Análisis de bodegas españolas mediante técnicas de clustering

Andrés Mituca Mituca

Contents

1 Resumen Ejecutivo

Este proyecto aplica técnicas de **clustering no supervisado** para segmentar bodegas españolas a partir de variables económicas clave. Tras un proceso riguroso de limpieza de datos, análisis exploratorio y validación estadística, se identifican seis perfiles económicos claramente diferenciados, que reflejan distintos modelos de negocio dentro del sector vitivinícola.

El análisis combina métodos jerárquicos y de partición, evaluados mediante múltiples **criterios de validación interna** (Hopkins, Silhouette, estabilidad), y se apoya en PCA y gráficos de perfiles para la interpretación económica de los resultados. Los hallazgos permiten entender la heterogeneidad estructural del sector, con aplicaciones potenciales en análisis financiero, diseño de políticas públicas, segmentación de clientes o benchmarking empresarial.

2 Introducción

El sector vitivinícola español es uno de los más relevantes a nivel europeo tanto por volumen de producción como por diversidad empresarial. Conviven en él bodegas de carácter familiar con grandes grupos empresariales, lo que genera una alta heterogeneidad en tamaño, estructura financiera, rentabilidad y eficiencia operativa.

En este contexto, resulta especialmente útil aplicar técnicas de aprendizaje no supervisado, como el **clustering**, que permiten identificar patrones latentes y perfiles homogéneos sin imponer una clasificación previa. Este enfoque facilita una visión más estructurada del sector y aporta valor para distintos agentes económicos (empresas, entidades financieras, analistas o administraciones públicas).

Este proyecto tiene como finalidad segmentar bodegas españolas a partir de variables económicas, evaluando distintos métodos de clustering y seleccionando el más adecuado mediante criterios estadísticos y de interpretabilidad económica.

Las técnicas de clustering permiten identificar patrones latentes en los datos sin una clasificación previa [1], [5].

3 Objetivos del Proyecto

El objetivo general de este trabajo es identificar y caracterizar perfiles económicos homogéneos de bodegas españolas mediante técnicas de clustering.

De forma específica, se persigue:

1) Preparar y limpiar una base de datos económica garantizando calidad y consistencia.
2) Analizar la existencia de estructura de agrupamiento en los datos.
3) Comparar distintos algoritmos de clustering (jerárquicos y de partición).
4) Seleccionar el método óptimo mediante criterios de validación interna.
5) Interpretar los clusters desde una perspectiva económica y de negocio.
6) Analizar la relación de los clusters con variables externas no utilizadas en su construcción.

4 Metodología

El análisis se estructura en las siguientes etapas:

1) Limpieza y preprocesado de datos

- Tratamiento de valores faltantes.
- Identificación y eliminación de valores extremos.
- Selección de variables relevantes.

2. Estandarización

- Escalado de las variables económicas para evitar dominancias por magnitud.

3. Análisis de tendencia al clustering

- Aplicación del estadístico de Hopkins.

4. Clustering

- Métodos jerárquicos (Ward y media).
- Métodos de partición (k-medias y k-medoides).

5. Validación

- Coeficiente de Silhouette.
- Medidas de estabilidad y compacidad (clValid).

6. Interpretación

- Análisis de componentes principales (PCA).
- Gráficos de perfiles medios.
- Relación con variables externas.

Este enfoque garantiza tanto rigor estadístico como interpretabilidad económica.
El análisis fue realizado en el software **R**, haciendo uso de diferentes librerías como: [8], [6], [4] y [7].

```r
datos_bodegas <- read_csv("datos/datos_bodegas_limpio_FINAL.csv")
```

5 Limpieza de datos y decisiones metodológicas

Agruparemos las bodegas en función de su contenido económico en los distintos parámetros incluidos en el bloque económico para luego relacionar los clusters obtenidos con la variable *esta_trip* y *valoracion* o con otras variables no consideradas en el análisis. Además, escalaremos los datos, puesto que cada variable está medida en diferentes unidades y magnitudes y no queremos que esto interfiera en la agrupación.

```r
datos_bodegas_economicos = datos_bodegas[,5:12]
datos_bodegas_economicos = scale(datos_bodegas_economicos, center = TRUE, scale = TRUE)
```

5.1 Decisiones clave de preprocesado

Durante el análisis exploratorio se identificaron bodegas con valores extremos en variables económicas clave. Estas observaciones distorsionaban significativamente las medidas de tendencia central y dispersión, comprometiendo la calidad del clustering.

Se optó por eliminar aproximadamente el **5%** de las bodegas identificadas como anómalas mediante consenso de distintos métodos estadísticos, en lugar de aplicar transformaciones logarítmicas. Esta decisión se fundamenta en que:

- Los valores extremos parecían errores de registro o casos no representativos del sector.
- La pérdida de información fue limitada.
- Se mantuvo la interpretabilidad económica directa de las variables.

Adicionalmente, se imputaron valores faltantes en el bloque económico y se seleccionaron los 30 stems con mayor coeficiente de variación en el bloque digital, priorizando variables con mayor capacidad discriminante; aunque no serán relevantes para el trabajo presente.

---

6 Medida de distancia y tendencia de agrupamiento

Con el objetivo de analizar el comportamiento económico y observar si existe tendencia de agrupamiento de las bodegas, procedemos a comparar dos tipos de distancias diferentes: la **distancia de Manhattan** y la **distancia euclidiana**.

Tras observar los gráficos obtendidos mediante el análisis (gráficos 5 y 6 del 'Anexo'), obtenemos las siguientes conclusiones:

Hacemos uso de la *distancia de Manhattan* en el clustering jerárquico (excepto en Ward) porque con variables económicas tan diferentes (euros, porcentajes, personas) es más robusta ante posibles outliers que hayan quedado.

Además, el gráfico que muestra la distancia de Manhattan (segundo gráfico) muestra bandas horizontales intensas en la parte inferior, lo que indica grupos de bodegas con perfiles económicos similares; y las líneas naranjas más marcadas sugieren separación más clara entre bodegas pequeñas y bodegas grandes.

Para posterior análisis de k-means y demás, preferimos mantener la *distancia euclidiana* que viene por defecto, ya que al trabajar con datos escalados funciona perfectamente y es la más usada en la literatura. Asimismo, ya la habíamos usado en el PCA para detectar anomalías, así que mantener la misma métrica da coherencia a todo el trabajo en su conjunto.

La tendencia al clustering se evaluó mediante el **estadístico de Hopkins** [2].

Los valores del estadístico de Hopkins nos confirman una tendencia de agrupamiento, puesto que ha sido calculado para diferentes valores de m (n en la función) y con diferentes semillas aleatorias, y sus valores oscilan entre 0.88 y 0.92.

Para entrar más en profundidad, el Hopkins alrededor del 0,9 deja claro que los datos no están dispersos al azar, sino que tienden a agruparse de manera bastante marcada. Además, como todos los valores están muy juntos, el patrón no parece fruto del ruido, sino algo consistente en toda la muestra. En general, este resultado apunta a que **sí hay una estructura real de clusters** y que aplicar técnicas de agrupación tiene sentido porque los datos "se dejan" segmentar.

Nota: pese a poder haber reducido el número de semillas para facilitar el cálculo del estadístico, hemos decidido seguir con nuestro análisis inicial para asegurar una mayor aleatoriedad.

---

7 Modelos jerárquicos

La calidad de los clusters se evaluó mediante el **coeficiente de Silhouette** [9].

En primer lugar, aplicaremos modelos jerárquicos; y cabe destacar que, aún conociendo los distintos métodos que existen para el cálculo de las distancias entre clusters, utilizaremos el método de Ward y el método de de la media para reducir la extensión, aunque somos conscientes de la recomendación de uso de varios métodos diferentes.

7.1 Método de Ward

Para obtener el número de clusters óptimo, combinaremos los criterios del coeficiente de Silhouette medio y la varianza intra-cluster, y fijaremos en 10 el máximo número de clusters permitido (gráfico 7 del 'Anexo').

Teniendo los datos ya escalados y centrados, podemos fijarnos únicamente en la estructura real de los grupos sin que una variable domine al resto. Aunque el silhouette alcanza su máximo en k=2, a partir de ese punto se observa una mejora sostenida hasta valores cercanos a 5–6, donde se estabiliza sin caídas bruscas. El método del codo también muestra un cambio notable en esa zona, indicando que añadir más clusters luego apenas reduce la variabilidad. Por lo tanto, $k \approx 6$ ofrece un equilibrio razonable entre detalle e interpretabilidad, permitiendo distinguir perfiles económicos distintos entre bodegas sin fragmentar en exceso la muestra.

```
## grupos1
##   1   2   3   4   5   6
## 228  37 120  36  60  17
```

Al cortar en 6 clusters con ward.D2 obtenemos grupos de tamaños muy distintos, lo cual encaja con la realidad económica del sector: muchas bodegas pequeñas (grupo 1) y pocas muy grandes con estructuras financieras específicas (grupos 2 y 6). La presencia de grupos intermedios refleja distintos modelos de negocio y niveles de rentabilidad y activos. En conjunto, la partición en 6 parece capturar bien esa *heterogeneidad* sin forzar divisiones artificiales.

7.2 Método de la media

Para obtener el número de clusters óptimo, de nuevo combinaremos los criterios del coeficiente de Silhouette medio y la varianza intra-cluster, y fijaremos en 10 el máximo número de clusters permitido (gráfico 9 del 'Anexo').

En el gráfico del *silhouette* se observa un máximo claro en k=2, pero a partir de ahí los valores se mantienen relativamente altos hasta $k \approx 6$, lo que sugiere que aumentar clusters aporta diferenciación útil sin perder calidad. Por su parte, el método del codo muestra una caída brusca hasta k=4 y luego una reducción progresiva hasta estabilizarse en torno a k=6, indicando que más allá apenas mejora la compactación. Esto encaja con la idea de que existen distintos perfiles económicos dentro del sector y no solo dos grandes grupos extremos. Por tanto, *kalrededorde*6 parece un punto razonable si buscamos detalle sin sobreajustar.

```
## grupos2
##   1   2   3   4   5   6
## 475   8  11   2   1   1
```

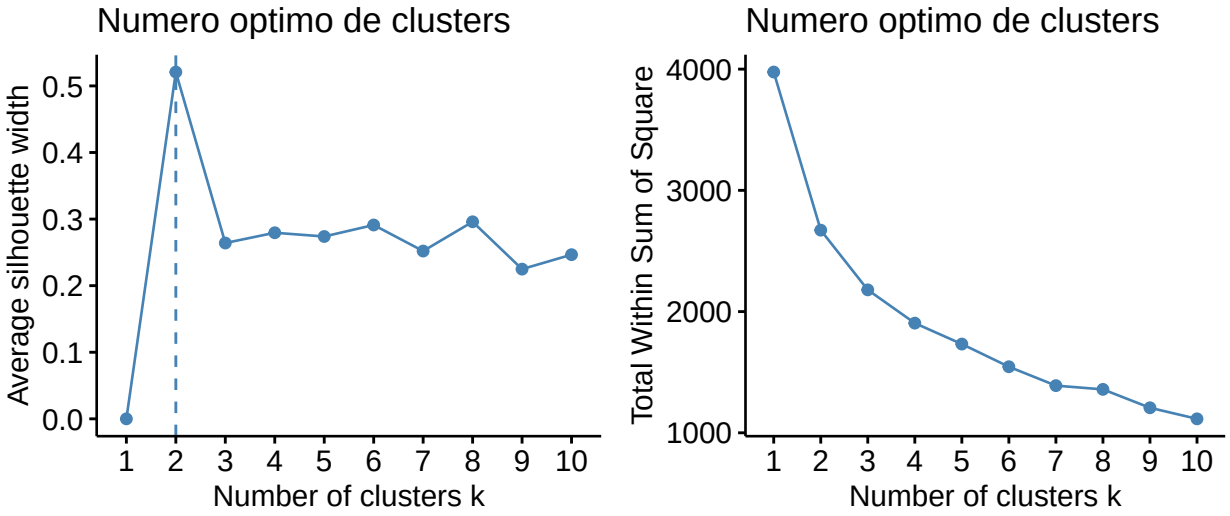
Los resultados del método de la media generan un *grupo dominante* con 475 bodegas y cinco grupos muy pequeños, algunos casi residuales. Esto indica que el algoritmo apenas encuentra diferencias claras y termina forzando clusters muy desequilibrados, señal de que este método no separa bien los perfiles económicos de nuestra muestra. Probablemente la estructura real sea más rica y el promedio no capte bien la variabilidad interna. Por eso, estos resultados respaldan que **Ward** ofrece una segmentación más útil y coherente.

---

8 Métodos de partición

8.1 K-medias

Antes de aplicar el método de k-medias, procedemos a determinar el número de clusters (gráfico 10 del 'Anexo').



Aunque los métodos automáticos señalan que el número óptimo de clusters es 2, para el análisis nos interesa trabajar con $\mathbf{k = 6}$ porque permite distinguir mejor los distintos perfiles económicos de las bodegas. Con más grupos podemos identificar diferencias en tamaño, rentabilidad o volumen de actividad que quedarían

demasiado simplificadas con solo dos clusters. En este caso, k = 6 proporciona una segmentación más útil para interpretar la realidad del sector y tomar decisiones más ajustadas.

De esta manera, aplicamos, el método de k-medias con 6 clusters.

```
##
##   1   2   3   4   5   6
## 234  15  81  47  98  23
```

8.2 K-medoides

Probaremos como última opción el método de los k-medoides que, en teoría, sería más robusto frente a los valores atípicos.

En este algoritmo también es necesario determinar a priori el número de clusters (gráfico 11 del 'Anexo').

Como conclusión del apartado presente, determinamos que según ambos criterios, podemos fijar en *6* el número de clusters óptimo para nuestro estudio.

```
##
##   1   2   3   4   5   6
## 202  63  46  75  94  18
```

---

9 Selección y validación del método de clustering

Teniendo en cuenta los resultados anteriores, procedemos a utilizar el criterio del coeficiente de Silhouette y la variabilidad intra-cluster para validar los resultados y decidir con cuáles nos quedamos.

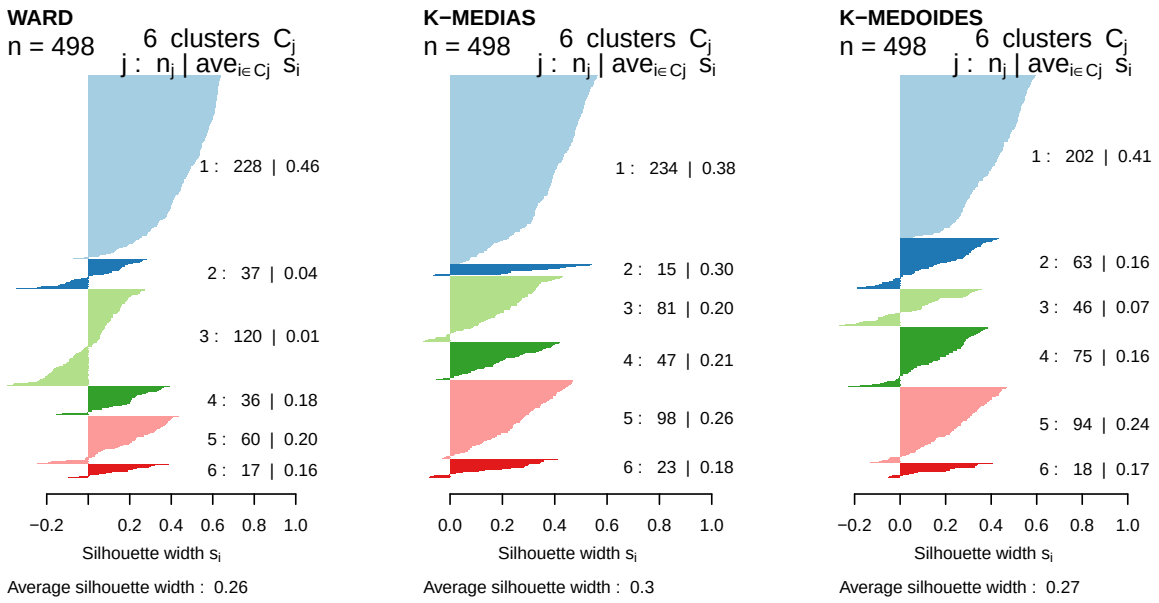


Figure 1: Comparación de métodos

En nuestro caso, el gráfico de **k-medias** es el que muestra una estructura más equilibrada: mantiene un *silhouette* cercano a **0,3**, con menos observaciones en la zona negativa y una forma más limpia que facilita interpretar los seis grupos. El método de **k-medoides** sería la segunda mejor opción, aunque presenta más dispersión dentro de algunos clusters. Por eso optamos por **k-medias**, que ofrece una segmentación más estable y útil para el análisis económico de las bodegas.

Asimismo, decidimos observar a continuación otros métodos de validación del clustering (gráfico 12 del 'Anexo').

```
##                     Score       Method Clusters
## APN            0.14008434       kmeans        6
## AD             2.04599565       kmeans        9
## ADM            0.38352706       kmeans        6
## FOM            0.71610473          pam        9
## Connectivity 136.04246032 hierarchical        6
## Dunn           0.02391116       kmeans        8
## Silhouette     0.28793276       kmeans        6
```

De esta manera, observamos que **k-medias** es el método que eligen la mayoría de los criterios. El número de clusters no parece estar tan claro, aunque es de *k=6* en la mayoría de los citerios (4/7), por lo que usaremos 6 clusters para el análisis.

---

10 Interpretación de los resultados del clustering

Una vez seleccionado el método y número de clusters que mejor parece funcionar en nuestros datos; en nuestro caso, *k-medias* y *k=6* clusters, procedemos con la interpretación o caracterización de los clusters a partir de las variables utilizadas para crearlos. Finalmente, estudiaremos la relación de los clusters con otras variables que no han intervenido en la definición de los clusters.

10.1 PCA

Para facilitar la interpretación de los clusters se utilizó el **Análisis de Componentes Principales (PCA)** [3].

En primer lugar, vamos a utilizar el PCA para ver cuáles de las variables utilizadas en el análisis clustering han contribuido más a la determinación de los clusters obtenidos mediante k-medias. Utilizaremos los clusters como variable suplementaria para poder después colorear las observaciones según el cluster al que pertenecen (gráfico 13 del 'Anexo').

Pese a que soy conocedor de que sería suficiente elegir 3 componentes principales, exploraré las 4 primeras por si aportan información adicional, ya que es posible representar la cuarta sin incrementar el número de gráficos.

En el plano principal (Dim1 y Dim2) se ve que la primera dimensión separa sobre todo por **tamaño económico**, ya que variables como *ingresos*, *valor añadido*, *trabajadores* o *activo total* apuntan claramente en esa dirección. Los grupos 2 y 6 se alejan más del centro, lo que encaja con bodegas de mayor escala o con resultados más extremos. La segunda dimensión recoge diferencias ligadas a *rentabilidad* (ROE y ROA), que ayudan a distinguir a algunos grupos más dispersos. En los ejes 3 y 4 aparecen matices secundarios, con menos peso, pero que confirman que resultado del ejercicio y activos siguen marcando diferencias entre clusters. En conjunto, el PCA respalda la segmentación: los grupos se ordenan principalmente por tamaño y, en segundo lugar, por rentabilidad.

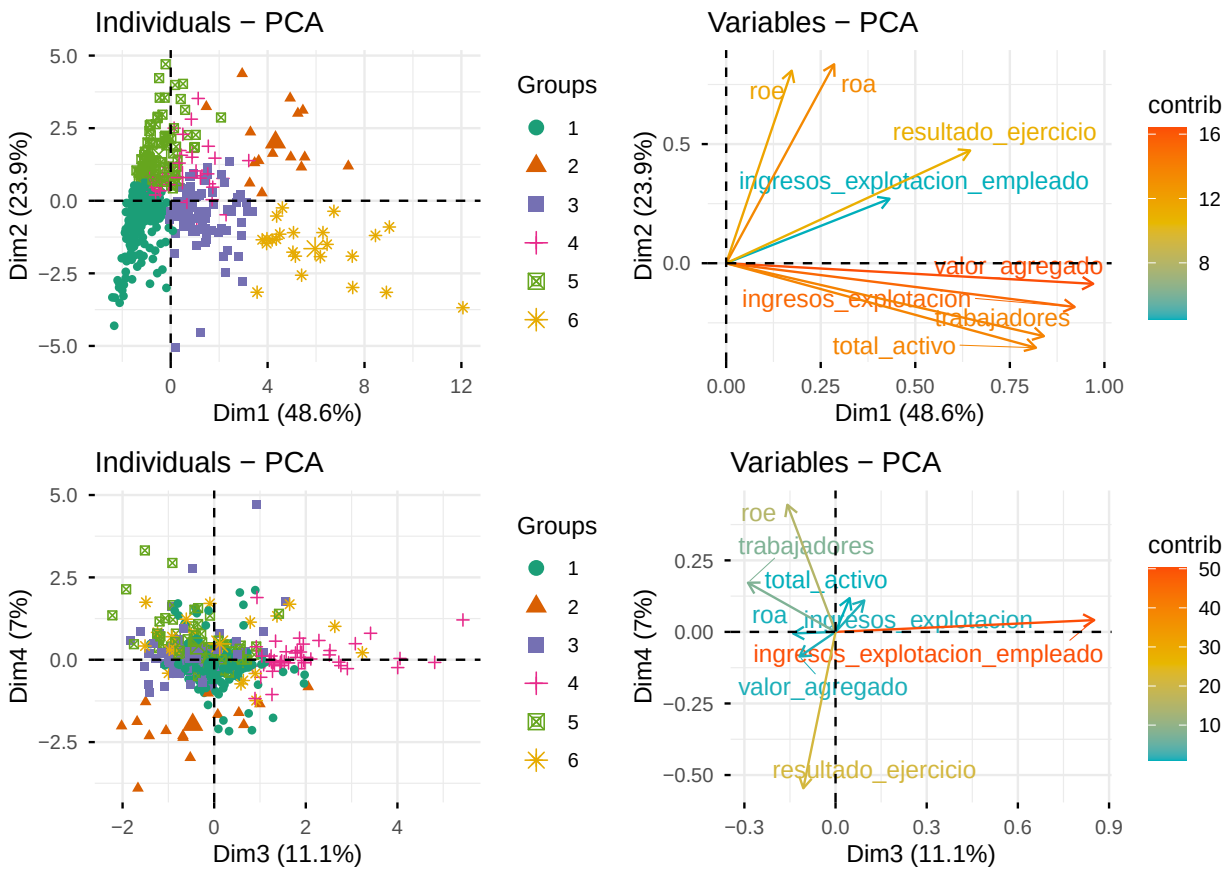


Figure 2: PCA para clustering

10.2 Gráfico de perfiles

El PCA nos ha generado hipótesis sobre las características de cada cluster de bodegas, más concretamente, las características económicas. Vamos a generar ahora un gráfico alternativo que nos puede ayudar también a caracterizar cada cluster y/o corroborar lo observado en el PCA. Se trata de representar el perfil medio de cada cluster para observar las diferencias entre ellos. Para ello, calcularé en primer lugar la media de cada variable para cada cluster y después representaremos estos perfiles medios en una única gráfica.

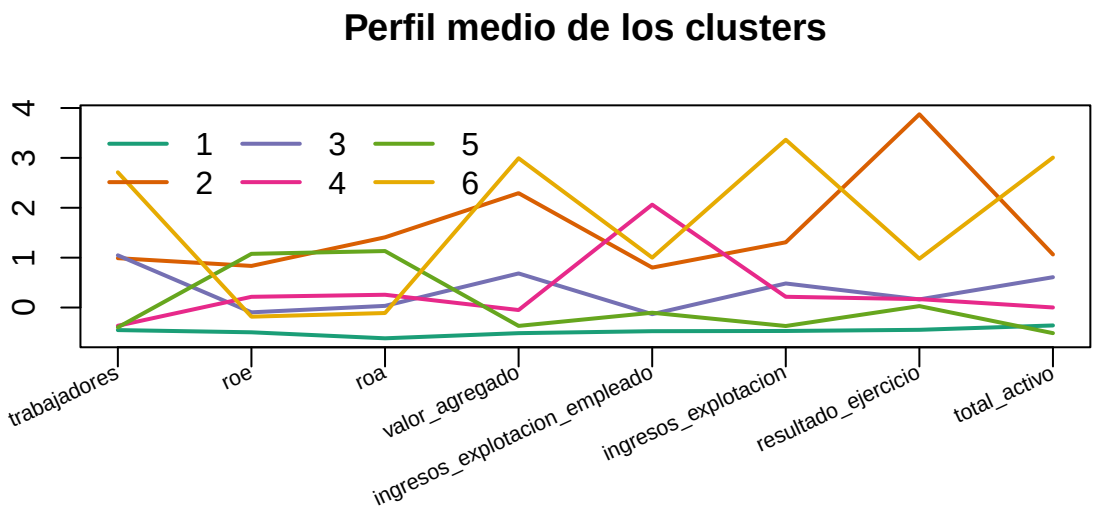


Figure 3: Gráfico de perfiles medios

El gráfico de perfiles medios confirma la existencia de seis segmentos bien diferenciados:

- **Cluster 1**: bodegas pequeñas, con bajos niveles de ingresos, activos y rentabilidad.
- **Clusters 2 y 6**: grandes bodegas con estructuras económicas complejas y elevado volumen de actividad.
- **Cluster 4**: bodegas con alta eficiencia y rentabilidad relativa.
- **Clusters 3 y 5**: perfiles intermedios con combinaciones específicas de tamaño y rentabilidad.

Esta segmentación refleja la diversidad estructural del sector vitivinícola y valida la utilidad del clustering aplicado.

10.3 Relación de los clusters con otras variables

Con el PCA y el gráfico de perfiles hemos caracterizado cada cluster, entendiendo qué variables han contribuido más a definir cada cluster y cómo lo han hecho.

Para finalizar, trataremos de encontrar la relación de los clusters con otras variables no utilizadas el hacer el clusterir.

Comenzaremos con la variable valoracion. Además, como esta variable es continua, la representaremos mediante un gráfico de cajas y bigotes y también haremos un posterior **ANOVA**.

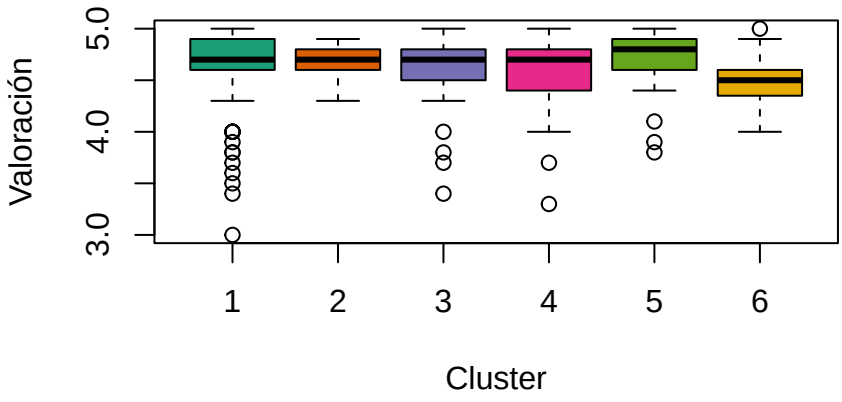


Figure 4: Boxplot de Valoración y clustering

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## misclust      5   1.01 0.20151   2.121 0.0625 .
## Residuals   344  32.69 0.09502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 148 observations deleted due to missingness

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = datos_bodegas$valoracion ~ misclust)
##
## $misclust
##             diff         lwr          upr     p adj
## 2-1  0.01375000 -0.27420680  0.301706795 0.9999933
## 3-1 -0.03458333 -0.16831633  0.099149666 0.9766326
## 4-1 -0.06314655 -0.24143975  0.115146646 0.9127776
## 5-1  0.06180556 -0.06356069  0.187171800 0.7192379
## 6-1 -0.17203947 -0.38640384  0.042324889 0.1968292
## 3-2 -0.04833333 -0.35007568  0.253409009 0.9974352
## 4-2 -0.07689655 -0.40086011  0.247067011 0.9840445
## 5-2  0.04805556 -0.25007298  0.346184087 0.9973572
## 6-2 -0.18578947 -0.53092134  0.159342393 0.6368141
## 4-3 -0.02856322 -0.22835759  0.171231150 0.9985124
## 5-3  0.09638889 -0.05803268  0.250810456 0.4743135
## 6-3 -0.13745614 -0.37001023  0.095097950 0.5367332
## 5-4  0.12495211 -0.06934141  0.319245621 0.4394880
## 6-4 -0.10889292 -0.36963293  0.151847090 0.8382297
## 6-5 -0.23384503 -0.46169056 -0.005999501 0.0404509
```

Por último, procederemos a observar si los clusters tienen relación con la variable *esta_trip*.

```
##    misclust
```

```
##         1     2     3     4     5     6
##   0 79.49 80.00 62.96 87.23 75.51 47.83
##   1 20.51 20.00 37.04 12.77 24.49 52.17
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table(datos_bodegas$esta_trip, misclust)
## X-squared = 21.949, df = NA, p-value = 0.002499
```

Con el *Pearson's Chi-squared test* podemos observar contrastes fuertes (por ejemplo, el cluster 6 tiene más de un 50% de bodegas en esta_trip, mientras que el 4 apenas llega al 13%). El χ^2 confirma esta intuición: con un p-valor de 0.0025 podemos decir que la distribución no es aleatoria y que los clusters sí guardan relación con esta_trip. En la práctica, esto sugiere que ciertos perfiles económicos están más presentes dentro de esta_trip, mientras que otros aparecen sobre todo fuera, lo que da pie a interpretar diferencias estratégicas entre grupos.

11 Conclusiones y aplicaciones prácticas

Este proyecto demuestra que el uso de técnicas de clustering permite identificar perfiles económicos claros y consistentes dentro del sector vitivinícola español. La segmentación obtenida aporta una visión estructurada que va más allá de simples clasificaciones por tamaño.

Los resultados pueden tener aplicaciones prácticas en:

- Análisis financiero y de riesgo.
- Diseño de políticas públicas sectoriales.
- Segmentación de clientes y benchmarking.
- Estrategias de digitalización diferenciadas.

Desde un punto de vista metodológico, el trabajo evidencia la importancia de combinar validación estadística, interpretabilidad económica y rigor en el preprocesado, elementos clave en cualquier proyecto de análisis de datos aplicado.

12 Anexo metodológico

12.1 Lectura de datos inicial y selección y escalado de variables

```
datos_bodegas <- read_csv("datos/datos_bodegas_limpio_FINAL.csv")
```

```
datos_bodegas_economicos = datos_bodegas[,5:12]
datos_bodegas_economicos = scale(datos_bodegas_economicos, center = TRUE, scale = TRUE)
```

12.2 Medida de distancia y tendencia de agrupamiento

```r
## Euclídea
midist <- get_dist(datos_bodegas_economicos, stand = FALSE, method = "euclidean")
fviz_dist(midist,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"),
          show_labels = TRUE, lab_size = 5)
```

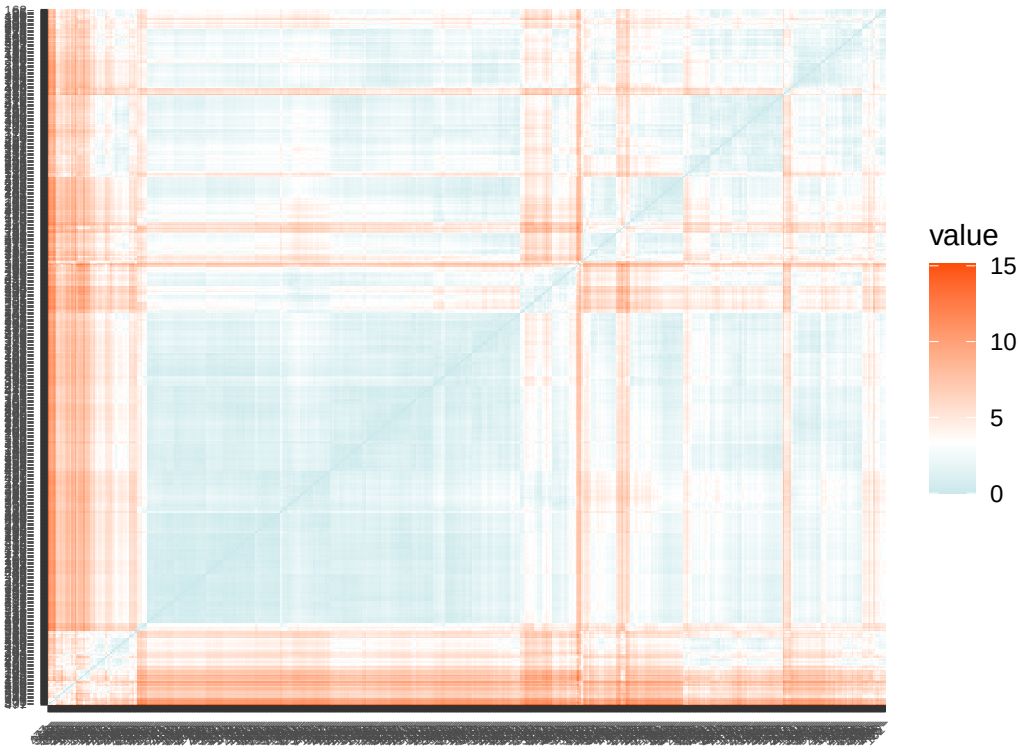


Figure 5: Distancia euclídea

```r
## Manhattan
midistM <- get_dist(datos_bodegas_economicos, stand = FALSE, method = "manhattan")
fviz_dist(midistM,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"),
          show_labels = TRUE, lab_size = 5)
```

12.3 Métodos jerárquicos

12.3.1 Método de Ward

```r
fviz_dend(clust1, k=6, cex = 0.4)
```

```r
library(gridExtra)
p1= fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = hcut, hc_method = "ward.D2",
             method = "silhouette",
```

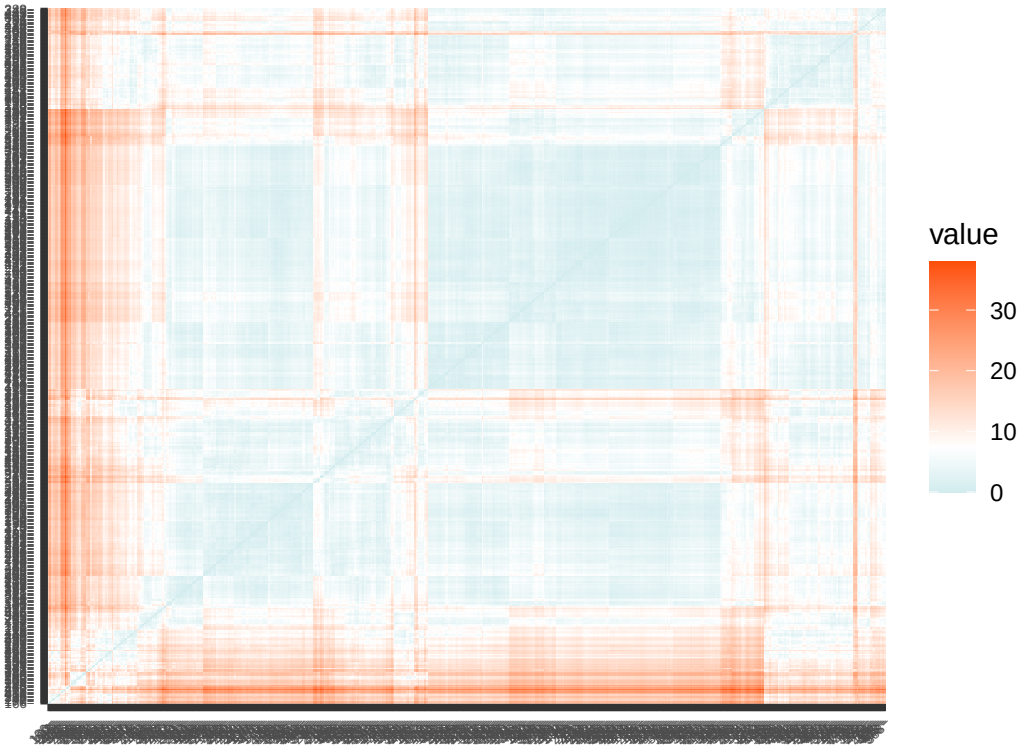


Figure 6: Distancia de Manhattan

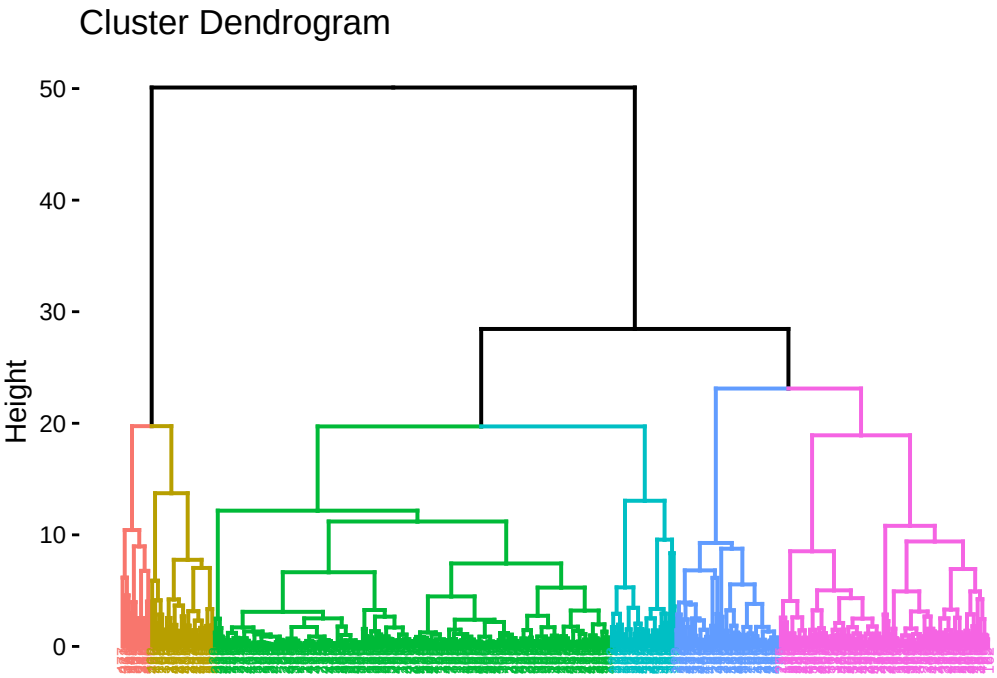


Figure 7: Dendrograma

```
              k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
p2 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = hcut, hc_method = "ward.D2",
              method = "wss",
              k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
grid.arrange(p1, p2, nrow = 1)
```

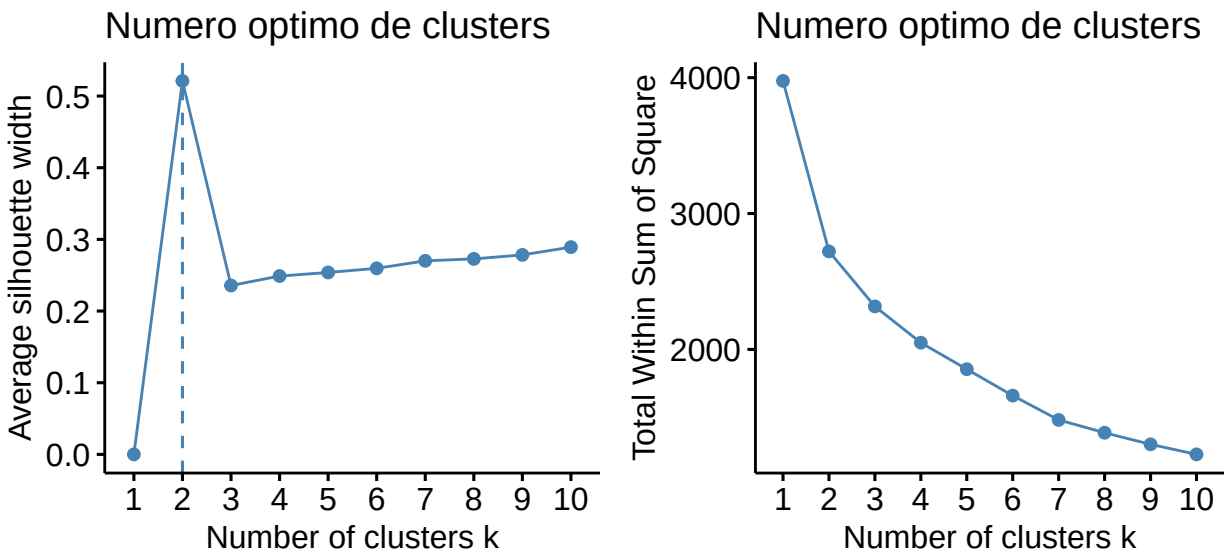


Figure 8: Método Ward

12.3.2 Método de la media

```
p1= fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = hcut, hc_method = "average",
              method = "silhouette",
              k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
p2 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = hcut, hc_method = "average",
              method = "wss",
              k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
grid.arrange(p1, p2, nrow = 1)
```

12.4 Métodos de partición

12.4.1 K-medias

```
p1 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = kmeans, method = "silhouette",
              k.max = 10, verbose = FALSE) +
```

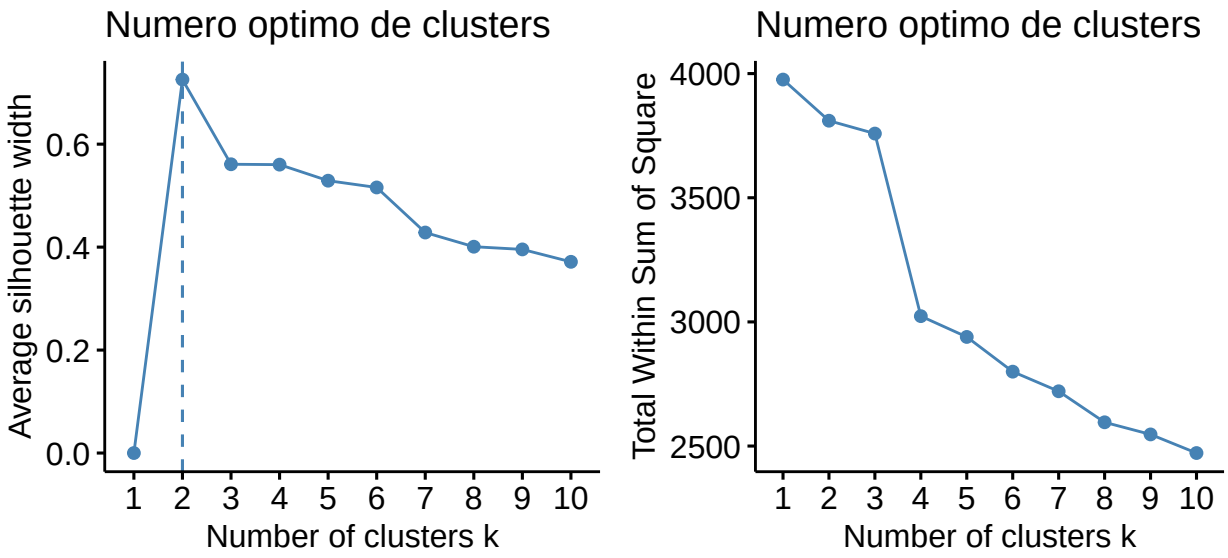


Figure 9: Método de la media

```
  labs(title = "Numero optimo de clusters")
p2 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = kmeans, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
grid.arrange(p1, p2, nrow = 1)
```

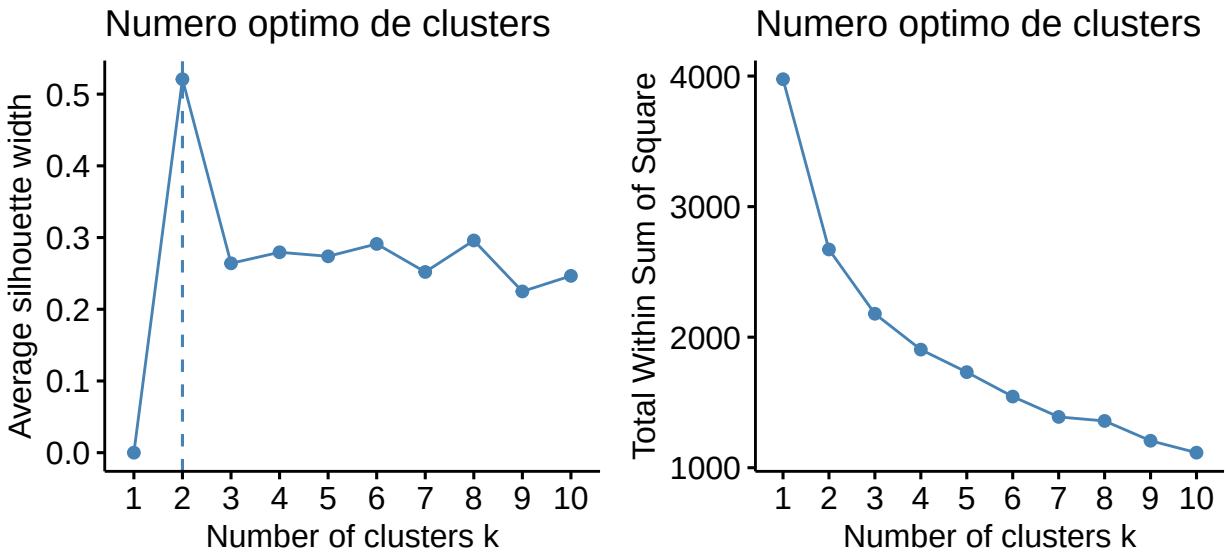


Figure 10: Método de K-medias

12.4.2 K-medoides

```
p1 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = pam, method = "silhouette",
             k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
p2 = fviz_nbclust(x = datos_bodegas_economicos, FUNcluster = pam, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Numero optimo de clusters")
grid.arrange(p1, p2, nrow = 1)
```

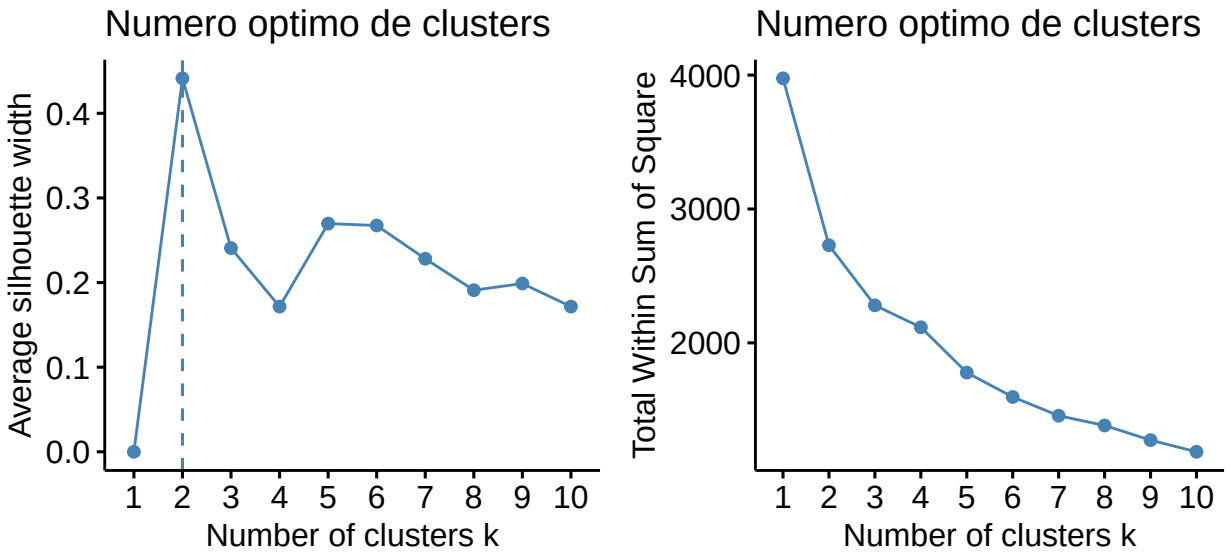


Figure 11: Método de K-medoides

12.5 Selección del método clustering

12.5.1 Comparación de métodos según criterio 'Silhouette'

```
par(mfrow = c(1,3))
library(RColorBrewer)
colores = brewer.pal(6, name = "Paired")
plot(silhouette(grupos1, midist), col=colores[1:6], border=NA, main = "WARD")
plot(silhouette(clust3$cluster, midist), col=colores[1:6], border=NA, main = "K-MEDIAS")
plot(silhouette(clust4$clustering, midist), col=colores, border=NA, main = "K-MEDOIDES")
```

12.5.2 Selección de número de clusters y métodos según diferentes criterios

```
metodos = c("hierarchical","kmeans","pam")
validacion = suppressMessages(clValid(datos_bodegas_economicos, nClust = 6:9, metric = "euclidean",
                      clMethods = metodos,
```

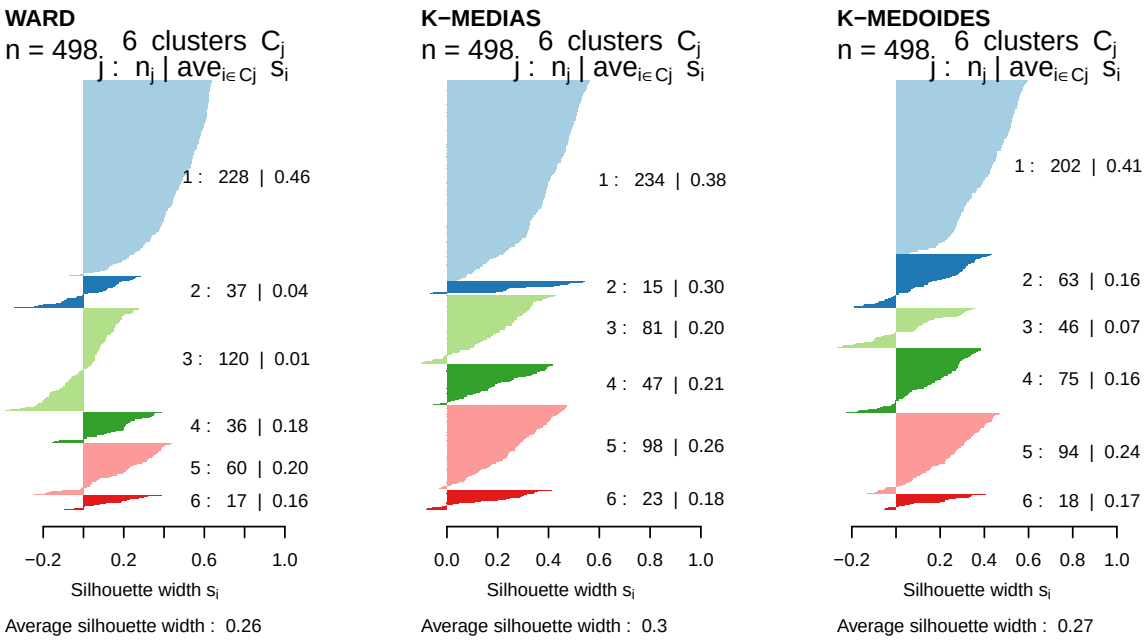


Figure 12: Comparación de métodos

```
                    validation = c("internal", "stability"),
                    method = "ward"))
summary(validacion)
```

```
##
## Clustering Methods:
##  hierarchical kmeans pam
##
## Cluster sizes:
##  6 7 8 9
##
## Validation Measures:
##                                  6        7        8        9
##
## hierarchical APN            0.2766   0.2837   0.3105   0.3252
##              AD             2.3249   2.2779   2.1859   2.1328
##              ADM            0.7414   0.8728   0.8028   0.8035
##              FOM            0.7709   0.7563   0.7501   0.7492
##              Connectivity 136.0425 153.0012 158.3980 160.8746
##              Dunn           0.0173   0.0173   0.0174   0.0174
##              Silhouette     0.2083   0.2057   0.2134   0.2085
## kmeans       APN            0.1401   0.2127   0.1973   0.2349
##              AD             2.1696   2.1835   2.0674   2.0460
##              ADM            0.3835   0.7216   0.5521   0.6413
##              FOM            0.7491   0.7378   0.7383   0.7380
##              Connectivity 163.3948 173.5960 172.7147 183.2750
##              Dunn           0.0160   0.0208   0.0239   0.0183
##              Silhouette     0.2879   0.2352   0.2473   0.2506
## pam          APN            0.3003   0.2670   0.3158   0.3257
```

```
##              AD              2.3128   2.1791   2.1524   2.0715
##              ADM             0.8066   0.6645   0.7794   0.7242
##              FOM             0.7518   0.7445   0.7387   0.7161
##              Connectivity  186.9218 210.2528 217.8492 232.8758
##              Dunn            0.0238   0.0124   0.0124   0.0131
##              Silhouette      0.2674   0.2281   0.1910   0.1989
##
## Optimal Scores:
##
##              Score    Method       Clusters
## APN            0.1401 kmeans       6
## AD             2.0460 kmeans       9
## ADM            0.3835 kmeans       6
## FOM            0.7161 pam          9
## Connectivity 136.0425 hierarchical 6
## Dunn           0.0239 kmeans       8
## Silhouette     0.2879 kmeans       6
```

12.6 PCA

12.6.1 Selección de componentes principales

```
misclust = factor(clust3$cluster)
miPCA = PCA(datos_bodegas_economicos, scale.unit = FALSE, graph = FALSE)
eig.val <- get_eigenvalue(miPCA)
VPmedio = 100 * (1/nrow(eig.val))
fviz_eig(miPCA, addlabels = TRUE) +
  geom_hline(yintercept=VPmedio, linetype=2, color="red")
```

12.6.2 Gráfico de componentes principales

```
misclust = factor(clust3$cluster)
miPCA = PCA(datos_bodegas_economicos, scale.unit = FALSE, graph = FALSE)

colores = brewer.pal(6, name = "Dark2")
p1 = fviz_pca_ind(miPCA, geom = "point", habillage = misclust, addEllipses = FALSE,
                  palette = colores)
p2 = fviz_pca_var(miPCA, repel = TRUE, col.var = "contrib",
                  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
p3 = fviz_pca_ind(miPCA, geom = "point", habillage = misclust, addEllipses = FALSE,
                  axes = 3:4, palette = colores)
p4 = fviz_pca_var(miPCA, axes = 3:4, repel = TRUE, col.var = "contrib",
                  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
grid.arrange(p1,p2,p3,p4, nrow = 2)
```

12.7 Gráfico de perfiles de clusters

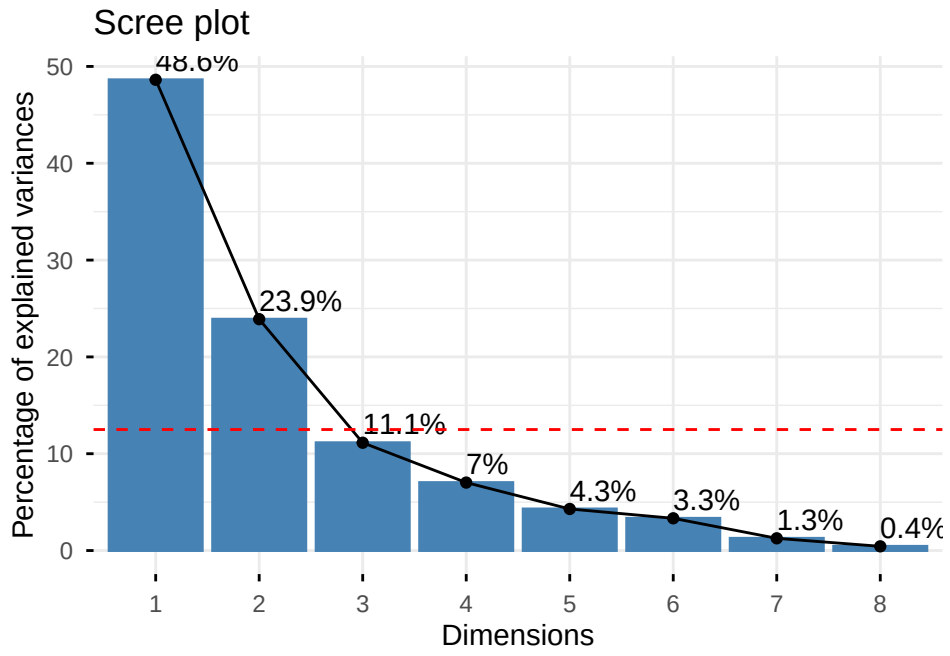


Figure 13: Elección de número de PCA's

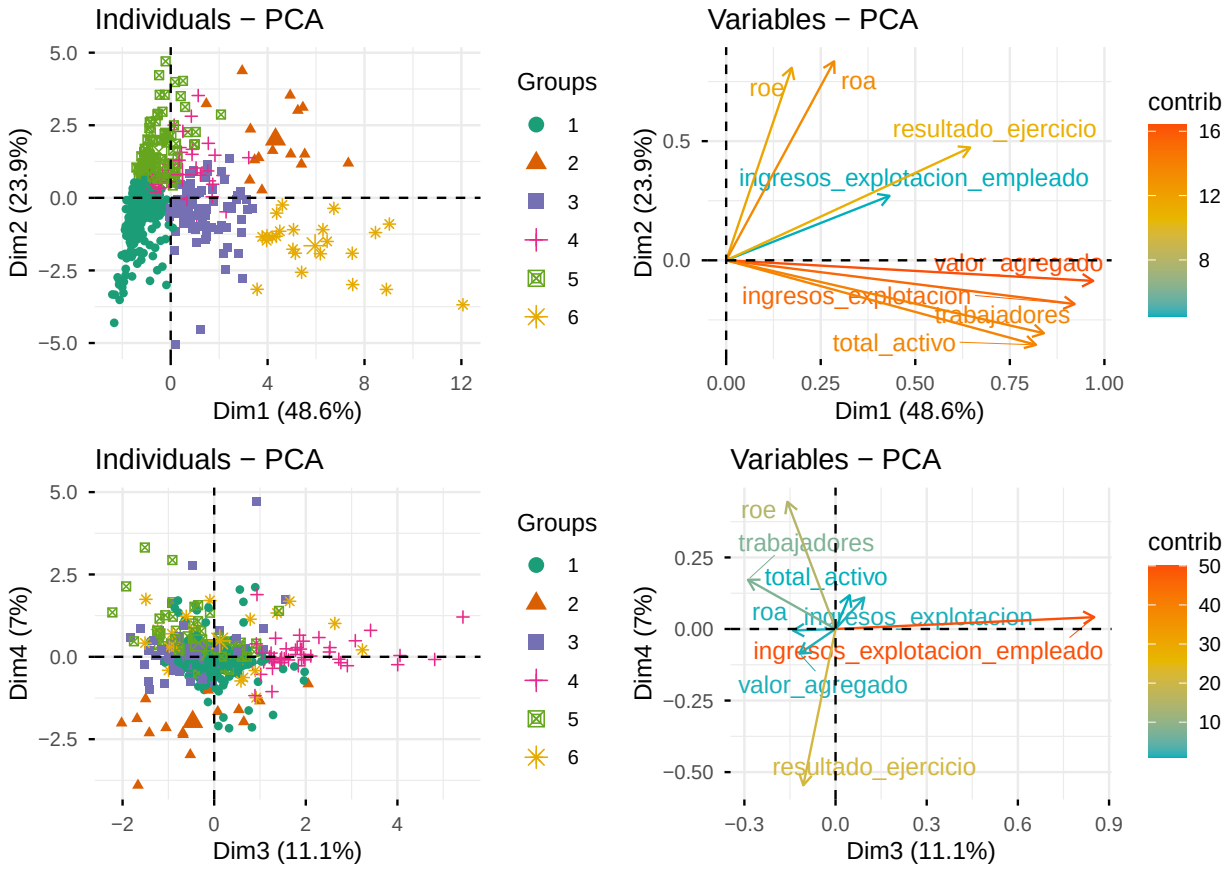


Figure 14: PCA para clustering

```r
mediasCluster = aggregate(datos_bodegas_economicos,
                          by = list("cluster" = misclust), mean)[,-1]
rownames(mediasCluster) = paste0("c",1:6)

matplot(t(mediasCluster), type = "l", col = colores, ylab = "", xlab = "",
        lwd = 2, lty = 1, main = "Perfil medio de los clusters", xaxt = "n")

# Añadir eje SIN etiquetas
axis(1, at = 1:ncol(datos_bodegas_economicos), labels = FALSE)

# Añadir etiquetas en diagonal
text(x = 1:ncol(datos_bodegas_economicos),
     y = par("usr")[3] - 0.5,    # posición bajo del eje
     labels = colnames(datos_bodegas_economicos),
     srt = 25,                   # ROTACIÓN
     adj = 1,                    # alineación
     xpd = TRUE,                 # permitir ir fuera del marco
     cex = 0.7)

legend("topleft", as.character(1:6), col = colores, lwd = 2, ncol = 3, bty = "n")
```

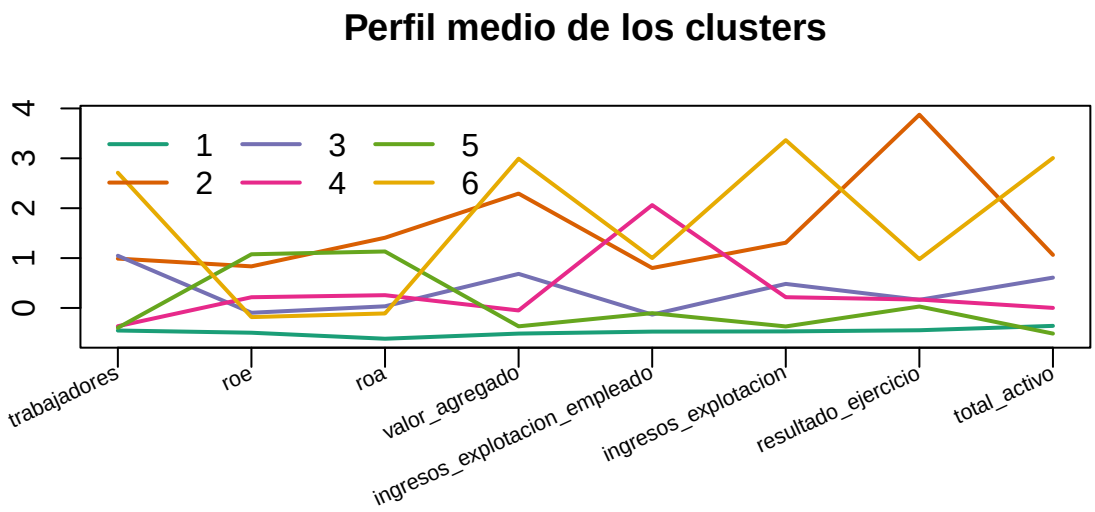


Figure 15: Gráfico de perfiles medios

12.8 Relación de los clusters con otras variables

12.8.1 Relación con variable 'valoración'

```r
boxplot(datos_bodegas$valoracion ~ misclust, col = brewer.pal(6, "Dark2"),
        xlab = "Cluster", ylab = "Valoración")
```

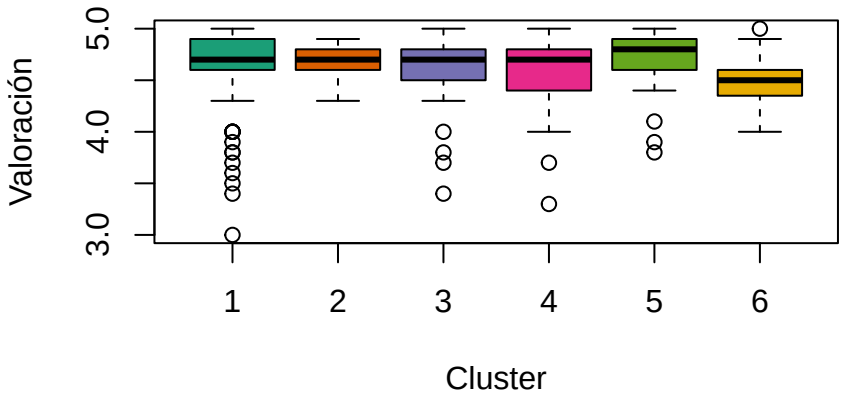


Figure 16: Boxplot de Valoración y clustering

```
mianova = aov(datos_bodegas$valoracion ~ misclust)
summary(mianova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## misclust      5   1.01 0.20151   2.121 0.0625 .
## Residuals   344  32.69 0.09502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 148 observations deleted due to missingness
```

```
TukeyHSD(mianova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = datos_bodegas$valoracion ~ misclust)
##
## $misclust
##            diff         lwr          upr     p adj
## 2-1  0.01375000 -0.27420680  0.301706795 0.9999933
## 3-1 -0.03458333 -0.16831633  0.099149666 0.9766326
## 4-1 -0.06314655 -0.24143975  0.115146646 0.9127776
## 5-1  0.06180556 -0.06356069  0.187171800 0.7192379
## 6-1 -0.17203947 -0.38640384  0.042324889 0.1968292
## 3-2 -0.04833333 -0.35007568  0.253409009 0.9974352
## 4-2 -0.07689655 -0.40086011  0.247067011 0.9840445
## 5-2  0.04805556 -0.25007298  0.346184087 0.9973572
## 6-2 -0.18578947 -0.53092134  0.159342393 0.6368141
## 4-3 -0.02856322 -0.22835759  0.171231150 0.9985124
## 5-3  0.09638889 -0.05803268  0.250810456 0.4743135
## 6-3 -0.13745614 -0.37001023  0.095097950 0.5367332
```

```
## 5-4  0.12495211 -0.06934141  0.319245621 0.4394880
## 6-4 -0.10889292 -0.36963293  0.151847090 0.8382297
## 6-5 -0.23384503 -0.46169056 -0.005999501 0.0404509
```

12.8.2 Relación con variable 'esta_trip'

```
round(100*proportions(table(datos_bodegas$esta_trip, misclust), 2),2)
```

```
##    misclust
##         1     2     3     4     5     6
##   0 79.49 80.00 62.96 87.23 75.51 47.83
##   1 20.51 20.00 37.04 12.77 24.49 52.17
```

```
chisq.test(table(datos_bodegas$esta_trip, misclust), simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table(datos_bodegas$esta_trip, misclust)
## X-squared = 21.949, df = NA, p-value = 0.002499
```

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009.

[2] Brian Hopkins and John G. Skellam. "A new method for determining the type of distribution of plant individuals". In: *Journal of Ecology* 42.1 (1954), pp. 228–236.

[3] Ian T. Jolliffe. "Principal Component Analysis". In: *Springer Series in Statistics* (2002).

[4] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package. 2020. URL: https://cran.r-project.org/package=factoextra.

[5] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

[6] Sébastien Lê, Julie Josse, and François Husson. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*. R package. 2008. URL: http://factominer.free.fr.

[7] Martin Maechler et al. *cluster: Cluster Analysis Basics and Extensions*. R package. 2023.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[9] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.