Algoritmo CURE

Iker Meneses Sales

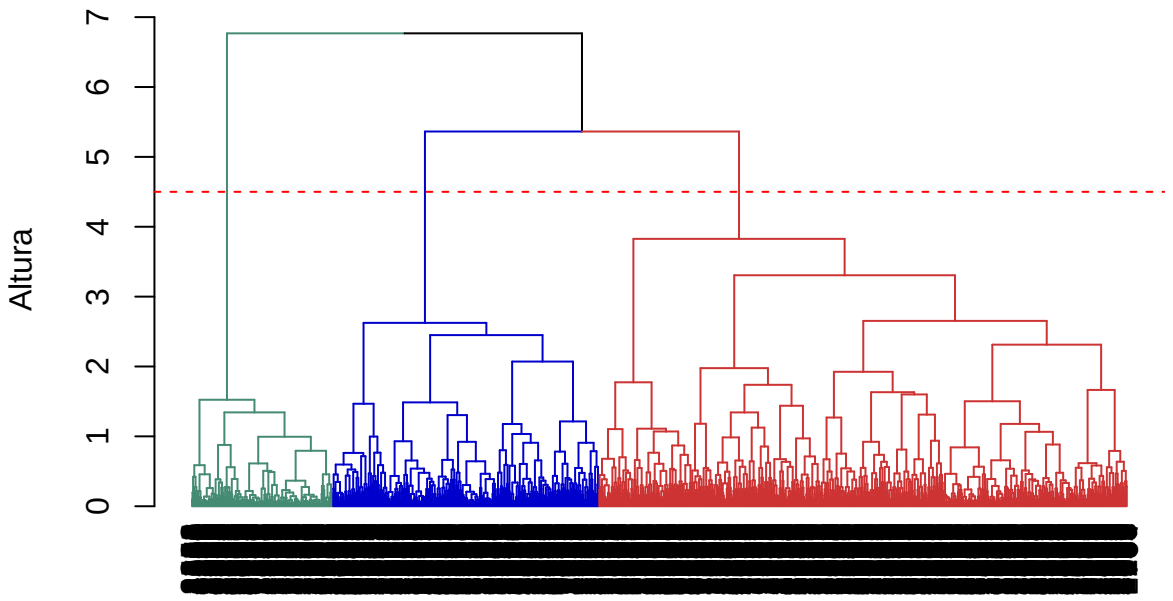
2023-10-29

Algoritmo CURE

Siguiendo con los algoritmos de clusterización para bases de datos grandes, es momento de realizar el CURE. CURE (Clustering Using REpresentatives) es un algoritmo de clustering para base de datos grandes en el cual se gestiona, inicialmente, una muestra de la base de datos a partir de la cual se realiza un clustering jerárquico (usando la distancia euclídea y el método de agregación simple) y se sacan un número pequeño de puntos (representantes) de cada cluster. Entonces, se acercan esos representantes hacia el centroide del cluster un 20% y, a partir de estos representantes acercados, se busca cuál es el que se encuentra más cercano de cada punto de la base de datos restante. Finalmente, una vez se encuentra el representante más cercano a cada individuo, se asigna el individuo al cluster al que pertenece el representante.

En este caso, como la base de datos escogida dispone de datos numéricos y categóricos, se ha decidido modificar las reglas del CURE y usar la distancia de Gower y el método de agregación de Ward en la construcción inicial del clustering. Así pues, realmente se podría afirmar que se está realizando un pseudoCURE en este caso.

Inicialmente, para este caso, se ha decidido escoger una muestra significativa y grande para evitar problemas en la construcción de los clusters iniciales. Así, se ha usado una muestra de $n = 2000$ con el objetivo de realizar el primer cluster a partir del cual se elegirán los representantes. El dendograma resultante reporta la siguiente estructura:



Tras analizar los resultados, se puede apreciar que el número de clusters óptimo es $k = 3$. De esta forma, la partición inicial de la muestra en cada cluster se puede apreciar en la parte inferior:

Table 1: Distribución inicial de individuos por cluster CURE

Cluster	Observaciones
1	569
2	1129
3	302

Ahora, a partir de este clustering jerárquico inicial, se escogerán los representantes. Para ello, se busca aquellos puntos más alejados entre sí y, a la vez, más alejados del centroide de cada cluster. Para este paso, se han elegido exactamente 5 representantes por cluster. Una vez se tienen seleccionados, el siguiente paso es acercarlos al centro. En este caso, se ha decidido aproximarlos un 20% hacia el centroide del cluster al que pertenecen.

Por último, se analiza cada punto y se busca el representante más cercano. Una vez se tiene esa información, se le asigna al individuo el cluster al que pertenece el representante más cercano. Para este paso, se ha procedido a procesar los datos de 500 en 500, para así evitar problemas con la capacidad de gestión de datos del ordenador. Así, el resultado del clustering final se presenta en la taba inferior:

Table 2: Distribución inicial de individuos por cluster CURE

1	1351
2	2275
3	1374

Profiling del CURE

A partir del CURE resultante, se analizan las características que diferencia cada cluster encontrado para así identificar las diferencias significativas más relevantes entre grupos. Aquí se muestran los gráficos de las medias por grupo para cada cluster resultante:

Antes de empezar a analizar cada variable, es importante destacar qué variables son sigificativas para diferenciar clusters. Para ello, se realizará un test de Chi-cuadrado para las variables categóricas y, por otro lado, un test F para las variables numéricas, a través de una tabla ANOVA:

Significación de las variables categóricas

En la siguiente tabla se puede apreciar cada variable con su p-valor asociado a la prueba de Chi-cuadrado correspondiente:

```
## Warning in chisq.test(cross_table): Chi-squared approximation may be incorrect

## Warning in chisq.test(cross_table): Chi-squared approximation may be incorrect
```

Table 3: P-valor asociado a cada variable categórica

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

Como se puede apreciar, en este caso, todas las variables son significativas, es decir, existen diferencias entre al menos un par de clusters. De esta forma, todas las variables categóricas serán tenidas en cuenta para el perfilamiento de los clusters.

Significación de las variables numéricas

Seguidamente, se seguirá el mismo procedimiento para las variables numéricas. Esta vez, sin embargo, se usarán los test F resultantes de la tabla ANOVA:

Table 4: P-valor asociado a cada variable numérica

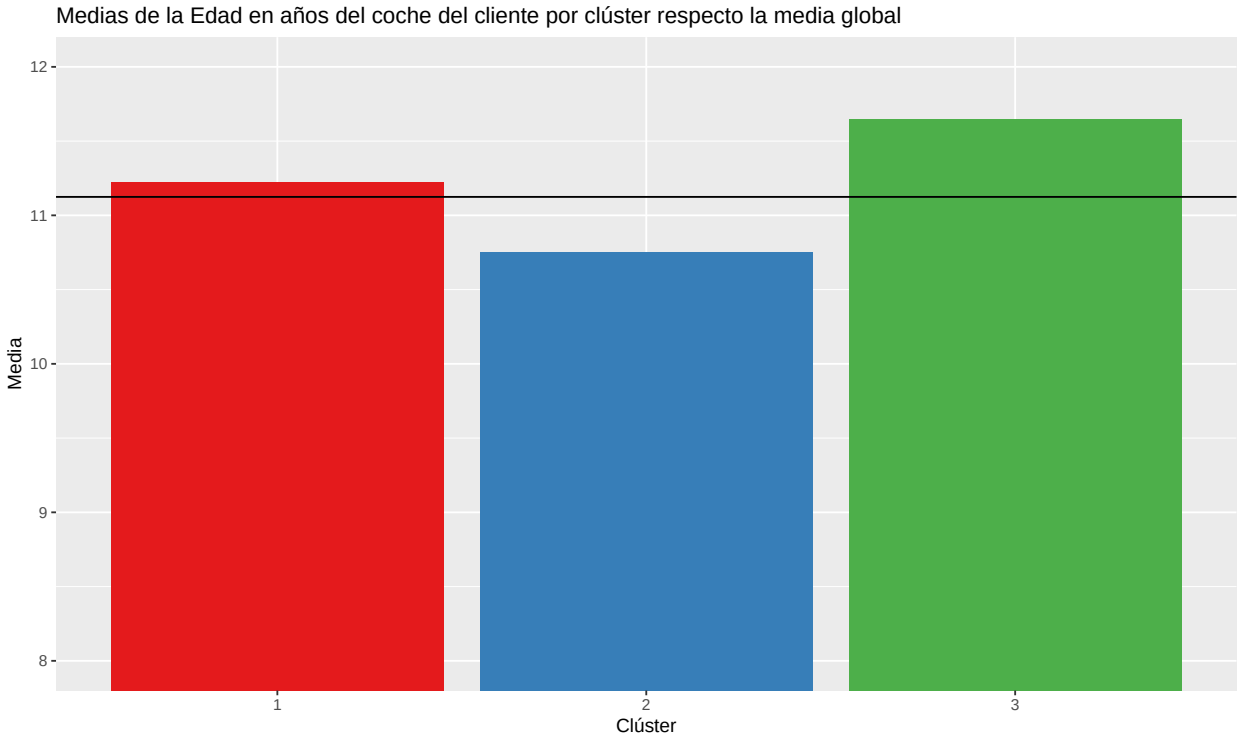
OWN_CAR_AGE	0.0044490
CNT_FAM_MEMBERS	0.0000000
log_AMT_INCOME_TOTAL	0.0000000
log_AMT_CREDIT	0.0000000
AGE_YEARS	0.0000000
RATIO_CREDIT_INCOME	0.0000379
RATIO_ANNUITY_CREDIT	0.0433127
DTI_RATIO	0.0014386

Nuevamente, como se puede apreciar, todas las variables son significativas, es decir, existen diferencias entre al menos un par de clusters. De esta forma, todas las variables categóricas serán tenidas en cuenta para el perfilamiento de los clusters.

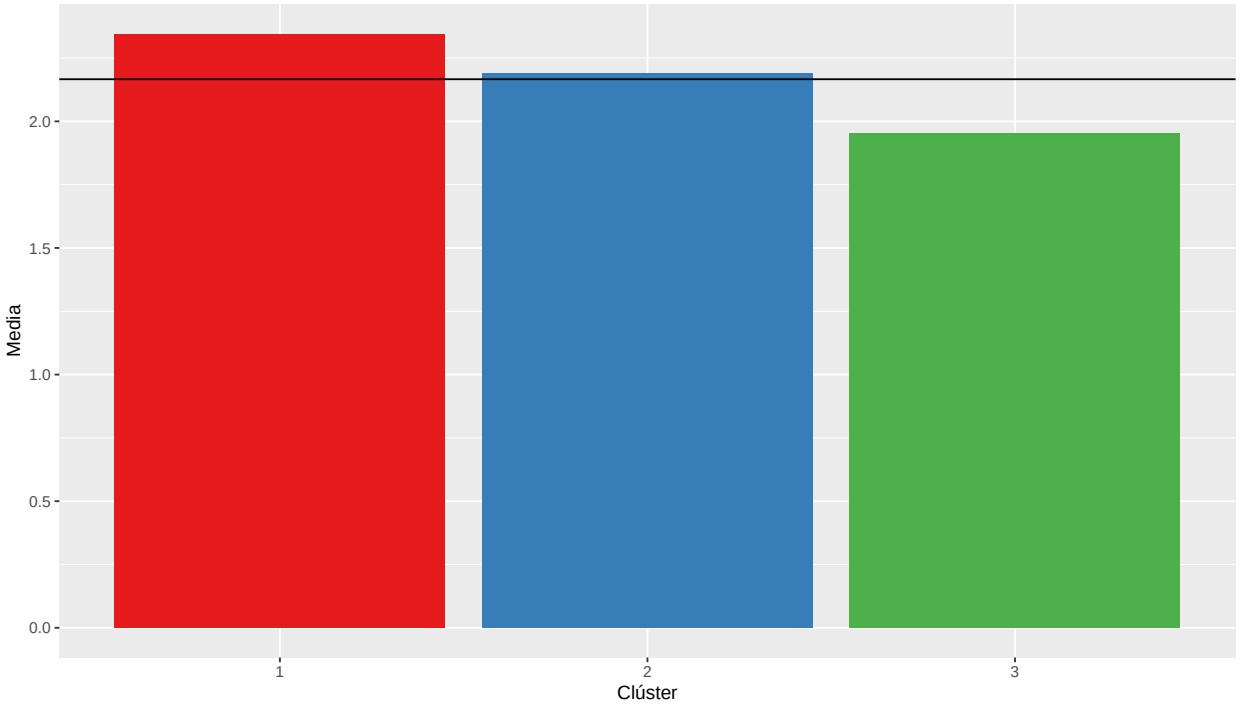
Análisis del profiling

Una vez ya hemos presenciado que todas las variables se usarán en el proceso de profiling de cada cluster, se ha procedido a realizar gráficos para cada variable, para así ver las características de cada grupo:

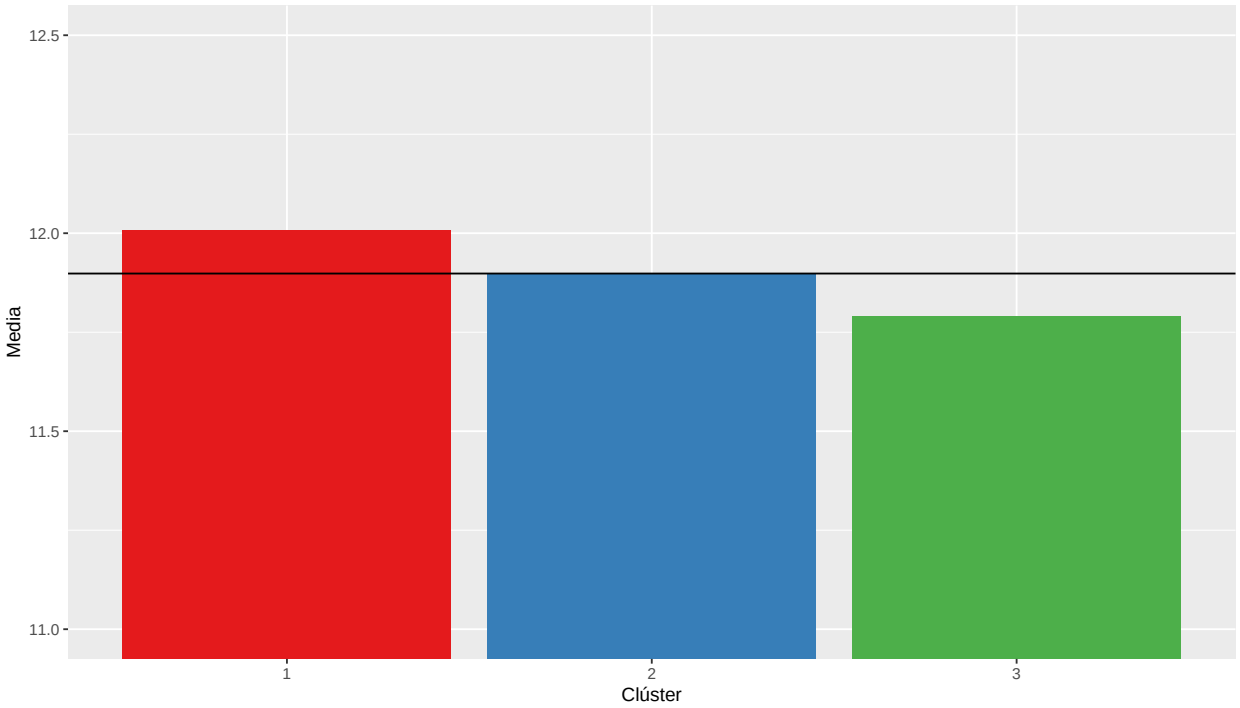
Variables numéricas



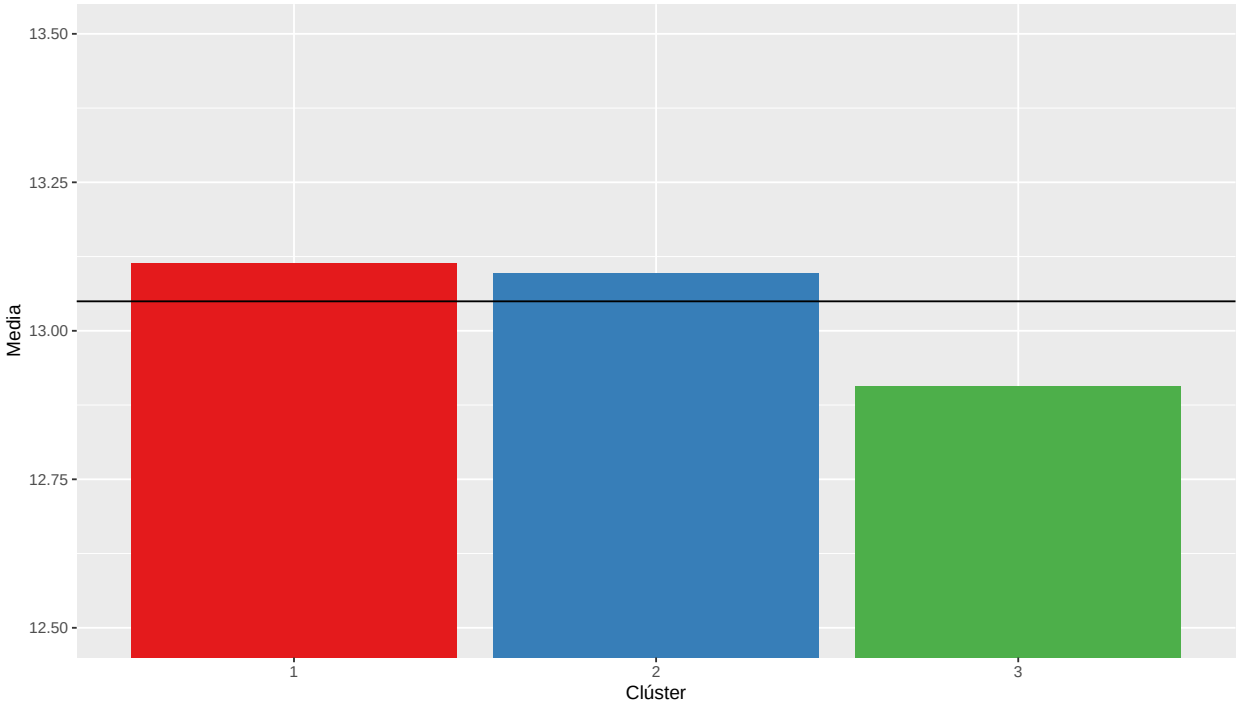
Medias del Número de familiares del cliente por clúster respecto la media global



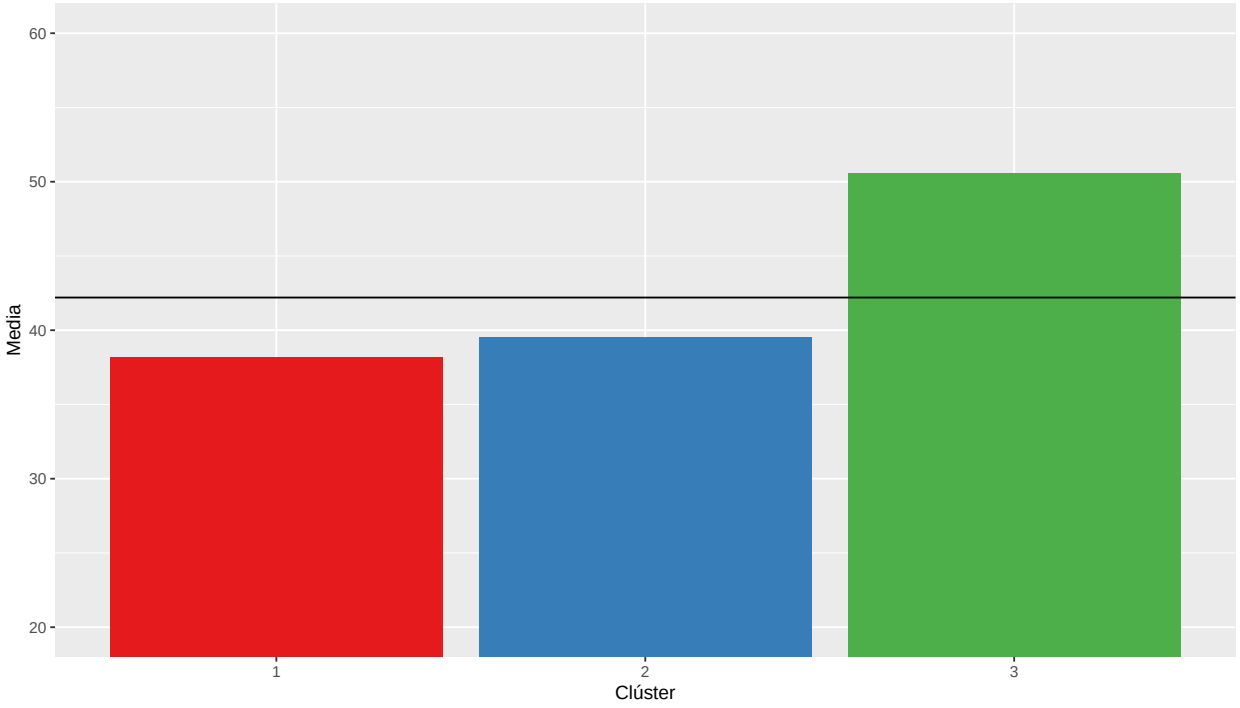
Medias del logaritmo de los Ingresos totales del cliente por clúster respecto la media global

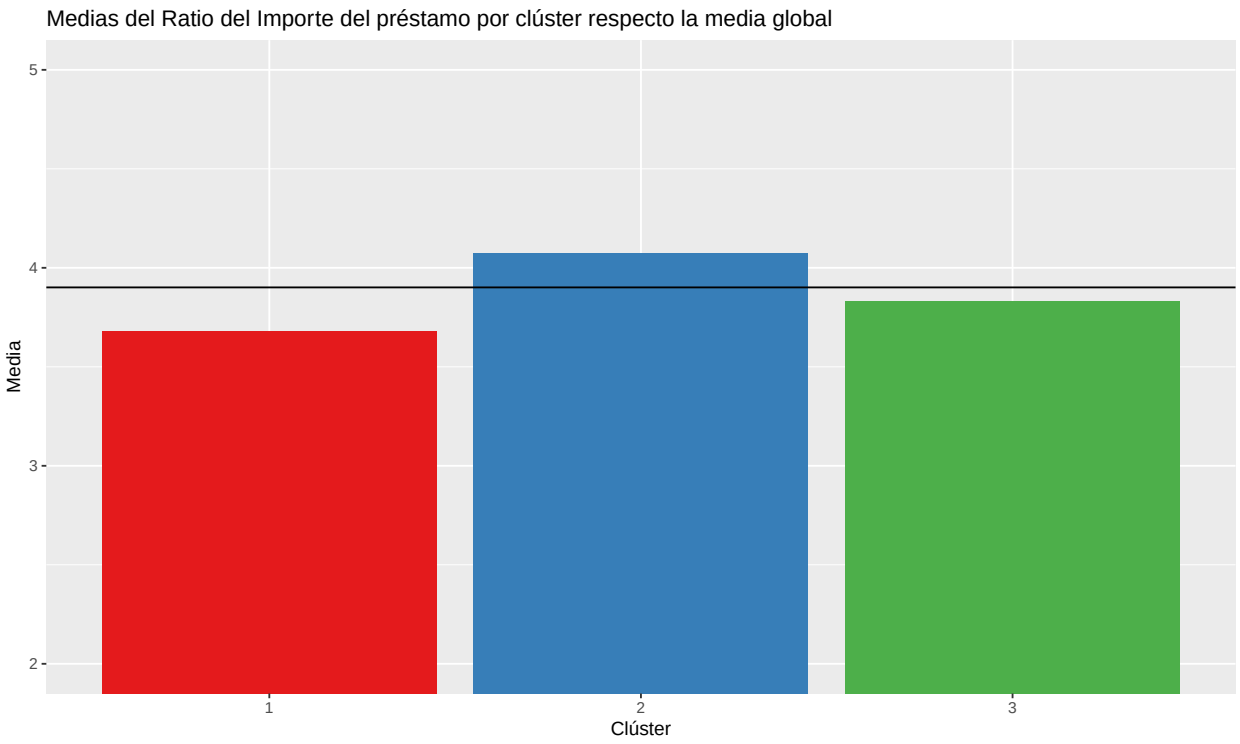


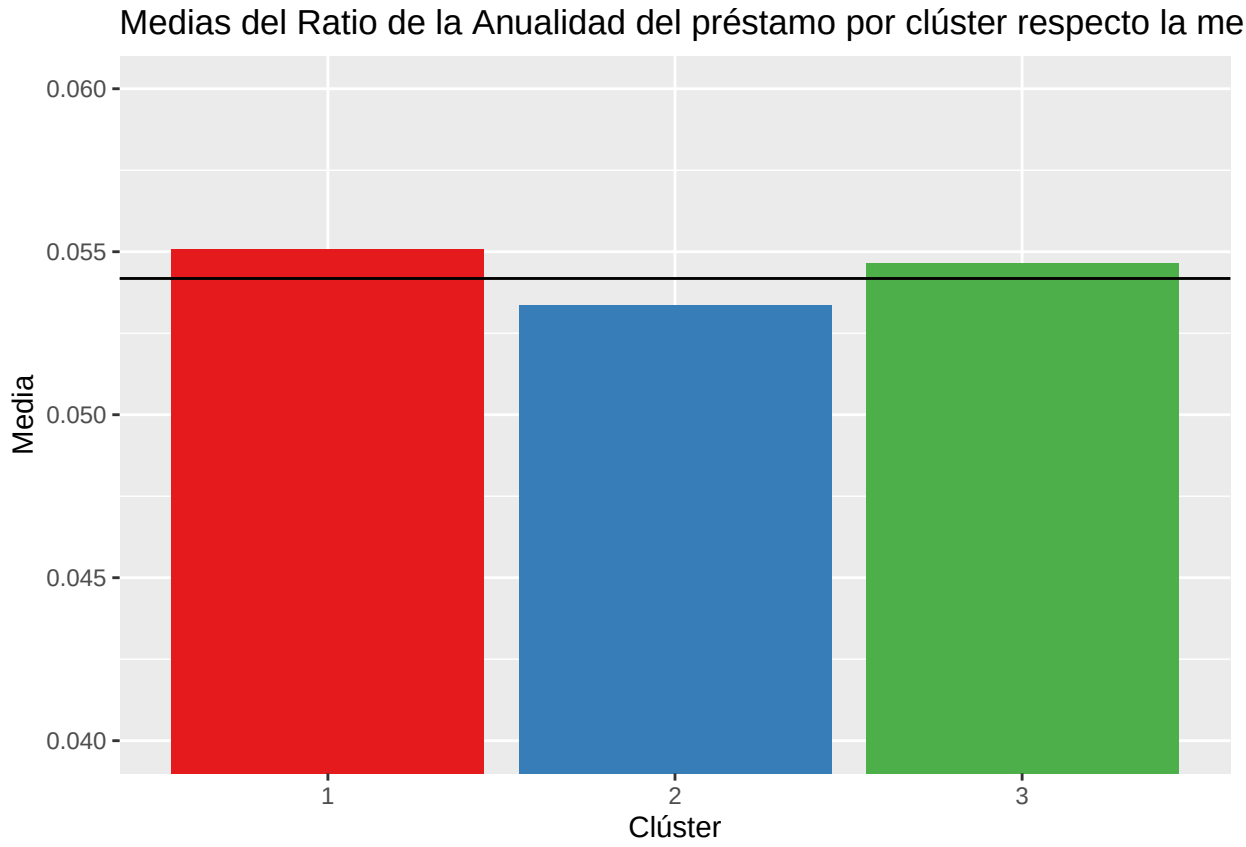
Medias del Importe de crédito del préstamo por clúster respecto la media global



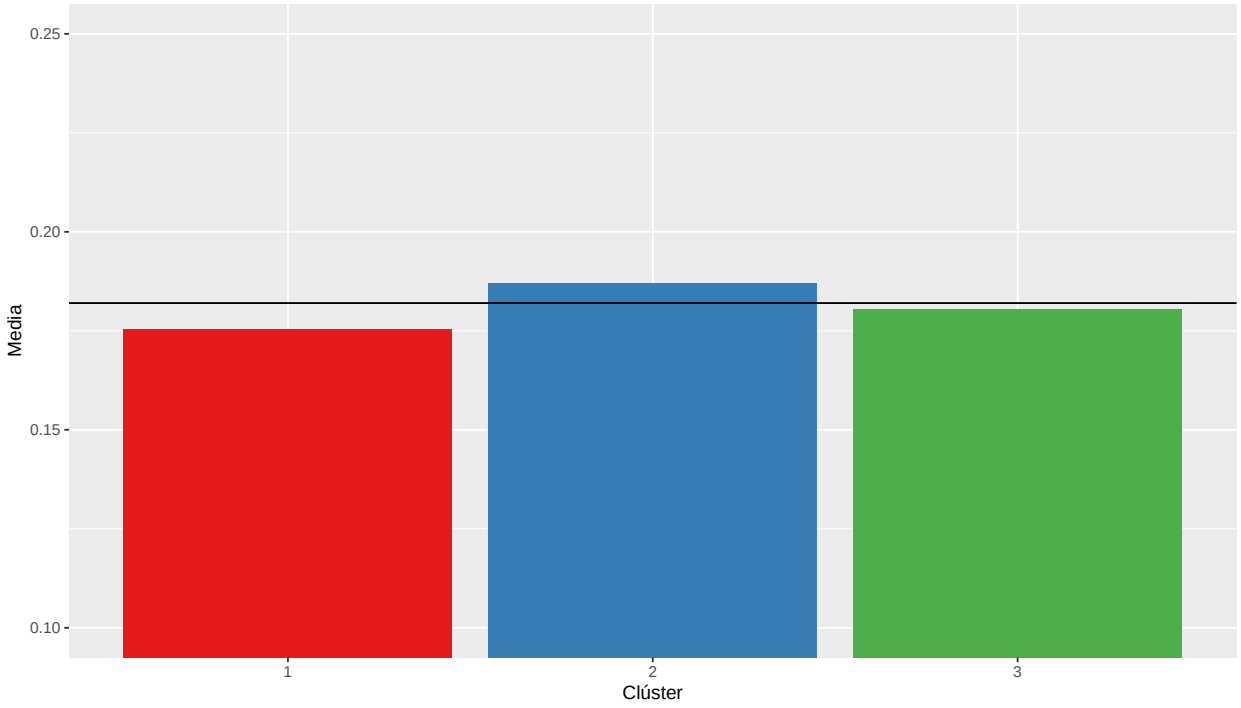
Medias de la Edad por clúster respecto la media global




Medias del Ratio del Importe del préstamo por clúster respecto la media global
5
4
3
2
Media
1
2
3
Clúster


Medias del Ratio de la Anualidad del préstamo por clúster respecto la me
0.060
0.055
0.050
0.045
0.040
Media
1
2
3
Clúster

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



Variables categóricas

Gráfico de la distribución del Género respecto el Clúster

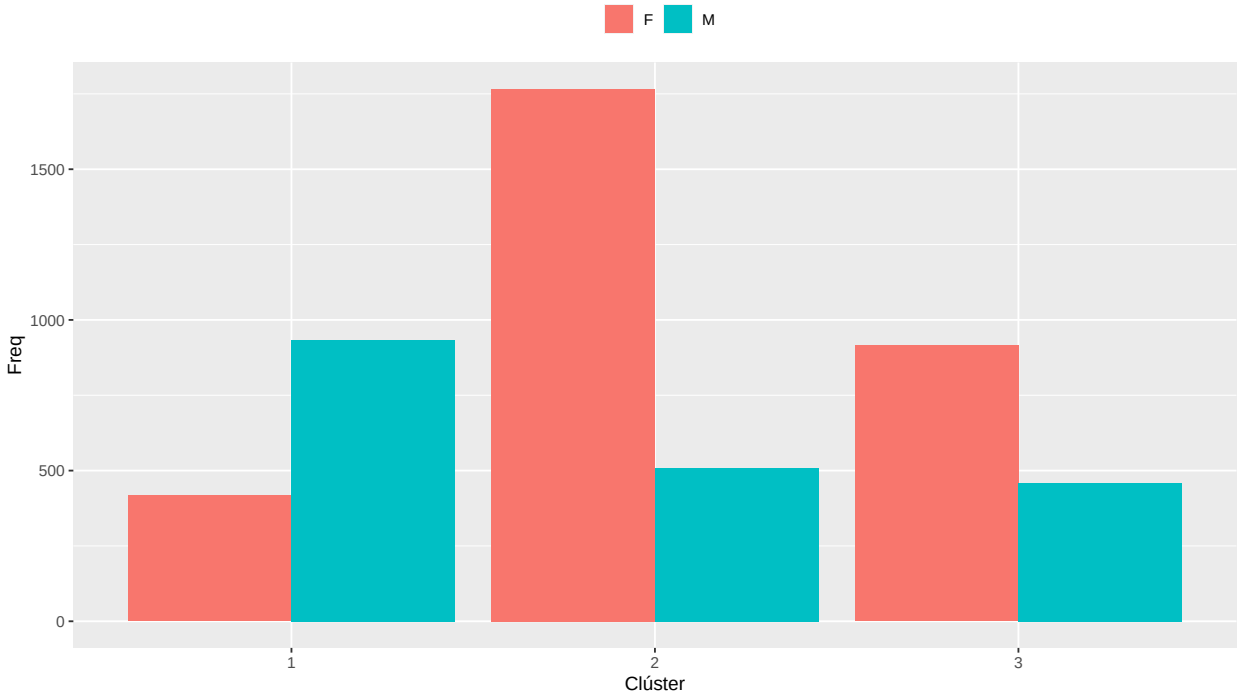


Gráfico de la distribución del Tipo de ingresos respecto el Clúster

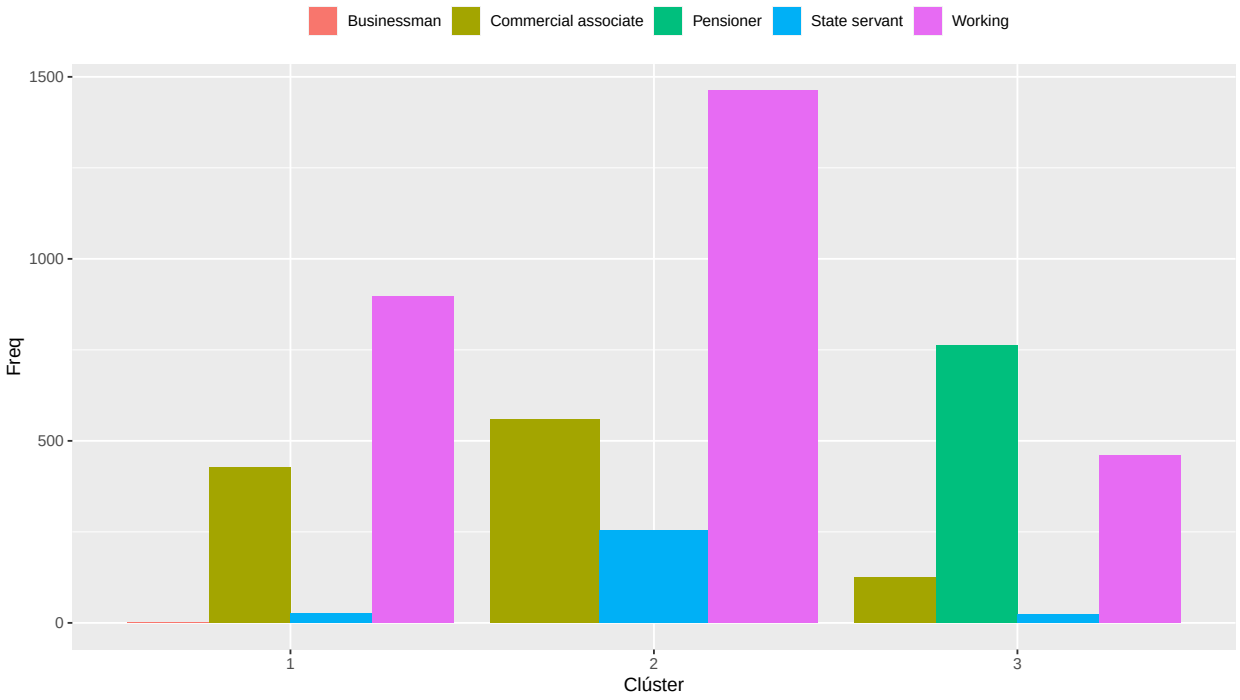


Gráfico de la distribución del Nivel de estudios del cliente respecto el Clúster

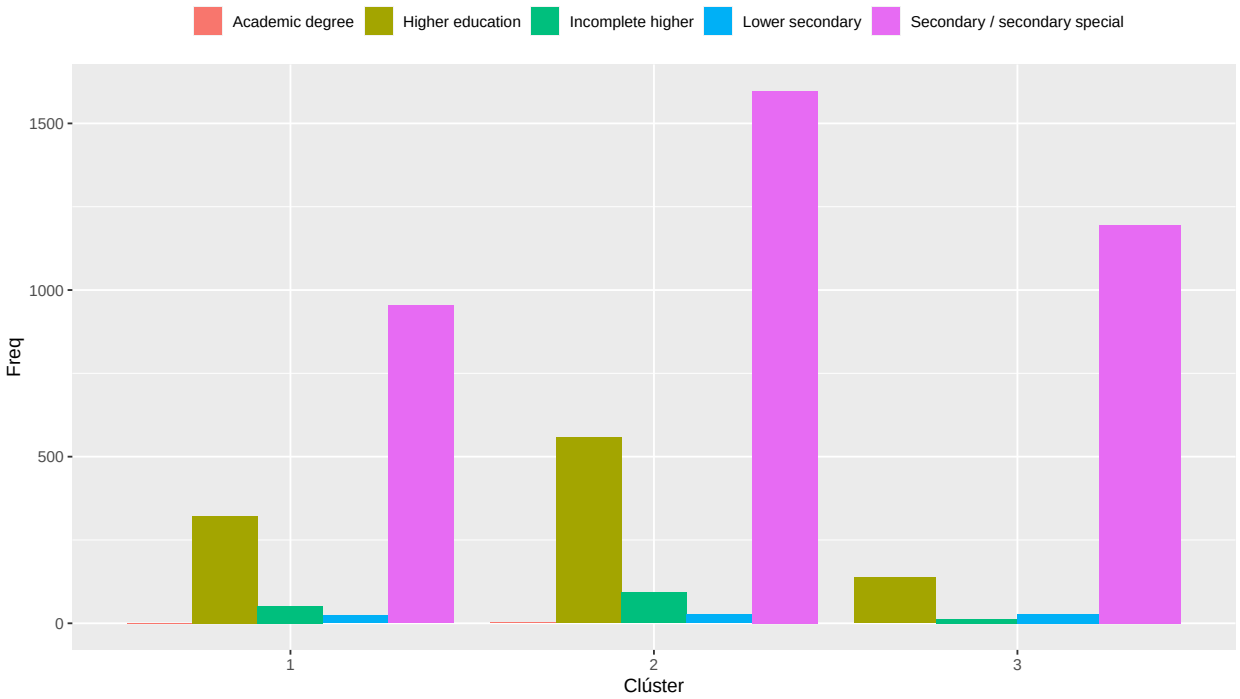


Gráfico de la distribución del Estado civil respecto el Clúster

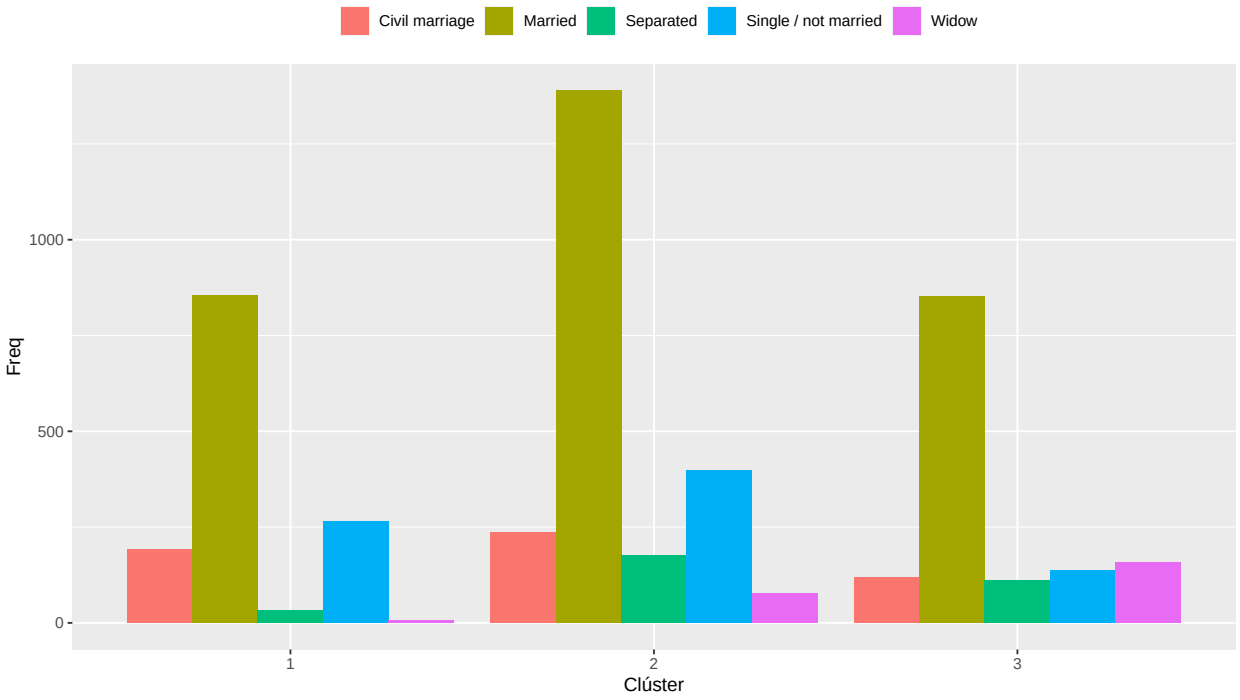


Gráfico de la distribución de la Actividad laboral respecto el Clúster

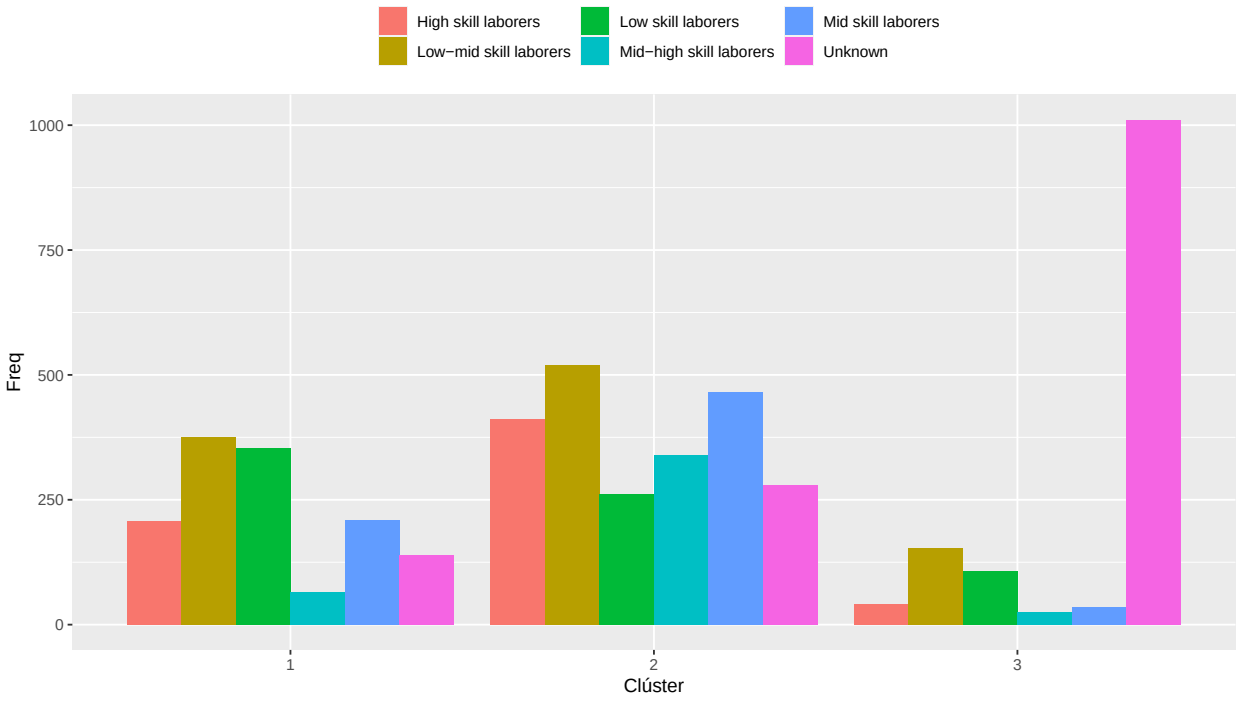


Gráfico de la distribución del Tipo de organización donde trabaja el cliente respecto el Clúster

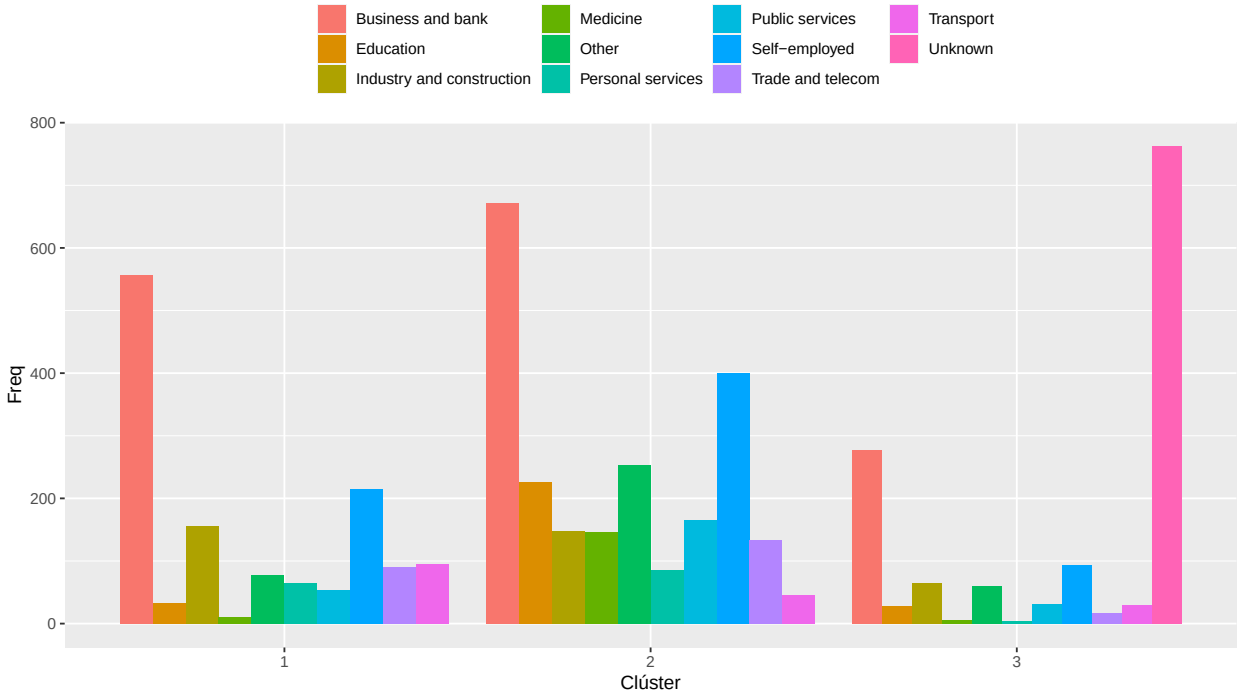
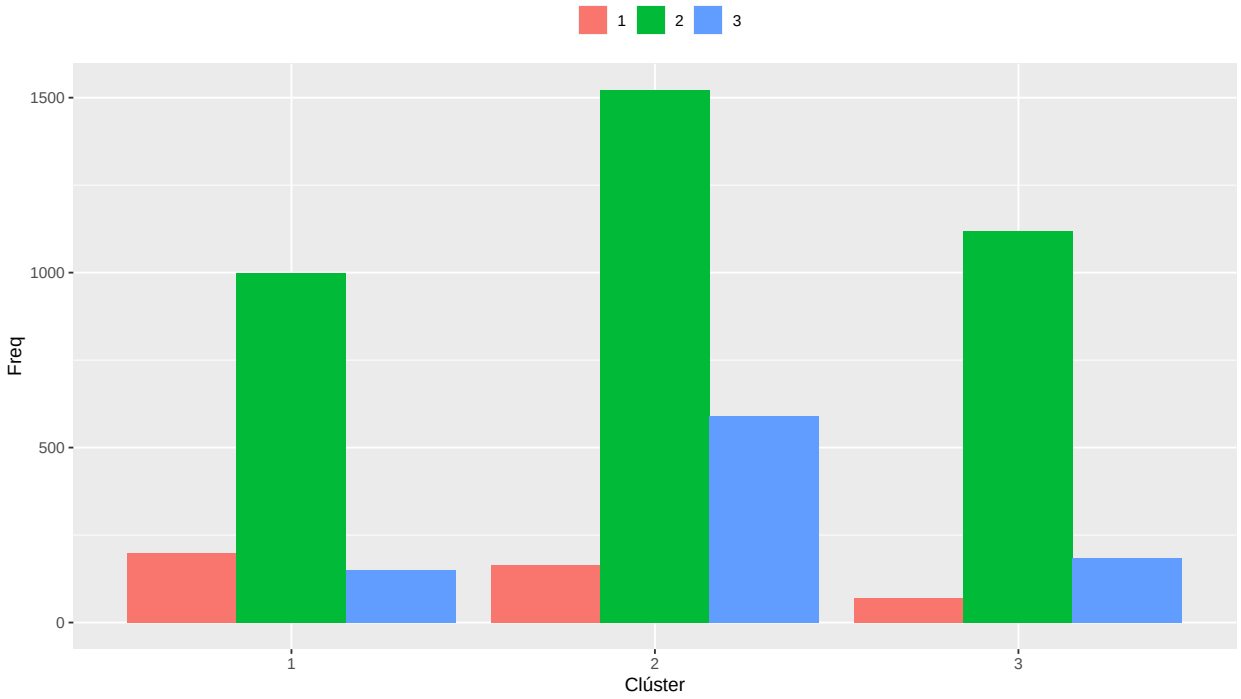


Gráfico de la distribución de la Calificación de la región donde vive el cliente respecto el Clúster



Así pues, tras haber analizado atentamente el resultado ofrecido por el profiling realizado, se han extraído las siguientes conclusiones para cada cluster:

- Cluster 1: Este cluster está formado por la gente con un ratio credit/income menor, además de un DTI ratio más bajo que los otros dos clusters encontrados. Esto nos indica que son personas más responsables con sus cuentas y que piden crédito cuando su situación financiera es positiva, ya que

tienden a endeudarse menos. Además, apreciando el análisis de las variables categóricas, se aprecia que este cluster presenta una mayor concentración de hombres (solteros en su mayoría) y dedicados principalmente al sector de la banca, en su mayoría como comerciales de este mismo sector. De esta forma, es lógico pensar que hagan una buena gestión de sus finanzas personales.

- Cluster 2: Este grupo se caracteriza principalmente por tener un ratio credit/income más alto y un ratio annuity/credit menor. Es decir, es gente que pide préstamos por una cantidad elevada en relación a sus ingresos, pero que generalmente deciden pagarlo a largo plazo. Est hehco, además, va relacionado con que la edad del coche media sea la menor entre los tres grupos: tal vez una parte del préstamo solicitado se ha destinado al coche. Entrando en el análisis de las variables categóricas, se aprecia que en su mayoría son mujeres que ocupan trabajos de gran capital humano (state servant) y residen en lugares con un buen rating por la empresa.

- Cluster 3: Por último, en este cluster se sitúan aquellos ciudadanos con una edad superior, en su gran mayoría pensionistas. Además, poseen coches con más años que el resto y presentan núcleos familiares más reducidos. En este cluster se encuentran la gran mayoría de personas viudas y, en general, el nivel educativo que más predomina es la secundaria.