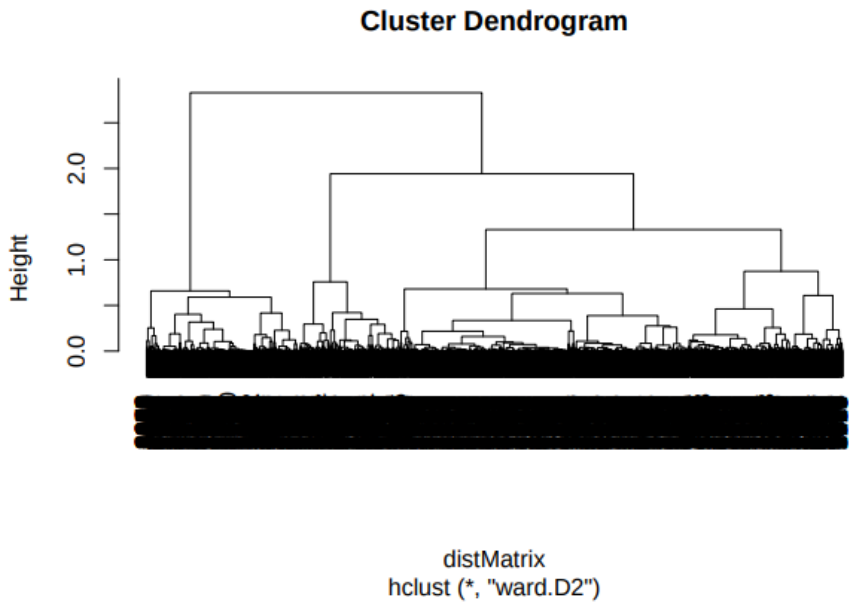
```r
h1 <- hclust(distMatrix,method="ward.D2")  # NOTICE THE COST
#versions noves "ward.D" i abans de plot: par(mar=rep(2,4)) si se quejara de los margenes del plot

plot(h1)
```



Qué muestra:
El dendrograma expone la estructura jerárquica de agrupamiento donde cada "hoja" corresponde a un cliente. Las líneas verticales representan la unión de dos subgrupos, y la altura a la que se unen indica la disimilitud entre ellos (distancia de Gower). El eje vertical cuantifica esa disimilitud, mientras que el eje horizontal muestra todos los individuos ordenados.

Significado:
Este gráfico nos permite visualizar cómo se forman los clusters de forma jerárquica y nos ayuda a determinar el número de grupos óptimo al observar niveles de corte con un incremento drástico de la distancia. Permite identificar un corte natural donde las uniones pasan de ser muy parecidas a mucho más disímiles.

Interpretación:

- Un "salto" grande en la altura de unión indica que estamos forzando a unir grupos muy distintos, sugiriendo un número de clusters menor al nivel de corte.
- Al elegir k = 4, se sitúa una línea horizontal justo antes de un gran aumento de altura, separando 4 grupos con mínima disimilitud interna.
- Cluster muy compactos aparecen como subárboles con un bajo rango de alturas.

3. Gráficos de medias por cluster (barplot)

```r
for (var in nombres_numericos) {
  barplot(tapply(dd[[var]], c2, mean),
          main = paste("Means of", var, "by Cluster"))
  abline(h = mean(dd[[var]]), col = "red")
}
```

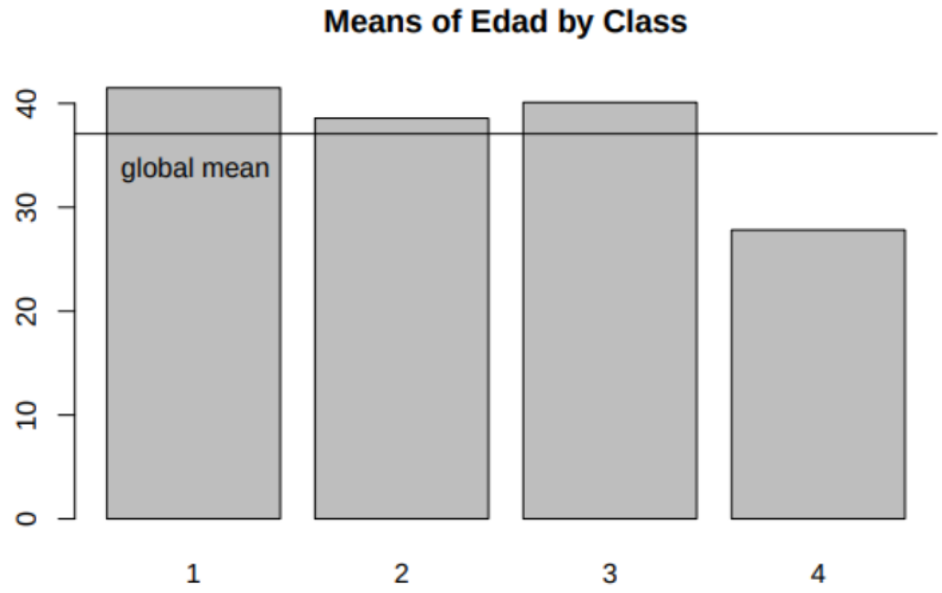
(Snippet: barplot + abline del promedio global)

Qué muestra:
Un barplot de la media de cada variable numérica para cada cluster (4 barras), acompañado de una línea horizontal roja que indica el valor medio global de esa variable sobre todo el dataset.

Significado:
Resalta qué clusters están por encima o por debajo del promedio general de cada variable, facilitando la detección de segmentos con valores extremos respecto al conjunto total.

Interpretación:

- Barras muy por encima de la línea roja marcan clusters "destacados" en esa variable (e.g., cluster 2 con gasto medio muy superior).
- Barras por debajo indican segmentos con menor actividad (e.g. cluster 4 con ahorro medio bajo).
- Ayuda a priorizar variables clave: si Ingresos muestra grandes diferencias de media entre clusters, es una variable discriminante.



4. **Matriz de dispersión coloreada por cluster (pairs)**

```
pairs(dcon[, 1:7], col = c2)
```

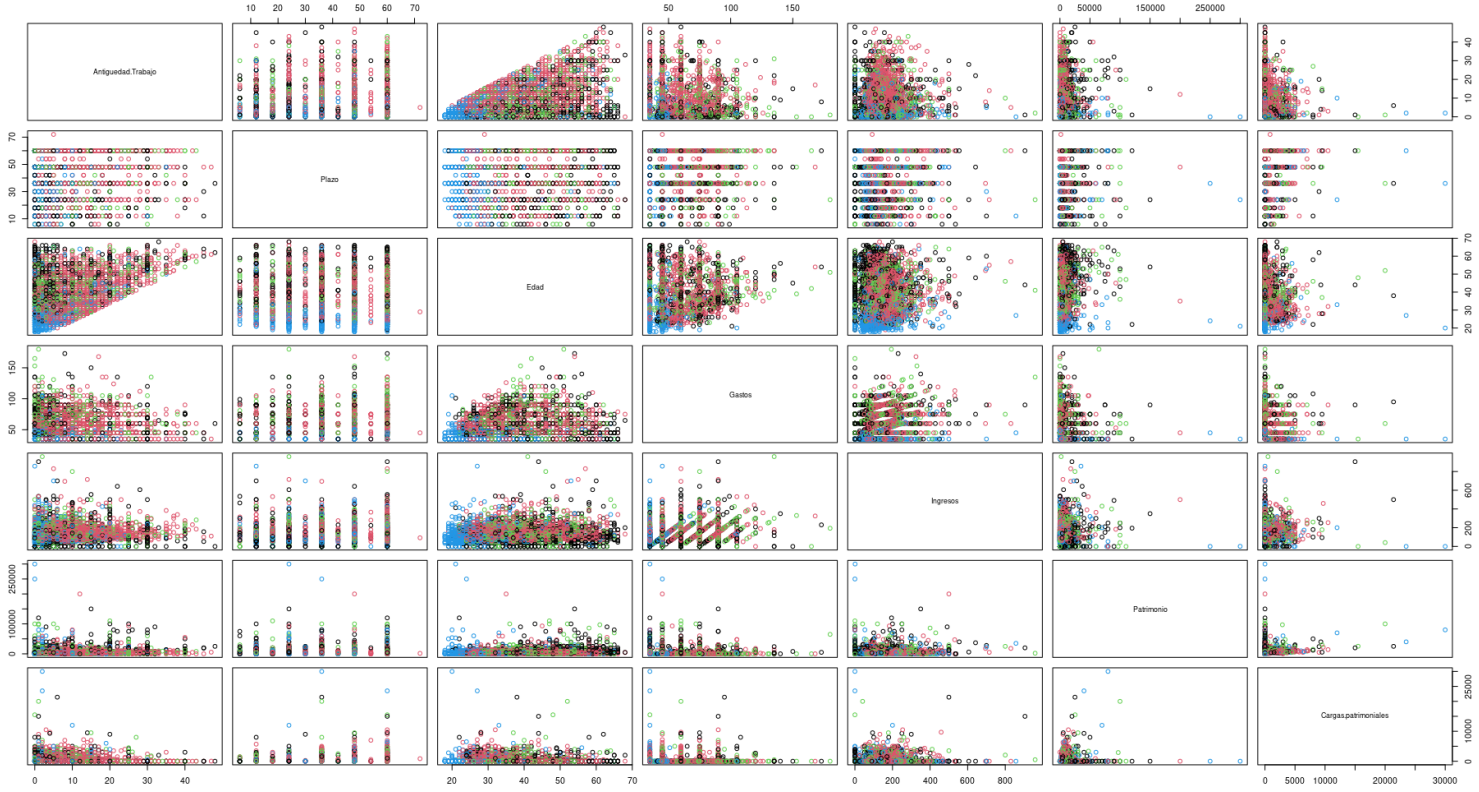
(Snippet: pairs de primeras 7 variables numéricas coloreado)

Qué muestra:
Grilla de scatterplots de cada par de variables numéricas (Edad vs Ingresos, Edad vs Gastos, etc.), donde cada punto está coloreado según el cluster al que pertenece.

Significado:
Permite observar relaciones bivariadas y comprobar la separabilidad de clusters en diferentes combinaciones de variables. Se aprecia la forma y densidad de cada grupo en el espacio multidimensional proyectado a dos dimensiones.

Interpretación:

- Paneles donde clusters aparecen muy bien agrupados y separados sugieren que esas variables combinadas discriminan eficazmente.
- Solapamientos indican variables con menor poder de separación.
- Direcciones de correlación: si Ingresos y Gastos muestran una nube con pendiente positiva, confirma relación directa entre ambas.



```r
plot(RatiFin,Estalvi,col=c1,main="Clustering of credit data in 3 classes")
legend("topright",c("class1","class2","class3"),pch=1,col=c(1:3), cex=0.6)
```

5. **Scatterplot de RatioFin vs Ahorro**

```r
plot(RatiFin, Estalvi, col = c2,
     main = "Clustering of credit data in 4 clusters")
```

(Snippet: plot de RatioFin vs Estalvi coloreado)

Qué muestra:

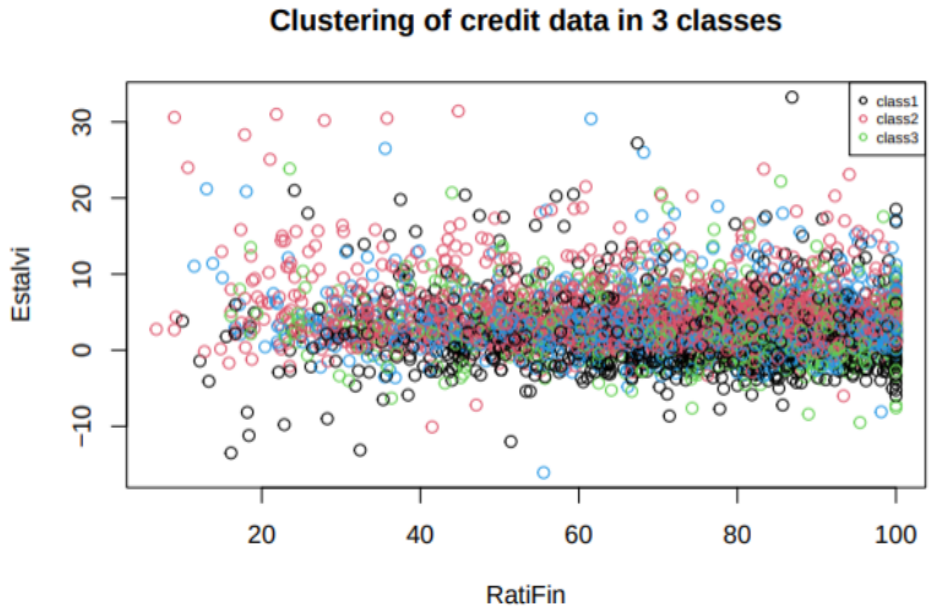
Cada punto es un cliente posicionado según su RatioFin (eje X) y su nivel de Ahorro (eje Y), coloreado por cluster.

Significado:

Visualiza la relación específica entre dos variables clave y cómo se distribuyen los clusters en este plano. Identifica patrones de comportamiento financiero.

Interpretación:

- Clusters con RatioFin alto y Ahorro bajo se ubican en la zona inferior derecha, marcando clientes con alta carga financiera y poco ahorro.
- Segmentos con puntos agrupados en la esquina superior izquierda (bajo ratio–alto ahorro) representan perfiles conservadores.
- Superposición mínima indica fuerte diferenciación en este par de variables.



7. **Gráficas de proporción de variables cualitativas**

```r
marg <- prop.table(table(c2, dd[[cat]]), 1)
plot(marg[1, ], type = "l", ylim = c(0, 1),
     main = paste("Prop. by cluster of", cat))
for (j in 1:ncol(marg)) lines(marg[, j], lty = j)
legend("topright", colnames(marg), lty = 1:ncol(marg))
```

(Snippet: plot + lines de proporciones)

Qué muestra:

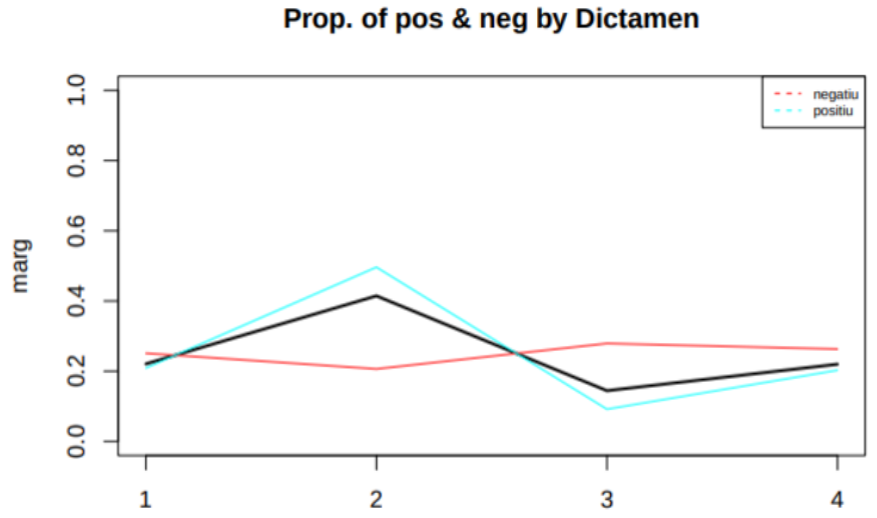
Para cada categoría de la variable cualitativa (Vivienda, Estado.civil, Tipo.trabajo...), se dibuja su proporción relativa dentro de cada cluster, normalizada a 1 en total por cluster.

Significado:

Revela la composición interna de cada cluster en términos porcentuales, permitiendo comparar la distribución de categorías independientemente del tamaño del cluster.

Interpretación:

- Líneas ascendentes o descendentes indican categorías más o menos representadas al pasar de un cluster a otro.
- Picos en la proporción de una categoría muestran clusters donde esa característica es dominante.
- Útil para entender perfiles cualitativos (e.g. clusters con mayor proporción de solteros vs casados).




```
## [1] "Categories=" "negatiu"     "positiu"
```

2. **Boxplots de variables numéricas por cluster**

```
for (var in nombres_numericos) {
  boxplot(dd[[var]] ~ c2, horizontal = TRUE,
          main = paste("Boxplot of", var, "vs Cluster"))
}
```

(Snippet: boxplot en bucle)

Qué muestra:

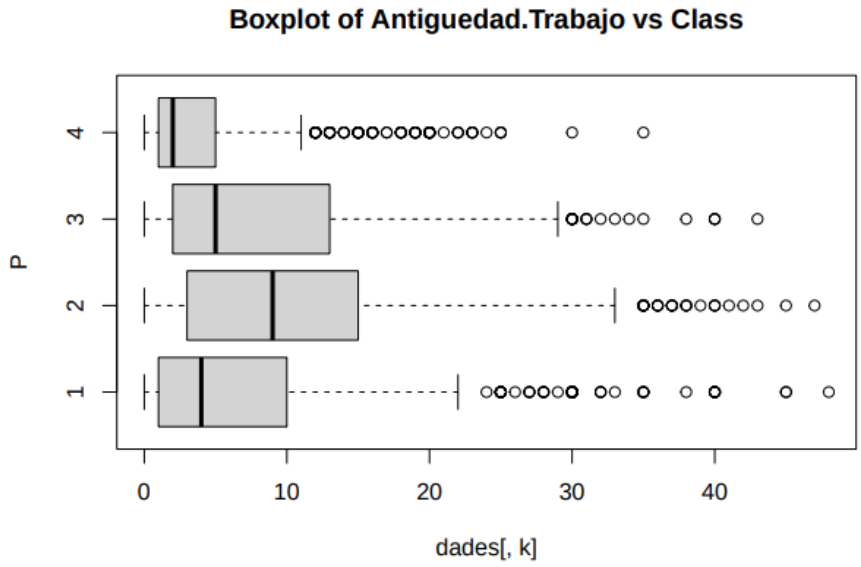
Para cada variable numérica (Edad, Ingresos, Gastos, Patrimonio, Ahorro, RatioFin, etc.), se dibujan cajas que representan la mediana, los cuartiles (Q1 y Q3) y los "bigotes" que cubren 1.5 veces el rango intercuartílico, con posibles puntos aislados como outliers, por cada uno de los 4 clusters.

Significado:

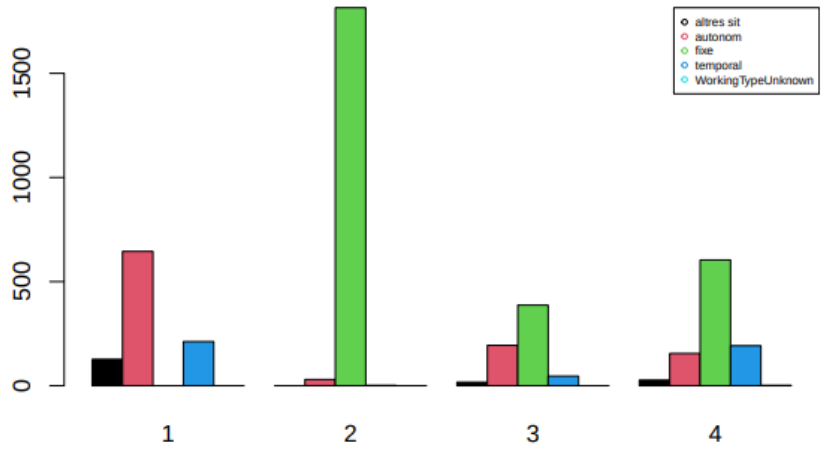
Los boxplots permiten comparar la distribución completa de cada variable entre clusters, destacando diferencias en mediana, dispersión y la presencia de valores atípicos. Es útil para detectar clusters con heterogeneidad alta o baja, y centrarse en variables que mejor distinguen los segmentos.

Interpretación:

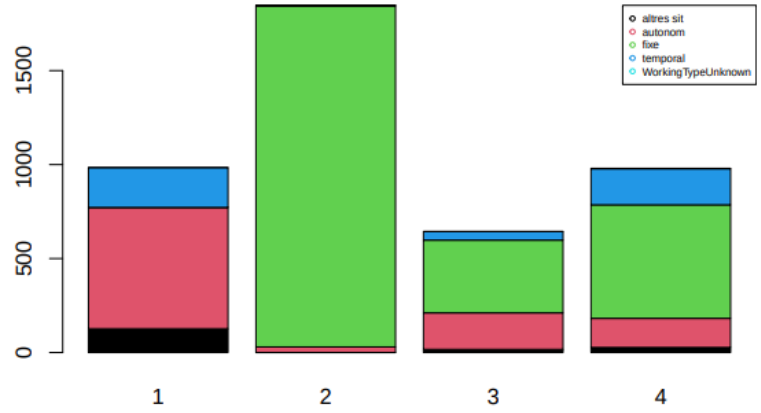
- Un cluster con mediana de Ingresos alta y rangos estrechos indica un grupo homogéneo de clientes con altos ingresos.
- Si en Gastos un cluster presenta varios outliers hacia valores muy altos, pueden identificarse clientes excepcionales con gastos extraordinarios.
- Comparar la posición de las cajas: si la caja de Patrimonio del cluster 3 está completamente por encima de las de otros clusters, revela que ese segmento posee activos sustancialmente superiores.




```
barplot(table(Tipo.trabajo, c2), beside=TRUE,col=c(1:length(levels(Tipo.trabajo))) )
legend("topright",levels(Tipo.trabajo),pch=1,cex=0.5, col=c(1:length(levels(Tipo.trabajo))))
```




```
barplot(table(Tipo.trabajo, c2), beside=FALSE,col=c(1:length(levels(Tipo.trabajo))) )
legend("topright",levels(Tipo.trabajo),pch=1,cex=0.5, col=c(1:length(levels(Tipo.trabajo))))
```



1. **Análisis de Componentes Principales (PCA) – Scatterplot de individuos**

```r
# Selección de ejes
# eje1 <- 1; eje2 <- 2
plot(Psi[, eje1], Psi[, eje2], type = "n")
# Configuración de ejes
axis(side = 1, pos = 0, labels = FALSE, col = "cyan")
axis(side = 3, pos = 0, labels = FALSE, col = "cyan")
axis(side = 2, pos = 0, labels = FALSE, col = "cyan")
axis(side = 4, pos = 0, labels = FALSE, col = "cyan")
# Dibujar etiquetas de individuos
text(Psi[, eje1], Psi[, eje2], labels = iden, cex = 0.5)
```

(Snippet: plot de individuos cite turn3file18)

Qué muestra:

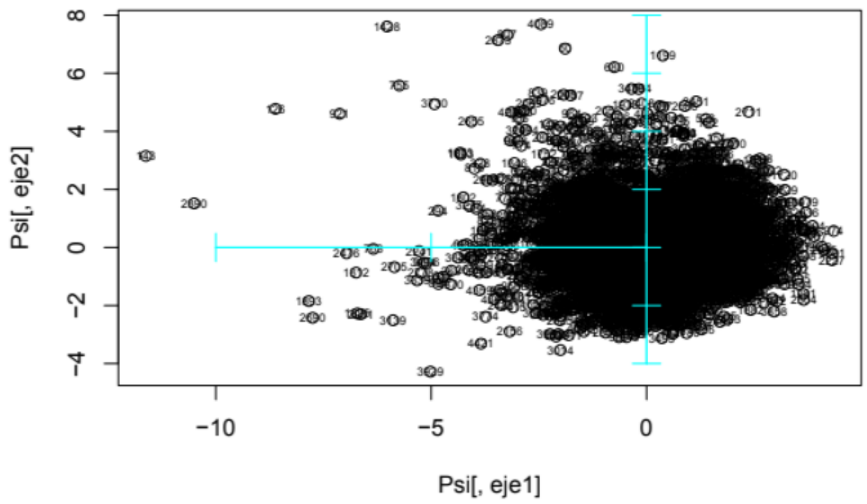
Este gráfico traza cada observación de tu dataset en un espacio bidimensional definido por el primer y segundo componente principal (PC1 vs PC2). Cada punto está etiquetado con su identificador original (`iden`), lo cual facilita rastrear casos particulares. Los ejes cruzan en el origen (0,0), representando el punto de media de todas las variables tras la transformación. No se muestran ejes numéricos para centrar la atención en la distribución geométrica.

Significado:

La PCA reduce la dimensionalidad del dataset conservando la máxima varianza. En este scatterplot, la proximidad de puntos indica similitud en las variables originales: observaciones cercanas comparten patrones de comportamiento. Además, la orientación de la nube de puntos muestra las direcciones de varianza máxima, permitiendo identificar el eje más informativo (PC1) y el segundo más relevante (PC2).

Interpretación:

- **Agregación de individuos:** Grupos densos de puntos señalan subconjuntos de clientes o casos con perfiles muy parecidos. Por ejemplo, si identificas un cúmulo a la derecha, podría corresponder a clientes con características extremadamente positivas en los componentes.
- **Identificación de outliers:** Puntos aislados a gran distancia del origen o de los grupos principales sugieren observaciones atípicas que pueden requerir revisión (posibles errores de datos o segmentos nicho).
- **Direccionalidad:** La orientación del conjunto de puntos sigue las direcciones de máxima varianza. Una nube alargada en diagonal indica correlación entre las variables originales que más contribuyen a PC1 y PC2.



2. **Proyección de variables numéricas – Diagrama de flechas**

```r
# Correlación variables originales vs componentes
Phi <- cor(dcon, Psi)
X <- Phi[, eje1]
Y <- Phi[, eje2]
plot(Psi[, eje1], Psi[, eje2], type = "n")
arrows(0, 0, X, Y, length = 0.07, col = "blue")
text(X, Y, labels = etiq, col = "darkblue", cex = 0.7)
```

Qué muestra:

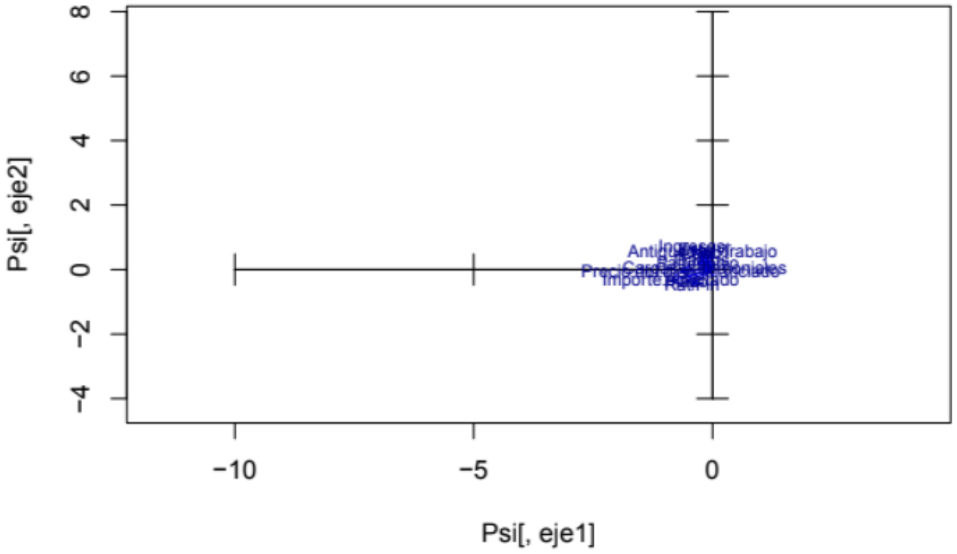
Se dibujan flechas (vectores) desde el origen en el plano de componentes, donde cada flecha indica la correlación entre una variable original y los componentes principales PC1 y PC2. La longitud de la flecha coincide con la magnitud de la correlación y la dirección señala si la relación es positiva o negativa.

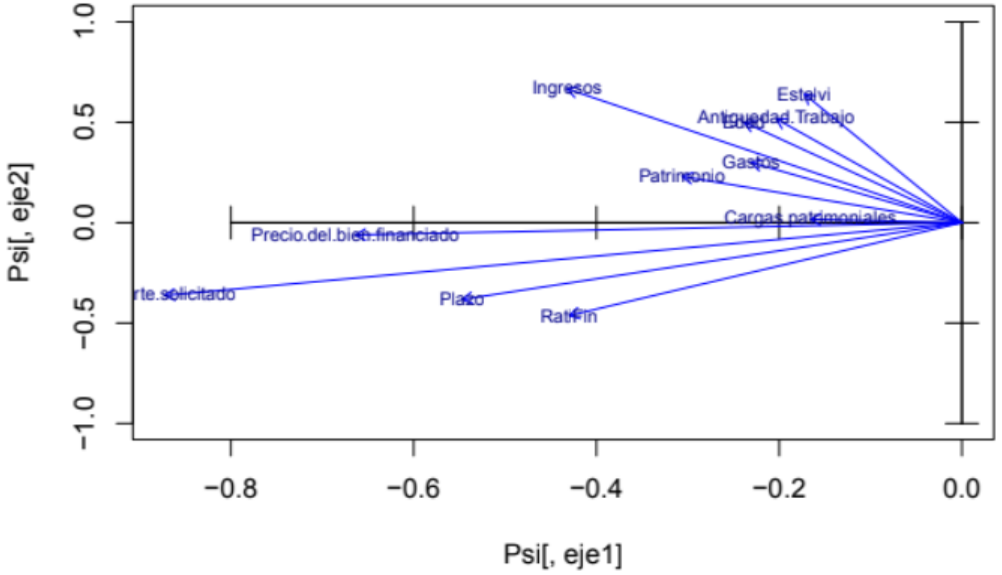
Significado:

Este diagrama factorial permite ver qué variables aportan más a la varianza explicada por los ejes. Variables con flechas más largas están mejor representadas en estos dos ejes, mientras que las flechas cortas tienen menor contribución. El ángulo entre dos flechas refleja la correlación entre las variables originales: ángulos pequeños implican correlación positiva, cercanos a 90° sugieren ausencia de correlación, y mayores o próximos a 180° indican correlación negativa.

Interpretación:

- **Variables clave:** Las variables cuyos vectores superan un cierto umbral de longitud (por ejemplo, >0.7) son las principales responsables de la separación de datos en PC1 y PC2.
- **Agrupación de variables:** Flechas que apuntan en direcciones similares y con ángulos pequeños corresponden a variables altamente correlacionadas (e.g., `Ingresos` y `Gastos`).
- **Análisis de direcciones opuestas:** Flechas opuestas señalan variables con comportamientos contrarios (e.g., `Ahorro` vs. `RatioFin`).
- **Selección de variables:** Sirve para decidir qué variables mantener o descartar en análisis posteriores basada en su contribución efectiva.





3. **Zoom de proyección de variables**

```
plot(Psi[, eje1], Psi[, eje2], type = "n",
     xlim = c(min(X, 0), max(X, 0)), ylim = c(-1, 1))
arrows(0, 0, X, Y, length = 0.07, col = "blue")
text(X, Y, labels = etiq, col = "darkblue", cex = 0.7)
```

(Snippet: zoom variables cite turn3file5)

Qué muestra:
Reproduce el diagrama de flechas anterior, pero ajusta los límites de los ejes para centrar la vista en el rango más relevante de valores, evitando espacios vacíos y mejorando la legibilidad de flechas de longitud intermedia.

Significado:
Cuando algunas flechas son excesivamente largas, las de longitud moderada pueden quedar invisibles cerca del origen. Este zoom ayuda a comparar y distinguir variables con aportaciones moderadas que de otro modo quedarían solapadas.

Interpretación:

- **Detalle de variables intermedias:** Permite ver con claridad variables con correlaciones en torno a 0.3–0.5, identificando potenciales comportamientos secundarios o interacciones.
- **Balance de representación:** Ayuda a evaluar si variables de aporte medio pueden aportar valor explicativo suficiente para análisis futuros, comparándolas con las más prominentes.

4. **Centroides de categorías nominales en el mapa factorial**

```
# Variables cualitativas nominales
dcat <- c(1, 3, 6:7)
colors <- rainbow(length(dcat))
plot(Psi[, eje1], Psi[, eje2], type = "n")
for (i in seq_along(dcat)) {
  k <- dcat[i]
  fdic1 <- tapply(Psi[, eje1], dd[, k], mean)
  fdic2 <- tapply(Psi[, eje2], dd[, k], mean)
  text(fdic1, fdic2, labels = levels(dd[, k]),
       col = colors[i], cex = 0.6)
}
legend("bottomleft", names(dd)[dcat], pch = 1,
       col = colors, cex = 0.6)
```

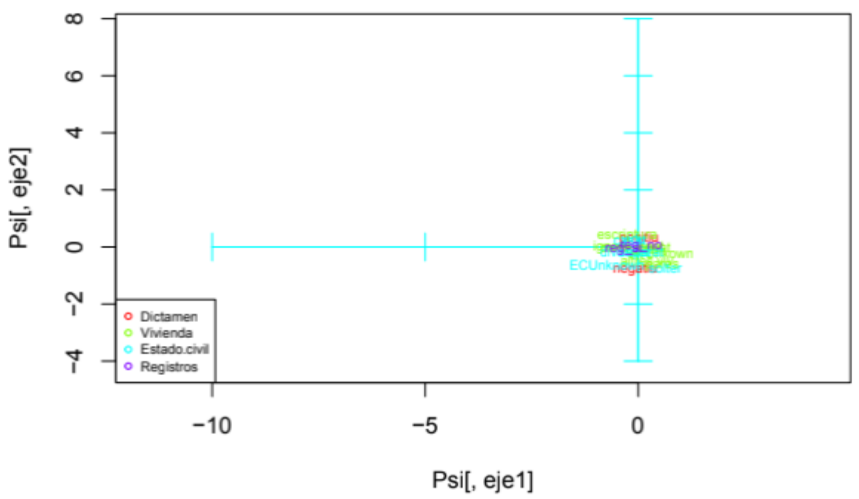
(Snippet: nominal cite turn3file16)

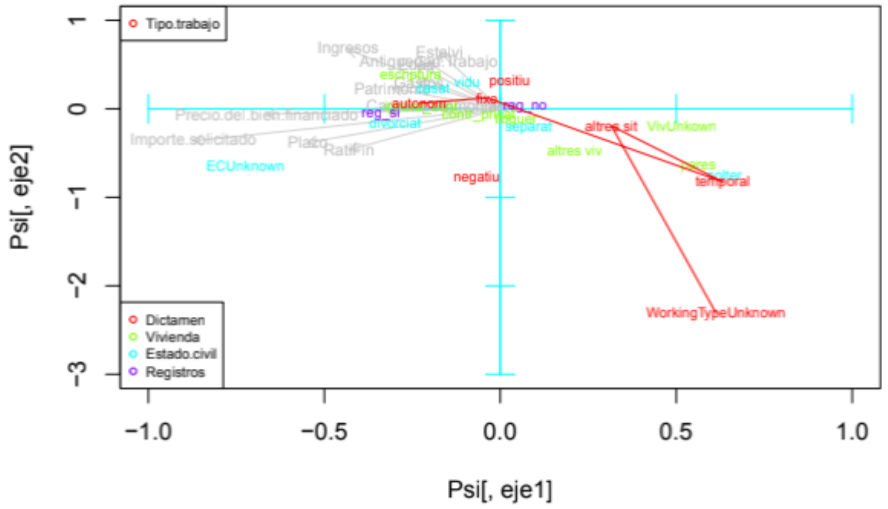
Qué muestra:
Coloca un texto para cada nivel de las variables cualitativas seleccionadas en el punto medio del grupo de individuos que comparten esa categoría. Cada etiqueta se sitúa en la coordenada media (centroide) de PC1 y PC2 de los casos de esa categoría.

Significado:
Al proyectar categorías nominales en el espacio factorial, se observa cómo se distribuyen los grupos cualitativos en relación con las componentes cuantitativas. Esto revela qué categorías están asociadas a perfiles específicos de variables continuas.

Interpretación:

- **Proximidad de etiquetas:** Categorías cuyos centroides están cerca comparten características cuantitativas similares. Por ejemplo, niveles de educación altos pueden situarse juntos.
- **Dirección de influencia:** Si un grupo nominal (e.g., "Hipoteca") aparece en la dirección de PC1 positiva, sugiere que esas observaciones tienden a tener valores altos en las variables que definen PC1.
- **Segmentación:** Ayuda a crear perfiles cualitativos sólidos: combinando posiciones de centroides con distancias, puedes segmentar clientes en función de atributos nominales.





5. **Trayectorias de categorías ordinales**

```r
# Variable ordinal
dordi <- c(8)
levels(dd[, dordi]) <- c("WorkingTypeUnknown", "altres sit", "temporal", "fixe", "autonom")
colors <- rainbow(length(dordi))
plot(Psi[, eje1], Psi[, eje2], type = "n")
fdic1 <- tapply(Psi[, eje1], dd[, dordi], mean)
fdic2 <- tapply(Psi[, eje2], dd[, dordi], mean)
lines(fdic1, fdic2, col = colors)
text(fdic1, fdic2, labels = levels(dd[, dordi]),
     col = colors, cex = 0.6)
legend("topleft", names(dd)[dordi], lty = 1,
       col = colors, cex = 0.6)
```

(Snippet: ordinal cite turn3file8)

Qué muestra:

Dibuja una línea conectando los centroides de cada nivel de una variable ordinal, siguiendo el orden natural de las categorías.

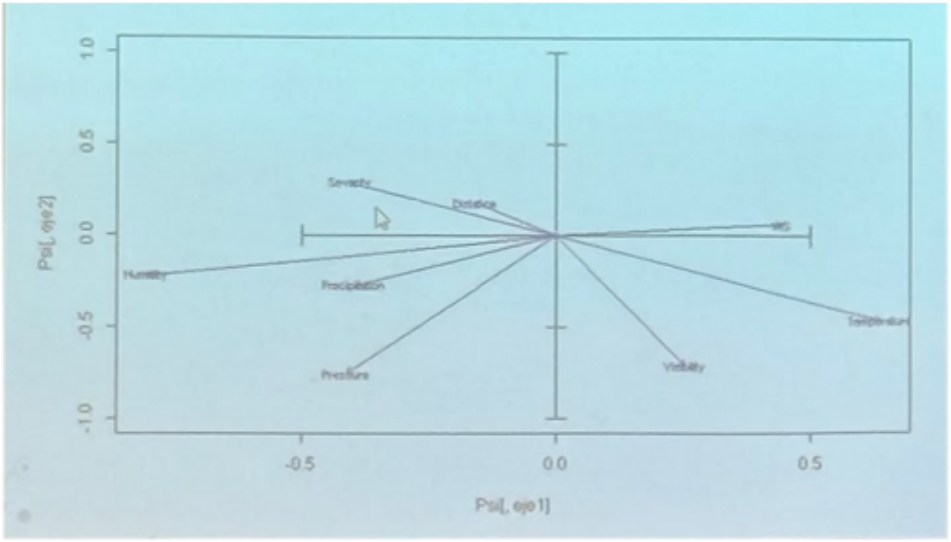
Significado:

Permite visualizar la progresión continua y ordenada de categorías, evaluando si efectivamente las componentes principales capturan la jerarquía implícita en la variable ordinal.

Interpretación:

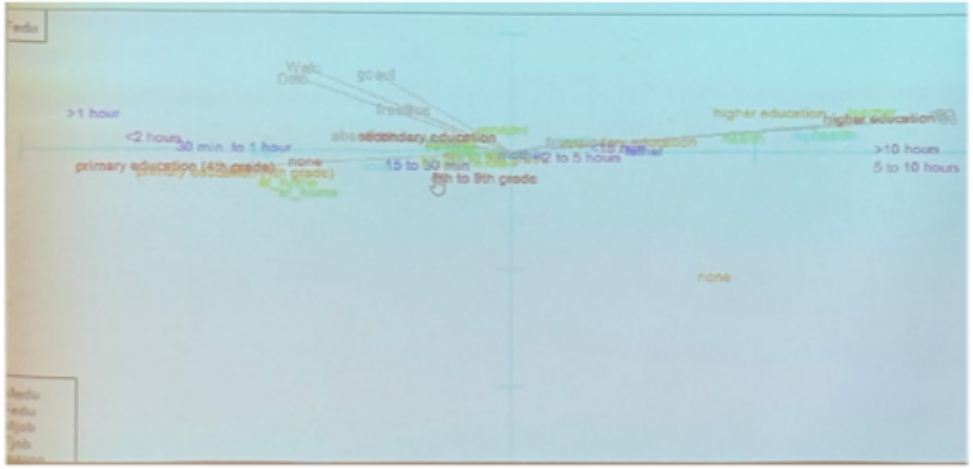
- **Linealidad:** Una trayectoria aproximadamente recta sugiere que la PCA respeta el orden original de la variable ordinal, capturando su variación de manera coherente.
- **Curvaturas:** Desviaciones significativas de la recta pueden indicar que las diferencias entre ciertos niveles no son homogéneas o que el componente seleccionado no refleja bien algunas transiciones.
- **Magnitud de salto:** La separación entre puntos consecutivos muestra la distancia relativa entre categorías adyacentes; saltos grandes pueden señalar categorías que diferencian mucho en el espacio factorial.

1. **What are the goals of PCA?**
    a. To perform classification
    b. To reduce dimensionality
    c. To Increase the number of features
    d. To handle missing data
    e. To find latent variables
    f. To identify variable relationships
2. **In PCA, the Inertia of a dimension can be observed in:**
    a. Eigenvalues
    b. Eigenvectors
    c. The original database
    d. The distance matrix
3. **Which components are necessary to calculate the new components?**
    a. Eigenvalues
    b. Eigenvectors
    c. The original database
    d. The distance matrix
4. **Choose the correct statements based on Figure 1:**



    a. Temperature and severity are negatively correlated.
    b. Temperature and severity are positively correlated.
    c. Temperature and severity are not correlated.
    d. Humidity and Visibility are the variables with the highest contribution to the first axis.
    e. WS has a high contribution to the first axis.
    f. Pressure and Visibility are negatively correlated.
    g. Pressure and Visibility are positively correlated.
    h. Pressure and Visibility are not correlated.
    i. If PCA Is being used to remove non-significant variables as a feature selection method, Distance would be the first variable to be excluded.

5. What conclusions can you get from Figure 2? If necessary, look up to Metadata.pdf for additional information.



- "Higher education" s'associa amb treballs qualificats i més hores de consum de mitjans.
- "Primary education" i "none" s'associen amb baixa activitat o feina no qualificada.
- Hi ha associació entre nivell educatiu i altres modalitats com temps dedicat o tipus de feina.
- Les modalitats s'agrupen, mostrant perfils diferenciats.

6. **The graphical result of hierarchical Clustering is:**
    a. A dendrogram
    b. A bar plot
    c. A pie chart
    d. A box plot
    e. A violin plot
7. **Before applying ascendant hierarchical clustering, you should choose:**
    a. Aggregation criteria
    b. Distance
    c. As many variables as possible
    d. Only numerical variables
    e. Variables with the highest eigenvalues
8. **Which techniques can be used in the profiling phase?**
    a. Multiple bar plots
    b. Multiple box plots
    c. ANOVA test
    d. χ^2 test
    e. Snake plot