Clustering

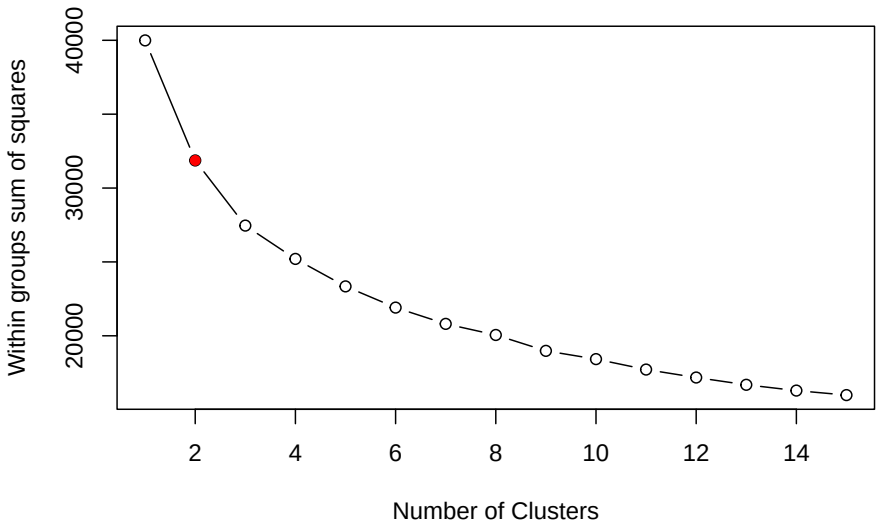
En esta sección, se emplearán diversos algoritmos de clasificación, específicamente el k-means y el Jerárquico. El propósito es asignar cada observación a un grupo correspondiente con el fin de llevar a cabo un perfilado. Este proceso implica etiquetar cada grupo con sus características más significativas, proporcionando así una descripción detallada y distintiva de cada perfil dentro de nuestros datos.

K-means

El algoritmo K-means solamente permitirá utilizar las variables numéricas. Por ello se separarán los datos numéricos de la base de datos preprocesada.

Antes de aplicar el propio algoritmo, se necesita seleccionar el número óptimo de clústeres. Para realizar esto, existen múltiples métodos, uno de ellos es el método del codo. Este consiste en aplicar el K-means para un rango de valores k y luego graficar la suma de los cuadrados de las distancias intraclúster en función de k. Para encontrar el óptimo con este método, sencillamente hace falta encontrar el "codo" del gráfico.

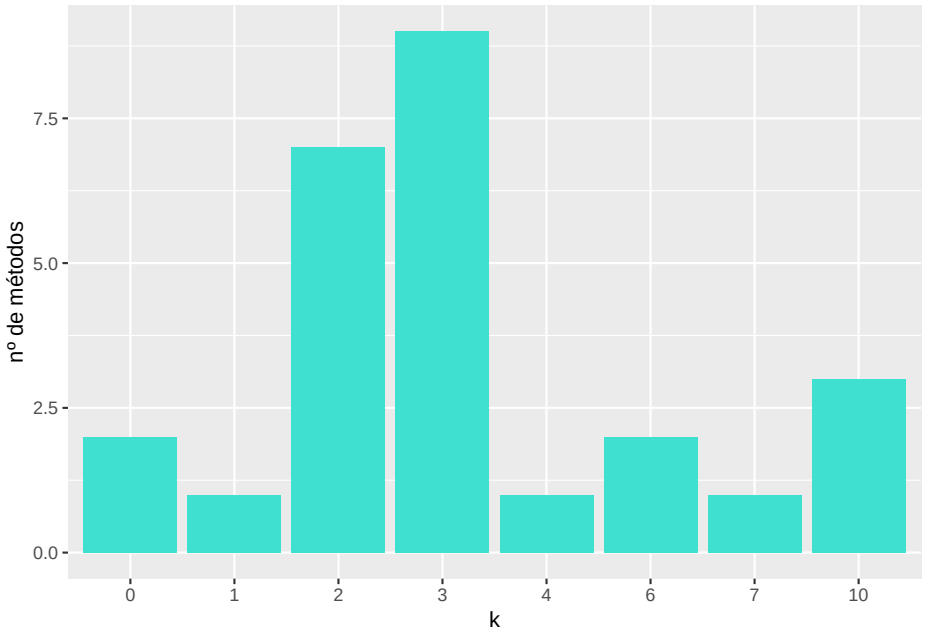
Figura 78: Método del codo



Como se puede apreciar en el gráfico, según el método del codo, el número óptimo de clústeres para el K-means de nuestra base de datos seria k=2.

Por lo que sigue, como existen muchos otros criterios para la selección de la k óptima, se usará la función NbClust, que permite aplicar una cantidad de 26 criterios para la selección de k, de esta manera se sabrá con mayor seguridad cuál es el óptimo real. Se grafican los resultados obtenidos por NbClust.

Figura 79: Número óptimo de clústers para el K-means

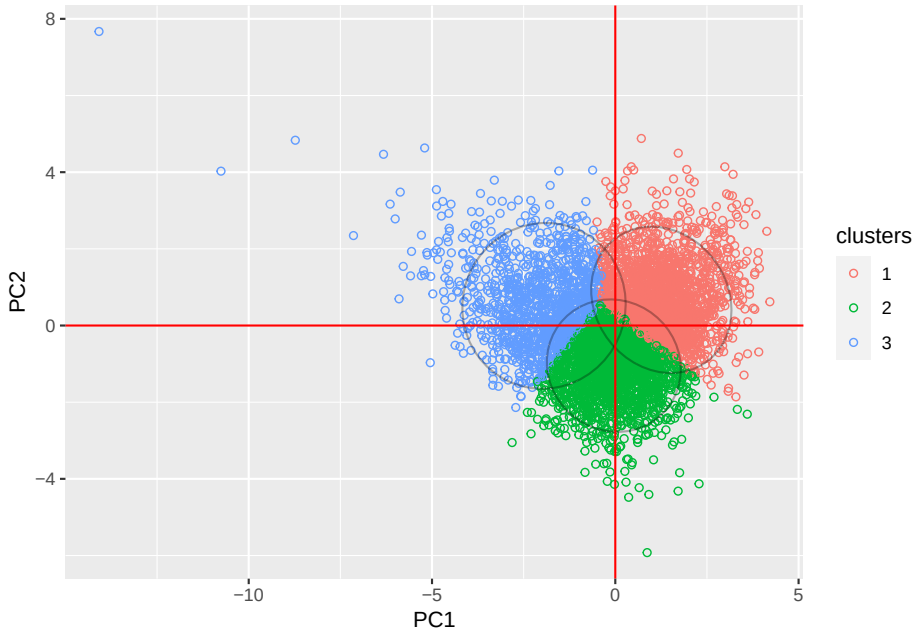


Como se puede apreciar en el histograma, tras haber utilizado todos los criterios, el número óptimo de clústeres que más métodos han escogido es k=3.

Como el óptimo se encuentra en k=3, el siguiente paso es realizar el K-means con esa k.

Después de aplicar el algoritmo y conseguir el grupo de cada individuo, se muestra el gráfico de los individuos pintados según su clase en el plano factorial de las dos primeras dimensiones del PCA, acompañado de cada una de las elipses de las clases.

Figura 80: Representación de las clases en PC1-PC2



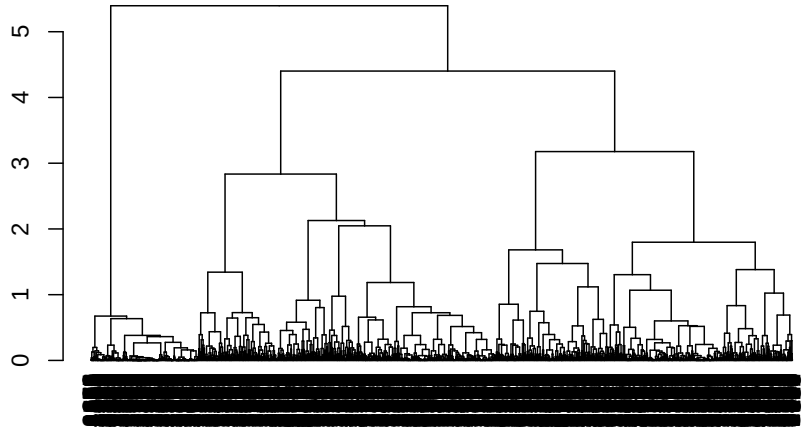
Ahora bien, como se puede ver en el gráfico, no se distinguen muy bien las tres clases, ya que están unas encima de las otras. Esto puede ser consecuencia de que la clasificación se ha hecho considerando solamente las variables numéricas, es por eso que es necesario realizar un clustering jerárquico.

Clustering jerárquico

En primera instancia, para realizar el clustering jerárquico se debe hacer primeramente un dendrograma con el método de Ward con la distancia de *Gower*^{2}. En el k-means solo se puede trabajar con las variables numéricas y en la base de datos hay variables tanto numéricas como cualitativas. La distancia de *Gower*^{2} nos permitirá calcular las distancias tanto de las variables numéricas como de las categóricas.

Así, se calcula dicha distancia y se grafica un dendrograma:

Figura 81: Dendograma



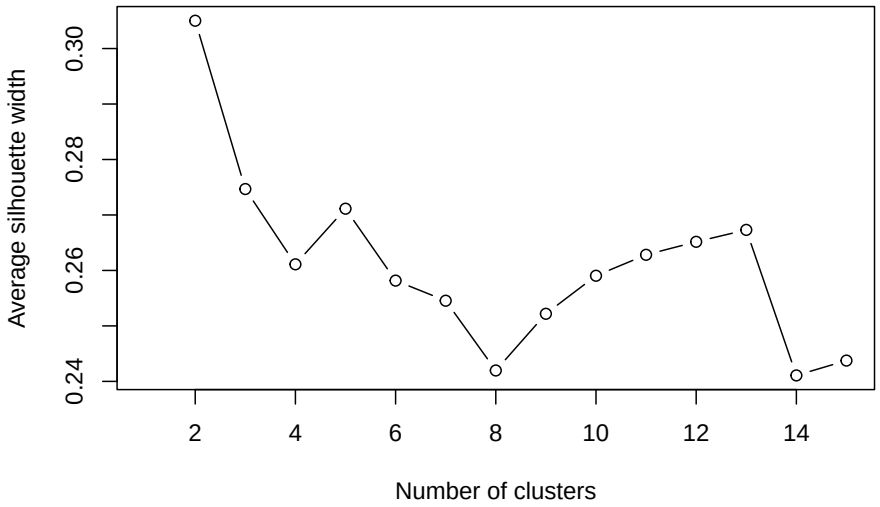
A primera vista se puede apreciar que el corte óptimo parece ser 2 clústeres. Esta cantidad de clústeres puede quedarse pequeña para los objetivos del trabajo. Consiguientemente, trataremos de tomar la decisión analíticamente, usando coeficientes que ayudan a decidir cuál es la mejor cantidad de clústeres.

Uno de ellos es el Coeficiente de Silhouette:

Los valores que retorna el Coeficiente de Silhouette van del 1 al -1. Generalmente, tomarán valores entre 1 y 0, siendo el 1 el mejor valor y 0 indicando la sobreposición de clústeres. Los valores negativos indicarían la asignación incorrecta de la muestra a los clústeres.

Lo que se hace es calcular el Coeficiente de Silhouette para diferentes cantidades de clúster y graficarlo, de manera que se cogerá el mayor valor como el número de clústeres según este criterio de Silhouette.

Figura 82: Silhouette



Como se puede ver, según el criterio de Silhouette, el número de clústeres óptimo es 2. No obstante, como existen muchos otros criterios, se usará -análogamente al K-means- la función NbClust, pero esta vez con la distancia de Gower, de este modo se consideran todas las variables.

Figura 83: Número óptimo de clústers para el Jerárquico

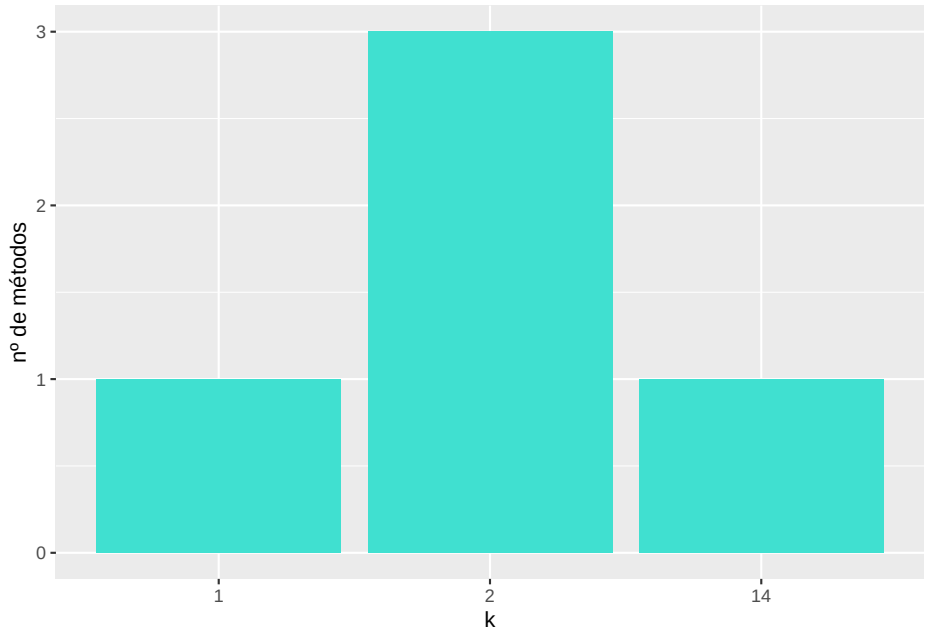
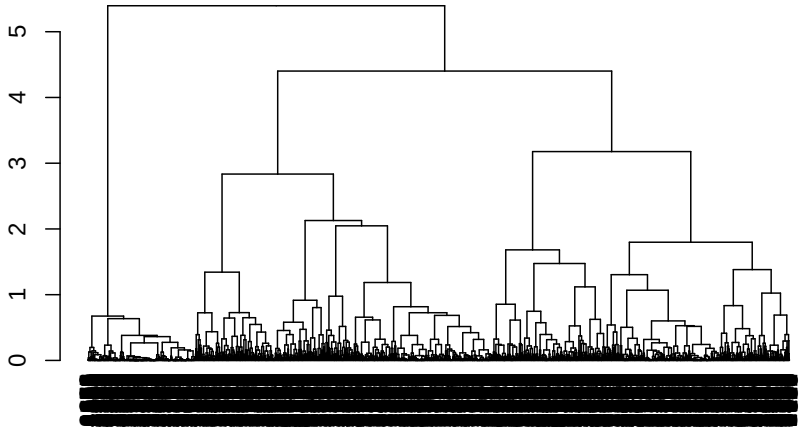
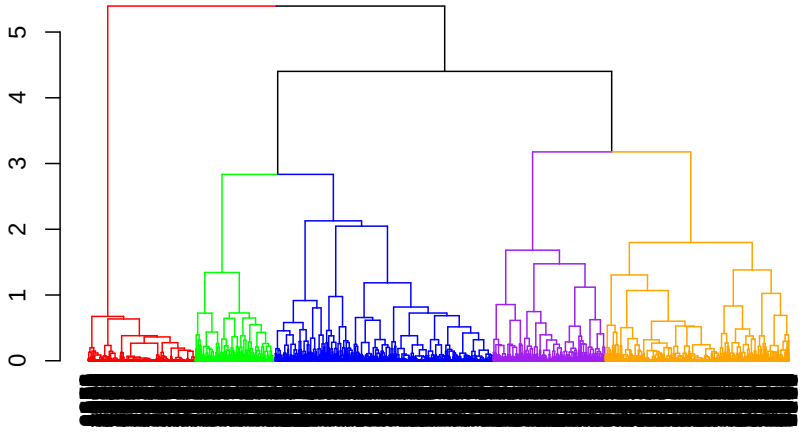


Figura 84: Dendograma



Con el dendrograma anterior, se confirma que el mejor corte (después de k = 2) es k = 3 y k = 5. Se divide el mismo dendrograma en k = 5 grupos:

Figura 85: Dendograma con la clasificación por clúster



Se escoge k=5 para hacer un perfilamiento de grupos detallado.

Profiling K-means

Con el objetivo de perfilar los grupos conseguidos mediante el algoritmo K-means primero veremos la significación de las variables para los grupos y después se graficarán para identificar las características definitorias de cada grupo.

A continuación se muestran los p-valores para evaluar la significación de cada variable. Primeramente de las variables categóricas y seguidamente las numéricas.

Cuadro 25: Significación de las categóricas

Variable	P_Value
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	0
ORGANIZATION_TYPE	0
REGION_RATING_CLIENT	0

Cuadro 26: Significación de las numéricas

OWN_CAR_AGE	c(Cluster = 0.000114855796849977)
CNT_FAM_MEMBERS	c(Cluster = 4.27829589973067e-06)
log_AMT_INCOME_TOTAL	c(Cluster = 0.942503452038247)
log_AMT_CREDIT	c(Cluster = 0)
AGE_YEARS	c(Cluster = 1.546029214791e-21)
RATIO_CREDIT_INCOME	c(Cluster = 0)
RATIO_ANNUITY_CREDIT	c(Cluster = 0)
DTI_RATIO	c(Cluster = 0)

Elaboramos una tabla donde se indica con 1 si se considera variable significativa para el clúster y 0 en caso contrario.

Cuadro 27: Significancia de p-valores para variables numéricas:

	x
OWN_CAR_AGE	1
CNT_FAM_MEMBERS	1
log_AMT_INCOME_TOTAL	0
log_AMT_CREDIT	1
AGE_YEARS	1
RATIO_CREDIT_INCOME	1
RATIO_ANNUITY_CREDIT	1
DTI_RATIO	1

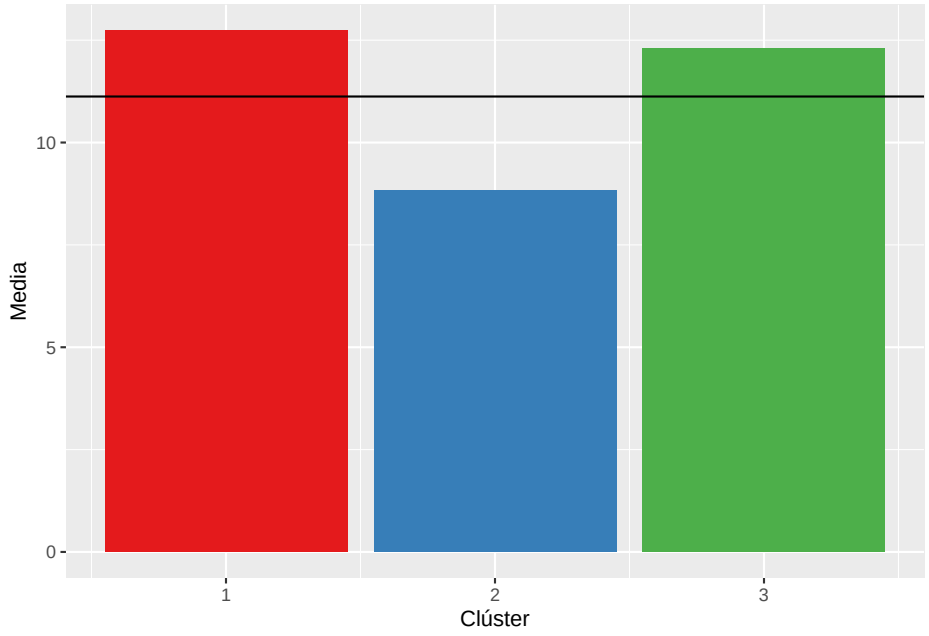
Cuadro 28: Significancia de p-valores para variables categóricas:

	x
CODE_GENDER	1
NAME_INCOME_TYPE	1
NAME_EDUCATION_TYPE	1
NAME_FAMILY_STATUS	1
OCCUPATION_TYPE	1
ORGANIZATION_TYPE	1
REGION_RATING_CLIENT	1

Vemos como solo nos descarta 1 variable, pero gráficamente muy pocas aportan información que muestren diferencias grandes entre clúster.

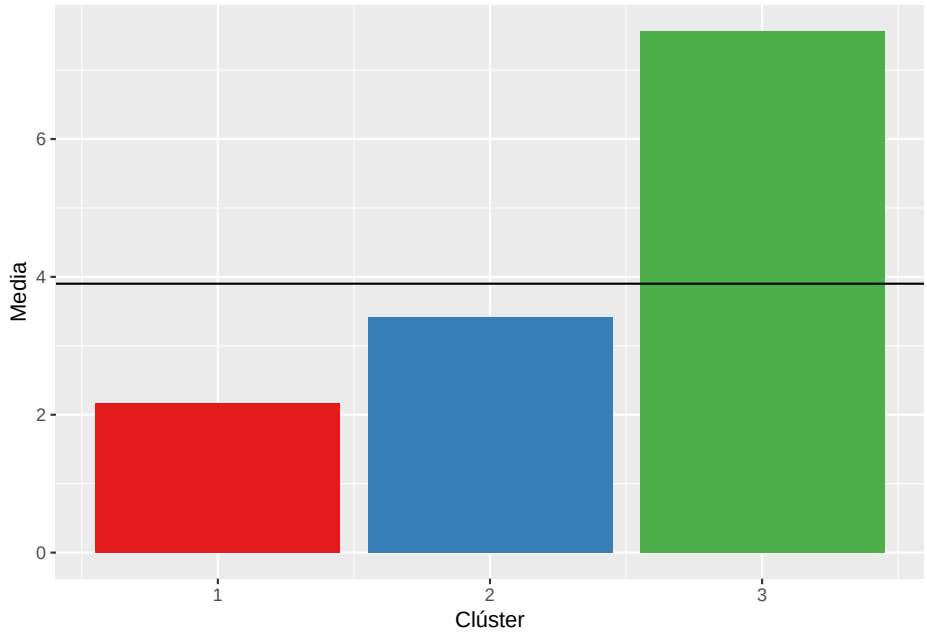
Se grafican las variables según clúster. Para las variables numéricas se mostrará la media grupal y la media global; para las variables categóricas se mostrarán las cantidades de cada nivel de la variable categórica por clúster.

Figura 86: Medias de la Edad en años del coche del cliente por clúster respecto la media global



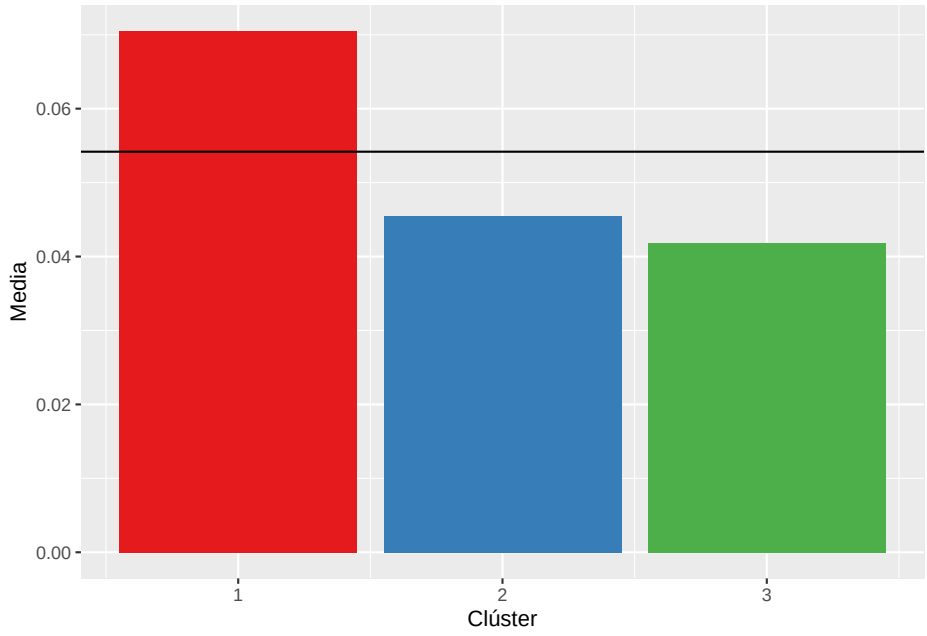
A partir del gráfico se puede observar como el clúster 2 es el que tiene los coches más nuevos.

Figura 87: Medias del Ratio del Importe del préstamo por clúster respecto la media global



Se ve como el clúster 1 es el que menos años tarda en devolver el préstamo, en concreto dos años. Por el contrario, el clúster 3 es el que más tarda en devolverlo, alrededor de 7 años.

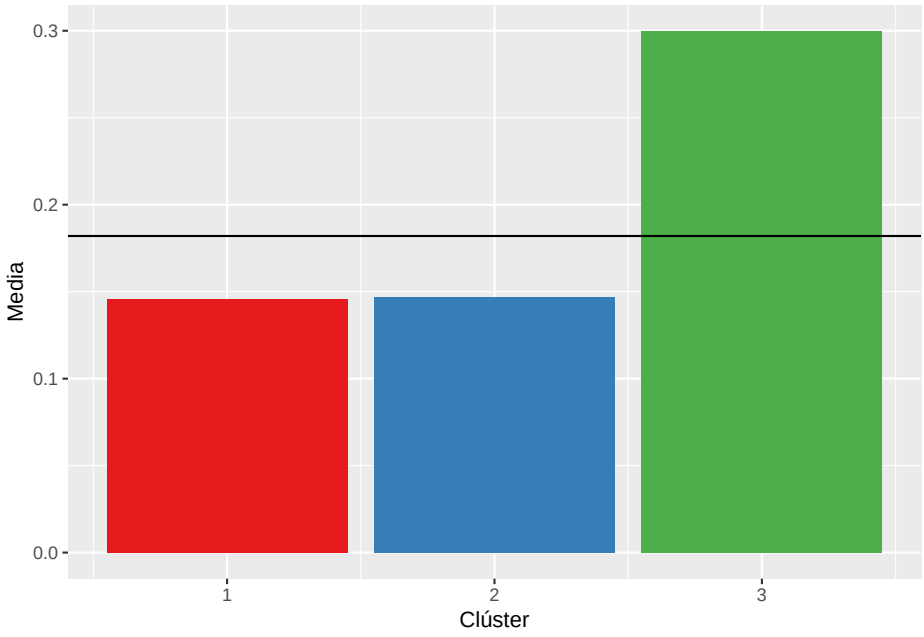
Figura 88: Medias del Ratio de la Anualidad del préstamo por clúster respecto la media global



Como se ha visto anteriormente, el clúster 1 es el que tiene una anualidad más alta, por otro lado el 3 es el que menos. Esto puede explicar el tiempo que se demoran en devolver el préstamo los individuos de cada

clúster.

Figura 89: Medias de la Capacidad de cliente para pagar la annuity con sus ingresos por clúster respecto la media global



Vemos como el clúster 3 tiene menos capacidad para pagar el préstamo.

Conclusiones

Clúster 1 se distingue por tener la anualidad más elevada y una menor demora en la devolución del préstamo.

Clúster 2 se caracteriza por incluir individuos con los coches más recientes.

Clúster 3 presenta una tendencia a tardar más en devolver el préstamo y exhibe una menor capacidad para hacerlo.

Es relevante notar que los patrones observados en los clústeres según las variables categóricas siguen la misma dinámica descrita. Esto podría deberse, en gran medida, a que el método k-means se centra exclusivamente en datos numéricos. Los gráficos asociados, aunque incluidos en el anexo, carecen de interpretación informativa directa sobre las características específicas de cada clúster.

Profiling Jerárquico

Con el objetivo de perfilar los grupos conseguidos mediante el algoritmo Jerárquico primero veremos la significación de las variables para los grupos y después se graficarán para identificar las características definitorias de cada grupo.

A continuación se muestran los p-valores para evaluar la significación de cada variable. Primeramente de las variables categóricas y seguidamente las numéricas.

Cuadro 29: Significación de las categóricas

Variable	P_Value
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	0
ORGANIZATION_TYPE	0
REGION_RATING_CLIENT	0

Cuadro 30: Significación de las numéricas

OWN_CAR_AGE	c(Cluster = 3.29268185193439e-09)
CNT_FAM_MEMBERS	c(Cluster = 0.324319480327695)
log_AMT_INCOME_TOTAL	c(Cluster = 3.15217179279226e-59)
log_AMT_CREDIT	c(Cluster = 0.0242641506823843)
AGE_YEARS	c(Cluster = 1.35532523003757e-214)
RATIO_CREDIT_INCOME	c(Cluster = 8.70167189360433e-25)
RATIO_ANNUITY_CREDIT	c(Cluster = 2.4213002726531e-05)
DTI_RATIO	c(Cluster = 6.138172493865e-22)

Elaboramos una tabla donde se indica con 1 si se considera variable significativa para el clúster y 0 en caso contrario.

Cuadro 31: Significancia de p-valores para variables numéricas:

	x
OWN_CAR_AGE	1
CNT_FAM_MEMBERS	0
log_AMT_INCOME_TOTAL	1
log_AMT_CREDIT	1
AGE_YEARS	1
RATIO_CREDIT_INCOME	1
RATIO_ANNUITY_CREDIT	1
DTI_RATIO	1

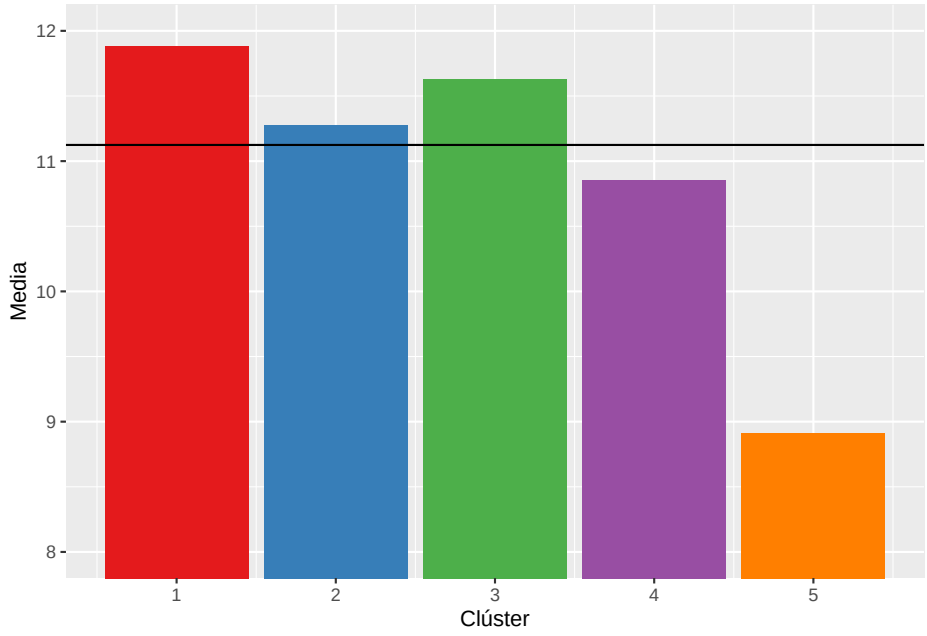
Cuadro 32: Significancia de p-valores para variables categóricas:

	x
CODE_GENDER	1
NAME_INCOME_TYPE	1
NAME_EDUCATION_TYPE	1
NAME_FAMILY_STATUS	1
OCCUPATION_TYPE	1
ORGANIZATION_TYPE	1
REGION_RATING_CLIENT	1

Podemos observar que solo se descarta una variable, pero al analizar gráficamente, podremos identificar qué variables son las que realmente aportan información significativa.

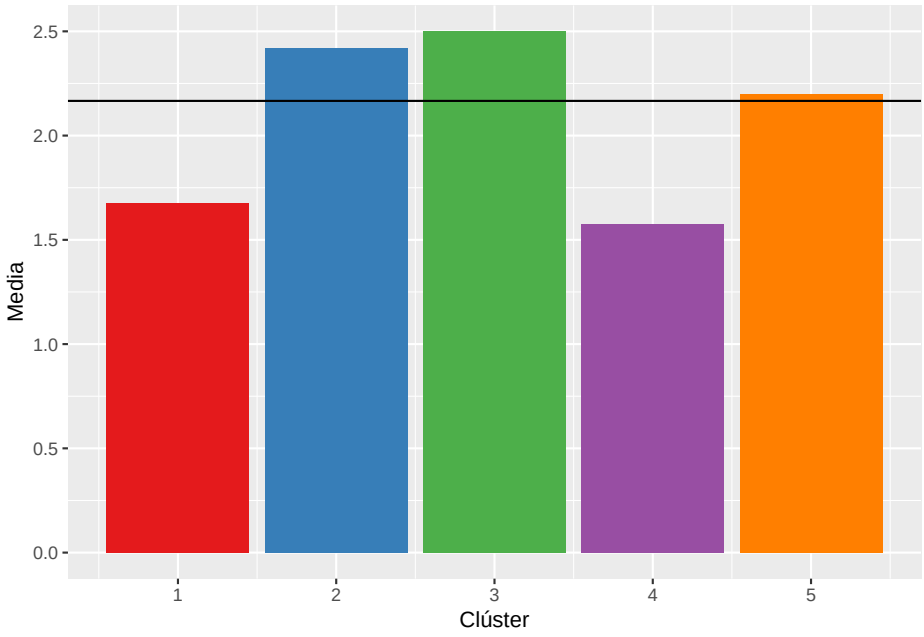
Se grafican las variables según clúster. Para las variables numéricas se mostrará la media grupal y la media global; para las variables categóricas se mostrarán las cantidades de cada nivel de la variable categórica por clúster.

Figura 90: Medias de la Edad en años del coche del cliente por clúster respecto la media global



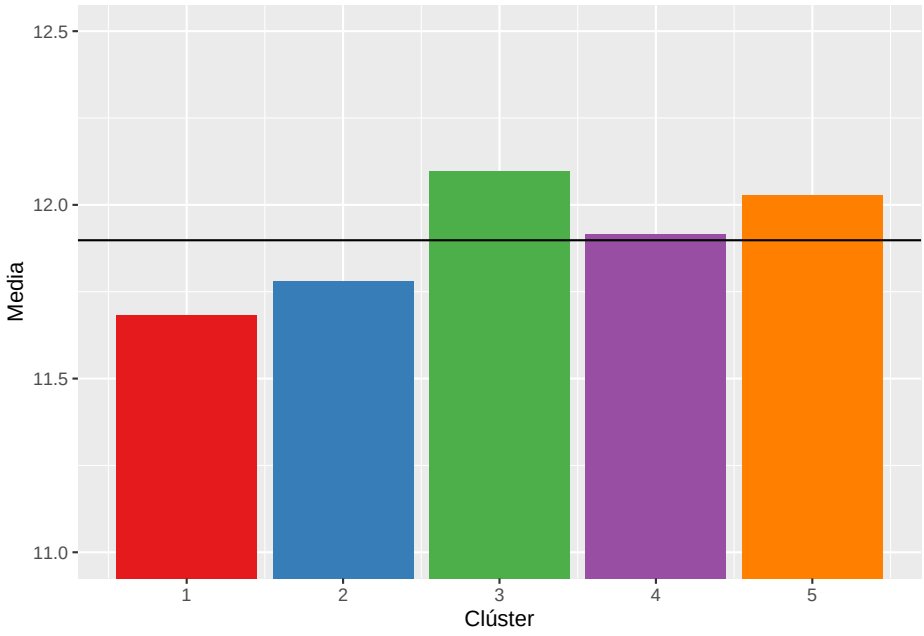
En lo que a la edad del coche para cada cliente respecta, vemos como el clúster 5 se caracteriza por tener una media de edad de coche mucho menor que los otros clústeres. Tiene los coches más nuevos, es decir, con menos años.

Figura 91: Medias del Número de familiares del cliente por clúster respecto la media global



Los clústeres 1 y 4 se caracterizan por tener el menor número de familiares. Por otro lado, el 2 y 3 tienen el mayor número de familiares.

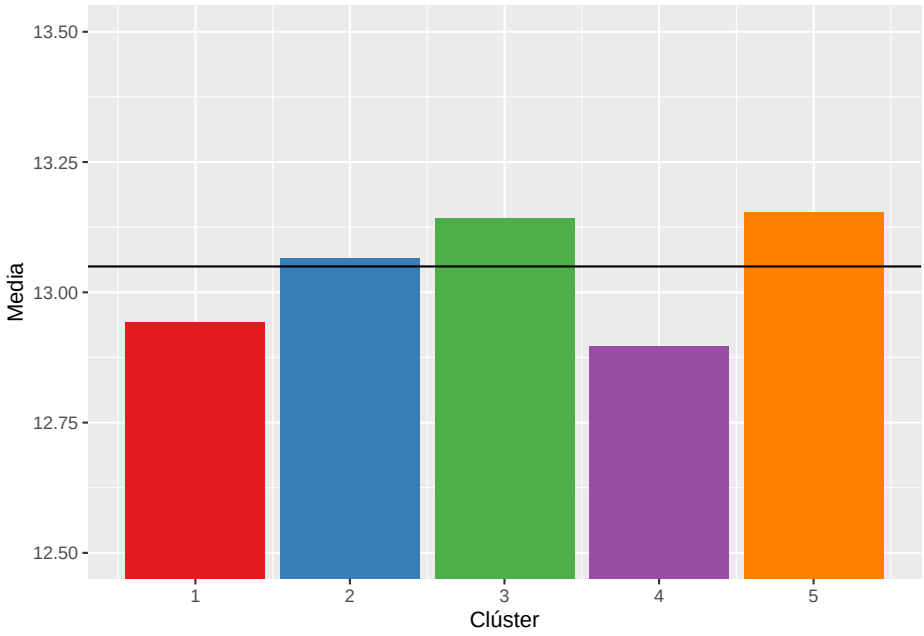
Figura 92: Medias del logaritmo de los Ingresos totales del cliente por clúster respecto la media global



A partir del gráfico, se observa que en los ingresos totales el clúster 1 se caracteriza por tener el menor número de ingresos y el clúster 3 el que mayor los tiene. No obstante, no hay diferencias muy grandes entre

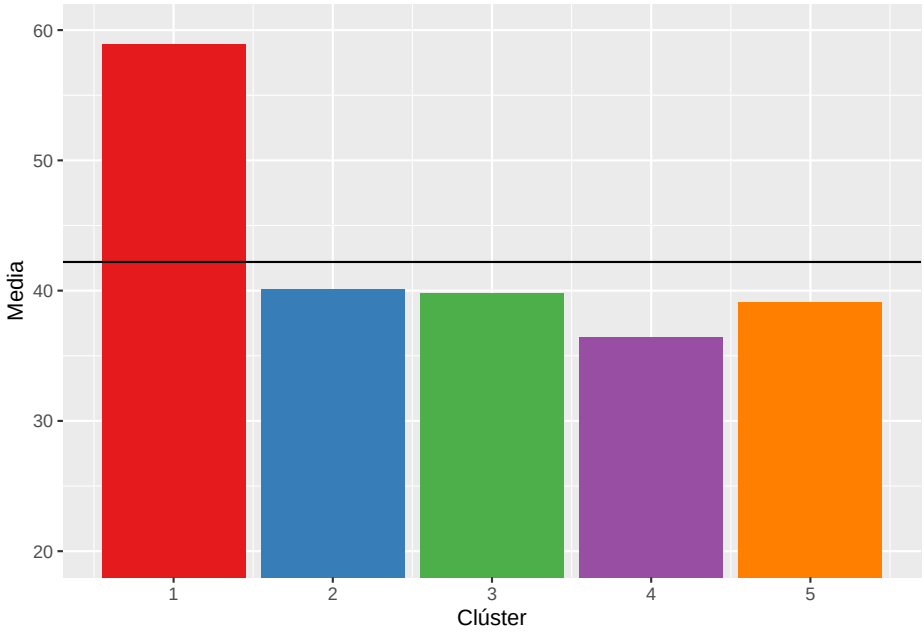
ellos debido a que es el logaritmo de estos ingresos y la interpretación queda afectada.

Figura 93: Medias del Importe de crédito del préstamo por clúster respecto la media global



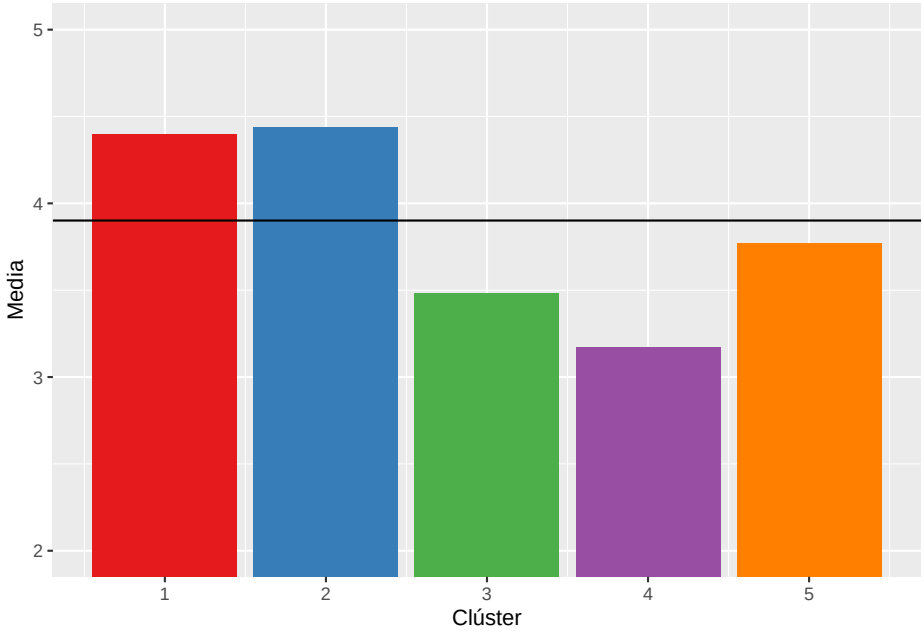
En el importe de crédito por préstamo se aprecia como el clúster 4 y 1 son los que más se diferencian, teniendo un importe menor. En cambio, el clúster 3 y 4 tienen un importe mayor que el resto.

Figura 94: Medias de la Edad por clúster respecto la media global



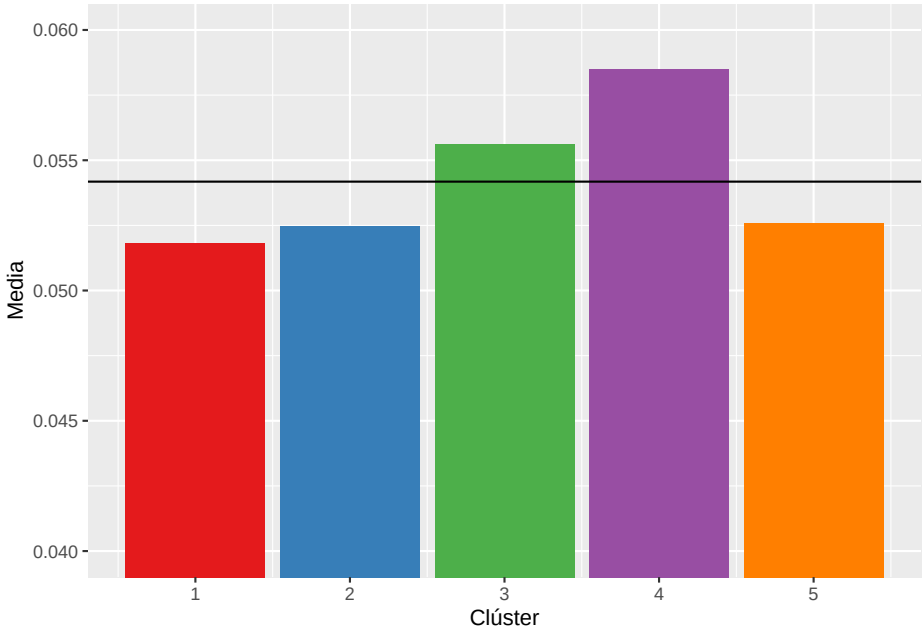
El clúster 1 se caracteriza por ser el grupo con individuos de más edad, en otras palabras, es el clúster con los individuos más mayores.

Figura 95: Medias del Ratio del Importe del préstamo por clúster respecto la media global



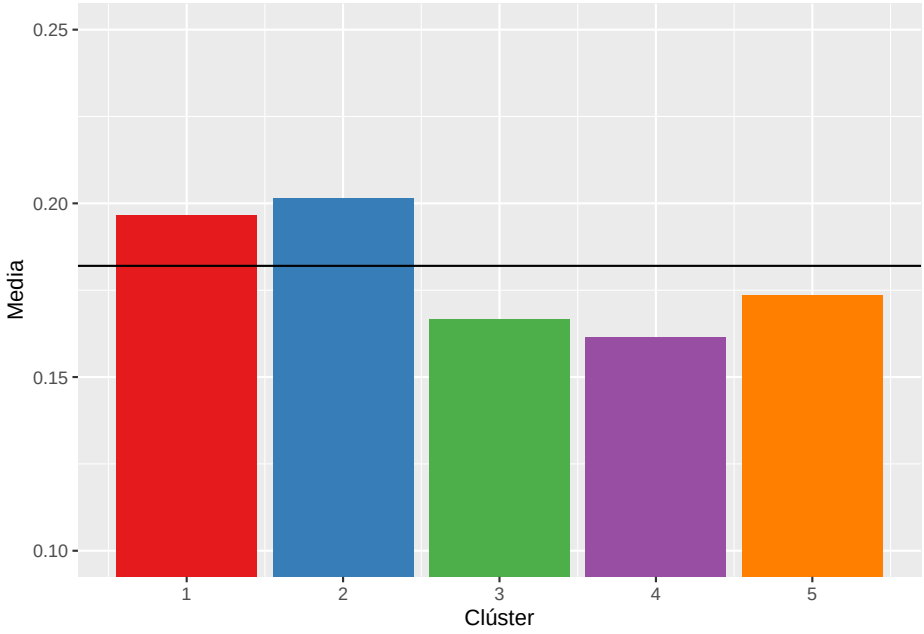
En el número de años que se tarda en devolver el crédito, se aprecia como los más rápidos son los sujetos del clúster 4. Contrariamente, el grupo 1 y 2 son los que más se demoran.

Figura 96: Medias del Ratio de la Anualidad del préstamo por clúster respecto la media global



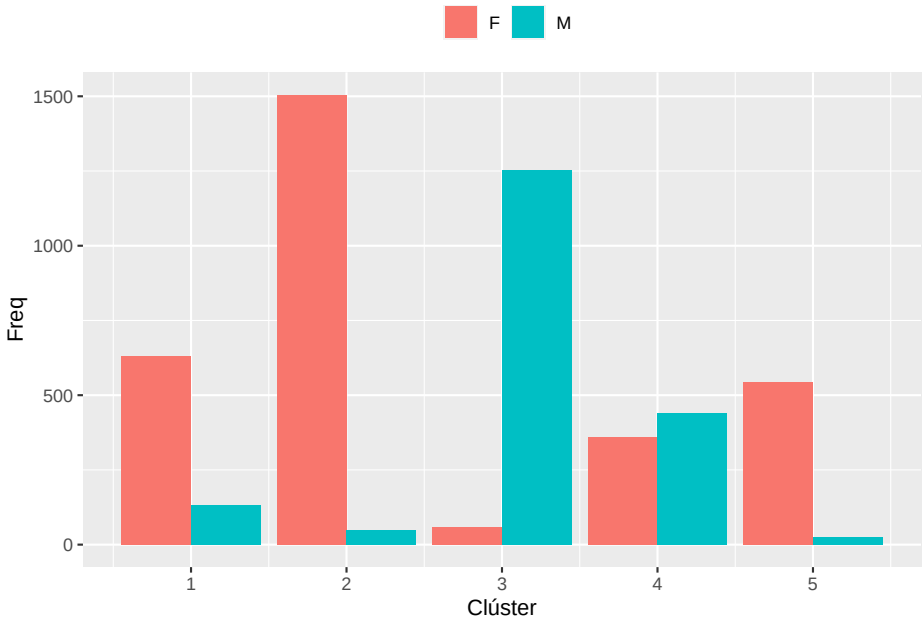
El clúster que se diferencia es el 4, con una Ratio entre la anualidad del préstamo y el crédito total solicitado.

Figura 97: Medias de la Capacidad de cliente para pagar la annuity con sus ingresos por clúster respecto la media global



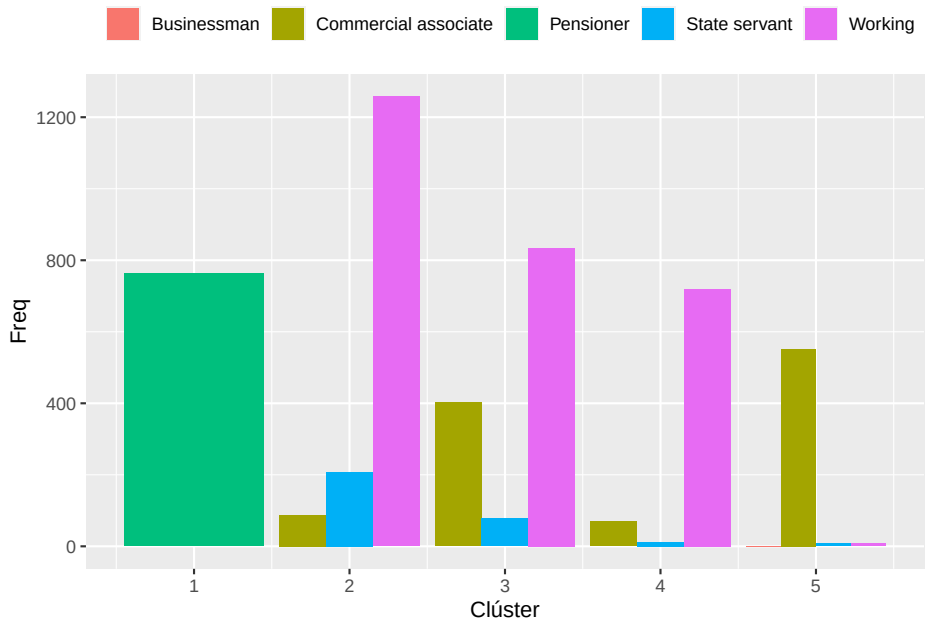
La ratio mide la capacidad del cliente para pagar la anualidad de su préstamo en relación con sus ingresos. Por ende, los menos capaces son los individuos del clúster 2.

Figura 98: Gráfico de la distribución del Género respecto el Clúster



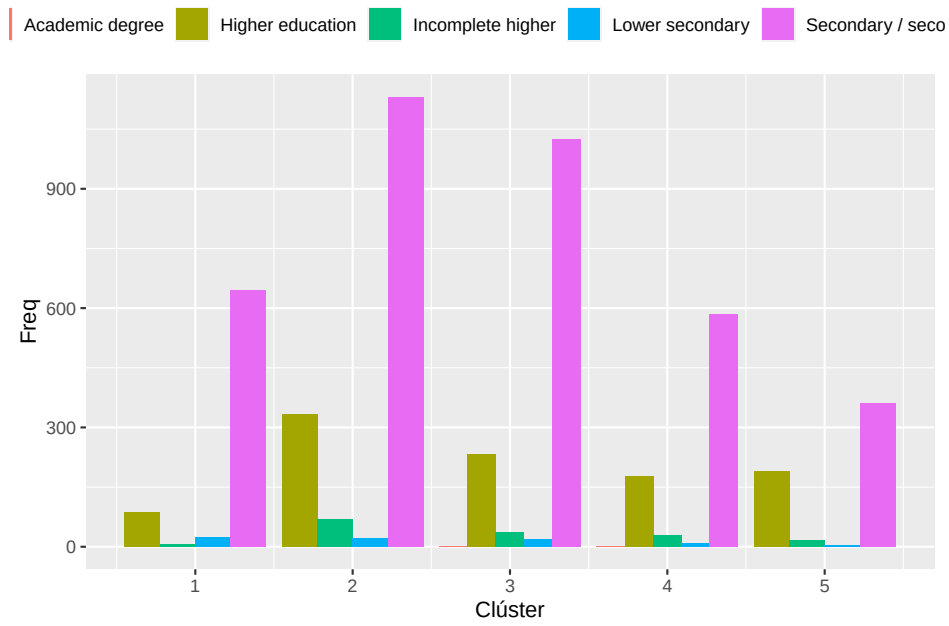
En la distribución del género de los individuos según el clúster, se observa como los grupos 1, 2 y 5 están formados por mujeres, el 3 por hombres y el 4 por hombres y mujeres a partes iguales.

Figura 99: Gráfico de la distribución del Tipo de ingresos respecto el Clúster



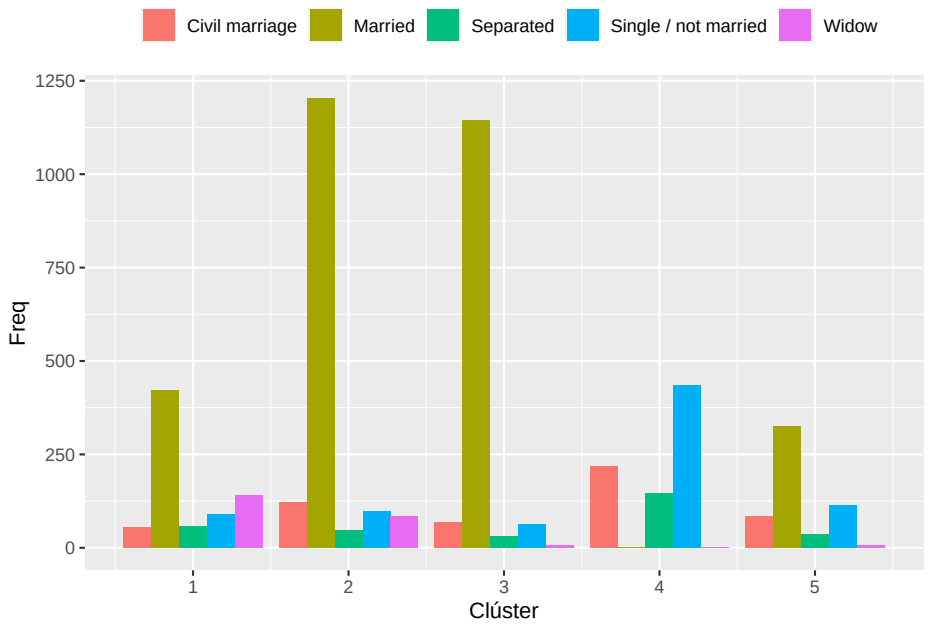
El primer grupo es el de los pensionistas y el quinto es el de los comerciantes asociantes.

Figura 100: Gráfico de la distribución del Nivel de estudios del cliente respecto el Clúster



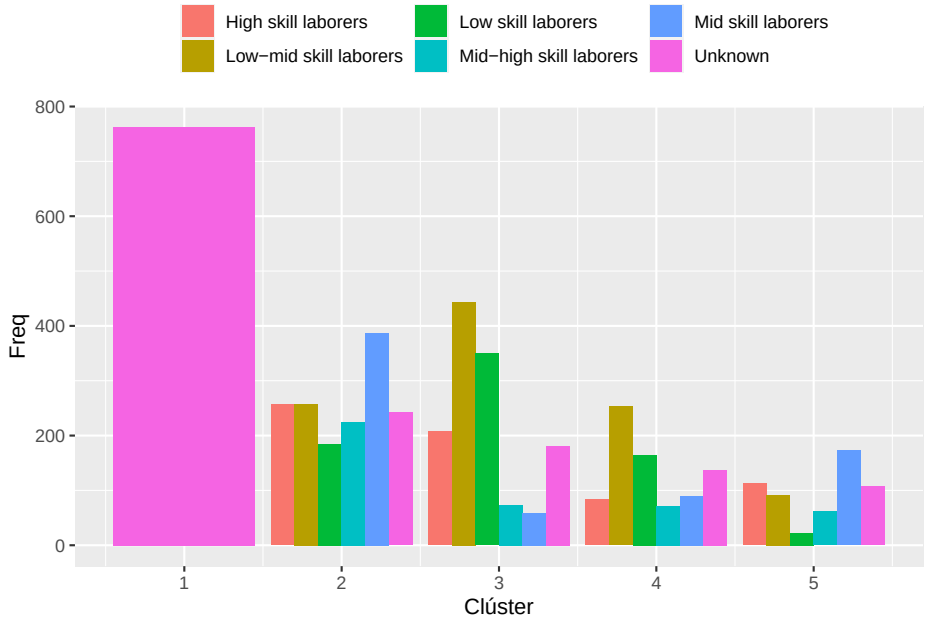
No hay diferencias con la distribución.

Figura 101: Gráfico de la distribución del Estado civil respecto el Clúster



En el Clúster 4 se incluye a personas que no están casadas o que tienen una unión civil.

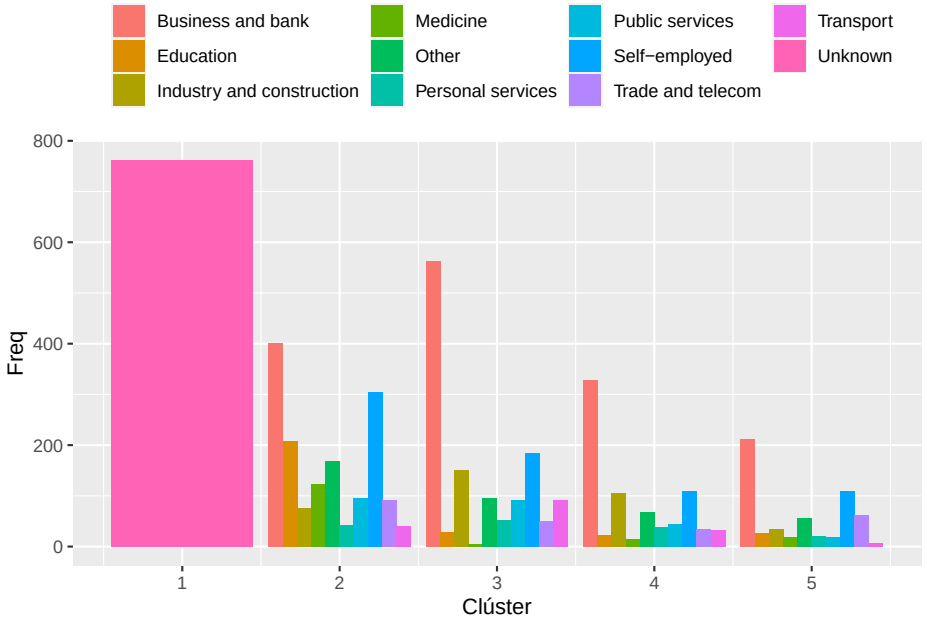
Figura 102: Gráfico de la distribución de la Actividad laboral respecto el Clúster



El grupo uno es el de los individuos que no trabajan, hecho que coincide con que también sea el grupo de

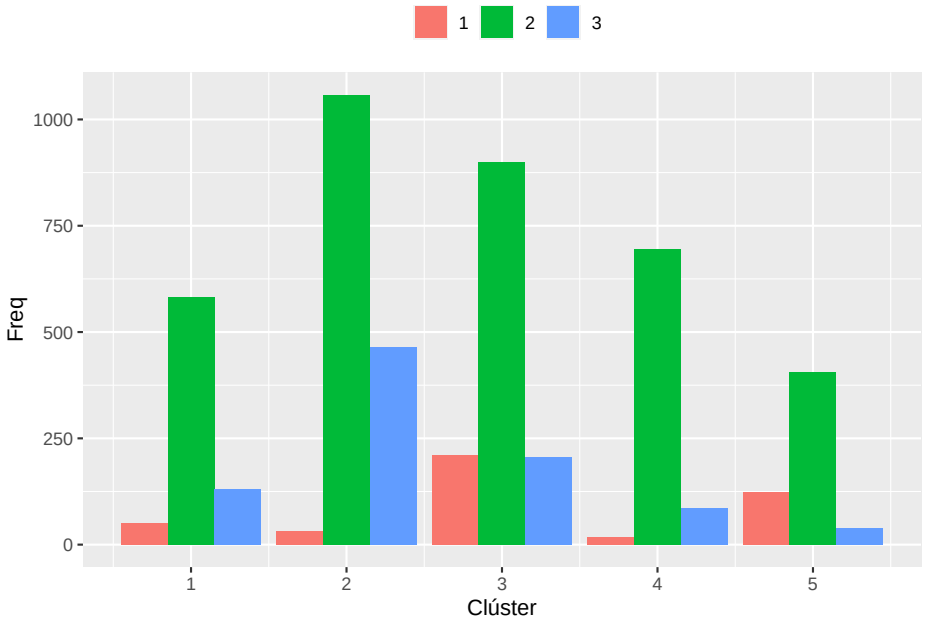
los pensionistas. Por otra parte, el grupo dos tiene más proporción de “mid skill workers”.

Figura 103: Gráfico de la distribución del Tipo de organización donde trabaja el cliente respecto el Clúster



Respecto al grupo uno no se puede saber a que tipo de organización pertenecen porque no trabajan, son los pensionistas. El resto sigue una distribución muy parecida aunque el grupo dos parece estar más dedicado a la educación que los otros.

Figura 104: Gráfico de la distribución de la Calificación de la región donde vive el cliente respecto el Clúster



No hay grandes diferencias en la distribución de la clasificación por región.

Conclusiones:

Clúster 1: Se caracteriza por individuos con pocos familiares y menor número de ingresos. Es el grupo de las mujeres de mayor edad pensionistas, las cuales tardan más en devolver el préstamo.

Clúster 2: Son mujeres, con una alta cantidad de familiares y mayor cantidad de importe del préstamo. Son los que peor capacidad de devolver el préstamo tienen, es decir, los que más se demoran en devolverlo. Proporcionalmente cuentan con más "mid skill workers".

Clúster 3: Es el grupo de los hombres con mayor número de ingresos y con más cantidad de familiares.

Clúster 4: En este clúster están los individuos con menor importe de crédito por préstamo, ratio de anualidad más grande además de ser los más rápidos en devolver el crédito. Se caracteriza por estar compuesto en la misma proporción tanto de hombres como de mujeres, con pocos familiares. Los individuos están solteros o casados civilmente.

Clúster 5: Este último grupo está formado por las mujeres con mayor número de crédito por préstamo. Caracterizadas por tener coches más nuevos y ser "commercial associates".

Comparación Profiling K-means y Jerárquico

Se destaca una diferencia en el número de clústers entre ambos métodos, con 3 clústers para K-means y 5 para el clústering jerárquico.

En el enfoque de clustering jerárquico, se logra obtener perfiles altamente específicos y fácilmente distinguibles para cada grupo, en contraste con los perfiles obtenidos a través de la metodología k-means.

En la metodología k-means, se observa que la explicación de los grupos se limita exclusivamente a variables numéricas, sin considerar ninguna variable categórica. Este enfoque numérico puro resulta en perfiles menos detallados, ya que no refleja la variabilidad explicada por las variables categóricas. Este aspecto contribuye a que los perfiles generados por k-means sean menos distintivos y caracterizados en comparación con los obtenidos mediante el clustering jerárquico.