tabla 3: Matriz de Correlación (método Pearson)

	Unidades Vendidas	Precio Venta prom	Costo Promedio	Venta c/IVA	Venta s/IVA	Margen AM	Cant. Mermas	Mermas	Utilidad Bruta	Costo Laboral	Bufet	Cant. Donaciones	Donacion Mercaderia
Unidades Vendidas	1.0000000	0.5407348	0.48241929	0.9863030	0.9795096	0.9353093	-0.20873521	-0.8516736	0.9368083	0.8956106	0.6328171	-0.5521958	-0.5808286
Precio Venta prom	0.5407348	1.0000000	0.89734976	0.6261526	0.6578695	0.7054932	-0.10860476	-0.6891136	0.7045369	0.6414465	0.3431779	-0.3372704	-0.3654815
Costo Promedio	0.4824193	0.8973498	1.00000000	0.5798386	0.5825410	0.5643223	0.01732851	-0.5093541	0.5654281	0.5064884	0.3058096	-0.2492675	-0.2696885
Venta c/IVA	0.9863030	0.6261526	0.57983857	1.0000000	0.9936052	0.9554893	-0.18649948	-0.8719771	0.9569345	0.9065758	0.6374460	-0.5397403	-0.5712246
Venta s/IVA	0.9795096	0.6578695	0.58254103	0.9936052	1.0000000	0.9801407	-0.20820444	-0.9060852	0.9811044	0.9293405	0.6145019	-0.5642288	-0.6003252
Margen AM	0.9353093	0.7054932	0.56432234	0.9554893	0.9801407	1.0000000	-0.27038792	-0.9486825	0.9999002	0.9433326	0.5834973	-0.5946734	-0.6358506
Cant. Mermas	-0.2087352	-0.1086048	0.01732851	-0.1864995	-0.2082044	-0.2703879	1.00000000	0.3422859	-0.2665285	-0.2135888	-0.2062028	0.3971030	0.4181522
Mermas	-0.8516736	-0.6891136	-0.50935415	-0.8719771	-0.9060852	-0.9486825	0.34228586	1.0000000	-0.9441196	-0.9370387	-0.5980921	0.5566299	0.5950892
Utilidad Bruta	0.9368083	0.7045369	0.56542810	0.9569345	0.9811044	0.9999002	-0.26652846	-0.9441196	1.0000000	0.9413567	0.5814490	-0.5949503	-0.6361505
Costo Laboral	0.8956106	0.6414465	0.50648843	0.9065758	0.9293405	0.9433326	-0.21358878	-0.9370387	0.9413567	1.0000000	0.6838449	-0.5613136	-0.5965817
Bufet	0.6328171	0.3431779	0.30580958	0.6374460	0.6145019	0.5834973	-0.20620280	-0.5980921	0.5814490	0.6838449	1.0000000	-0.2415524	-0.2603253
Cant. Donaciones	-0.5521958	-0.3372704	-0.24926750	-0.5397403	-0.5642288	-0.5946734	0.39710298	0.5566299	-0.5949503	-0.5613136	-0.2415524	1.0000000	0.9753569
Donacion Mercaderia	-0.5808286	-0.3654815	-0.26968849	-0.5712246	-0.6003252	-0.6358506	0.41815223	0.5950892	-0.6361505	-0.5965817	-0.2603253	0.9753569	1.0000000

En el gráfico 1 se observa la mitad superior de la matriz (ya que la mitad inferior es asimétrica). Con los colores se resaltan los patrones respecto a la correlación positiva, negativa y poca o nula correlación.

Gráfico 1: Representación gráfica de matriz de correlación



MÉTODO DE COMPONENTES PRINCIPALES (CP)

Con el objeto de reducir la información brindada por las variables originales seleccionadas, al aplicar el método de CP, se conservan los CP cuyas varianzas pueden explicar la mayor cantidad de varianza de dichas variables originales.

La metodología se implementó con dos bibliotecas diferentes en R, para verificar los resultados. En el presente informe se muestran los resultados más relevantes de cada una de ellas. El desarrollo completo de la metodología puede encontrarse en el script de R adjunto.

En principio se corrió la función de prcomp sin escalar las variables **(Anexo II - Gráfico 2)** y como resultado surgió que el primer CP tiene una varianza exorbitante y las demás, varianzas muy pequeñas.

Luego, se continuó con la función prcomp escalando las variables **(Anexo II - Gráfico 3)** y calculando los componentes principales. En el siguiente cuadro se observa la varianza explicada por cada uno de los CP, la porción respecto al total y la porción de la varianza acumulada **(Anexo II - Gráfico 4)**:

Tabla 4: Resumen estadístico – Función Summary

Summary STATS	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	2.96	1.25	1.01	0.86	0.66	0.48	0.24	0.20	0.17	0.13	0.06	0.02	1,48E-12
Proportion of Variance	0.68	0.12	0.08	0.05	0.03	0.02	0.004	0.003	0.002	0.001	0.0003	0.00005	0.000e+00
Cumulative Proportion	0.68	0.79	0.88	0.93	0.97	0.98	0.99	0.996	0.998	0.999	0.999	1.00	1.00+03

Las primeras tres componentes tienen las varianzas (autovalores) mayores que 1 y entre las tres recogen el 88% de la varianza de las variables originales. Se parte de 13 CP, pero si se seleccionan las tres primeras, se estaría perdiendo solo el 12% de la información. Si se agrega una cuarta componente se estaría explicando el 93% de la variabilidad total y se perdería el 7% de la información. En el **Anexo II - Gráfico 5 y 6** se muestra gráficamente el aporte de varianza de cada CP.

Matriz de rotación

La matriz de rotación nos brinda información sobre cómo cada variable original contribuye a cada componente principal. Cada fila corresponde a una variable original, y cada columna corresponde a un componente principal. La matriz de rotación se muestra en **Anexo II - Tabla 5.**

Carga factorial

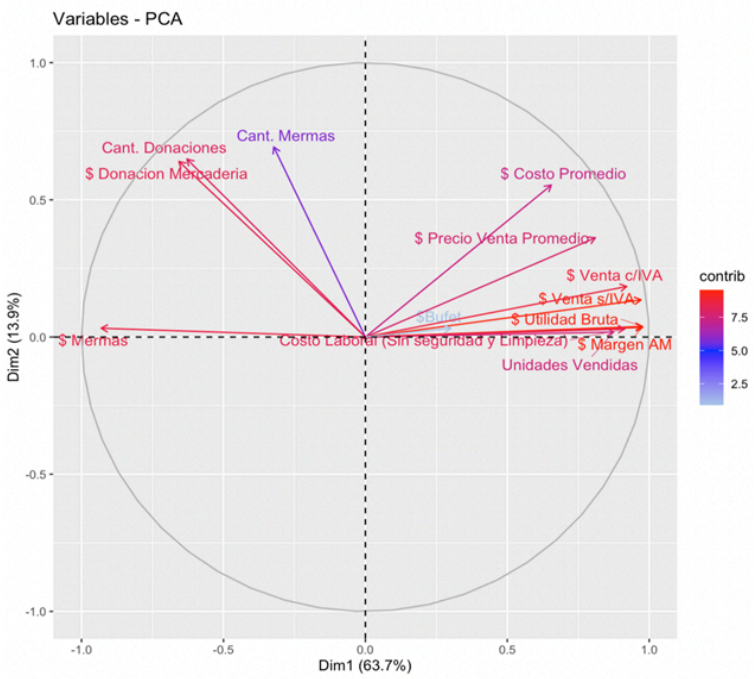
La siguiente tabla indica las variables que están mejor representadas en cada componente principal.

Tabla 6: Carga factorial

```
> pca_data.fm$var$cos2 # Para saber que variables estan mejor representadas
                                        Dim.1         Dim.2        Dim.3       Dim.4         Dim.5
Unidades Vendidas                  0.88696304 0.00002407529 0.0125064658 0.047578958 0.00403723925
$ Precio Venta Promedio            0.54704473 0.01650437834 0.2678016704 0.130785016 0.00006995612
$ Costo Promedio                   0.40454280 0.01718766975 0.4246224879 0.093712966 0.01745514230
$ Venta c/IVA                      0.93188013 0.00159227136 0.0003440334 0.027915303 0.00226567614
$ Venta s/IVA                      0.96283150 0.00066249521 0.0003504794 0.016945556 0.00701040919
$ Margen AM                        0.97032636 0.00065584724 0.0017024743 0.001291988 0.01553933860
Cant. Mermas                       0.06306715 0.03768083495 0.4153173086 0.469189028 0.00407260871
$ Mermas                           0.88487684 0.00788311705 0.0117047804 0.002368606 0.01116651791
$ Utilidad Bruta                   0.96951156 0.00051628909 0.0014567756 0.001571365 0.01567564335
Costo Laboral (Sin seguridad y Limpieza) 0.90703375 0.00223649918 0.0059967837 0.015600301 0.00104665029
$Bufet                             0.30926250 0.37065068307 0.0718076199 0.015283158 0.22968612579
Cant. Donaciones                   0.26526552 0.66277479355 0.0150968169 0.009814783 0.03229490747
$ Donacion Mercaderia              0.20349982 0.75117969115 0.0037765249 0.003215257 0.02362839740
> |
```

Este siguiente gráfico muestra la contribución de las variables al CP. Dependiendo la contribución se asigna un color a cada variable. Las variables más contributivas tendrán colores más intensos.

<u>Gráfico 7</u>: Representación gráfica de las Componentes Principales



INTERPRETACIÓN DE LAS NUEVAS VARIABLES

CP1: Unidades Vendidas - Margen - Venta C/IVA y Venta S/IVA -> Rendimiento Global

CP2: Donación de Mercadería - Cant. Donaciones -> Responsabilidad Socio-Empresarial

CP3: Costo Promedio - Cantidad de Mermas- Costo Promedio -> Eficiencia Operativa

CONCLUSIÓN

En conclusión, el análisis de componentes principales (PCA) aplicado a nuestro conjunto de datos, que inicialmente incluía 13 variables relacionadas con montos de ingresos y egresos, ha demostrado ser una herramienta valiosa para reducir la complejidad y destacar patrones fundamentales. Al seleccionar tres componentes principales que explican conjuntamente el 87,8 % de la variabilidad original, se ha logrado condensar la información de manera significativa.

Estas tres dimensiones se representan como *rendimiento global, responsabilidad socio-empresarial y eficiencia operativa.*

Bibliografía

- Aluja, T., Morineau, A. (1999). Aprender de los datos: el análisis de componentes principales, una aproximación desde el data mining. EUB Barcelona.
- Cuadras C. M. (2019). Nuevos Métodos de Análisis Multivariante. CMC Editions Barcelona.
- Johnson, D. E. (2000). Métodos multivariados aplicados al Análisis de Datos. International Thomson Editores. México.
- Joaquín Aldás (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo SA.