

Universidad Tecnológica Nacional FRRo



Minería de datos

Comisión 5E05

Trabajo Práctico Integrador

Entrega final

Grupo N° 5

Integrantes:

47066 - Gorosito, Adriel

47447 - Botali, Santiago

Índice

1ra etapa: contexto	5
Problema	5
Objetivos	5
Análisis exploratorio	6
Análisis univariante	6
Análisis de cada variable	7
Comprobicleta	7
EstadoCivil	8
Propietario	9
Género	9
TotalHijos	9
CantAutomoviles	10
Edad	11
IngresoAnual	12
Educación	12
Ocupación	13
Distancia	14
Región	15
Valores nulos	16
Rellenado de valores nulos	17
Variable objetivo	17
Valores atípicos	18
Análisis multivariante	19
Matriz S	19
Matriz R	20
Diagramas de dispersión estratificados	20
Variable de estratificación: Comprobicleta	20
Variable de estratificación: Propietario	21
Variable de estratificación: TotalHijos	21
Variable de estratificación: CantAutomoviles	21
Diagramas de caja estratificados	22
TotalHijos vs IngresoAnual	22
TotalHijos vs Edad	23
CantAutomoviles vs IngresoAnual	24
CantAutomoviles vs TotalHijos	25
CantAutomoviles vs Edad	26
Propietario vs CantAutomoviles	27

Propietario vs IngresoAnual	27
ComproBicicleta vs CantAutomoviles	28
Propietario vs Edad	29
ComproBicicleta vs IngresoAnual	29
ComproBicicleta vs TotalHijos	30
ComproBicicleta vs Edad	31
Técnicas predictivas	32
Árbol de decisión	33
Partición 65 y 35	34
Partición 70 y 30	35
Partición 75 y 25	35
Partición 80 y 20	36
Conclusión	37
Algoritmo de los vecinos más próximos (KNN)	38
Partición 65 y 35	39
Partición 70 y 30	40
Partición 75 y 25	40
Conclusión	41
Análisis discriminante lineal (LDA)	41
Supuestos requeridos	41
No multicolinealidad	41
Prueba de normalidad	42
Matrices de varianza y covarianza iguales	42
Conclusión	43
Partición 65 y 35	43
Partición 70 y 30	43
Partición 75 y 25	44
Conclusión	44
Predicción	44
Análisis de los resultados	45
2da etapa: contexto	51
Problema	51
Objetivos	51
Clustering k-medias	52
Proceso general	52
Variación del parámetro “k”	53
k = 3	53

k = 4	61
k = 5	67
Clustering jerárquico	68
Uso de Python	68
Todos los datos	69
Con estandarización	69
Criterio “distance”	69
Criterio “maxclust”	70
Con normalización	70
Criterio “distance”	71
Criterio “maxclust”	71
Muestra de 100 observaciones	72
Con estandarización	72
Criterio “distance”	72
Criterio “maxclust”	73
Con normalización	73
Criterio “distance”	74
Criterio “maxclust”	74
Conclusión	74
Uso de SPSS Statistics	75
3 clusters	75
4 clusters	76
5 clusters	77
Conclusión	77
Clustering bietápico	78
Todas las variables	79
Todas las variables categóricas	80
Todas las variables numéricas	81
Numéricas sin algunas variables	83
Análisis del mejor modelo	86
Mejor técnica de clustering	91
Detalles	91
Gráficas	92
Mercados objetivos	99
Estructura de datos	99
Análisis de componentes principales	100
Validación de hipótesis	105

1ra etapa: contexto

Problema

El gerente de una empresa que se dedica a la venta de productos para el hogar desea mantener un buen posicionamiento en el mercado. Para ello, elaboró una serie de estrategias comerciales para el siguiente año.

Una de las estrategias planteadas es aumentar la venta de bicicletas gracias a un convenio con otra empresa. Se fabricarán tres tipos de bicicletas:

- Bicicletas para niños (Kinder)
- Bicicletas estándares (Basic)
- Bicicletas deportivas (Sport)

Para poder llevar a cabo esto, el sector de marketing necesita ayuda en la campaña publicitaria por correo electrónico. Para ello, se posee un archivo de 1500 potenciales clientes, sobre los cuales hay que decidir si se le envía o no la publicidad y el contenido del correo.

Objetivos

- Establecer un criterio de elección de cliente potencial.
- Decidir a qué clientes mandarle la publicidad, teniendo en cuenta que es mejor mandarle la publicidad a alguien desinteresado que tener una pérdida de una venta.
- Clasificar a los clientes para determinar el tipo de bicicleta que le podría llegar a interesar, para luego poder realizar marketing personalizado.
- Seleccionar tres países adecuados para comercializar la nueva línea de bicicletas.

Análisis exploratorio

Con los datos que disponemos, a continuación realizaremos un análisis exploratorio para obtener información necesaria para el estudio de la información, utilizando nuestro criterio para reconocer aquellos datos que no son relevantes y aquellos que sí.

El análisis exploratorio está compuesto por dos tipos de análisis: el análisis univariante y el análisis multivariante.

Análisis univariante

El análisis univariante se centra en el estudio, análisis y descripción de una sola variable a la vez. En este enfoque, se analizan las características, distribución y medidas estadísticas de una variable individual, como la media, la mediana, la desviación estándar, los percentiles, etc. El análisis univariante proporciona información sobre una sola dimensión de los datos y no tiene en cuenta la relación con otras variables.

El objetivo principal del análisis univariante es comprender y describir las características de una variable en particular, identificar patrones, tendencias y posibles valores atípicos. Es útil para obtener información básica sobre una variable y resumir sus propiedades estadísticas.

Algunos ejemplos de técnicas utilizadas en este tipo de análisis incluyen histogramas, gráficos de barras, diagramas de caja, medidas de tendencia central y de dispersión, pruebas de hipótesis univariadas, entre otros.

Para comenzar con el análisis univariante, generamos una tabla que nos permita ver, de forma general, todas las variables. Esta tabla nos muestra, para cada variable, el nombre, la cantidad de entradas no nulas y el tipo de dato. Además, informa de manera general, el rango de los datos (el cual es de 6400) y la cantidad de tipos de datos.

RangeIndex: 6400 entries, 0 to 6399			
Data columns (total 21 columns):			
#	Column	Non-Null Count	Dtype
0	IdCliente	6400 non-null	int64
1	IdCiudad	6400 non-null	int64
2	Nombre	6400 non-null	object
3	Apellido	6400 non-null	object
4	FechaNacimiento	6400 non-null	object
5	EstadoCivil	6400 non-null	object
6	Genero	6400 non-null	object
7	Email	6400 non-null	object
8	IngresoAnual	6390 non-null	float64
9	TotalHijos	6400 non-null	int64
10	Educacion	6400 non-null	object
11	Ocupacion	6400 non-null	object
12	Propietario	6400 non-null	int64
13	CantAutomoviles	6400 non-null	int64
14	Direccion	6400 non-null	object
15	Telefono	6400 non-null	object
16	FechaPrimeraCompra	6400 non-null	object
17	Distancia	6400 non-null	object
18	Region	6400 non-null	object
19	Edad	6400 non-null	int64
20	ComproBicicleta	6400 non-null	int64

dtypes: float64(1), int64(7), object(13)

Posteriormente, generamos una matriz donde las columnas son las variables numéricas de interés y las filas son estadísticos.

	IngresoAnual	TotalHijos	Propietario	CantAutomoviles	Edad	ComproBicicleta
count	6390.000000	6400.000000	6400.000000	6400.000000	6400.000000	6400.000000
mean	57532.081377	1.894844	0.676562	1.547656	51.195469	0.394375
std	32331.969091	1.630993	0.467825	1.147060	11.517698	0.488754
min	10000.000000	0.000000	0.000000	0.000000	32.000000	0.000000
max	170000.000000	5.000000	1.000000	4.000000	102.000000	1.000000

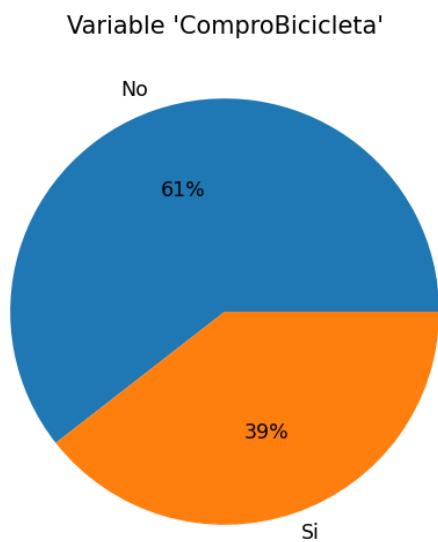
En las filas, **count** es el total de datos *no nulos* analizados, **mean** es la media de los datos, **std** es el desvío estándar (distancia promedio de los datos a la media), **min** es el valor mínimo de la columna (por ejemplo, el ingreso anual mínimo percibido es de \$10000) y **max** el valor máximo de la misma (el ingreso anual máximo percibido es de \$170000).

Esto nos brinda una información general de los clientes ya percibidos, que nos servirá para proceder con nuestro análisis.

Prosiguiendo con el análisis univariante, decidimos generar una gráfica para cada variable, a fin de obtener un poco más de información de cada una. La gráfica varía si se trata de una variable numérica o categórica.

Análisis de cada variable

ComproBicicleta

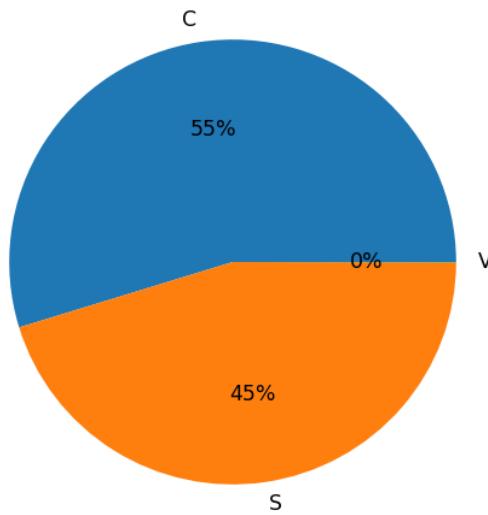


- Clientes que no compraron bicicleta (61%): 3876.
- Clientes que compraron bicicleta (39%): 2524.

Si bien los clientes que compraron bicicletas son muchos, siendo menores que los que no compraron. Casi el 40% compraron bicicletas, mientras que casi el 60% no lo hizo.

EstadoCivil

Variable 'EstadoCivil'

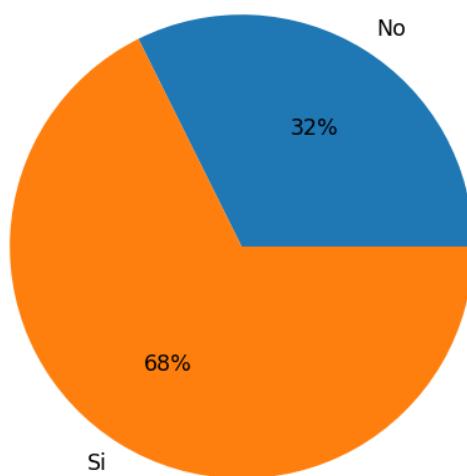


- Clientes con estado civil “Casado” (55%): 3504.
- Clientes con estado civil “Soltero” (45%): 2894.
- Clientes con estado civil “Viudo” (0%): 2.

Se puede observar que los casados son el grupo mayoritario con un 55%, luego con una diferencia del 10% se encuentran los clientes solteros y en una proporción casi inexistente encontramos a los clientes viudos.

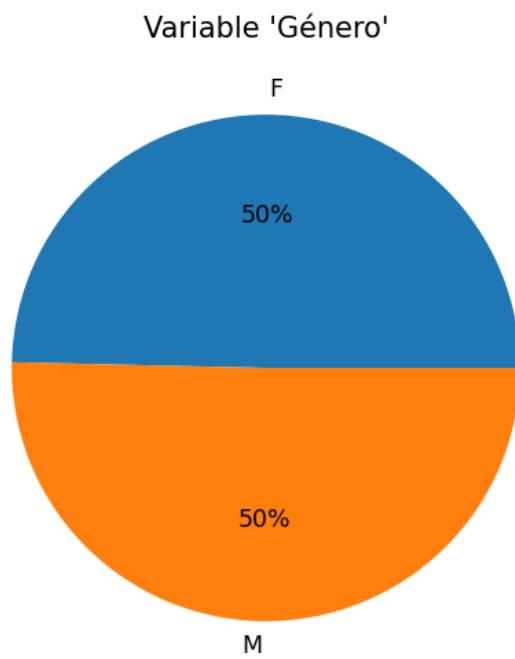
Propietario

Variable 'Propietario'



- Clientes que son propietarios (68%): 4330.
- Clientes que no son propietarios (32%): 2070.

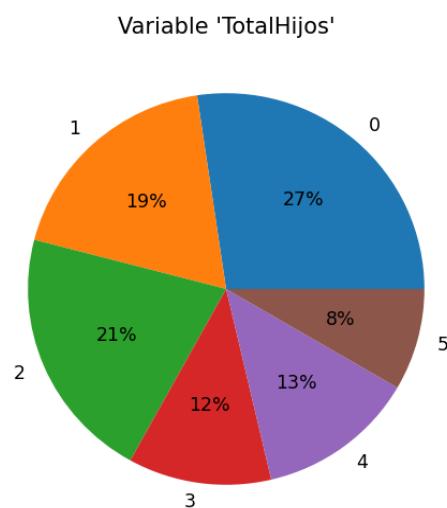
Género



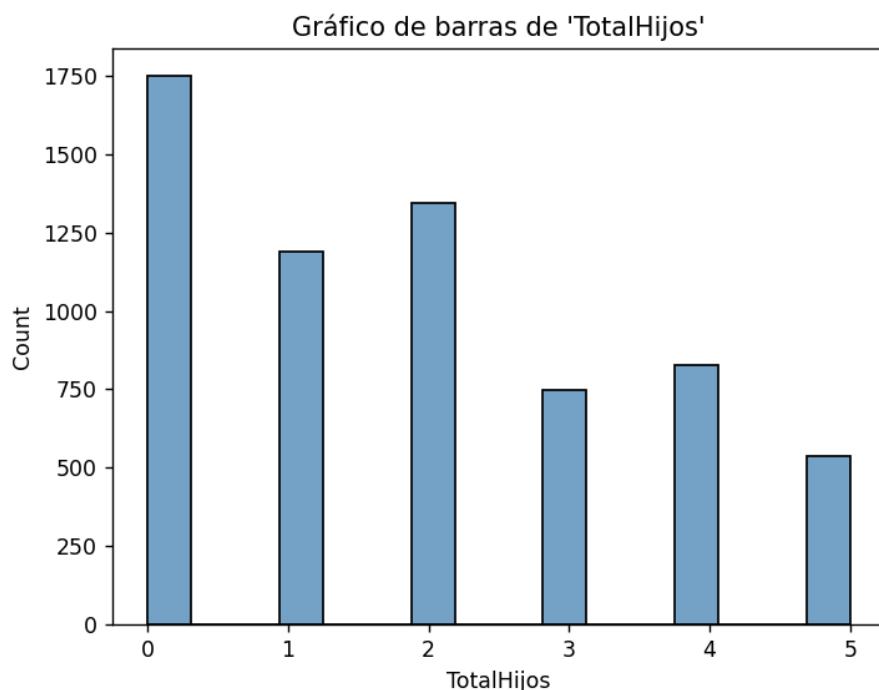
- Clientes de género “Masculino” (50%): 3223.
- Clientes de género “Femenino” (50%): 3117.

La única observación que vemos es que la cantidad de personas del género masculino y femenino son casi la misma.

TotalHijos

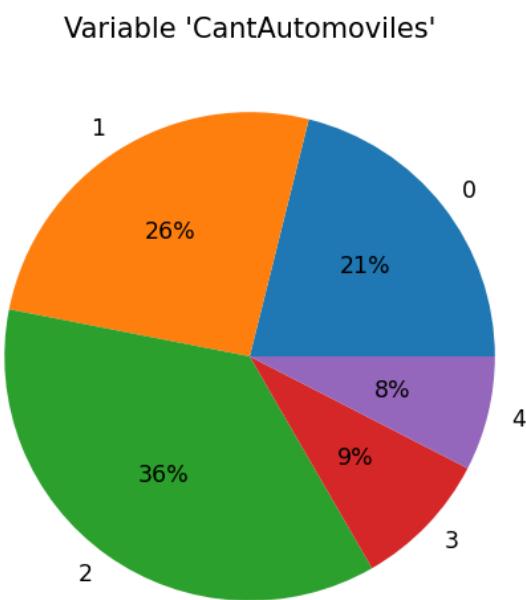


- Clientes sin hijos (27%): 1750.
- Clientes con un hijo (19%): 1191.
- Clientes con dos hijos (21%): 1345.
- Clientes con tres hijos (12%): 748.
- Clientes con cuatro hijos (13%): 828.
- Clientes con cinco hijos (8%): 538.

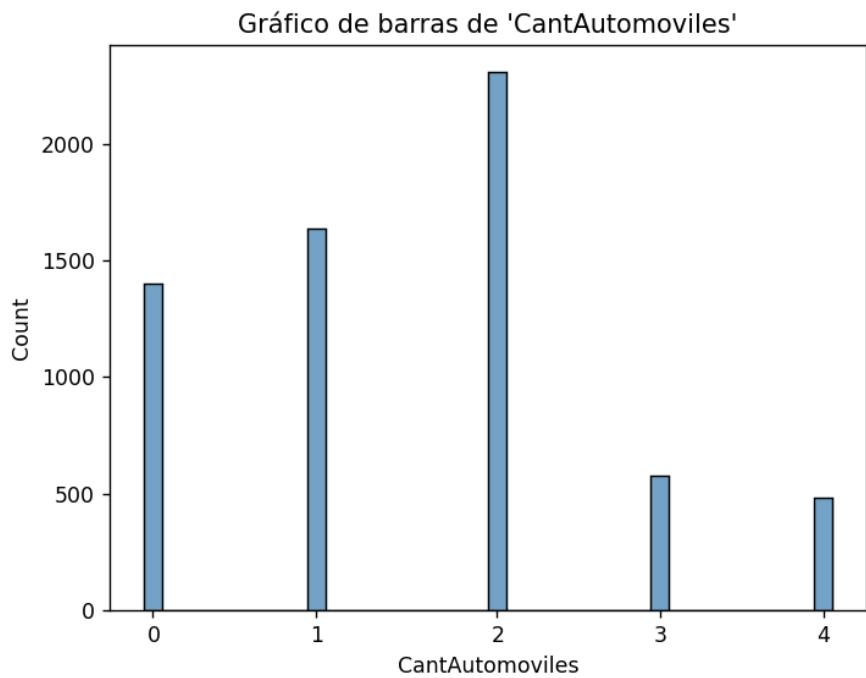


Haciendo un análisis rápido, observamos que el 27% de clientes no tienen hijos, mientras que el 73% tienen uno o más.

CantAutomoviles

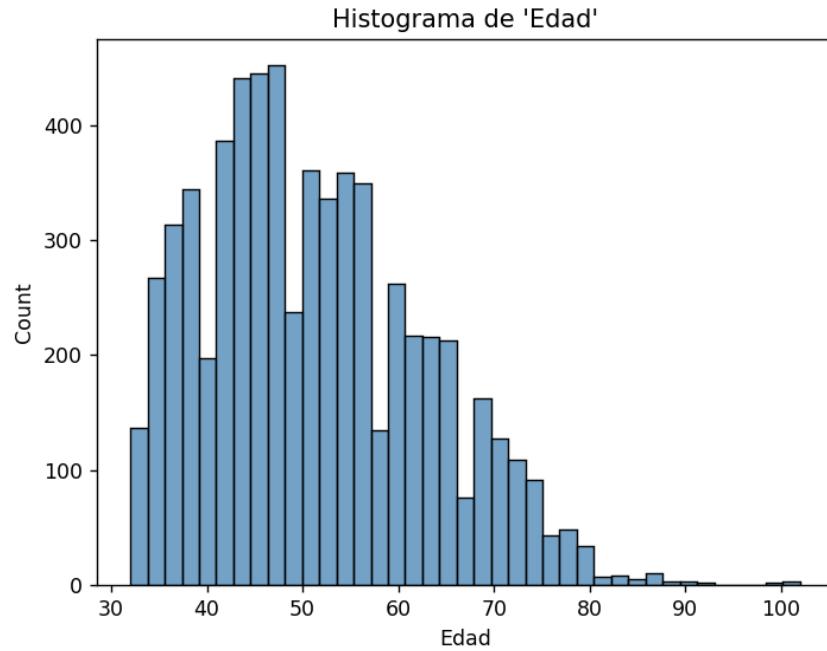


- Clientes con ningún auto (21%): 1399.
- Clientes con un auto (26%): 1635.
- Clientes con dos autos (36%): 2308.
- Clientes con 3 autos (9%): 578.
- Clientes con 4 autos (8%): 480.



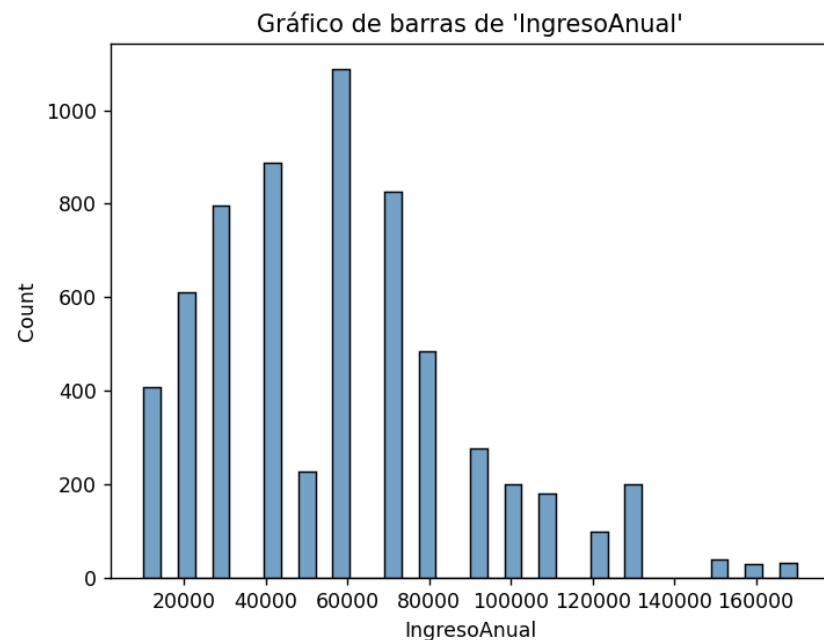
Observamos que hay más clientes con dos autos que clientes sin autos, lo que nos genera un poco de ruido, pero es algo que podría pasar en la realidad.

Edad



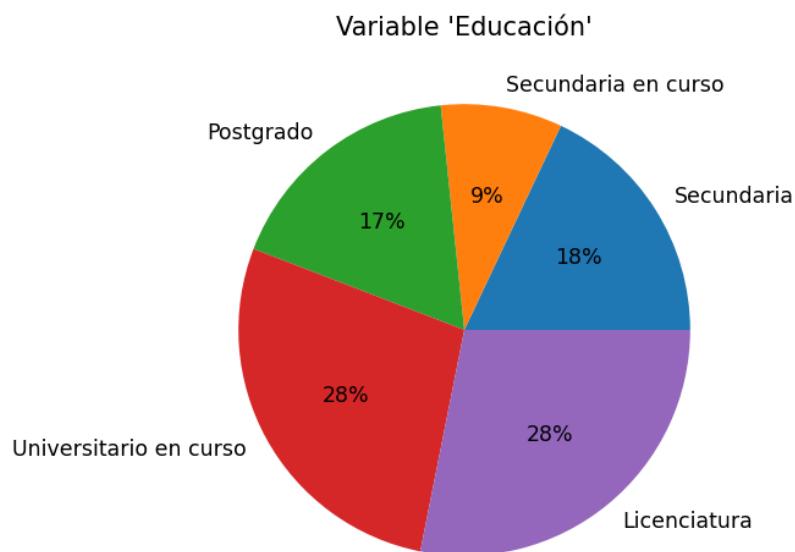
Viendo la gráfica, no se puede decir mucho, salvo que se puede notar que puede que hayan valores atípicos (luego se realizará el análisis correspondiente).

IngresoAnual



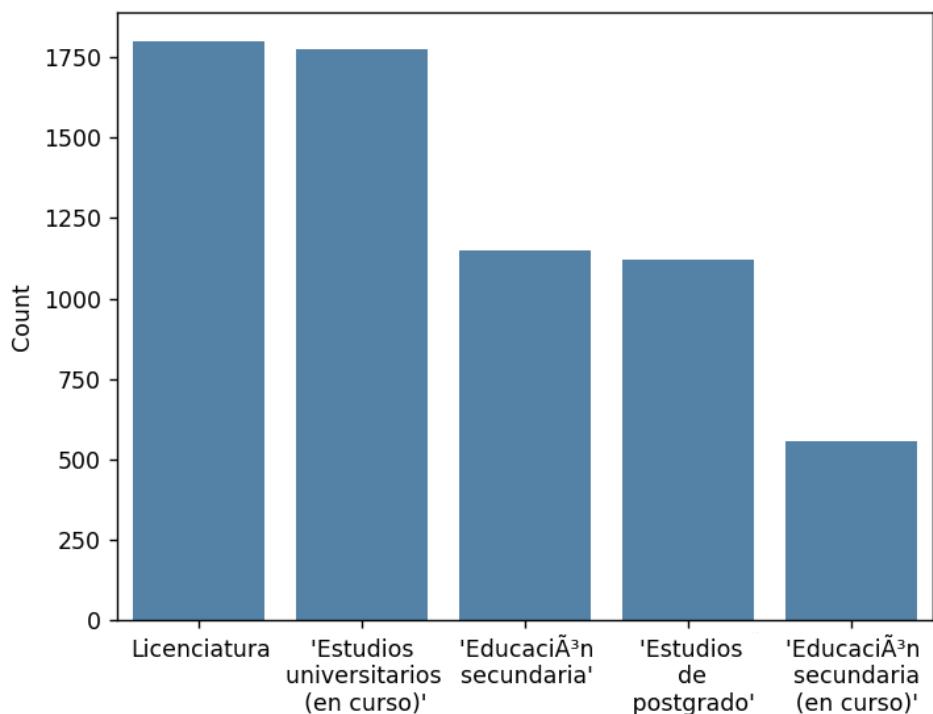
Vemos que, al igual que la Edad, puede que hayan valores atípicos.

Educación



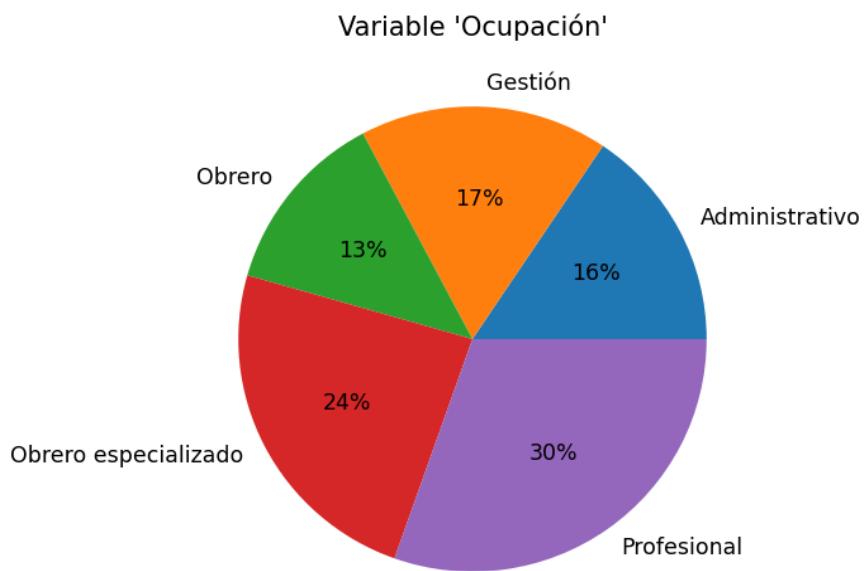
- Clientes con educación “Postgrado” (17%): 1119.
- Clientes con educación “Secundaria en curso” (9%): 557.
- Clientes con educación “Secundaria” (18%): 1150.
- Clientes con educación “Universitario en curso”(28%): 1774.
- Clientes con educación “Licenciatura” (28%): 1800.

Gráfico de barras de la educación

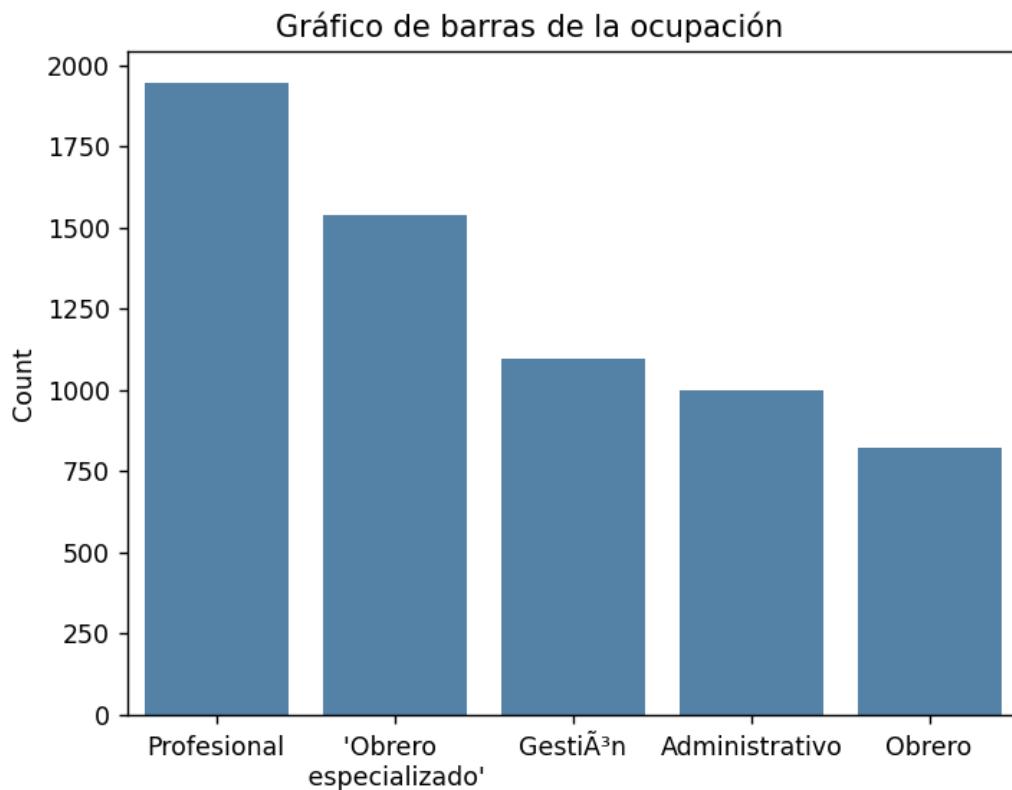


No observamos nada interesante.

Ocupación

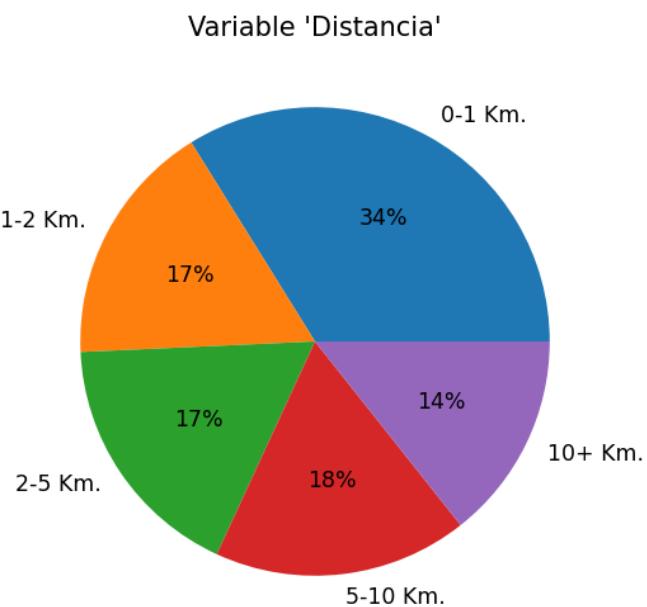


- Cliente con ocupación “Gestión” (17%): 1098.
- Cliente con ocupación “Obrero” (13%): 821.
- Cliente con ocupación “Obrero especializado” (24%): 1357.
- Cliente con ocupación “Administrativo” (16%): 998.
- Cliente con ocupación “Profesional” (30%): 1946.



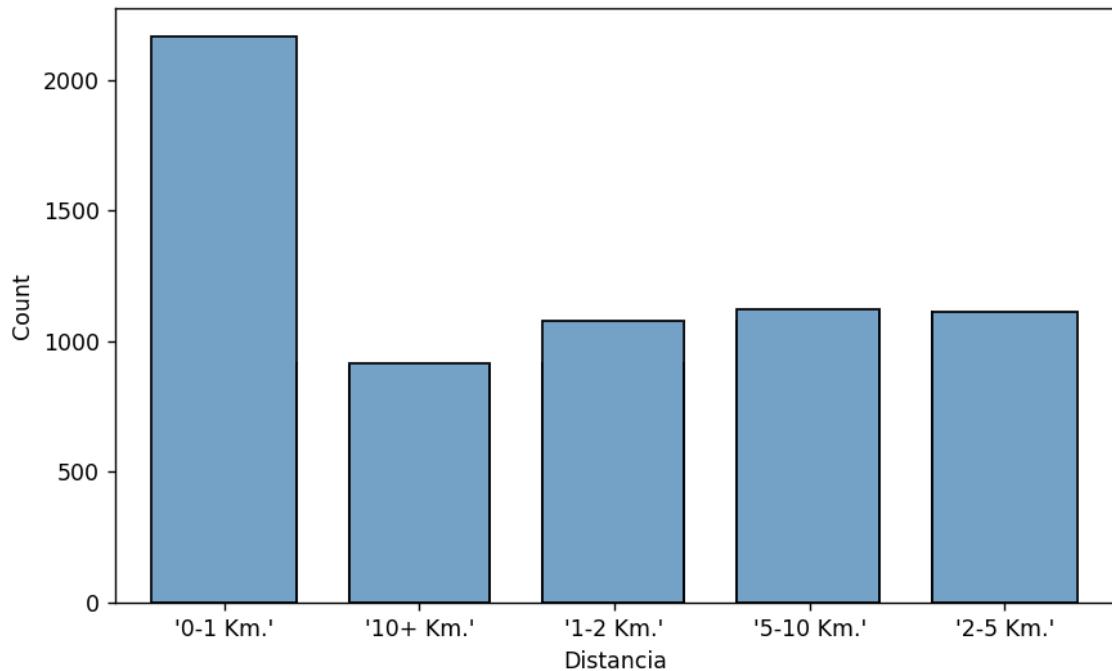
Al igual que para la educación, no encontramos nada interesante.

Distancia



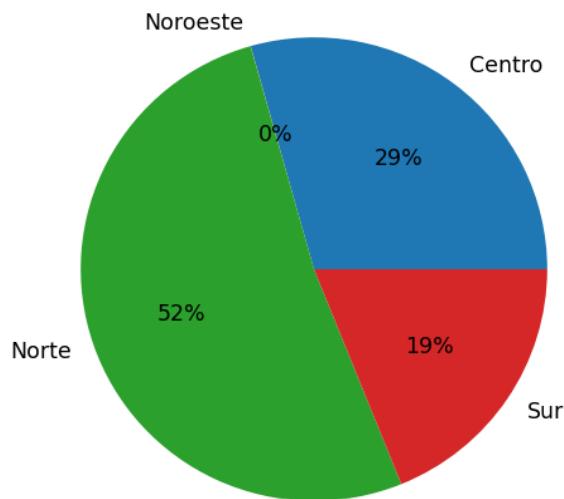
- Clientes que viven entre 0-1 Km de distancia (34%): 2166.
- Clientes que viven entre 1-2 Kms de distancia (17%): 1080.
- Clientes que viven entre 2-5 Kms de distancia (17%): 1114.
- Clientes que viven entre 5-10 Kms de distancia (18%): 1122.
- Clientes que viven a más de 10 Kms de distancia (14%): 918.

Gráfico de barras de 'Distancia'

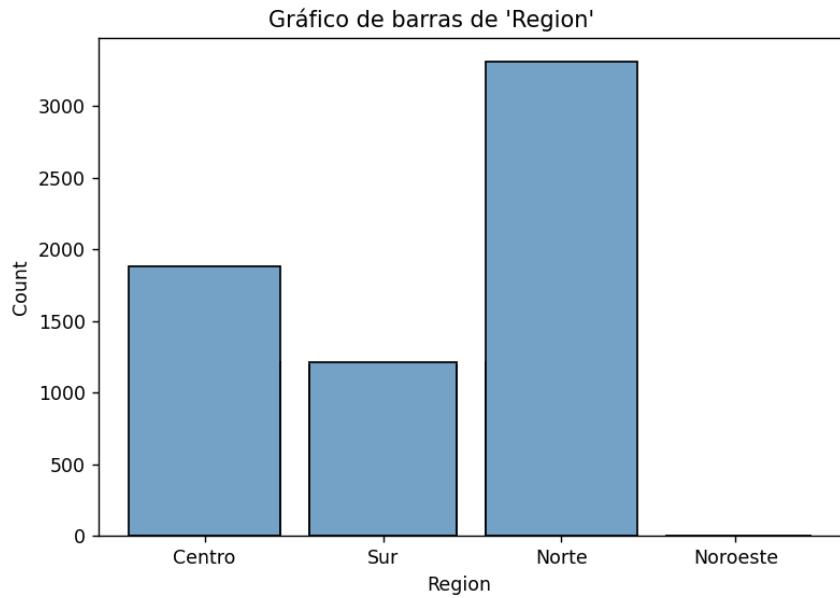


Región

Variable 'Región'



- Clientes que viven en la región “Norte” (52%): 3310.
- Clientes que viven en la región “Sur” (19%): 1209.
- Clientes que viven en la región “Centro” (29%): 1880.
- Clientes que viven en la región “Noroeste” (0%): 1.



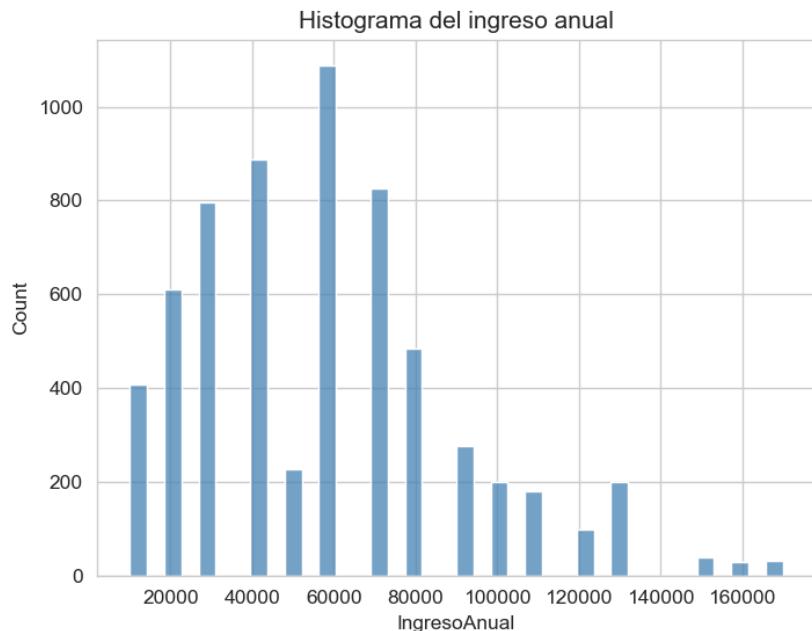
Se puede ver que el valor “Noroeste” es casi nulo (hay solamente un registro que adopta dicho valor) por lo que se podría considerar como una anomalía.

Valores nulos

El primer desperfecto que encontramos, es que en **IngresoAnual** hay 6390 datos no nulos (como el rango es de 6400, entonces hay 10 valores nulos). Para este caso, hay dos formas de proceder: dejar los valores nulos o rellenarlos. Nosotros elegimos rellenarlos debido a que deseamos que todas las variables tengan la misma cantidad de datos.

Rellenado de valores nulos

Para llenar estos valores, podemos utilizar la media o la mediana. En general, se recomienda utilizar la media si los datos siguen una distribución normal. Para averiguarlo, generamos un histograma para analizar la distribución de los datos.



Como podemos ver, la distribución no es normal, sino que es sesgada a la derecha. Luego, decidimos imputar los valores nulos con la mediana.

En las siguientes tablas se pueden ver el antes y después de la imputación:

	IngresoAnual	IngresoAnual
count	6390.00000	6400.00000
mean	57532.081377	57535.937500
std	32331.969091	32306.842992
min	10000.000000	10000.000000
max	170000.000000	170000.000000

Observamos que varía sólo la media y el desvío; el mínimo, máximo y los cuartiles siguen siendo iguales (esto es así ya que se rellena con la mediana). No consideramos que el hecho de imputar los datos haya sido factible debido a la casi insignificante variabilidad que ocasionó, pero si lo consideramos necesario para obtener un resultado más específico y completo del estudio.

Variable objetivo

Analizaremos la columna “ComproBicicleta”. Esta es la variable que interesa estudiar, ya que parecería informar si el cliente compró una bicicleta anteriormente o no.

Generando algunos de los valores de esta variable, nos encontramos que se trata de una variable binaria (es decir, toma solo dos valores, 0 si es falso y 1 si es verdadero).

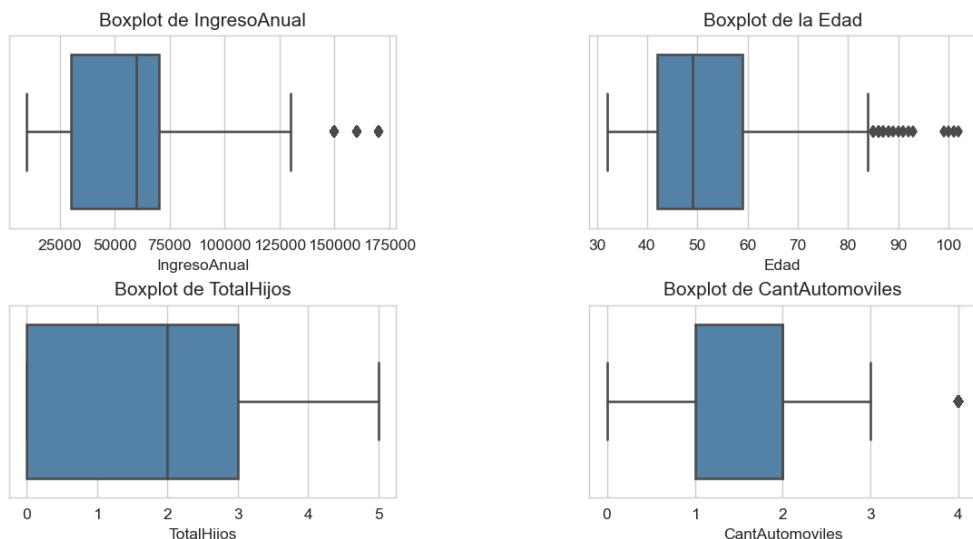
ComproBicicleta
0 3876
1 2524

Por lo tanto, según la tabla, 3876 personas no compraron bicicleta en el pasado y 2524 sí lo hicieron.

Observando la media e interpretando, el 39% de las personas compraron bicicletas (es decir, 2524 personas). Ya que se trata de una variable binaria, no tiene sentido analizar el desvío estándar. Esto es debido a que, al tomar solo dos valores, no hay variabilidad en los datos.

Valores atípicos

Para analizar los valores atípicos, generamos cuatro diagramas de caja y bigotes para las columnas de tipo numérico:



En los diagramas de “IngresoAnual”, “Edad” y “CantAutomoviles” observamos valores atípicos. Decidimos que los vamos a tener en cuenta para el análisis en lugar de imputarlos o eliminarlos, esto ya que estos valores nos van a dar información sobre situaciones excepcionales para el estudio. Además, al tenerlos en cuenta podemos ver cómo afectan la forma y la distribución de los datos. También, nuestra conclusión será más precisa al tener en cuenta todos los datos y obtendremos una imagen más precisa de lo que estamos buscando, en este caso potenciales clientes para venderles bicicletas.

Análisis multivariante

El análisis multivariante involucra el estudio simultáneo de dos o más variables para analizar las relaciones y las interacciones entre ellas. En este enfoque, se examina cómo las variables se relacionan entre sí y cómo influyen mutuamente.

El análisis multivariante permite comprender la complejidad de los datos al considerar múltiples dimensiones. Se utiliza para identificar patrones, asociaciones y dependencias entre las variables, así como para realizar predicciones y modelar fenómenos más complejos.

Ejemplos de técnicas utilizadas en el análisis multivariante incluyen regresión lineal y logística, análisis de componentes principales, análisis de clusters, análisis discriminante, análisis de correspondencias, análisis factorial, entre otros.

Matriz S

Decidimos comenzar el análisis multivariante generando una matriz S para ver la correlación entre las variables. La matriz S, también conocida como matriz de dispersión, es una matriz cuadrada que se utiliza para estudiar las relaciones entre pares de variables numéricas no binarias.

Para generar la matriz, decidimos no tener en cuenta las variables “Propietario” y “ComproBicicleta” (por ser binarias) y las variables “IdCliente” e “IdCiudad” por tratarse de identificadores (ya que cada valor es único y no se repite, no hay variabilidad en la variable, por lo que no aporta valor en términos de patrones). Tampoco tuvimos en cuenta las variables categóricas, por lo que terminamos trabajando con sólo cuatro variables: “IngresoAnual”, “TotalHijos”, “CantAutomoviles” y “Edad”.

	IngresoAnual	TotalHijos	CantAutomoviles	Edad
IngresoAnual	1.000000	0.222296	0.469289	0.153101
TotalHijos	0.222296	1.000000	0.272527	0.495425
CantAutomoviles	0.469289	0.272527	1.000000	0.169977
Edad	0.153101	0.495425	0.169977	1.000000

En la matriz S, un valor cercano a 1 o -1 indica una gran relación entre el par de variables.

- Si el valor es cercano a 1, significa que si una variable aumenta, la otra tiende a aumentar.
- Si el valor es cercano a -1, significa que si una variable aumenta, la otra tiende a disminuir.

Además, si el valor absoluto del valor obtenido es mayor o igual a 0.75, entonces se considera que las variables se correlacionan entre sí.

Las variables que más se relacionan entre sí son “Edad” con “TotalHijos” e “IngresoAnual” con “CantAutomoviles”. Sin embargo, los valores no son cercanos a 1 ni a -1 (0.495 y 0.469, respectivamente). Por lo tanto, se puede decir que no hay correlación directa entre las variables.

Matriz R

La matriz R se utiliza para calcular la matriz de covarianza entre las variables numéricas. A diferencia de la matriz S (que mide la correlación de las variables), la covarianza no está normalizada y su magnitud depende de las unidades de las variables involucradas. Por lo tanto, no se puede utilizar para comparar dos variables que tienen unidades distintas (como lo es este caso). Sin embargo, optamos por generarla igual.

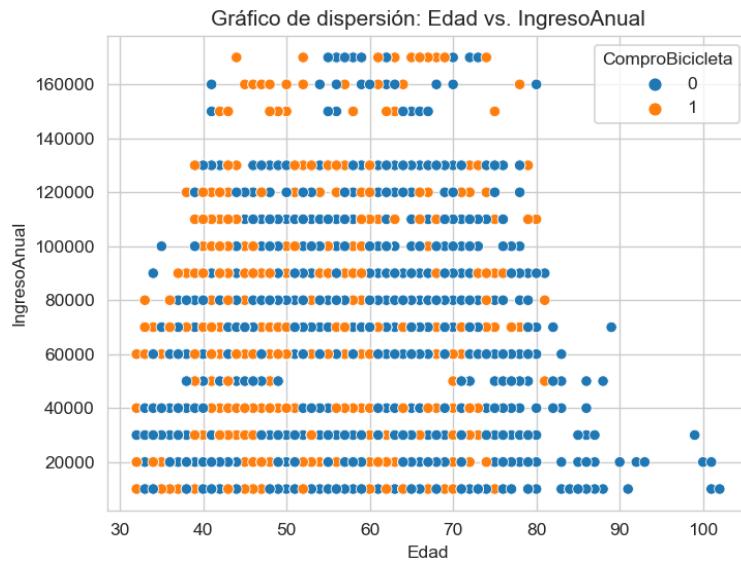
	IngresoAnual	TotalHijos	CantAutomoviles	Edad
IngresoAnual	1.045356e+09	11725.856084	17394.779662	57017.474038
TotalHijos	1.172586e+04	2.660139	0.509857	9.306696
CantAutomoviles	1.739478e+04	0.509857	1.315747	2.245645
Edad	5.701747e+04	9.306696	2.245645	132.657363

Cuando vemos los resultados obtenidos, confirmamos que para este conjunto de datos no sirve de mucho hacer esta matriz debido a que las unidades que participan poseen unidades diferentes. Concluimos que con esta matriz no podemos hacer ningún análisis que sirva.

Diagramas de dispersión estratificados

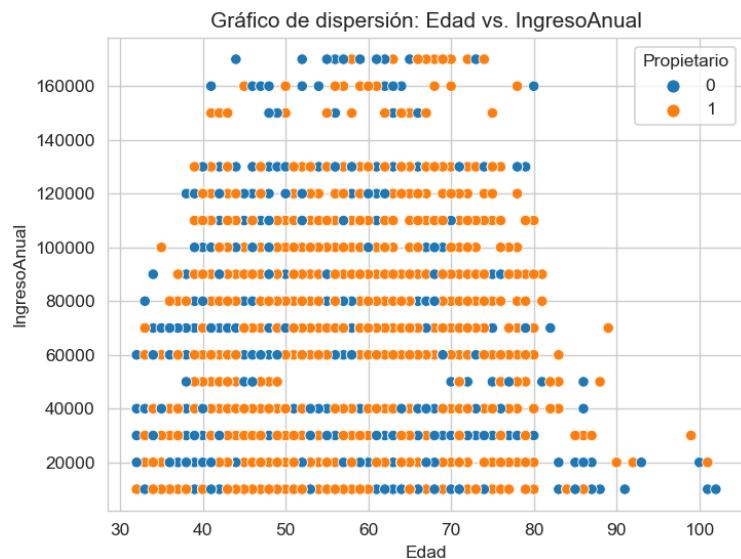
Los diagramas de dispersión estratificados se utilizan para variables continuas. Sin embargo, en nuestro conjunto de datos, las variables, en su mayoría, son de tipo discreto. Decidimos generar este tipo de diagrama solo para Edad vs. IngresoAnual, variando la variable de estratificación. A pesar de que la variable “Edad” no es continua, debido al rango de valores que adopta, sirve para generar las gráficas.

Variable de estratificación: Comprobicicleta



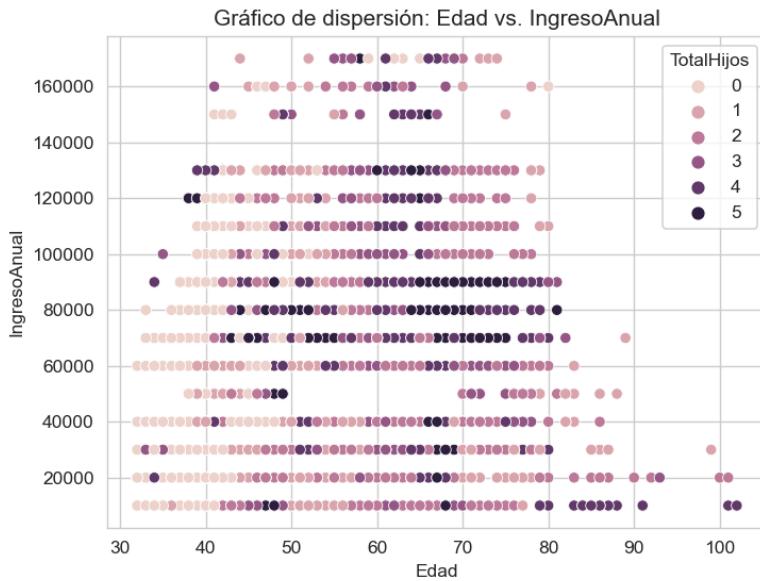
En esta gráfica realmente no vemos nada. Los datos se encuentran de forma muy dispersa. Podríamos decir que la mayoría de clientes que compraron bicicletas tienen menos de 80 años, pero siendo sinceros no es una información muy relevante.

Variable de estratificación: Propietario



En esta otra gráfica, al igual que la anterior, no observamos nada.

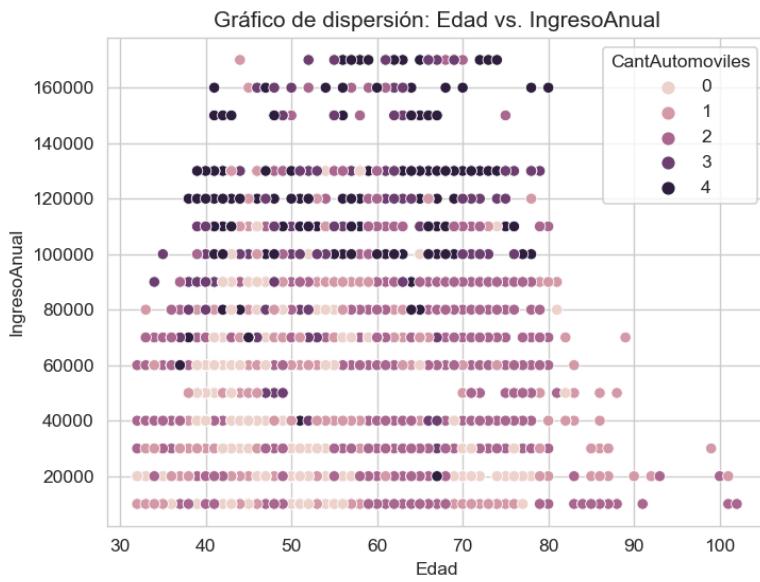
Variable de estratificación: TotalHijos



En esta gráfica vemos dos cosas:

- Las personas sin hijos tienen entre 30 y 50 años.
- Las personas que tienen 5 hijos tienen un ingreso anual que va desde los \$70.000 hasta los \$90.000.

Variable de estratificación: CantAutomoviles



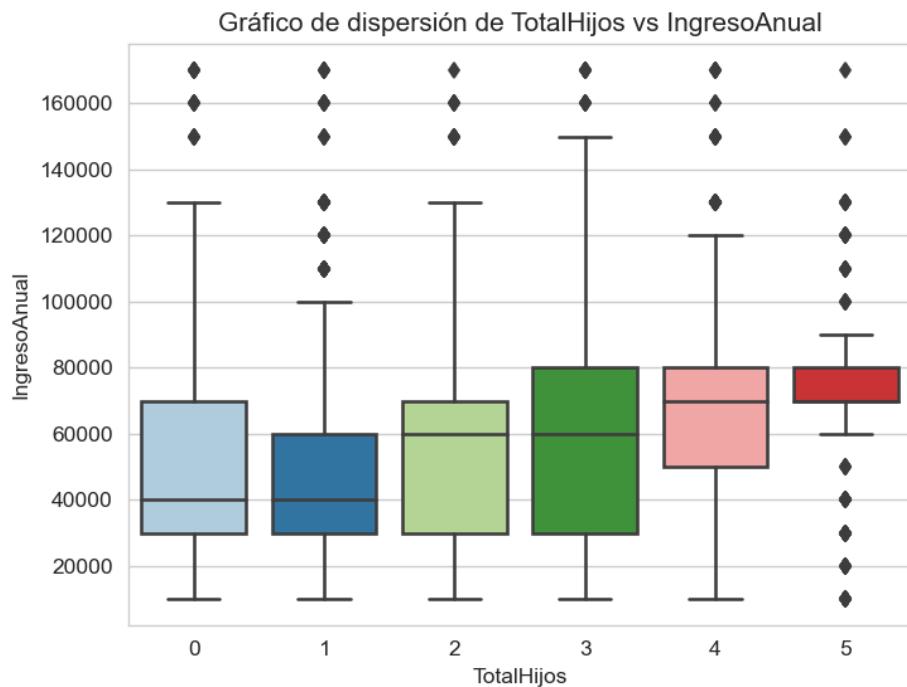
En esta gráfica vemos:

- Las personas que tienen 4 autos tienen un ingreso anual igual o mayor a \$100.000 y tienen entre 40 y 80 años.
- Las personas que tienen 2 hijos o menos tienen un ingreso anual menor a \$100.000.

Diagramas de caja estratificados

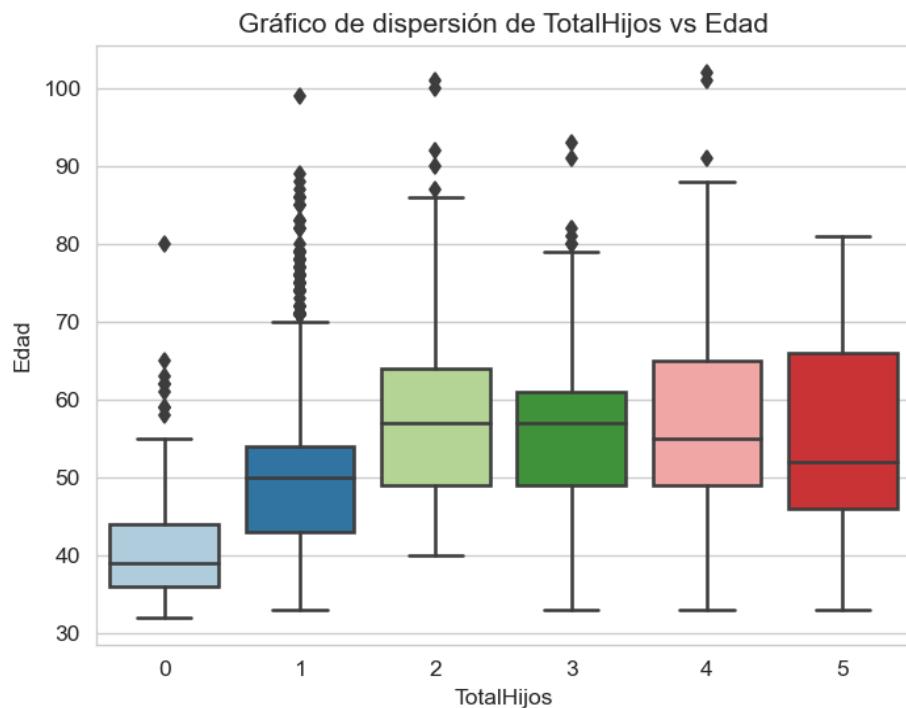
El diagrama de caja estratificado es una representación gráfica que se utiliza en estadística para comparar y analizar múltiples conjuntos de datos de manera simultánea. Se crean múltiples cajas para cada grupo o estrato de datos, lo que permite realizar comparaciones entre ellos.

TotalHijos vs IngresoAnual



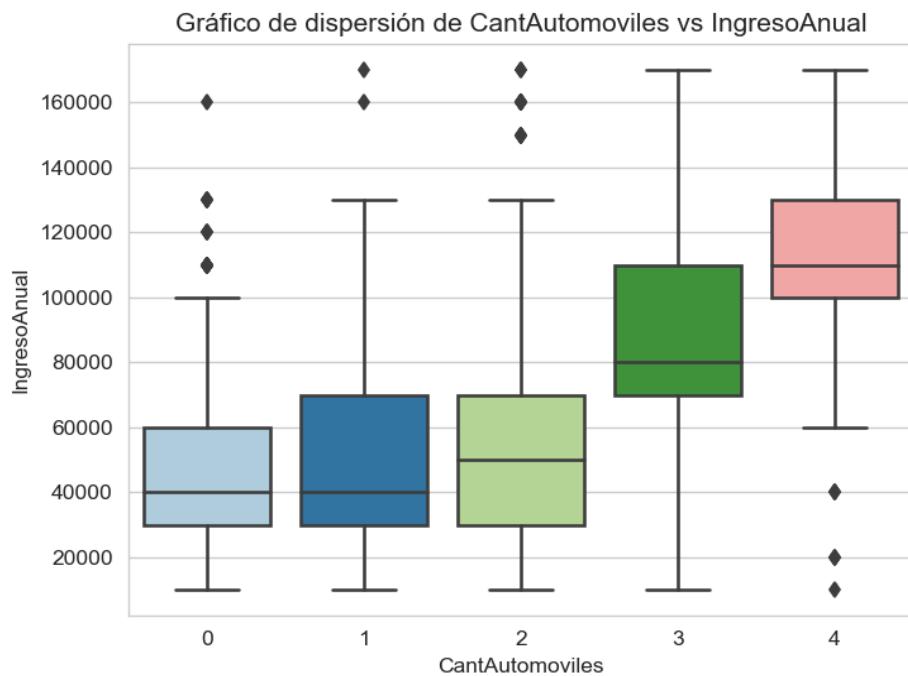
- La mediana es igual para los que tienen un hijo o ninguno
- La mediana aumenta para los que tienen 2 o 3 hijos
- La mediana aumenta más para los que tienen 4 o 5 hijos.
- Hay un rango considerablemente menor de ingresos anuales para los que tienen 5 hijos.
- Hay valores de ingreso anual por encima de la media para todas las cantidades de hijos.

TotalHijos vs Edad



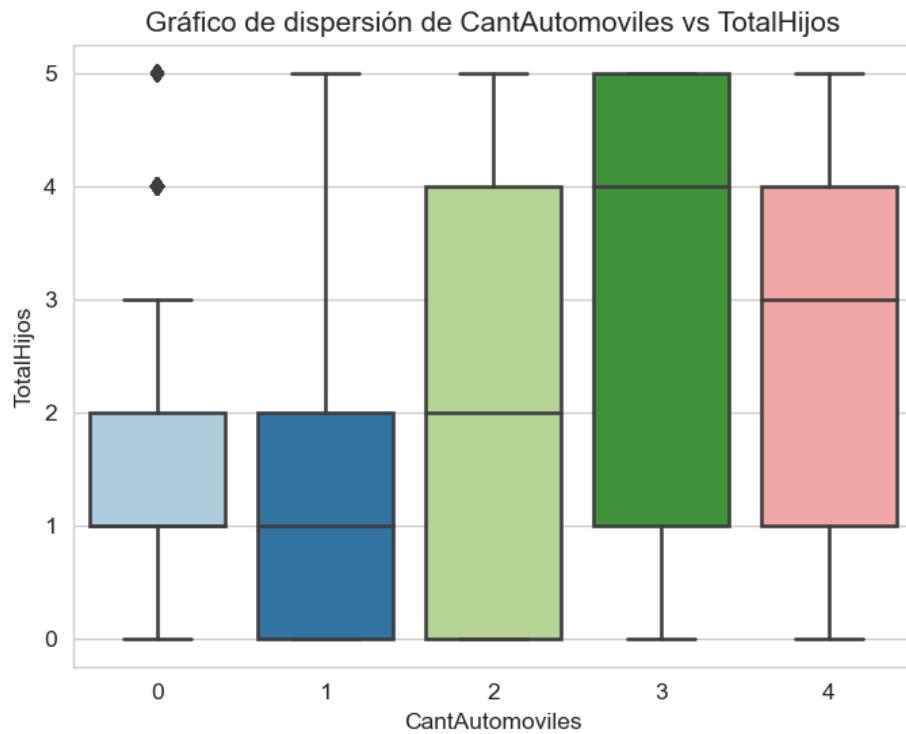
- Las edades que cubre la caja del boxplot van desplazándose hacia arriba desde 0 hasta 2 hijos.
- Las medianas de edad aumentan hasta los 2 hijos, luego se mantienen estables, incluso disminuyendo un poco en el caso de 5 hijos.
- No se observan valores atípicos para quienes tienen 5 hijos.
- Observamos que quienes no tienen hijos son en su mayoría más jóvenes que los demás.

CantAutomoviles vs IngresoAnual



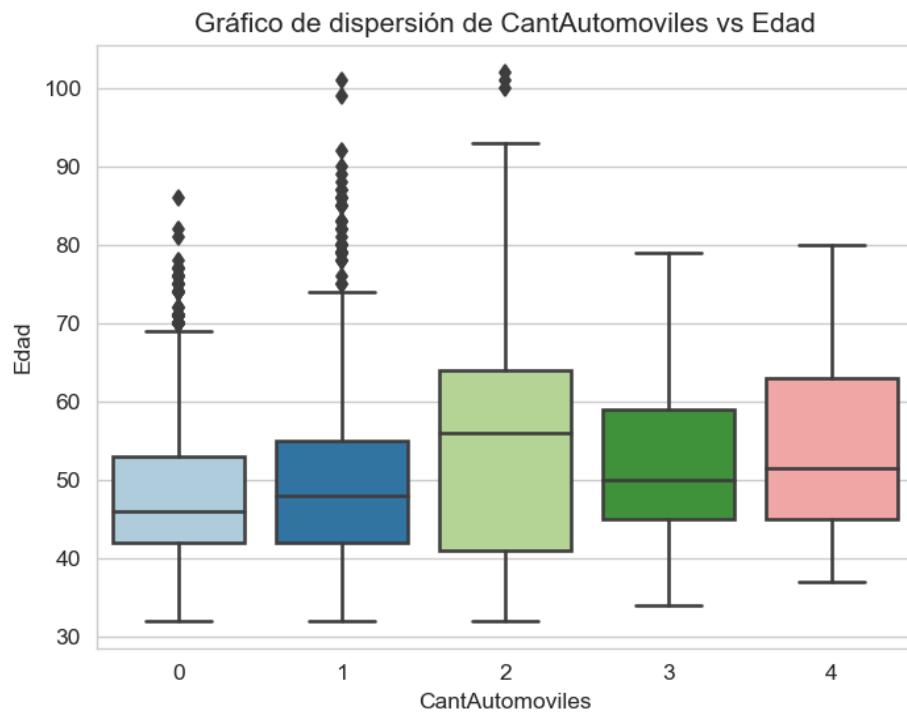
- La mediana de ingreso anual es proporcional a la cantidad de automóviles que poseen, siendo algo lógico.
- Observamos que existen varios valores atípicos, pero los más extraños se encuentran en 4 automóviles al hallar varios clientes con un ingreso anual menor a 40 mil (considerado bajo) que poseen 4 automóviles.

CantAutomoviles vs TotalHijos



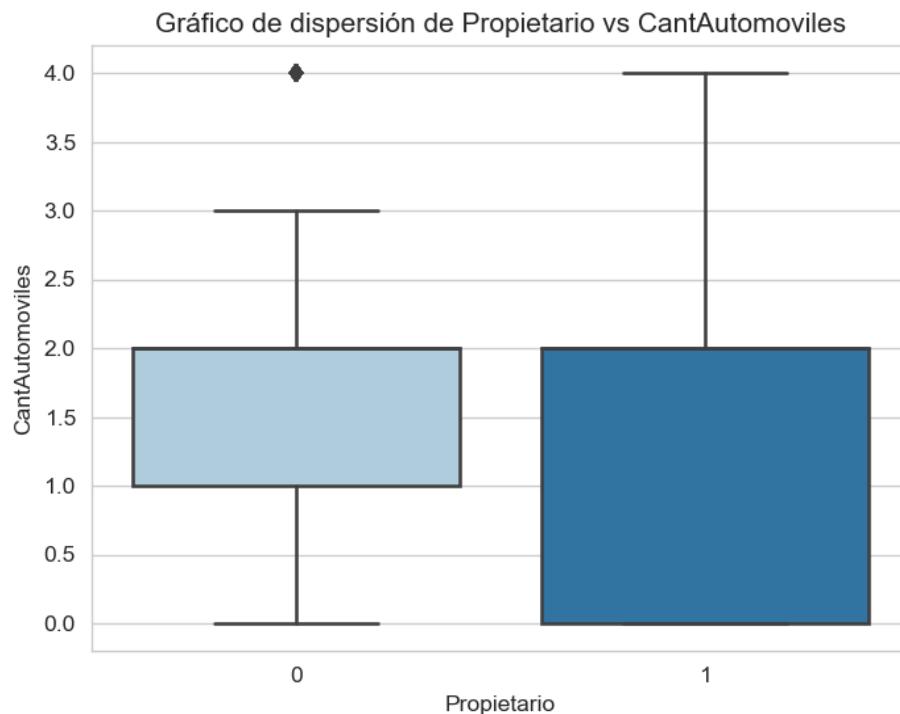
- Observamos que la mediana aumenta a medida que se incrementa la cantidad de automóviles. Existe una relación en la mayoría de los casos respecto a la cantidad de automóviles e hijos (a mayor cantidad de hijos, más automóviles).
- Para la cantidad de 2 automóviles existe una distribución casi uniforme.

CantAutomoviles vs Edad



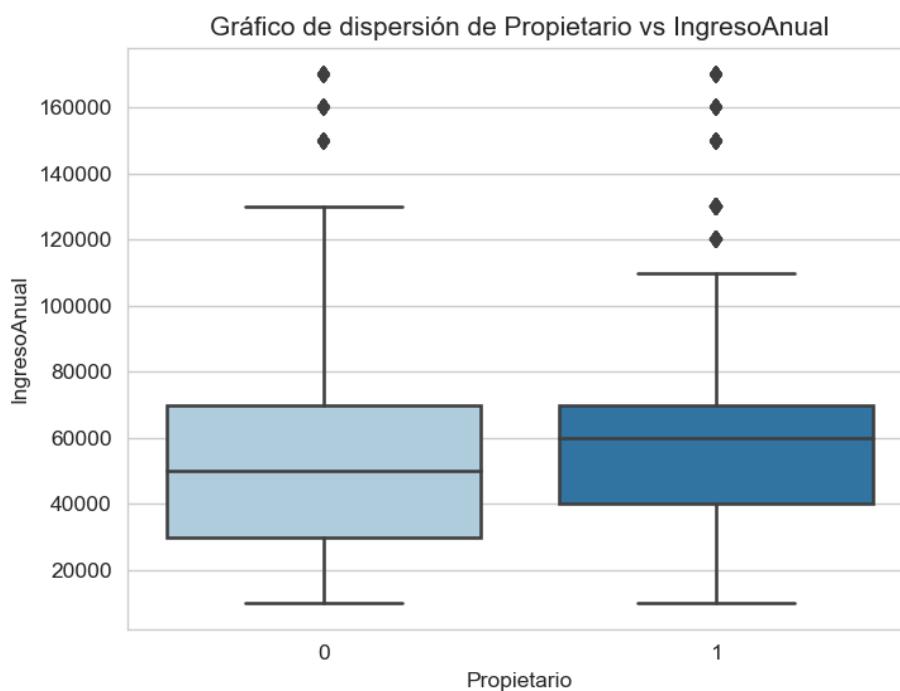
- Observamos que para toda la cantidad de automóviles la mediana de la edad varía poco (45-55).
- Para las personas que cuentan con más de 3 automóviles no se encuentran outliers.
- Para aquellas personas que cuentan con menos de 3 automóviles, observamos cierta uniformidad en su distribución pero contando con varios valores anómalos.

Propietario vs CantAutomoviles



- Observamos que el 75% de los clientes propietarios se encuentran entre 0 y 2 automóviles, mientras que los no propietarios tienen en su mayoría 1 o 2.

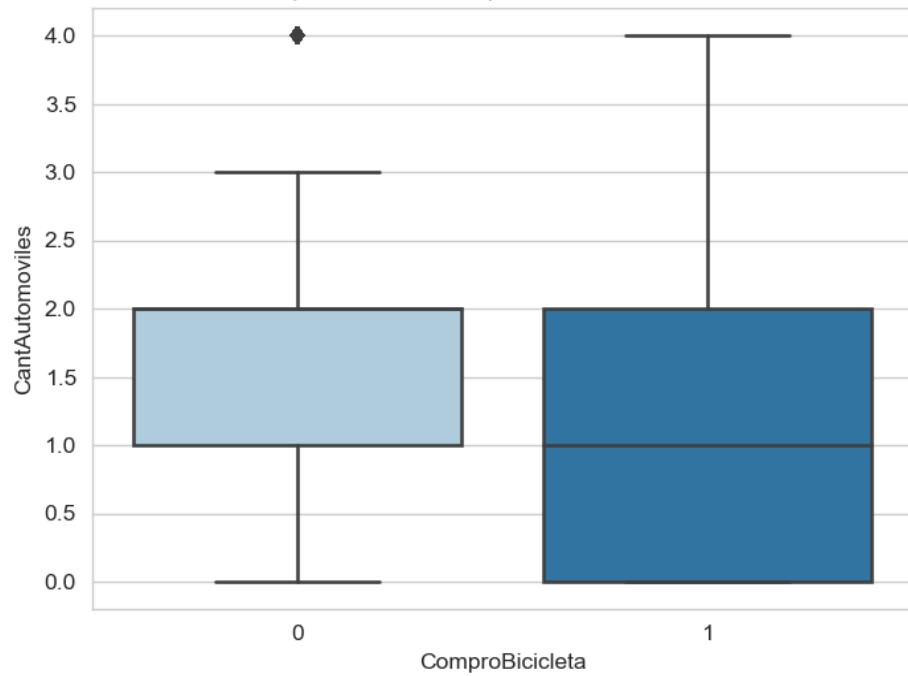
Propietario vs IngresoAnual



- Observamos que la caja de los clientes que son propietarios es más estrecha hacia arriba, por lo que los ingresos anuales menores de estos son más altos.

ComproBicicleta vs CantAutomoviles

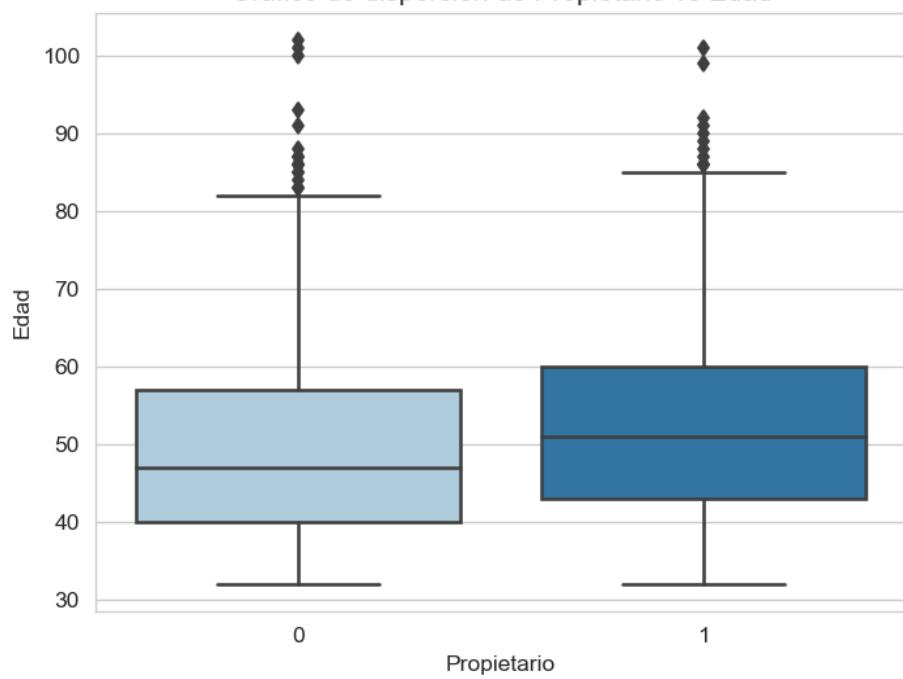
Gráfico de dispersión de ComproBicicleta vs CantAutomoviles



- Los clientes que compraron bicicletas en el pasado son aquellos que poseen de 0 a 2 automóviles.

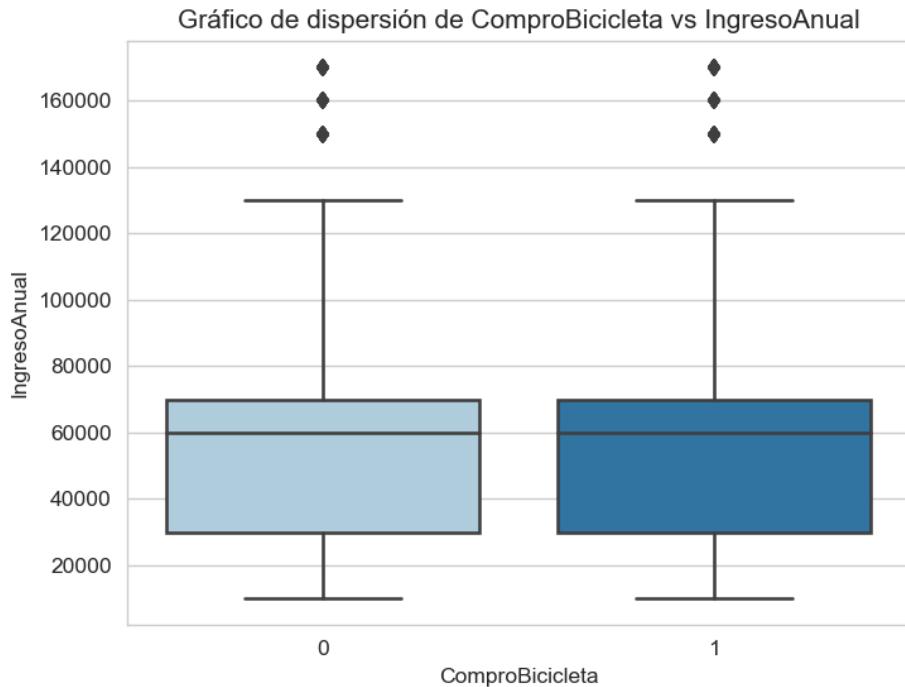
Propietario vs Edad

Gráfico de dispersión de Propietario vs Edad



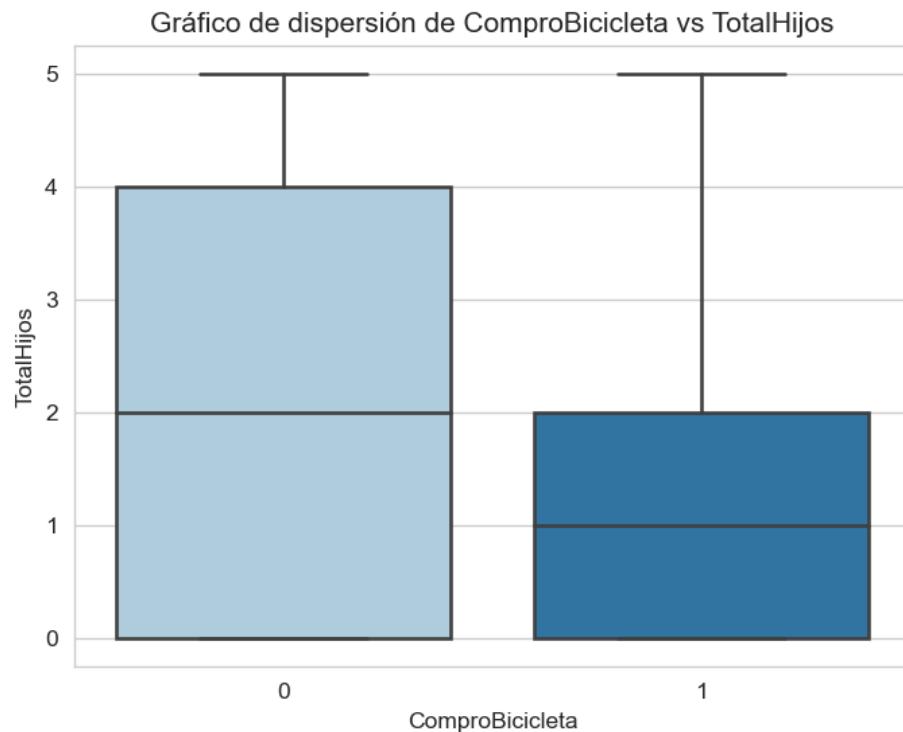
- Observamos que los cuartiles de los clientes no propietarios se encuentran entre 40 y 58 aproximadamente, mientras que los de los propietarios entre 43 aproximadamente y 60.
- La mediana de los no propietarios (poco menos de 50) es un poco menor que la de los propietarios (poco más de 50).

ComproBicicleta vs IngresoAnual



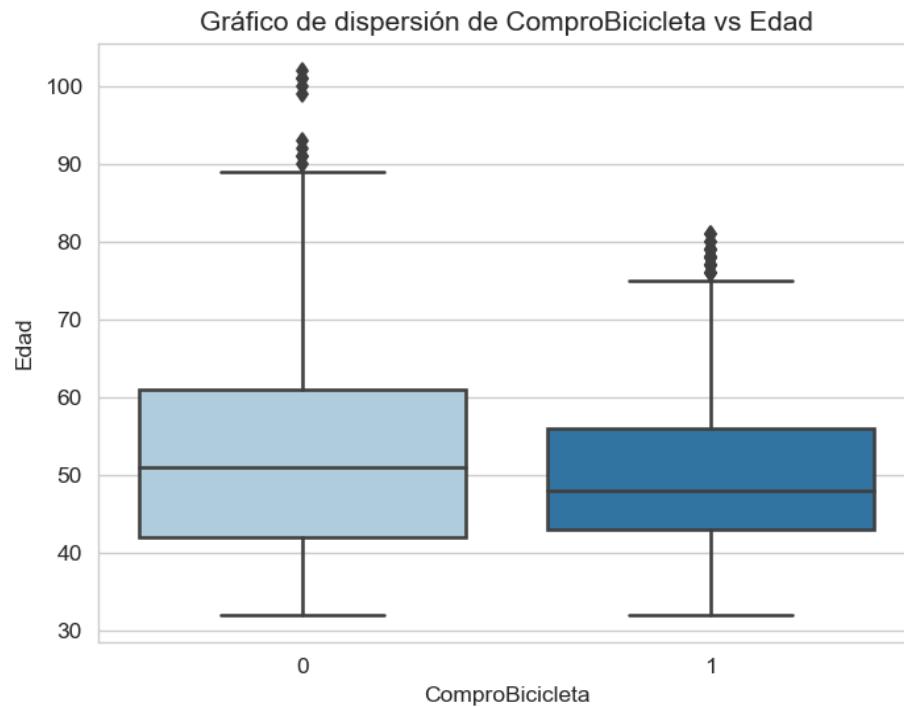
- Observamos que ambas gráficas son iguales, por esto y por lo demostrado en la matriz S, concluimos que ambas variables no están relacionadas entre sí.

ComproBicicleta vs TotalHijos



- Observamos que la mediana disminuye en el caso de los clientes que sí compraron bicicleta.
- El rango de los clientes que no compraron bicicletas es mayor al de los que sí compraron.
- Encontramos que por tener mayor cantidad de hijos, no será más probable la compra de una bicicleta para ellos.

ComproBicicleta vs Edad



- Observamos que ambos diagramas son muy parecidos por lo que comprendemos que las variables no están relacionadas.

Técnicas predictivas

Un modelo predictivo es un modelo analítico que se construye con el objetivo de predecir resultados o comportamientos futuros en función de los datos históricos y los patrones encontrados en ellos.

En este informe, utilizamos tres técnicas predictivas: árbol de decisión, algoritmo de los vecinos más próximos y análisis discriminante lineal. Para cada técnica, la variable a predecir fue “ComproBicicleta” y las variables predictoras fueron solo las numéricas (obviando IdCiudad e IdCliente).

Para poder realizar una comparación entre las tres técnicas y así decidir cual utilizar, generamos una matriz de confusión para cada una. Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

Con esta matriz se puede calcular la accuracy (precisión del modelo, evalúa la precisión general del modelo en todas las clases) y la class recall (tasa de verdaderos positivos, mide la capacidad del modelo para identificar correctamente los ejemplos positivos en una clase específica).

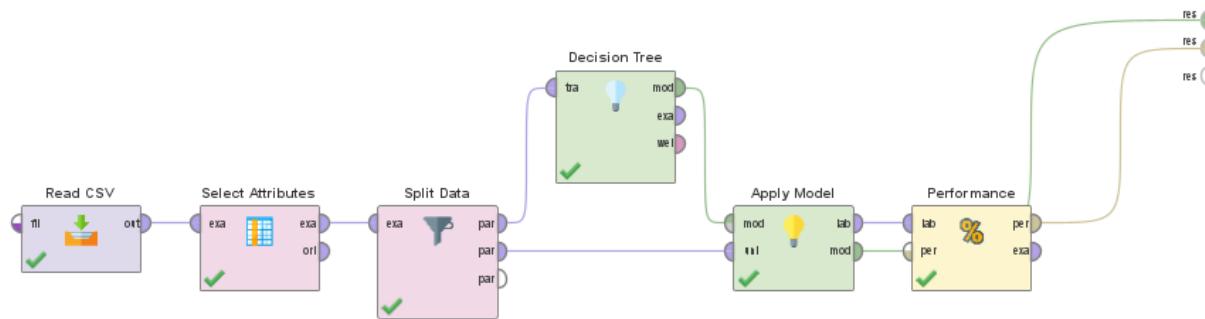
Interesa encontrar un modelo que tenga un buen class recall para los unos, manteniendo un balance con la accuracy. Sin embargo, en el contexto del problema de negocio, la métrica a optimizar es el class recall. El motivo es que el objetivo principal es identificar a los clientes que tienen más probabilidades de comprar una bicicleta en el futuro, es decir, maximizar la detección de los verdaderos positivos (clientes que se predicen como 1 y realmente lo son). En este caso, el envío de publicidad a algunos clientes erróneamente clasificados como positivos (falsos positivos) no sería un problema significativo. Resumiendo, optimizar la "class recall" permitiría identificar a la mayor cantidad posible de posibles clientes de bicicletas, aumentando así las oportunidades de éxito en la campaña publicitaria.

Árbol de decisión

El árbol de decisión es un algoritmo de aprendizaje supervisado que se utiliza para clasificar y predecir valores basados en reglas de decisión en forma de árbol.

Para poder utilizar el árbol de decisión tuvimos que asignarle, a la variable “ComproBicicleta”, el atributo “label” y transformarla a tipo nominal. Luego, utilizamos el muestreo automático de RapidMiner, el cual realiza un muestreo estratificado si la variable label es de tipo nominal o un muestreo aleatorio si no lo es. En este caso, la variable “ComproBicicleta” es de tipo nominal, por lo tanto, se utilizó un muestreo estratificado.

El modelado del problema en RapidMiner nos quedó de la siguiente manera:



Decidimos generar modelos variando las particiones, las variables y los parámetros del árbol.

- Utilizamos cuatro particiones distintas:
 - 65% datos de entrenamiento, 35% datos de testeo.
 - 70% datos de entrenamiento, 30% datos de testeo.
 - 75% datos de entrenamiento, 25% datos de testeo.
 - 80% datos de entrenamiento, 20% datos de testeo.
- El proceso en general con las variables fue:
 - Iniciamos siempre con las 5 variables numéricas: IngresoAnual, TotalHijos, Propietario, CantAutomoviles.
 - Agregamos las variables categóricas: Educación, Ocupación, Distancia, Región.
 - Quitamos una o varias variables hasta obtener el mejor modelo (no mostraremos todas las matrices de confusión generadas, nos limitaremos a mostrar el mejor obtenido).
- Los parámetros del árbol que variamos fueron:
 - Criterio: gain_ratio, information_ratio, gini_index y accuracy.
 - Profundidad máxima: entre 0 y 9, entre 10 y 19, entre 20 y 29 y mayor a 50.
 - Poda con una confianza: entre 0 y 0.2, entre 0.2 y 0.4 y mayor a 0.4.
 - En ninguno de los casos se aplicó prepoda.

Luego de varias pruebas (y para no mostrar todos los modelos generados), concluimos que los mejores resultados fueron quitando las variables Edad y Región del árbol. Además, en la mayoría de los casos, decidimos utilizar los parámetros “gain_ratio”, máxima profundidad igual a 50 y confianza igual a 0.15 para el árbol, debido a que fueron los que nos dieron mejores resultados de forma general.

Partición 65 y 35

Con las variables IngresoAnual, TotalHijos, Propietario, CantAutomoviles y Edad obtuvimos:

accuracy: 70.34%

	true 1	true 0	class precision
pred. 1	596	377	61.25%
pred. 0	287	979	77.33%
class recall	67.50%	72.20%	

La interpretación de la matriz de confusión es la siguiente:

- Se predijo 1 cuando el valor verdadero era 1 (correcto) un total de 596 veces.
- Se predijo 1 cuando el valor verdadero era 0 (incorrecto) un total de 377 veces.
- Se predijo 0 cuando el valor verdadero era 1 (incorrecto) un total de 287 veces.
- Se predijo 0 cuando el valor verdadero era 0 (correcto) un total de 979 veces.

(Cabe aclarar que las matrices de confusión que generamos utilizan los datos de testeo, en este caso, un 35%).

Si sumamos las variables categóricas y quitamos la variable Región:

accuracy: 73.07%

	true 1	true 0	class precision
pred. 1	635	355	64.14%
pred. 0	248	1001	80.14%
class recall	71.91%	73.82%	

Quitando Región y Edad aumentó bastante la precisión con respecto al anterior pero disminuyó el class recall:

accuracy: 76.55%

	true 1	true 0	class precision
pred. 1	616	258	70.48%
pred. 0	267	1098	80.44%
class recall	69.76%	80.97%	

Luego de probar agregando y quitando variables y variando los parámetros del árbol, concluimos que el mejor modelo obtenido para esta partición es el que tiene todas las variables menos Región. Dicho modelo posee el mayor class recall (71.91%) y además tiene un buen balance entre el susodicho y la accuracy (precisión del 72.89%).

Partición 70 y 30

Con las variables IngresoAnual, TotalHijos, Propietario, CantAutomoviles y Edad obtuvimos:

accuracy: 71.44%

	true 1	true 0	class precision
pred. 1	500	291	63.21%
pred. 0	257	871	77.22%
class recall	66.05%	74.96%	

Con únicamente las variables categóricas (sin las numéricas):

accuracy: 66.70%

	true 1	true 0	class precision
pred. 1	345	227	60.31%
pred. 0	412	935	69.41%
class recall	45.57%	80.46%	

Con todas las variables menos Edad:

accuracy: 75.77%

	true 1	true 0	class precision
pred. 1	538	246	68.62%
pred. 0	219	916	80.70%
class recall	71.07%	78.83%	

Luego de seguir quitando y agregando variables y modificando los parámetros del árbol, el modelo de la última matriz fue el mejor que obtuvimos. Sin embargo, este modelo no es mejor que el obtenido con la partición 65 y 35. Por lo tanto, decidimos descartarlo.

Partición 75 y 25

Todas numéricas:

accuracy: 72.42%

	true 1	true 0	class precision
pred. 1	421	231	64.57%
pred. 0	210	737	77.82%
class recall	66.72%	76.14%	

Todas categóricas:

accuracy: 66.48%

	true 1	true 0	class precision
pred. 1	283	188	60.08%
pred. 0	348	780	69.15%
class recall	44.85%	80.58%	

Todas menos Edad y Región:

accuracy: 76.36%

	true 1	true 0	class precision
pred. 1	440	187	70.18%
pred. 0	191	781	80.35%
class recall	69.73%	80.68%	

Quitando solo Región:

accuracy: 73.86%

	true 1	true 0	class precision
pred. 1	445	232	65.73%
pred. 0	186	736	79.83%
class recall	70.52%	76.03%	

Ninguno de los modelos generados superó al modelo de la partición 65 y 35.

Partición 80 y 20

Con todas las variables numéricas:

accuracy: 72.50%

	true 1	true 0	class precision
pred. 1	336	183	64.74%
pred. 0	169	592	77.79%
class recall	66.53%	76.39%	

Con todas las variables categóricas:

accuracy: 66.17%

	true 1	true 0	class precision
pred. 1	234	162	59.09%
pred. 0	271	613	69.34%
class recall	46.34%	79.10%	

Con todas las variables exceptuando Edad y con una confianza de 0.1:

accuracy: 75.78%

	true 1	true 0	class precision
pred. 1	345	150	69.70%
pred. 0	160	625	79.62%
class recall	68.32%	80.65%	

El mismo modelo (sin Edad) pero con una confianza de 0.5:

accuracy: 77.27%

	true 1	true 0	class precision
pred. 1	338	124	73.16%
pred. 0	167	651	79.58%
class recall	66.93%	84.00%	

Es el modelo de mayor precisión que obtuvimos, con un 77.27%. Sin embargo, el class recall no es muy bueno.

Conclusión

El mejor modelo que conseguimos lo obtuvimos a partir de:

- Partición 65 y 35.
- Todas las variables menos Edad.
- Para el árbol:
 - Criterio: gain_ratio.
 - Maximal depth: 50.
 - Confidence: 0.1.

La matriz de confusión del modelo en cuestión es:

accuracy: 73.07%

	true 1	true 0	class precision
pred. 1	635	355	64.14%
pred. 0	248	1001	80.14%
class recall	71.91%	73.82%	

Elegimos este modelo porque:

- Tiene el mayor class recall.
- No tiene la mayor precisión de todas pero aún así es bastante alta, logrando un balance con el class recall.

Continuaremos generando modelos con otros algoritmos intentando obtener uno mejor.

Algoritmo de los vecinos más próximos (KNN)

El algoritmo predictivo KNN (K-nearest neighbors) consiste en clasificar un nuevo caso, en función de su distancia con los casos vecinos. Se trabaja con un parámetro “k” que determina la cantidad de vecinos cercanos con los cuales se comparará la nueva observación.

```
K igual a 8:  
Accuracy de K-NN train: 0.85  
F1 de K-NN train: 0.84  
Accuracy de K-NN test: 0.74  
F1 de K-NN test: 0.71  
  
-----  
K igual a 9:  
Accuracy de K-NN train: 0.86  
F1 de K-NN train: 0.84  
Accuracy de K-NN test: 0.73  
F1 de K-NN test: 0.71  
  
-----  
K igual a 10:  
Accuracy de K-NN train: 0.86  
F1 de K-NN train: 0.84  
Accuracy de K-NN test: 0.74  
F1 de K-NN test: 0.71  
  
-----  
K igual a 11:  
Accuracy de K-NN train: 0.86  
F1 de K-NN train: 0.84  
Accuracy de K-NN test: 0.73  
F1 de K-NN test: 0.71  
  
-----  
K igual a 12:  
Accuracy de K-NN train: 0.86  
F1 de K-NN train: 0.84  
Accuracy de K-NN test: 0.73  
F1 de K-NN test: 0.71
```

En este caso, luego de realizar el respectivo análisis, utilizamos $k = 10$. Se puede ver que no hay muchas diferencias entre las distintas k , sin embargo, cuando $k = 10$ tenemos una accuracy para el test de K-NN de 0.74, el cual es levemente mayor que los demás.

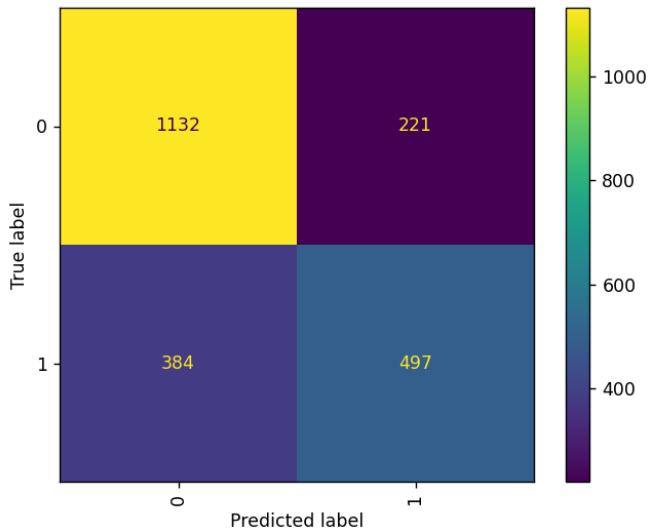
Partición 65 y 35

Al igual que para el árbol de decisión, realizamos las pruebas con varias particiones. En esta utilizamos el 65% de los datos para entrenar y el 35% para testear (para la matriz de confusión). Utilizamos el muestreo estratificado por medio del parámetro “stratify”.

Las variables que ingresaron al modelo fueron:

- IngresoAnual.
- TotalHijos.
- CantAutomoviles.
- Edad.

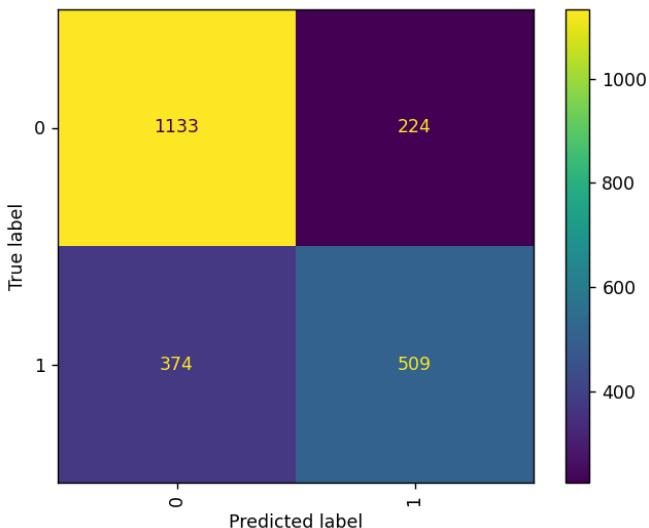
La matriz de confusión obtenida fue:



La precisión de la matriz de confusión no está en la gráfica, pero se puede calcular fácilmente: es el porcentaje de la suma de las predicciones acertadas sobre el total de valores: $Precisión = (Verdaderos Positivos + Verdaderos Negativos) / Total\ de\ casos$. Luego, la precisión de este primer modelo con el algoritmo KNN es de 72.9%.

Para calcular el class recall, se utiliza: $Class\ Recall = Verdaderos\ Positivos / Total\ de\ verdaderos$. Para esta matriz se tiene un class recall de 56.41%.

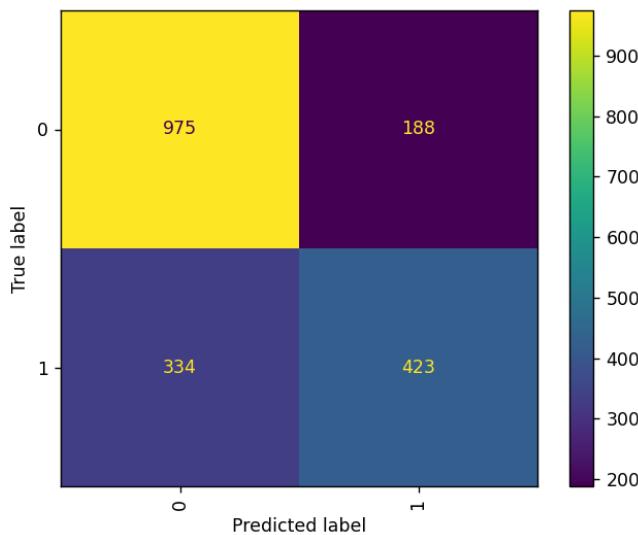
Luego de generar varios modelos quitando y añadiendo variables, decidimos agregar la variable “Propietario” a las cuatro del primer modelo y obtuvimos:



Este modelo fue el mejor que obtuvimos con KNN, con una precisión de 73.3% y un class recall de 57.64%. Por lo tanto, por ahora nos estaríamos quedando con el árbol.

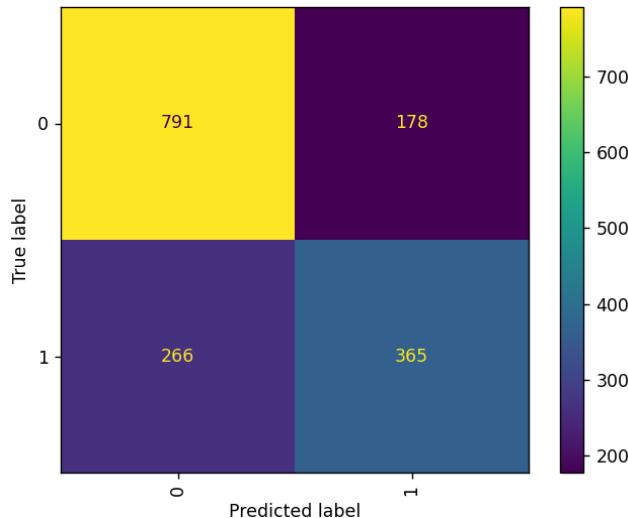
Partición 70 y 30

Luego de analizar, obtuvimos que $k = 10$ (nuevamente) era la mejor opción. La mejor matriz de confusión que generamos fue:



Esta matriz tiene una accuracy de 72.81% y un class recall de 55.87%. La precisión no es mala pero el class recall si lo es. Por lo tanto, descartamos este modelo y procedemos con la siguiente partición.

Partición 75 y 25



La precisión del mejor modelo es de 72.25% y un class recall de 57.84%.

Conclusión

Podemos ver que KNN no está generando muy buenos modelos (a comparación del árbol de decisión). Por ello, hasta acá llegamos con el algoritmo de los vecinos más próximos y procedemos con LDA en busca de mejores resultados.

Análisis discriminante lineal (LDA)

El Análisis discriminante es una técnica de aprendizaje supervisado utilizada para encontrar una combinación lineal de variables independientes que mejor discrimina entre clases o grupos distintos.

Supuestos requeridos

Los supuestos requeridos para LDA son:

- No multicolinealidad.
- Normalidad: distribución normal de las variables iguales.
- Las matrices de varianza y covarianza deben tener el mismo tamaño (tanto en ancho como en altura).

No multicolinealidad

Para analizar la multicolinealidad realizamos un diagnóstico de la misma, en el cual a través del índice de condición observaremos si existe o no multicolinealidad.

Si el valor del índice es mayor que 30, vamos a tener multicolinealidad, con un valor entre 20 y 30 será una multicolinealidad moderada y con un valor menor que 20 no hay multicolinealidad.

Modelo	Dimensión	Autovalor	Índice de condición	(Constante)	Edad	Proporciones de varianza			
						CantAutomóviles	Propietario	TotalHijos	IngresoAnual
1	1	4,972	1,000	,00	,00	,01	,01	,01	,01
	2	,382	3,607	,00	,00	,27	,41	,00	,04
	3	,313	3,988	,01	,00	,01	,03	,78	,04
	4	,176	5,308	,04	,04	,44	,52	,00	,02
	5	,137	6,018	,03	,02	,27	,01	,02	,89
	6	,020	15,779	,92	,93	,00	,01	,19	,01

a. Variable dependiente: ComproBicicleta

En nuestro caso, obtuvimos 15,779 en el Índice de condición, el cual es menor a 20, justificando la no multicolinealidad entre las variables para aplicarlas en el análisis.

Prueba de normalidad

Para analizar la normalidad de las variables, al contar con más de 50 observaciones realizamos la prueba estadística de Kolmogorov-Smirnov, la cual nos proporciona un dato "p" que indica si los datos se desvían significativamente de la normalidad. Si este valor es mayor a nuestro nivel de significancia determinado (0.05), se puede aceptar el supuesto de normalidad.

Como observamos, todas las variables tienen un valor $p < 0.05$, por lo que podemos justificar que las variables no están normalmente distribuidas. Por lo tanto, nuestros datos fallan el test de normalidad.

Pruebas de normalidad

Kolmogorov-Smirnov ^a			
Estadístico	gl	Sig.	
Edad	,079	6400	,000
IngresoAnual	,129	6400	,000
TotalHijos	,168	6400	,000
Propietario	,432	6400	,000
CantAutomoviles	,181	6400	,000

a. Corrección de significación de Lilliefors

Matrices de varianza y covarianza iguales

Para analizar la varianza y covarianza de las poblaciones, haremos la prueba de M de Box, obteniendo:

Resultados de prueba

M de Box	132.173
F	Aprox. 44.042
gl1	3
gl2	1516326772
Sig.	.000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

El valor de p de la prueba es 0, con lo cual rechazamos la hipótesis de que las matrices de varianza y covarianza de las poblaciones son iguales, concluyendo que no se cumple el supuesto de matrices de varianza y covarianza iguales.

Conclusión

Nuestros datos no superaron los supuestos requeridos para poder realizar el análisis discriminante lineal. Sin embargo, a pesar de no cumplirlos, generaremos modelos con el fin de probar la técnica de minería.

Partición 65 y 35

Para las tres particiones, las variables dependientes que utilizamos fueron:

- IngresoAnual.
- TotalHijos.
- CantAutomoviles.
- Edad.
- Propietario.

Sin embargo, luego fuimos variando agregando/quitando. Las matrices de confusión que mostraremos a continuación son las de los mejores modelos.

Resultados de clasificación^{a,b}

Original	Recuento	ComproBicicleta	Pertenencia a grupos pronosticada		Total
			0	1	
	0		820	528	1348
	1		326	565	891
%	0		60,8	39,2	100,0
	1		36,6	63,4	100,0

Para esta partición obtuvimos una precisión del 61.85% y un class recall de 63.41%.

Partición 70 y 30

Resultados de clasificación^{a,b}

Original	Recuento	ComproBicicleta	Pertenencia a grupos pronosticada		Total
			0	1	
	0		703	445	1148
	1		277	495	772
%	0		61,2	38,8	100,0
	1		35,9	64,1	100,0

Para la partición con 70% datos de entrenamiento y 30% datos de test, tenemos un modelo con una precisión de 62.39% y un class recall de 64.11% (un modelo un poco mejor que el anterior).

Partición 75 y 25

Resultados de clasificación^{a,b}

Original	Recuento	ComproBicicleta	Pertenencia a grupos pronosticada		Total
			0	1	
	0		581	380	961
	1		223	416	639
%	0		60,5	39,5	100,0
	1		34,9	65,1	100,0

El modelo tiene una precisión de 62.31% y un class recall de 65.1% (mejor que los dos anteriores).

Conclusión

Predicción

Luego de utilizar las tres técnicas de minería y de generar modelos (y sus respectivas matrices de confusión) ya sea para árbol de decisión, KNN y LDA, obtuvimos que el mejor de todos es el obtenido por medio del árbol con la partición de 65% datos de entrenamiento y 35% datos de prueba. Este modelo tiene una precisión del 73.07% y un class recall del 71.91%. Lo utilizaremos para realizar la predicción final (es decir, determinar a qué usuario se le debe enviar la publicidad o no).

accuracy: 73.07%

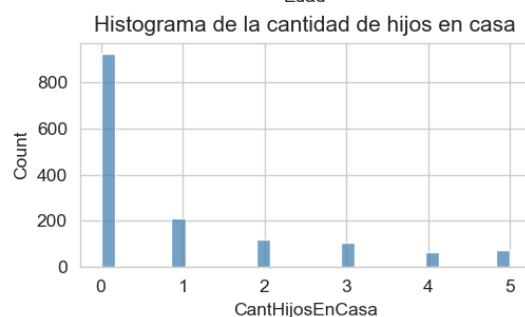
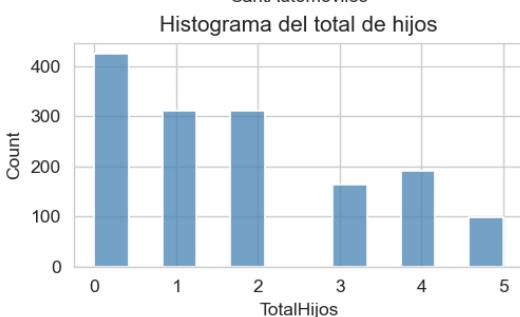
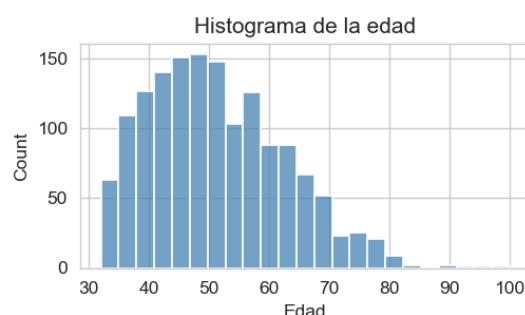
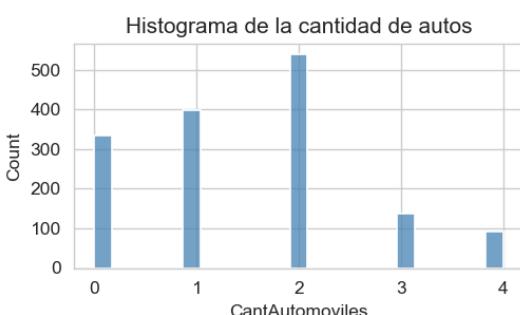
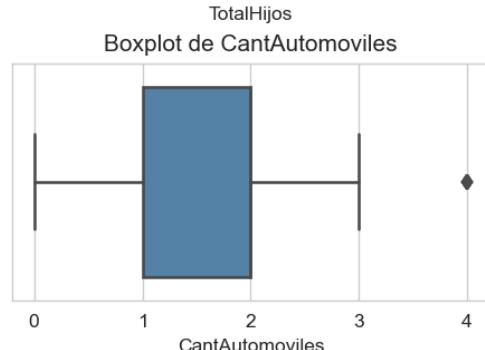
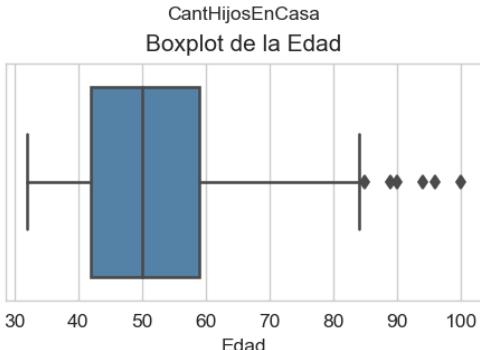
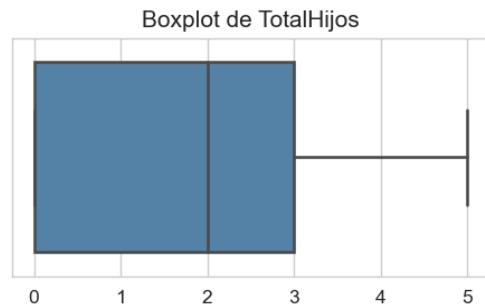
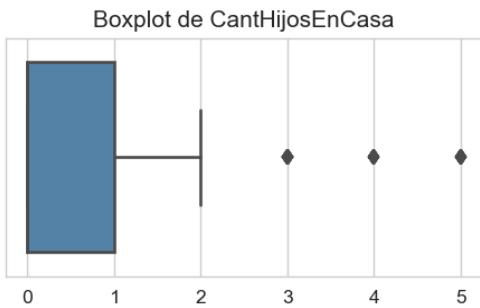
	true 1	true 0	class precision
pred. 1	635	355	64.14%
pred. 0	248	1001	80.14%
class recall	71.91%	73.82%	

La predicción obtenida con el modelo de árbol de decisión utilizado es la siguiente. De un total de 1500 destinatarios, el árbol predijo:

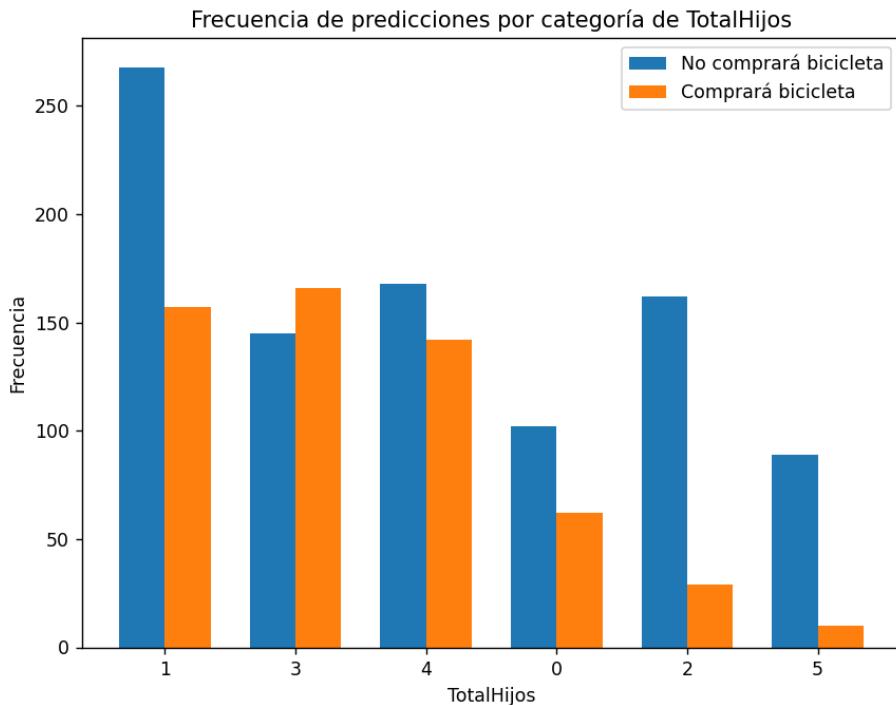
- No enviarle la publicidad a 934 personas.
- Enviarle la publicidad a 566 personas.

Análisis de los resultados

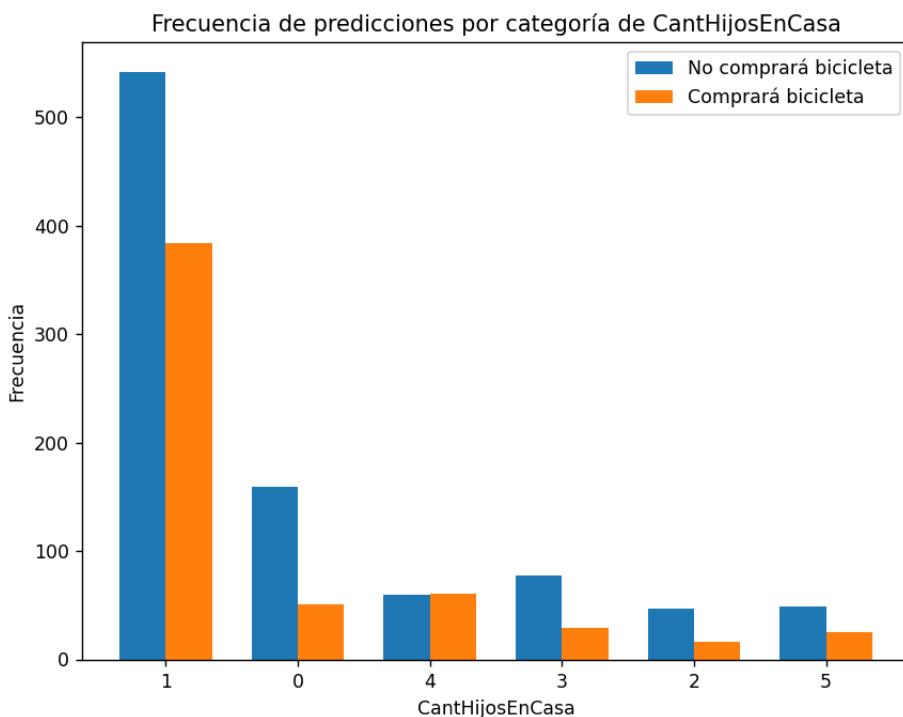
Decidimos generar algunas gráficas para el nuevo conjunto de datos.



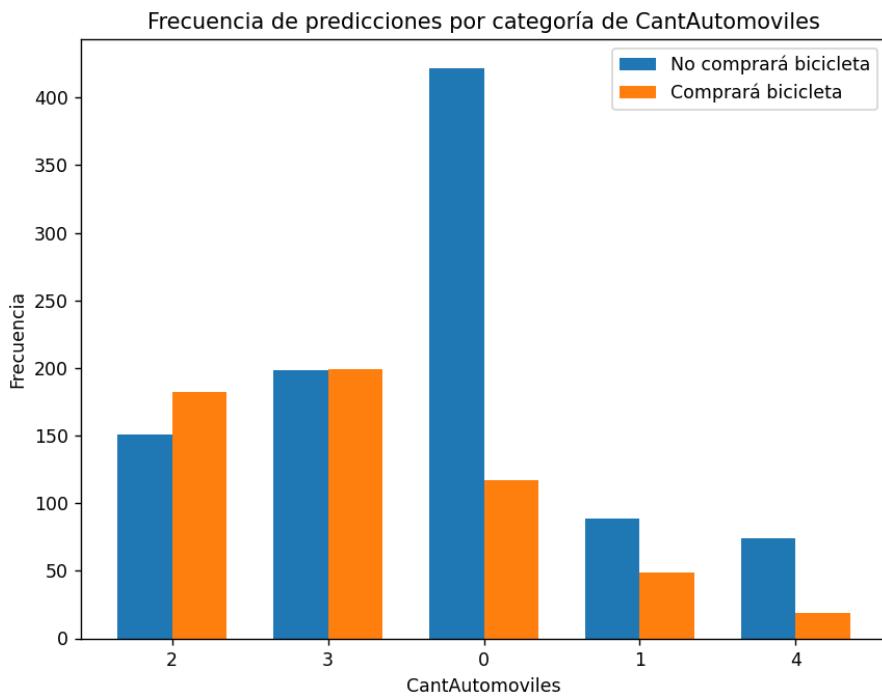
En estas gráficas podemos observar el rango y distribución de los datos de las variables seleccionadas, pudiendo observar también los valores anómalos. Sin embargo, son gráficas bastante generales. Interesa más ver las variables en relación a la variable en estudio. Para ello, generamos gráficas con respecto a la variable predicha “prediction(ComproBicicleta)”, la cual adquiere valores 0 y 1. Para cada gráfica, si adquiere el valor 0 pusimos “No comprará bicicleta”, mientras que si adquiere el valor 1 entonces “Comprará bicicleta”.



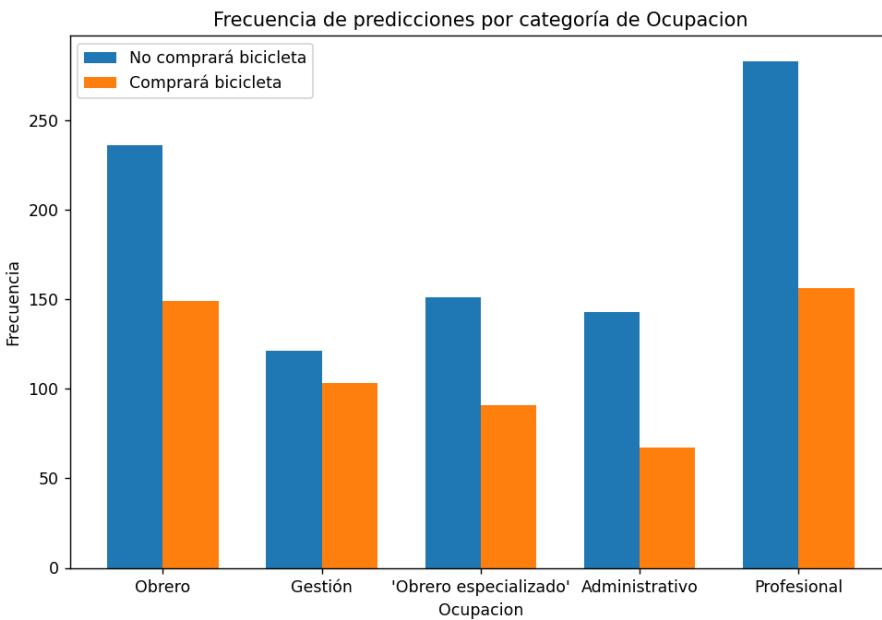
Observamos que la frecuencia de quienes no comprarán bicicletas es mayor en casi la totalidad de clientes dependiendo la cantidad de hijos. Solo aquellos clientes que tienen tres hijos comprarán más bicicletas respecto a los demás.



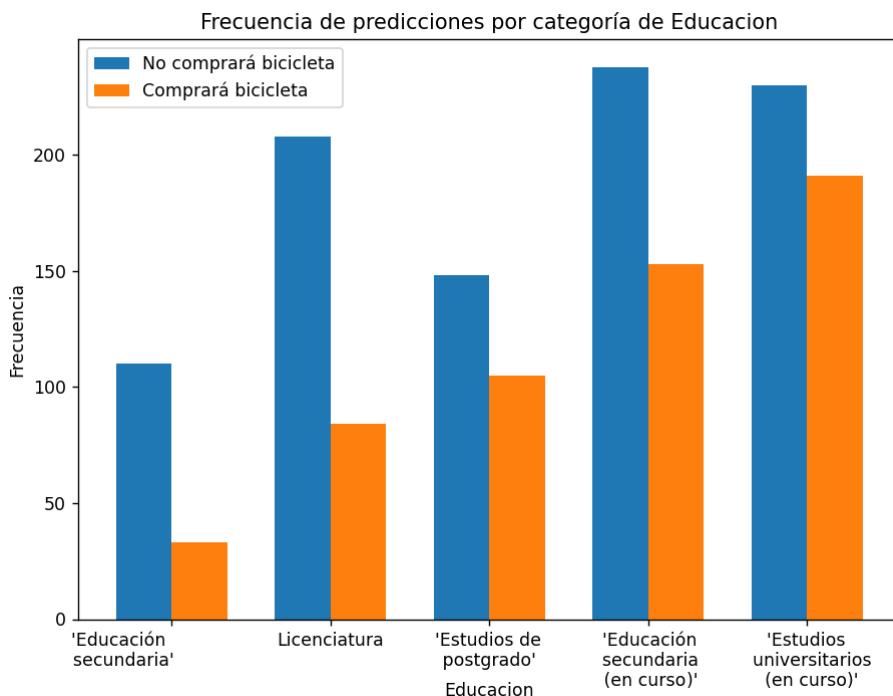
Los clientes que tienen un solo hijo en casa son los de mayor frecuencia, siendo muchos los que no comprarán bicicleta pero también muchos los que sí.



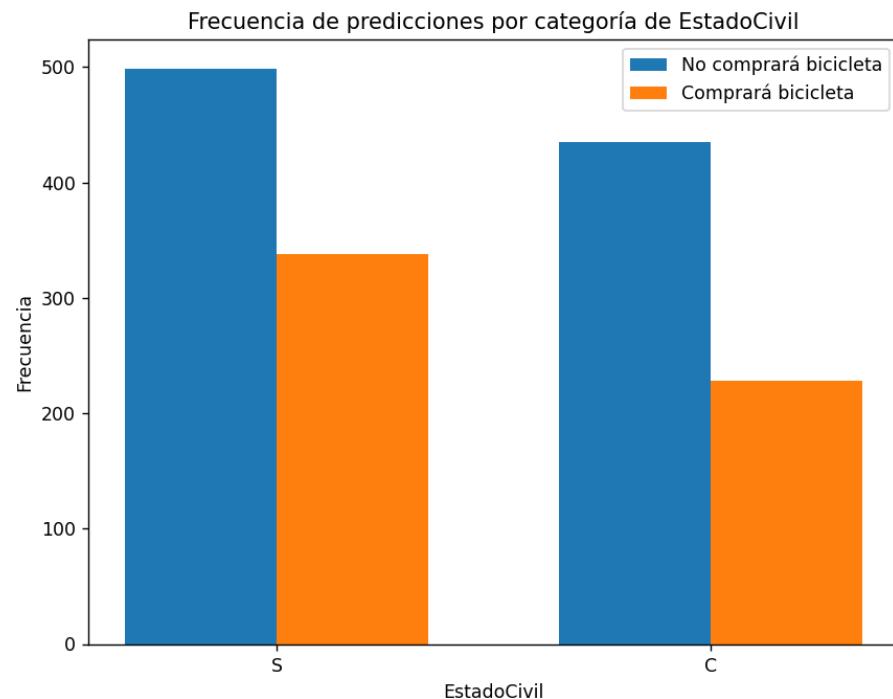
Vemos que la frecuencia de aquellos clientes que no tienen automóviles y no comprarán bicicleta es muy grande respecto a las demás. Encontramos una similitud de frecuencia en sí comprarán o no en los clientes que tienen tres automóviles.



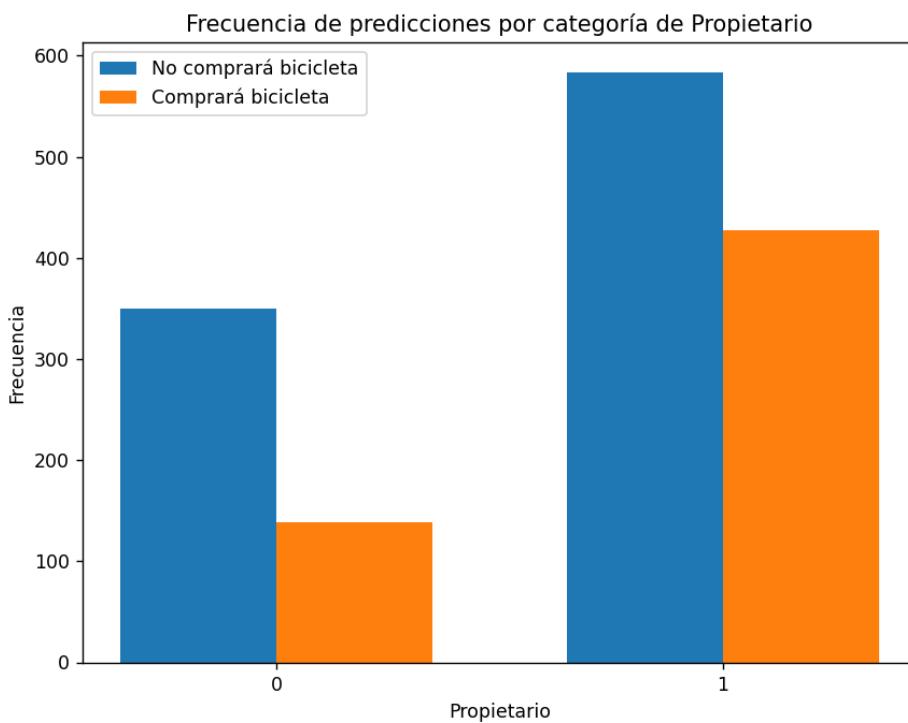
Es curioso que la mayor cantidad de clientes que comprarán bicicleta son profesionales y obreros, aunque también son los que menos comprarán.



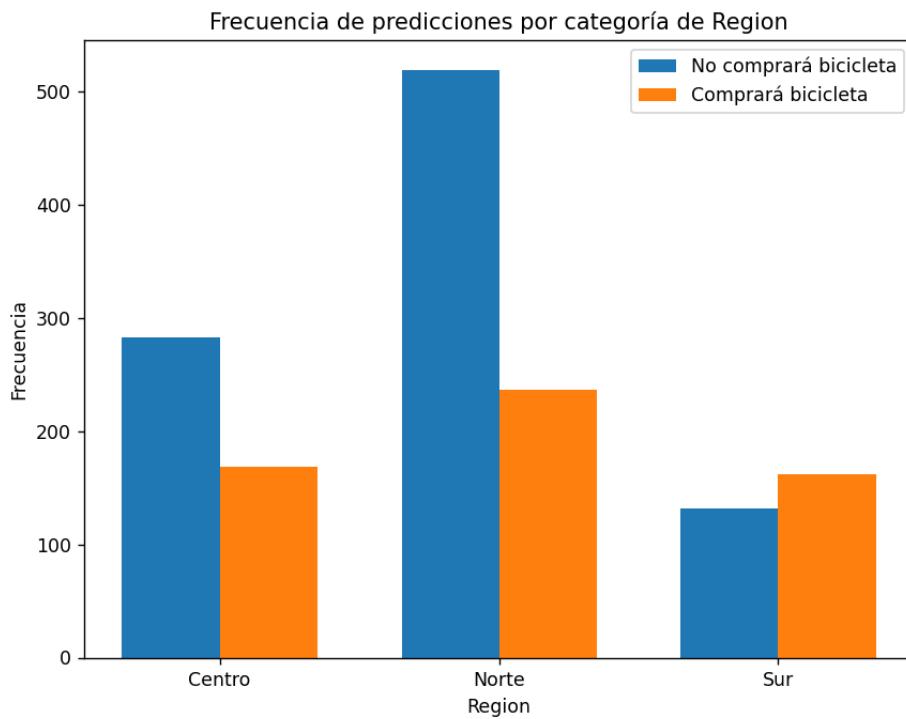
Aquellos clientes con estudios universitarios en curso son en mayor cantidad los que comprarán bicicletas, seguidos por los de educación secundaria, también en curso. En cuanto a los clientes que tienen licenciatura, encontramos una gran diferencia entre quienes no comprarán y quienes si.



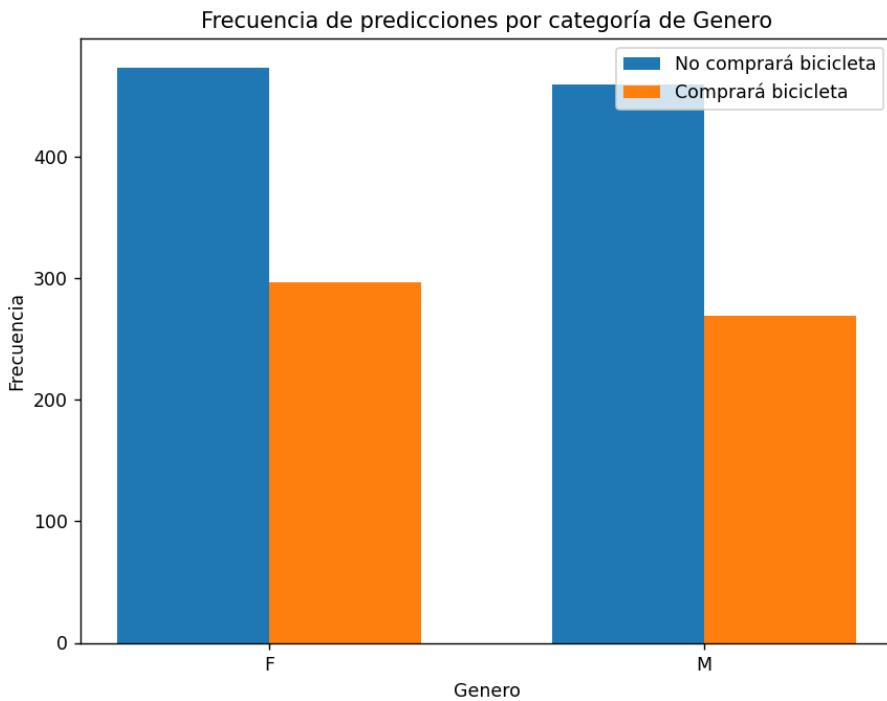
No vemos nada interesante.



Aquellos clientes que son propietarios son mucho más propensos a comprar que los que no son, aunque también son más propensos a no comprar.



Analizando por región, encontramos que en el sur la cantidad de clientes que sí comprará bicicletas es mayor que los que no. Sobre el norte podemos decir que la diferencia entre clientes que no comprarán y si lo harán es muy grande.



Respecto al género, encontramos similitud en las frecuencias, siendo en ambas mayor la cantidad de clientes que no comprarán bicicletas.

El análisis previo del nuevo conjunto de datos y de estas gráficas nos va a servir para abarcar de mejor forma la etapa N.º 2. Por ahora, es una simple vista previa a las nuevas variables (la mayoría son las del conjunto de datos original, las variables nuevas son “Región” y “CantHijosEnCasa”). Cuando tengamos que deliberar qué variables utilizar para predecir qué tipo de bicicleta enviarle a cada cliente, podría pasar que estas gráficas nos ayuden a decidir cuáles podríamos usar como predictoras y cuáles no.

2da etapa: contexto

Problema

El jefe de publicidad detectó la necesidad de caracterizar a sus clientes para determinar el tipo de producto que puede llegar a interesarle (bicicleta kínder, basic o sport) para realizar marketing personalizado.

Además, el gerente de ventas nos comentó que está evaluando la posibilidad de comercializar la nueva línea de bicicletas en mercados extranjeros. Está interesado en comenzar la campaña en 3 países de características sociales y económicas similares al nuestro.

Objetivos

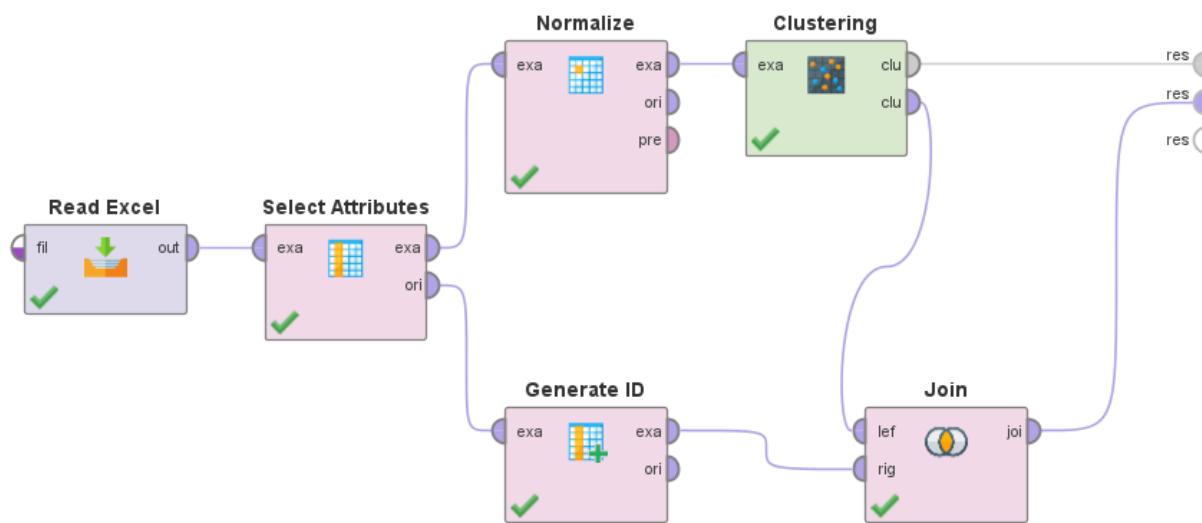
- Hallar un agrupamiento de los clientes que sirva para el negocio, teniendo en cuenta las características de los mismos. Tomando como base a las personas elegidas para enviarles la publicidad (es decir, los destinatarios), debemos encontrar una clasificación que nos indique qué tipo de publicidad enviarle a cada uno.
- Determinar tres mercados candidatos similares a Argentina para poder expandir el alcance del negocio y lanzar la campaña publicitaria en estos países.

Clustering k-medias

Proceso general

Hemos empleado el algoritmo de clustering k-medias proporcionado por el software RapidMiner para caracterizar a los destinatarios finales a quienes decidimos enviar publicidad de bicicletas. El objetivo es distinguir a los clientes en varios grupos y determinar qué tipo de publicidad enviarles (Kinder, Basic, Sport).

El RapidMiner nos quedó de la siguiente forma:



Comenzamos utilizando el nodo de lectura desde Excel para obtener los posibles clientes seleccionados previamente mediante el algoritmo de predicción en la etapa anterior.

Luego, utilizamos el nodo para seleccionar atributos y seleccionamos solamente los atributos numéricos. Esto es debido a que el algoritmo k-medias funciona mejor con atributos cuantitativos debido a su enfoque en distancias.

A continuación, normalizamos los datos en un rango de valores entre 0 y 1. Esto es importante y debe ser realizado ya que el atributo "Ingreso Anual" posee valores considerablemente más altos que los demás atributos utilizados en el análisis. Si no se realiza este paso, se podría afectar la variabilidad de los datos.

Luego de normalizar los datos, empleamos el nodo Clustering para realizar la técnica de minería de datos. Con este nodo aplicamos el algoritmo k-medias para armar "k" clusters (grupos).

Para poder utilizar el nodo Join y volver a unir los datos, tuvimos que utilizar el nodo Generate ID. Esto es debido a que excluimos el ID en la selección de atributos, ya que lo

consideramos irrelevante para el análisis. Entonces, generamos nuevas ID y unimos con las ID generadas por el cluster.

Variación del parámetro “k”

Realizamos el análisis con diferentes valores de k, comenzando por k = 3 (ya que necesitamos al menos tres grupos diferentes para los tipos de bicicletas). Luego, probamos con k = 4 y k = 5. No continuamos con valores mayores a 5, ya que la cantidad de clusters generados dificultaba la diferenciación de los clientes.

k = 3

Utilizando el parámetro k = 3 obtuvimos:

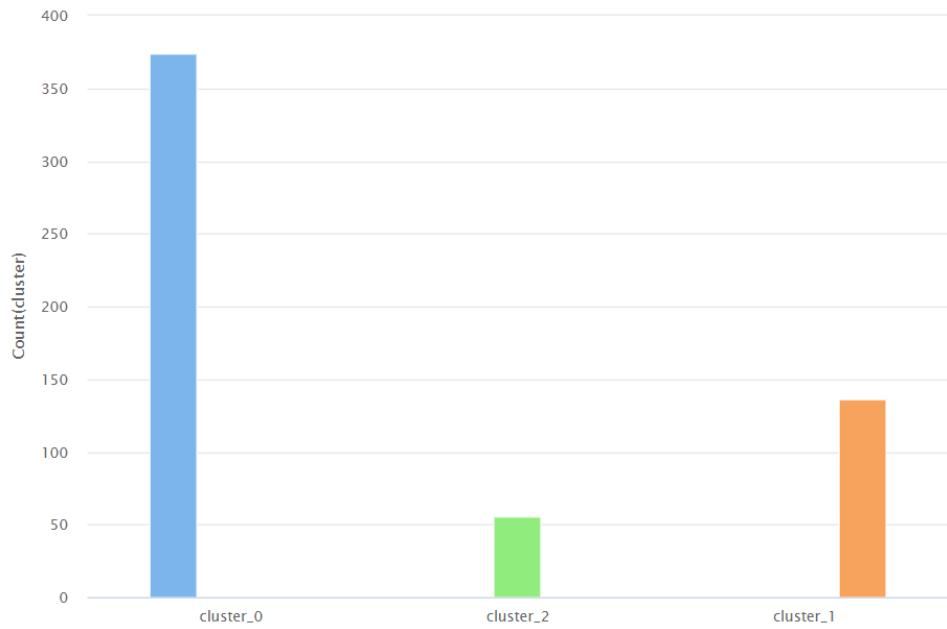
Cluster Model

```
Cluster 0: 374 items
Cluster 1: 136 items
Cluster 2: 56 items
Total number of items: 566
```

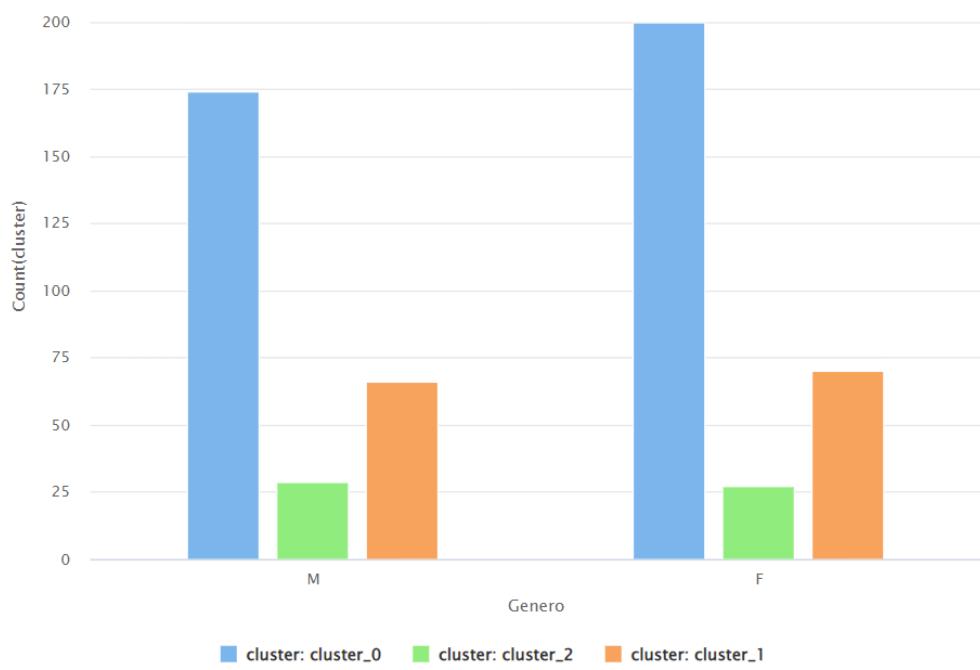
El cluster 0 tiene muchos más datos que los demás. Esto podría llegar a implicar un problema. Analizaremos detenidamente los datos de cada cluster.

Attribute	cluster_0	cluster_1	cluster_2
IngresoAnual	0.048	0.061	0.096
TotalHijos	0.287	0.229	0.389
CantHijosEnCasa	0.080	0.109	0.807
Propietario	1	0	0.946
CantAutomoviles	0.211	0.346	0.683
Edad	0.365	0.341	0.386

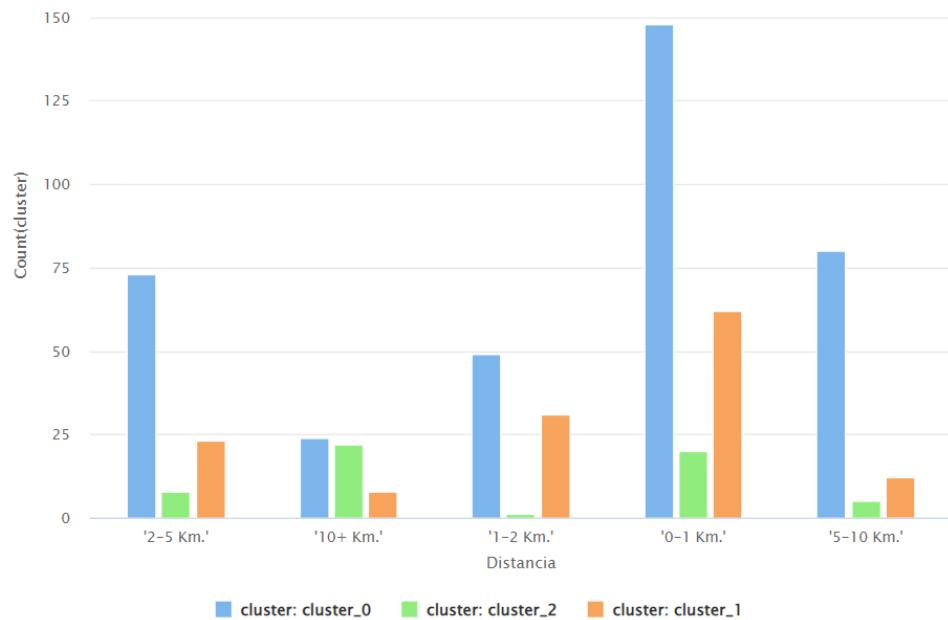
Para poder realizar un mejor análisis, decidimos generar gráficas para algunas variables. Acompañados de estas gráficas podremos realizar una conclusión y decidir qué tipo de bicicleta asignarle a cada cluster.



Primero observamos la distribución de datos por cluster, siendo el primero el que más observaciones tiene por mucho.



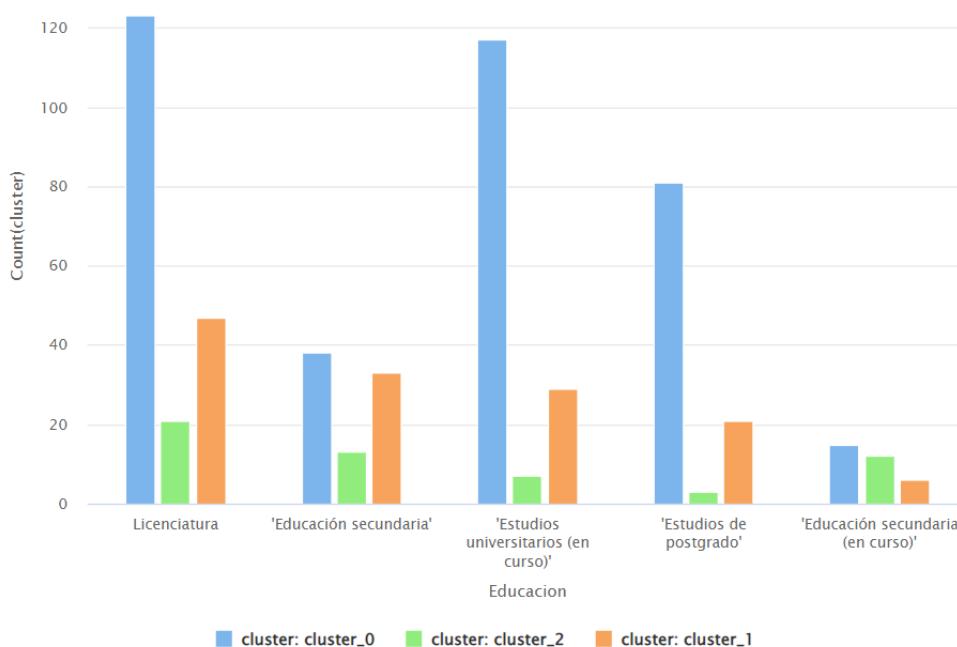
En cuanto al género, observamos que los 3 clusters se clasificaron casi igualitariamente, no encontrando una gran diferencia en algún caso como para que sea relevante.



En cuanto a la distancia:

- Cluster 0: la mayoría de los clientes de este cluster, se encuentran entre 0-1 km de distancia al trabajo, seguido por los clientes que se encuentran entre 5 y 10 km, y en un menor porcentaje, entre 2 y 5 km.
- Cluster 1: podemos observar que también los clientes pertenecientes a este cluster, en su mayoría se encuentran entre 0-1 km de distancia del trabajo, seguido por los clientes que se encuentran entre 1 y 2 km y luego con muy poca diferencia, en 2 y 5 km.
- Cluster 2: la mayoría de sus clientes se encuentran a más de 10 km de distancia o entre 0-1 km.

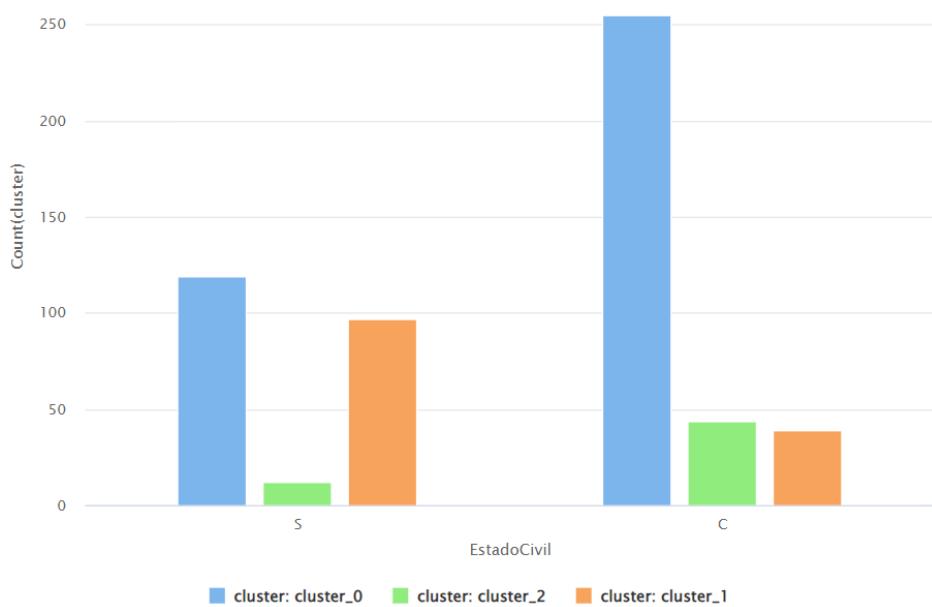
Observando esto, no podemos realizar ninguna conclusión.



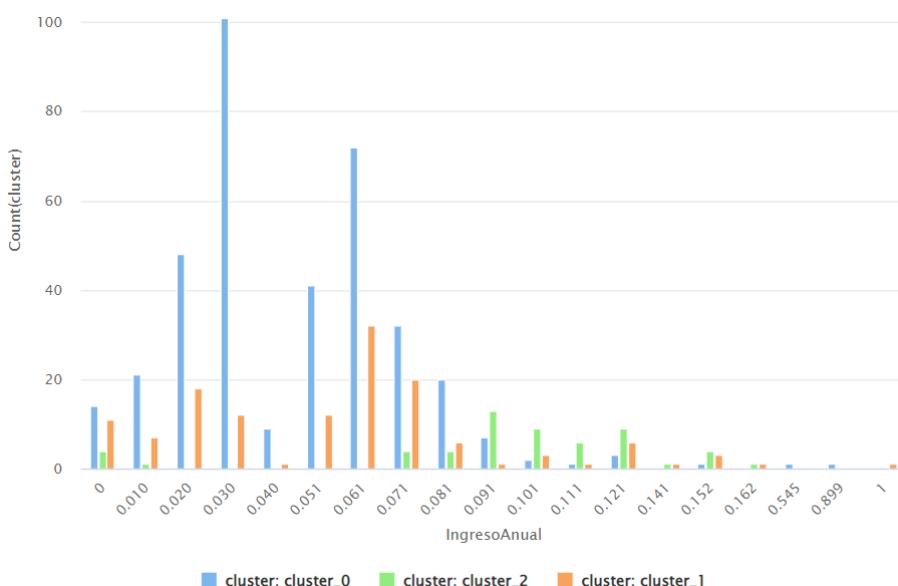
En cuanto a la educación:

- Cluster 0: la mayoría de sus clientes tienen una licenciatura, estudios universitarios (en curso) o estudios de postgrado.
- Cluster 1: no existe alguna diferencia tan grande como para que sea relevante.
- Cluster 2: observamos que la cantidad de clientes que se encuentran con educación secundaria en curso, es mayor que aquellos que tienen estudios de postgrado o estudios universitarios en curso.

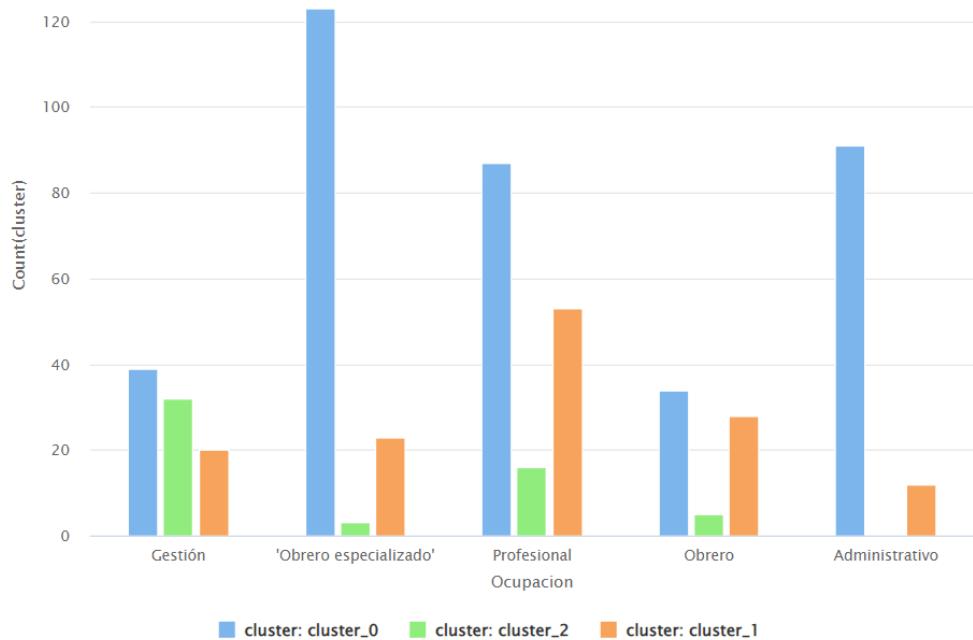
Como conclusión, podemos decir que los clientes pertenecientes al cluster 0 en su mayoría son personas con un título terciario o con estudios para realizar uno. De los otros clusters no se puede decir mucho.



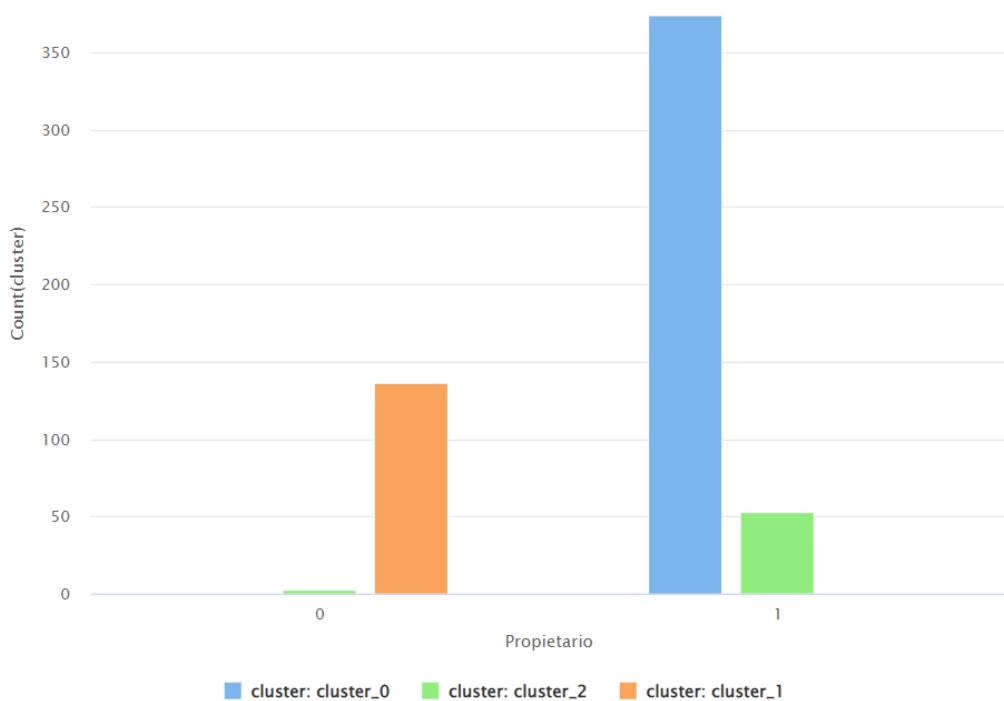
En cuanto al estado civil, podemos observar que en los clusters 0 y 2 se agruparon (en su mayoría) aquellos clientes con estado civil casado, mientras que en el cluster 1, la mayoría de sus clientes están solteros.



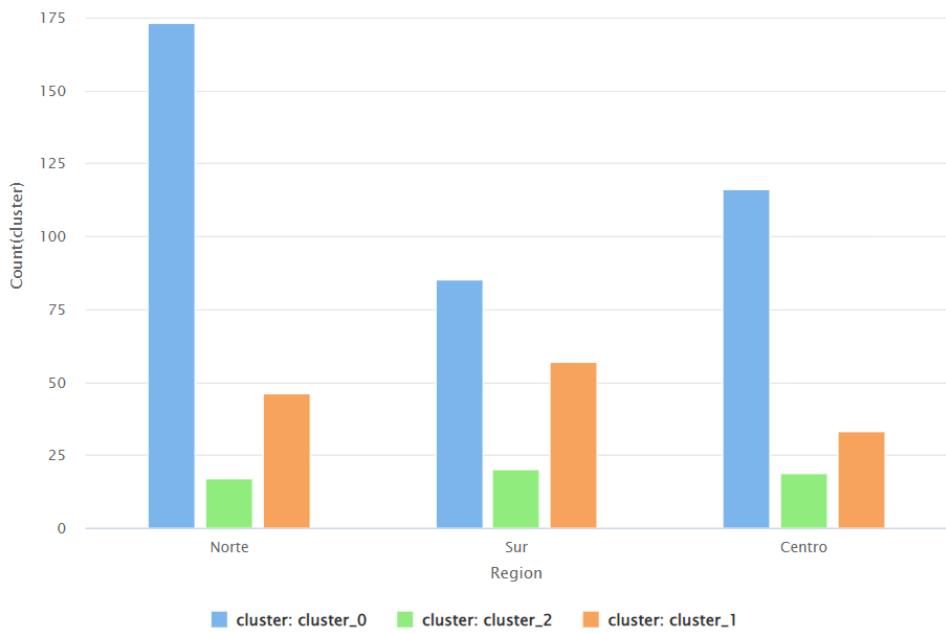
Si bien los valores verdaderos fueron normalizados entre 0 y 1, se puede ver que el cluster 2 tiene a las personas con un ingreso anual casi mayor a la media. En cambio, el cluster 0 y el cluster 1 tienen a los clientes con un ingreso anual no muy elevado.



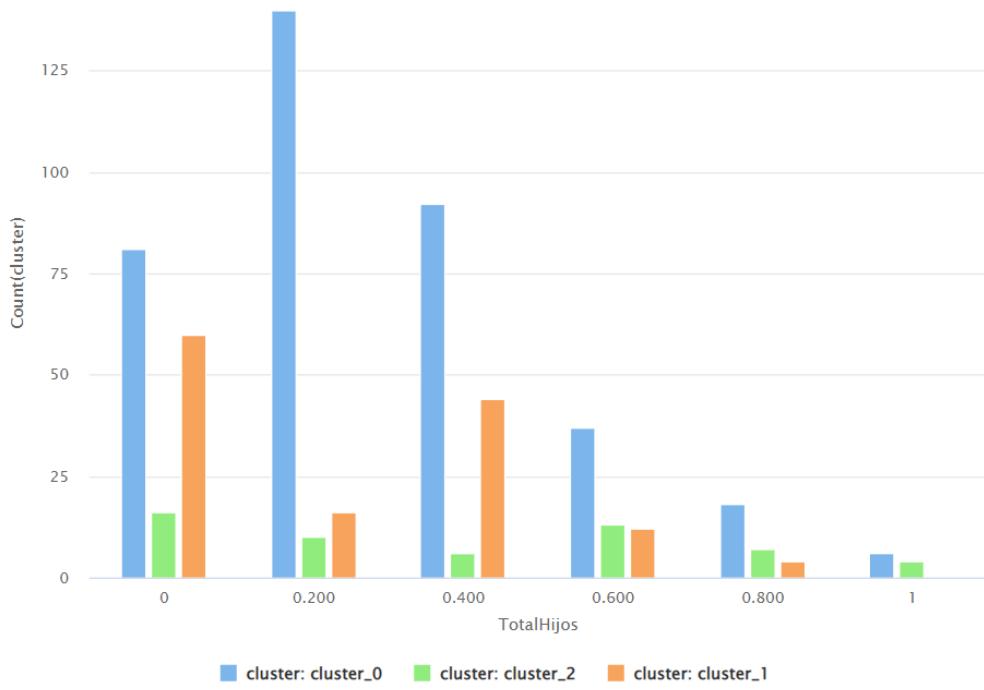
Para la ocupación, viendo la gráfica, no vemos nada interesante.



En cuanto a si el cliente es propietario o no, los clientes del cluster 0 son todos propietarios, mientras que los del cluster 1 no lo son. Los clientes pertenecientes al cluster 2 son casi en su totalidad propietarios.

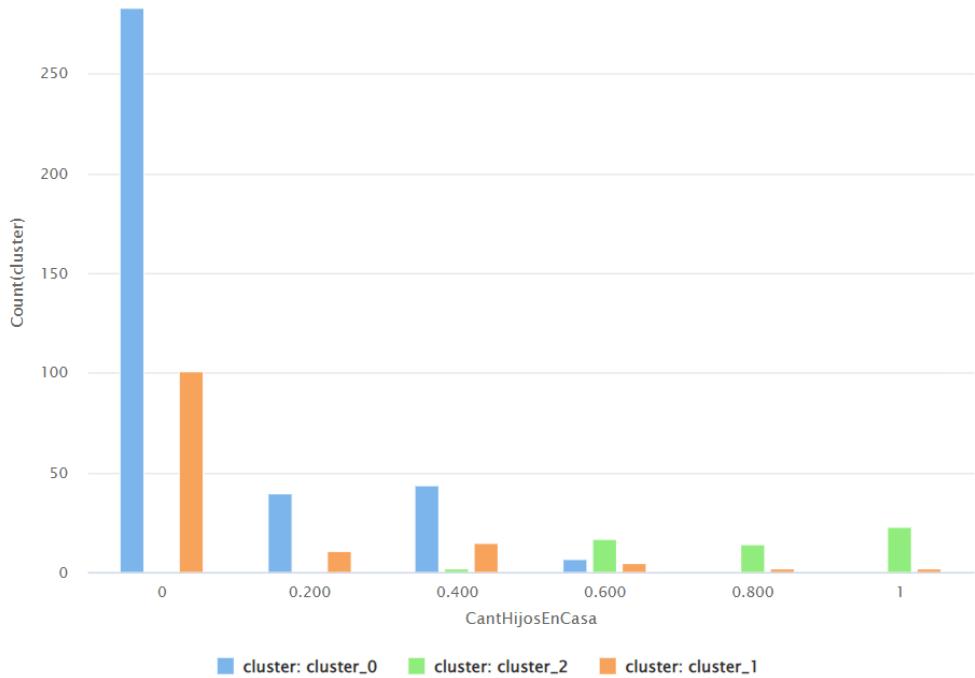


En el cluster 0 observamos una mayor cantidad de clientes de región norte, mientras que en los otros clusters no existe tanta diferencia entre sus regiones.

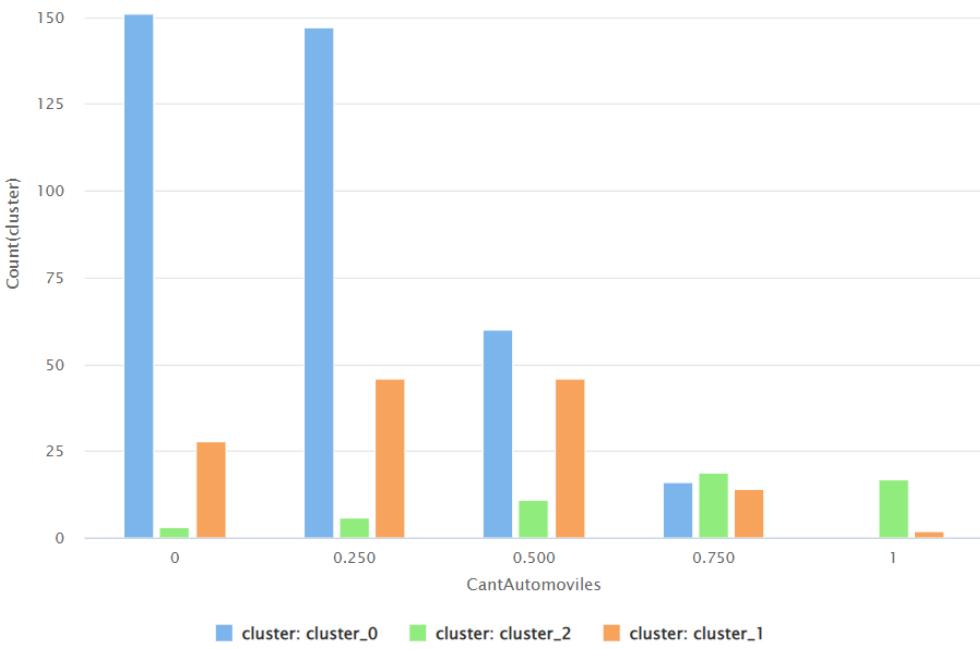


En el cluster 0, la mayoría de los clientes poseen 1 o más hijos. En el cluster 1, predomina la cantidad de clientes sin hijos. En cuanto al cluster 2, podemos decir que la mayoría de sus clientes tienen 1 o más hijos.

Creemos que esta información va a ser relevante para decidir qué publicidad enviar.



Podemos ver que en el cluster 0, la mayoría de los clientes no posee hijos en casa al igual que en el cluster 1. En el cluster 2 todos los clientes tienen mínimo un hijo en casa. Creemos que esta información también va a ser relevante debido a que nos interesa saber si los clientes son padres y cuántos hijos tienen (para ver si se les puede ofrecer la bicicleta kinder).



En el cluster 0, la mayoría no tienen automóvil o tienen solo uno. En el cluster 1 la mayoría tiene uno o más autos y en el cluster 2 la mayoría posee dos o más. Esta información puede llegar a ser relevante porque nos indica si el cliente tiene transporte propio.
Luego de analizar las gráficas anteriores, realizamos las siguientes conclusiones:

- Cluster 0: contiene un 66% de las observaciones (374).
 - En este cluster se encuentran aquellos clientes que son propietarios, poseen un ingreso anual medio/bajo, entre 0 y 2 hijos en su mayoría (y dentro de estos un porcentaje menor los tiene en su casa), en su mayoría no poseen automóvil, pero una gran parte posee solo 1 y ser de edad adulta.
En cuanto a estos datos analizados, creemos que aquellos que aportan más información son:
 - Todos los clientes de este cluster son propietarios de vivienda.
 - Aun siendo todos propietarios, el ingreso anual de estos clientes es el más bajo respecto a los demás clusters.
 - Es el cluster con mayor observaciones, siendo estas un 66% de las totales, pudiendo concluir que la mayoría de la población (mínimo un 66%) es propietario.
- Cluster 1: contiene un 24% de las observaciones (136).
 - En este cluster se encuentran aquellos clientes que no son propietarios, poseen un ingreso anual medio, mayormente entre 0 y 2 hijos, tienen en su mayoría entre 0 y 2 autos y son de edad adulta.
En cuanto a estos datos analizados, creemos que aquellos que aportan más información son:
 - Todos los clientes pertenecientes a este cluster no son propietarios de vivienda.
 - Sin ser propietarios, la cantidad de automóviles de los clientes de este cluster es mayor que la del primer cluster.
- Cluster 2: contiene un 9,9% de las observaciones (56).
 - En este cluster se encuentran aquellos clientes que en su mayoría (casi totalidad) son propietarios, poseen un ingreso anual medio, en su mayoría poseen un hijo o más (casi en su totalidad se encuentran en sus casas), poseen 3 automóviles como mínimo y se encuentran los de mayor edad respecto a los demás clusters.
En cuanto a estos datos analizados, creemos que aquellos que aportan más información son:
 - Este cluster posee casi un 100% de clientes propietarios, solo 3 de ellos no lo son.
 - Se observa que es el cluster con mayor ingreso anual, mayor cantidad de automóviles y además propietarios. Por ende, podemos concluir que en este cluster se encuentran aquellos clientes que mejor económicamente se encuentran. Esto, además se refleja en la cantidad de hijos y en el hecho de que la mayoría vive con ellos.
 - Se encuentran los clientes con mayor edad.

Las bicicletas que decidimos para cada cluster son:

- Cluster 0: basic o kinder.
- Cluster 1: sport.
- Cluster 2: kinder.

A pesar de haber obtenido resultados claros, estos son provisarios. No podemos asegurar la calidad de los mismos por lo que seguiremos intentando con distintos clusters de mayor "k".

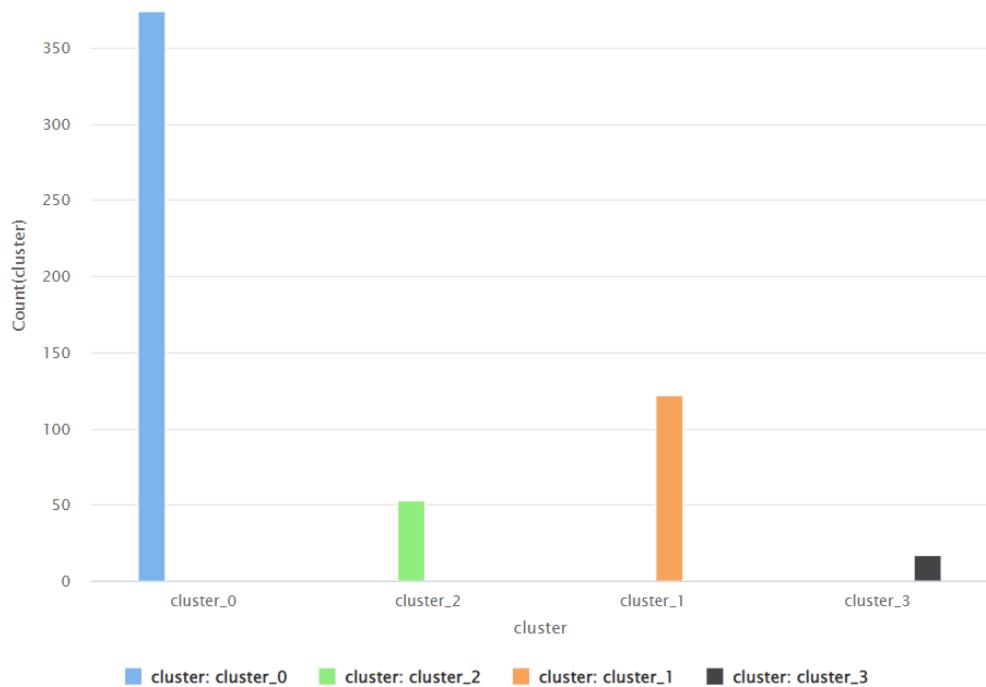
k = 4

Con el parámetro k = 4 obtuvimos:

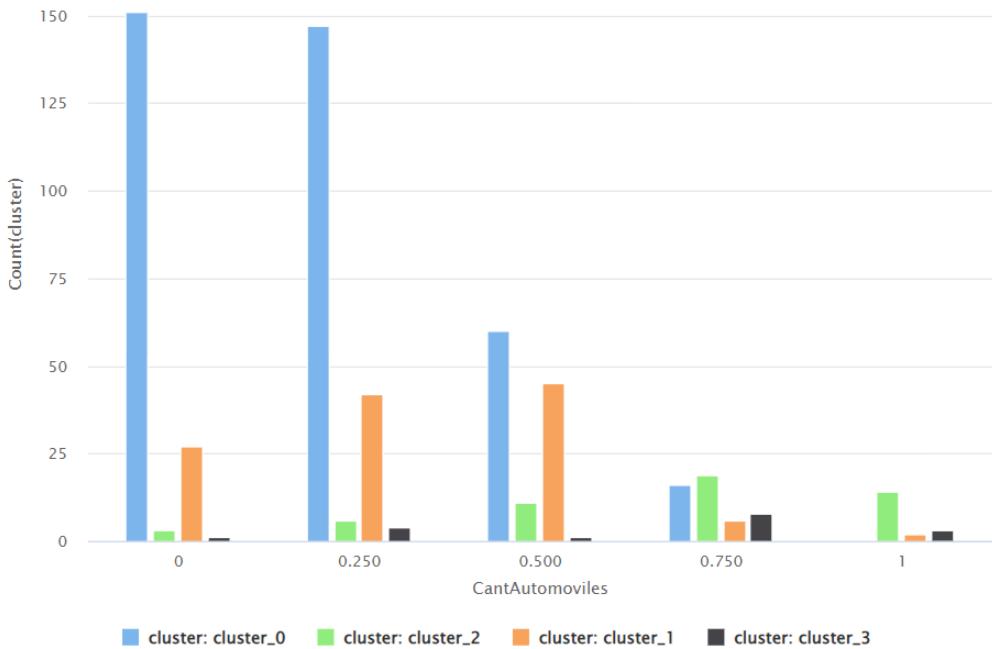
Cluster Model

```
Cluster 0: 374 items
Cluster 1: 122 items
Cluster 2: 53 items
Cluster 3: 17 items
Total number of items: 566
```

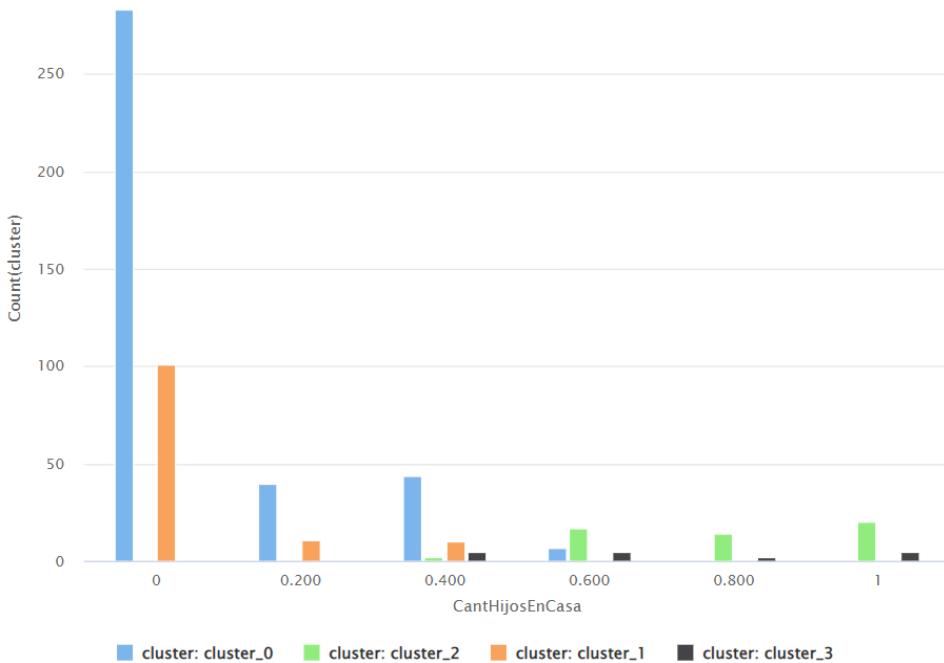
Attribute	cluster_0	cluster_1	cluster_2	cluster_3
IngresoAnual	0.048	0.056	0.093	0.112
TotalHijos	0.287	0.203	0.358	0.541
CantHijosEnCasa	0.080	0.051	0.796	0.682
Propietario	1	0	1	0
CantAutomoviles	0.211	0.324	0.665	0.618
Edad	0.365	0.328	0.387	0.439



El cluster 0 tiene muchos más datos que los demás. Analizaremos detenidamente los datos de cada cluster.

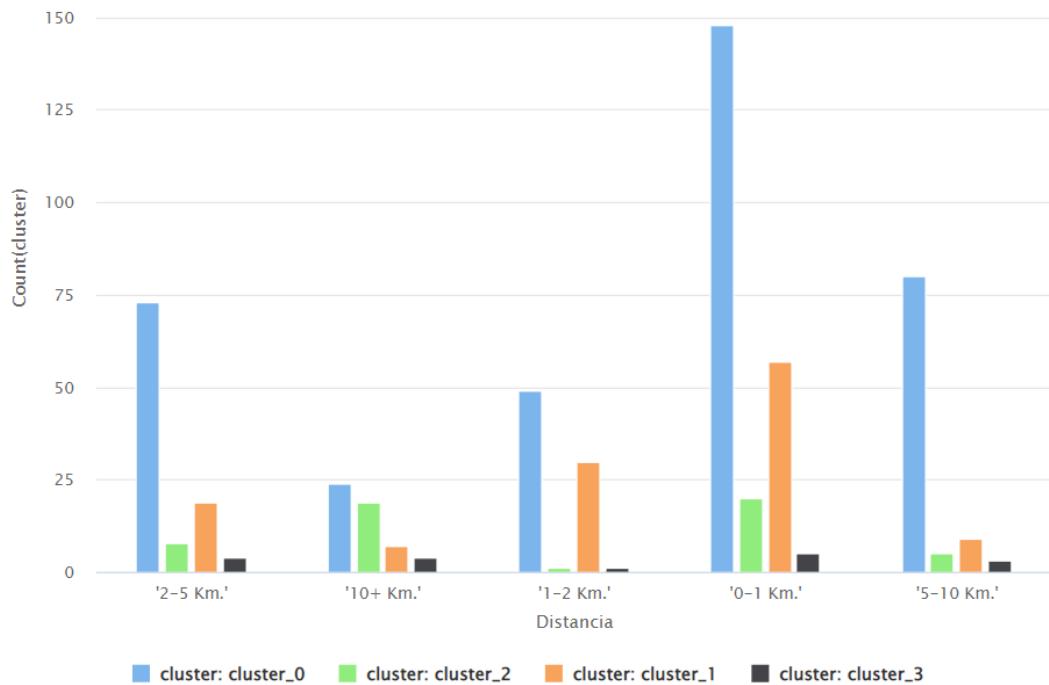


Observamos que en el cluster 0 la mayoría de los clientes no posee automóvil o solo posee uno, en el cluster 1 la cantidad de clientes sin automóvil es menor, siendo mayor la cantidad de clientes con 2 automóviles que 1. En el cluster 2 observamos que existe un mayor porcentaje de clientes con 3 o 4 autos, lo mismo para el cluster 3.

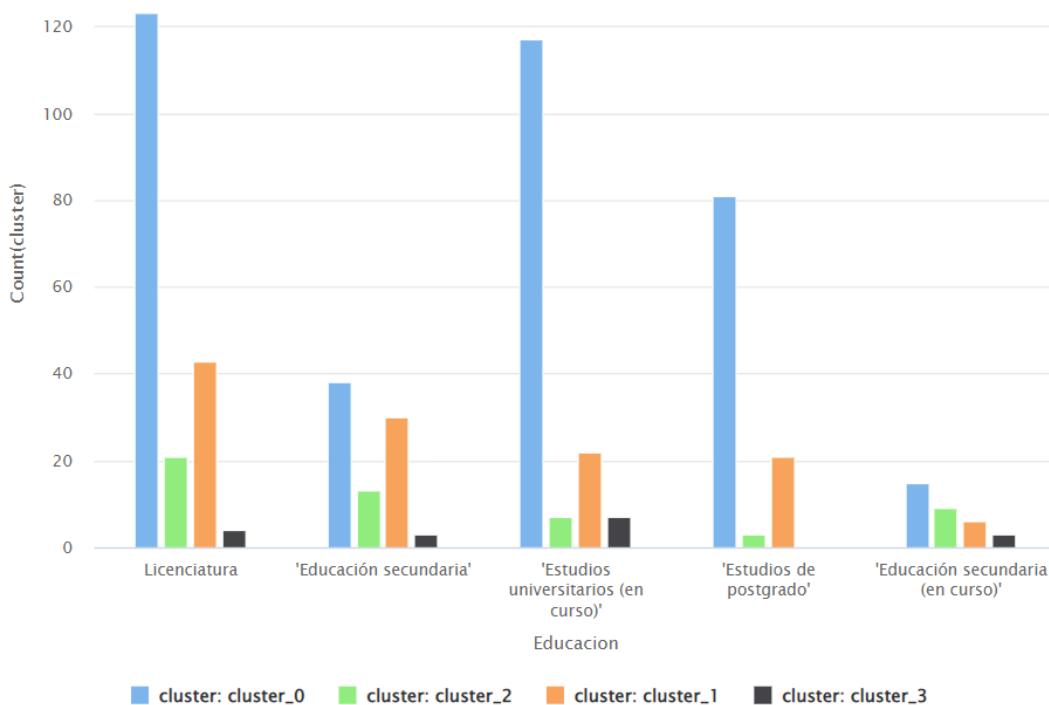


En cuanto a la cantidad de hijos en casa, observamos que en el cluster 0 y 1 la mayoría de los clientes no tienen ningún hijo en casa, mientras que en los clusters 2 y 3 los clientes tienen mínimo 2. Entonces, podríamos separar los 4 clusters en 2 categorías, siendo una

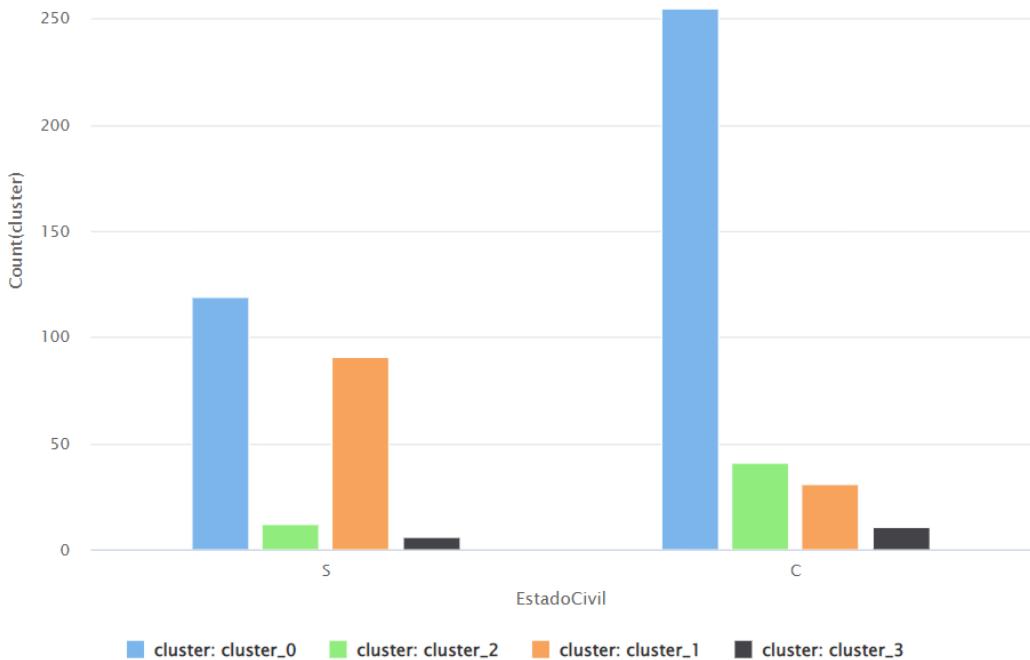
“Tiene hijos en casa” y otra “No tiene hijos en casa”, esta clasificación nos interesa para la categoría kinder de bicicletas.



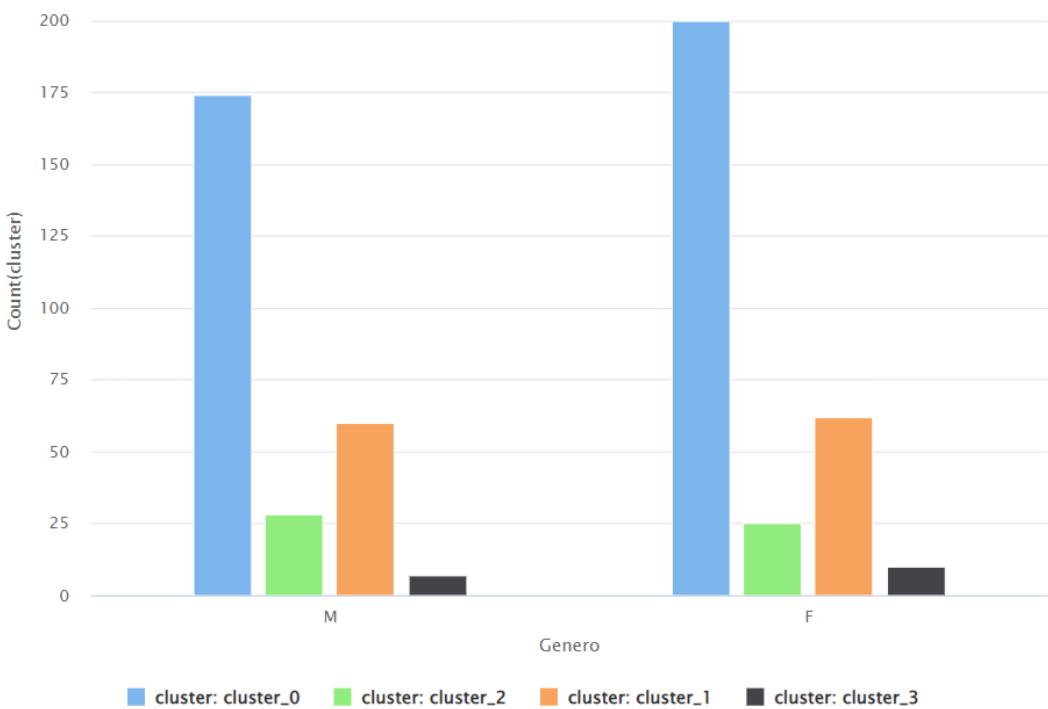
Observamos que en los clusters 0 y 1 la mayoría de los clientes se encuentran entre 0 y 1 km de distancia al trabajo. En el cluster 2 la mayoría de los clientes se encuentran entre 0 y 1 km o más de 10 kms, mientras que en el cluster 3 encontramos solo que la minoría de clientes se encuentra entre 1 y 2 kilómetros.



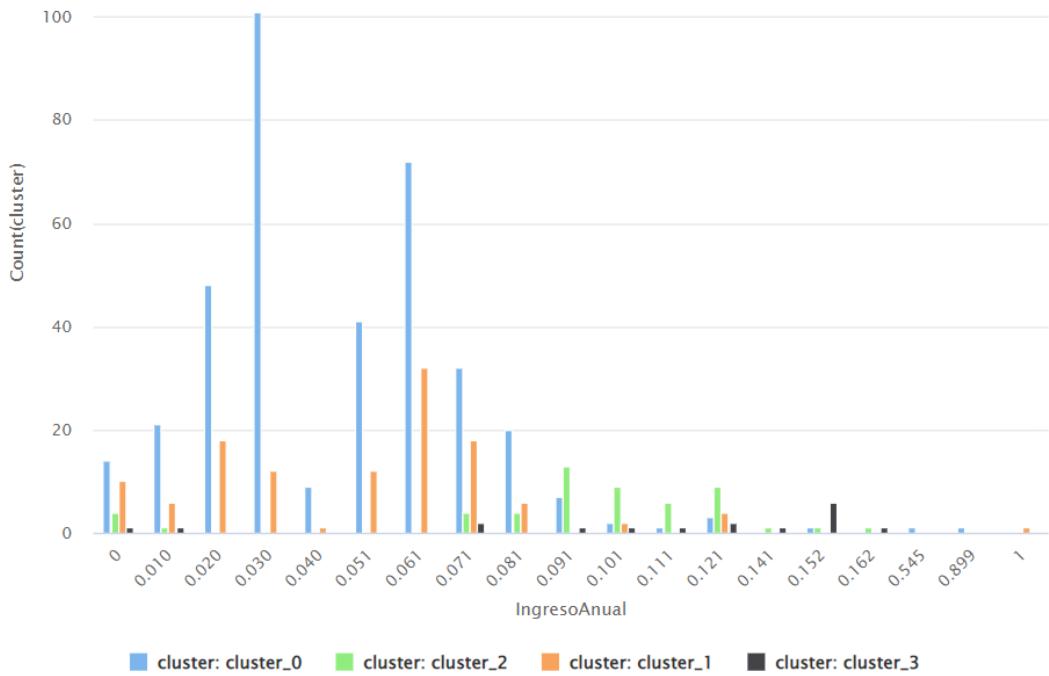
En cuanto a la educación, no observamos información relevante para el estudio.



En cuanto al estado civil, observamos que en el cluster 0 y en el cluster 2 los clientes están mayormente casados, mientras que en el cluster 1 son en su mayoría solteros. En el cluster 3 no hay una gran diferencia entre los solteros y casados.

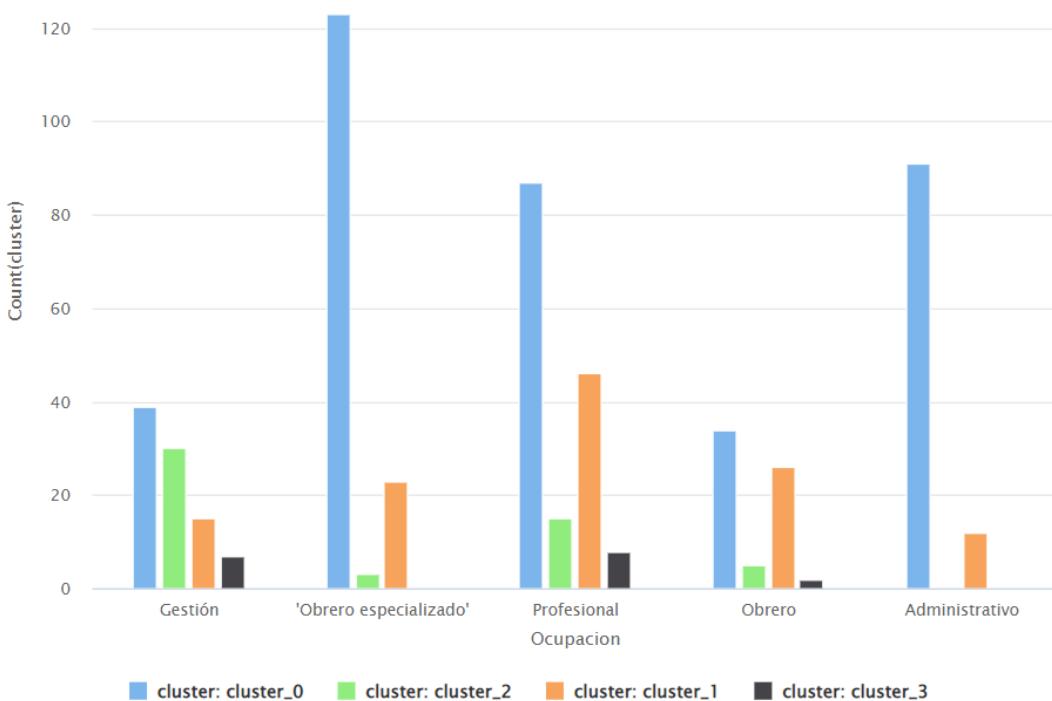


Sobre el género de los clientes, encontramos una distribución casi igualitaria en todos los clusters.

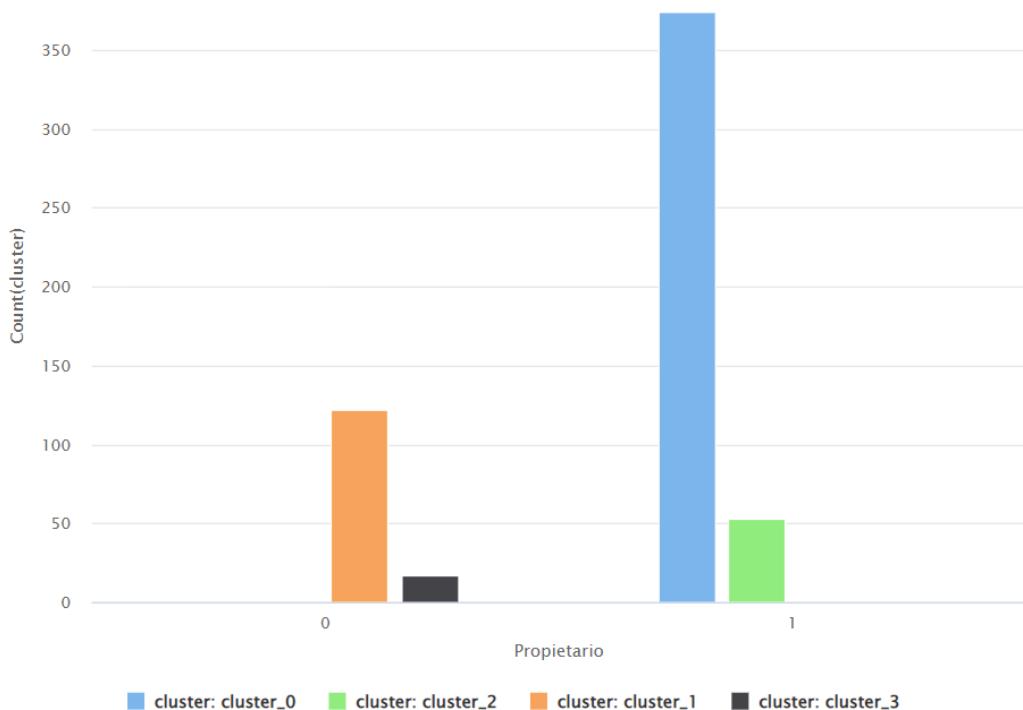


Acerca del ingreso podemos decir que en el cluster 0 la mayoría de los clientes tienen un ingreso medio/bajo, en el cluster 1 un ingreso medio, en el cluster 2 medio/alto y en el cluster 3 la mayoría de los clientes poseen un ingreso anual alto.

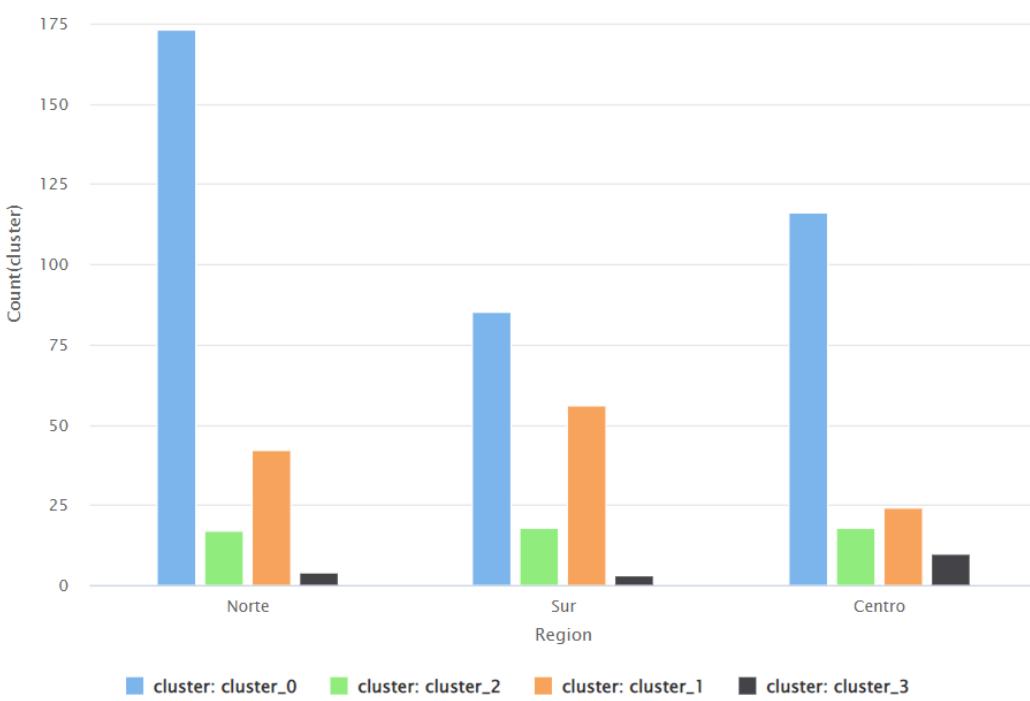
Nos interesa esta información para conocer el poder adquisitivo de los clientes y saber que clase de bicicleta ofrecer.



Observando la gráfica de ocupaciones de los clientes, no encontramos nada interesante.



En cuanto a si son propietarios o no, observamos que los clientes del cluster 0 y 2 si lo son, mientras que los del cluster 1 y 3 no.



De forma general, de los clusters obtenidos podemos decir:

- Cluster 0: Clientes con ingreso anual bajo, sin hijos en casa en su mayoría, propietarios, con 1 o ningún automóvil.
- Cluster 1: Clientes con ingreso anual medio/bajo, sin hijos en casa, no son propietarios, tienen 1 o más automóviles en su mayoría.

- Cluster 2: Clientes con ingreso anual medio, con hijos en casa en su mayoría, son propietarios y tienen varios automóviles.
- Cluster 3: Clientes con un ingreso anual alto, hijos en casa mayormente, no son propietarios y tienen varios automóviles.

Conclusiones de sobre los grupos formados (tipo de publicidad a enviar):

- Cluster 0: basic.
- Cluster 1: basic.
- Cluster 2: kinder - sport.
- Cluster 3: kinder - sport.

k = 5

Utilizando el parámetro k = 5 obtuvimos:

Cluster Model

```

Cluster 0: 156 items
Cluster 1: 122 items
Cluster 2: 227 items
Cluster 3: 44 items
Cluster 4: 17 items
Total number of items: 566

```

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
IngresoAnual	0.050	0.056	0.046	0.103	0.112
TotalHijos	0.536	0.203	0.134	0.282	0.541
CantHijosEnCasa	0.188	0.051	0.029	0.818	0.682
Propietario	1	0	1	1	0
CantAutomoviles	0.234	0.324	0.194	0.761	0.618
Edad	0.444	0.328	0.309	0.403	0.439

Sin embargo, luego de ver esta tabla, optamos por no generar las gráficas para cada variable ya que resulta muy complicado diferenciar los grupos. Por lo tanto, no seguiremos con el análisis de k-medias con k = 5.

Entonces, obtuvimos dos modelos con k-medias: uno para k = 3 y otro para k = 4. El análisis que realizamos con k = 3 es bastante convincente como para tomar una decisión. Sin embargo, procederemos a estudiar los demás tipos de clustering para obtener más modelos y así decidir cuál consideramos mejor. Procederemos ahora con cluster jerárquico.

Clustering jerárquico

La técnica de clustering jerárquico genera un árbol jerárquico en el que las observaciones se sitúan en las hojas, mientras que los nodos representan subconjuntos que pueden utilizarse como grupos. Para analizar el modelo generado, se genera un gráfico llamado “dendrograma”.

Utilizaremos Python y el software SPSS Statistics. Una vez obtenido el dendrograma, analizaremos los niveles de corte en los que las distancias entre elementos del mismo grupo sean mínimas y las distancias entre grupos sean mayores. Esto nos permitirá determinar la cantidad óptima de grupos.

Uso de Python

Para realizar el clustering jerárquico decidimos primero utilizar Python. Para ello, generamos clusters utilizando todos los datos y luego una muestra de 100 observaciones.

Nos dimos cuenta que, según la transformación de las variables, cambiaba el tamaño de los clusters generados. En otras palabras, no es lo mismo utilizar normalización que estandarización.

Además, también variamos el criterio de asignación de clusters entre los parámetros “distance” (distancia) y “maxclust” (cantidad máxima de clusters).

En resumen, los parámetros que variamos para generar los clusters fueron:

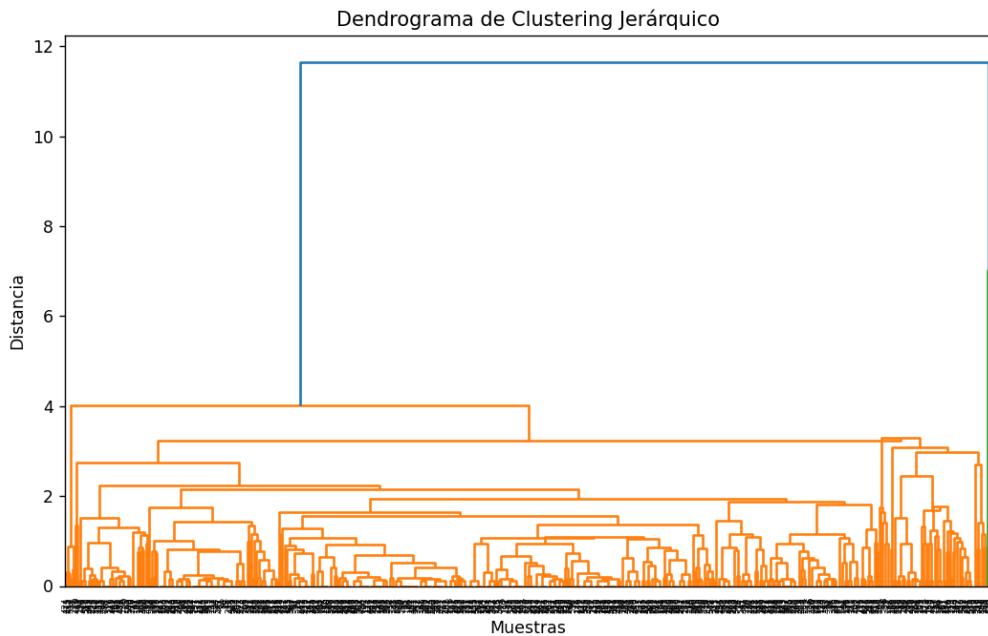
- Cantidad de observaciones: todas (566) y 100.
- Transformación de las variables numéricas: normalización (MinMaxScaler) y estandarización (standardScaler).
- Asignación de clusters: distance y maxclust.

Para cada normalización y estandarización generamos un dendrograma. Luego, según la asignación de clusters utilizada, generamos una gráfica que muestra la cantidad de clusters seleccionada por el parámetro y además, la cantidad de observaciones por cluster.

A continuación, procederemos a mostrar las gráficas.

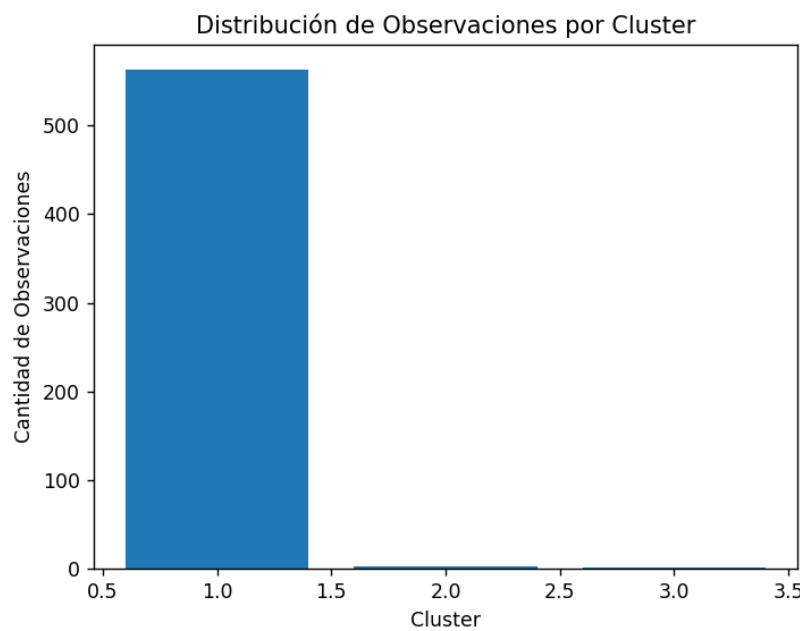
Todos los datos

Con estandarización



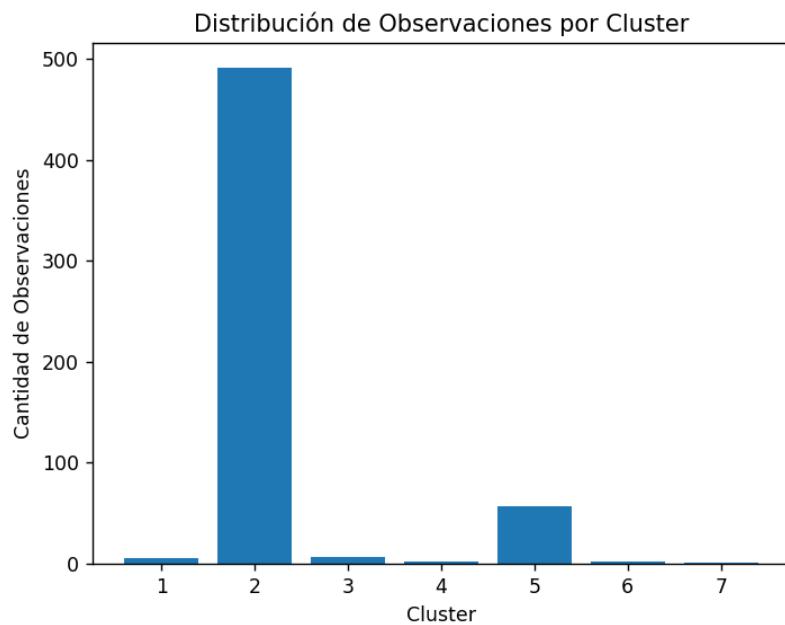
Con el dendrograma generado podríamos llegar a realizar un corte que genere 3 o 4 clusters. Sin embargo, habrá un cluster que tendrá casi el 95% de los datos. Veamos cuantos clusters asignan los parámetros.

Criterio “distance”



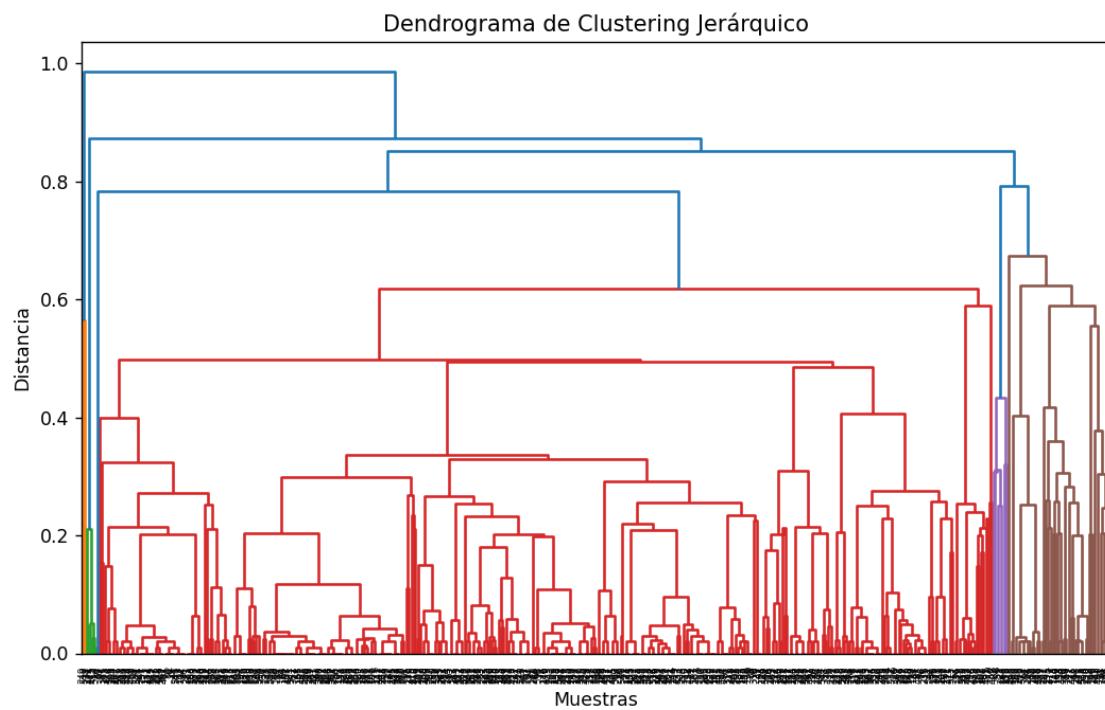
Como sospechábamos, se generó un cluster que contiene casi la totalidad de las observaciones.

Criterio “maxclust”



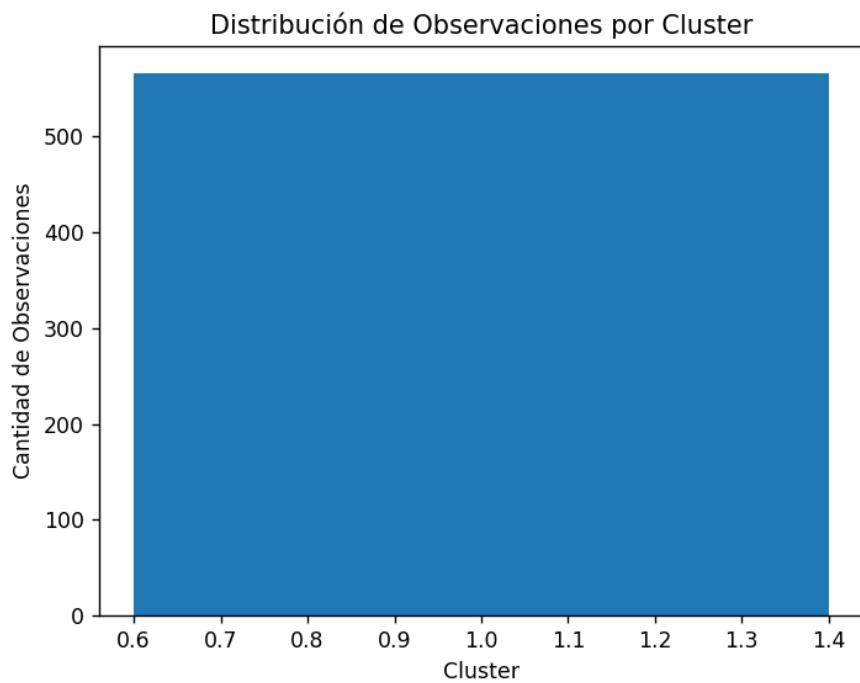
Se generaron siete clusters. Sin embargo, casi la totalidad de las observaciones se encuentran en tan solo dos.

Con normalización



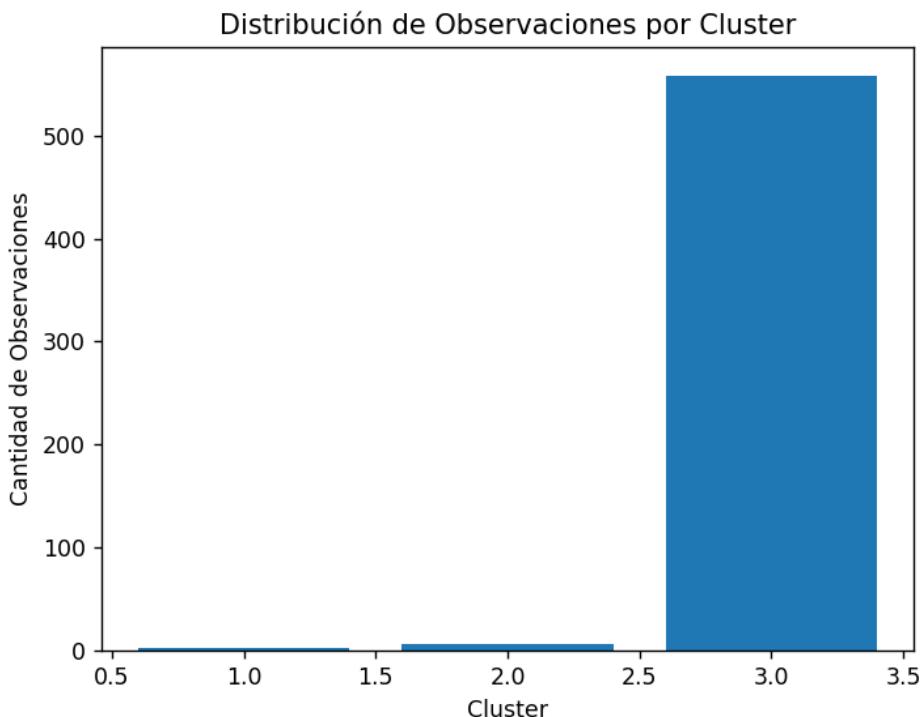
A simple vista podríamos decir que obtuvimos un mejor dendrograma que con la estandarización. Vamos a ver cuántos clusters se generan.

Criterio “distance”



Se generó un solo cluster. Claramente esto no sirve, lo descartaremos.

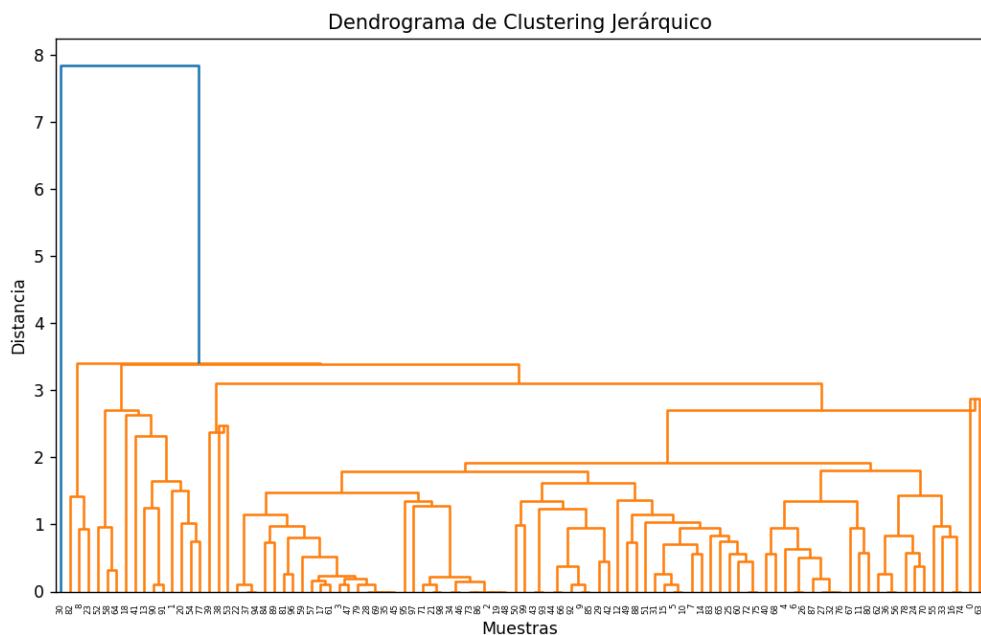
Criterio “maxclust”



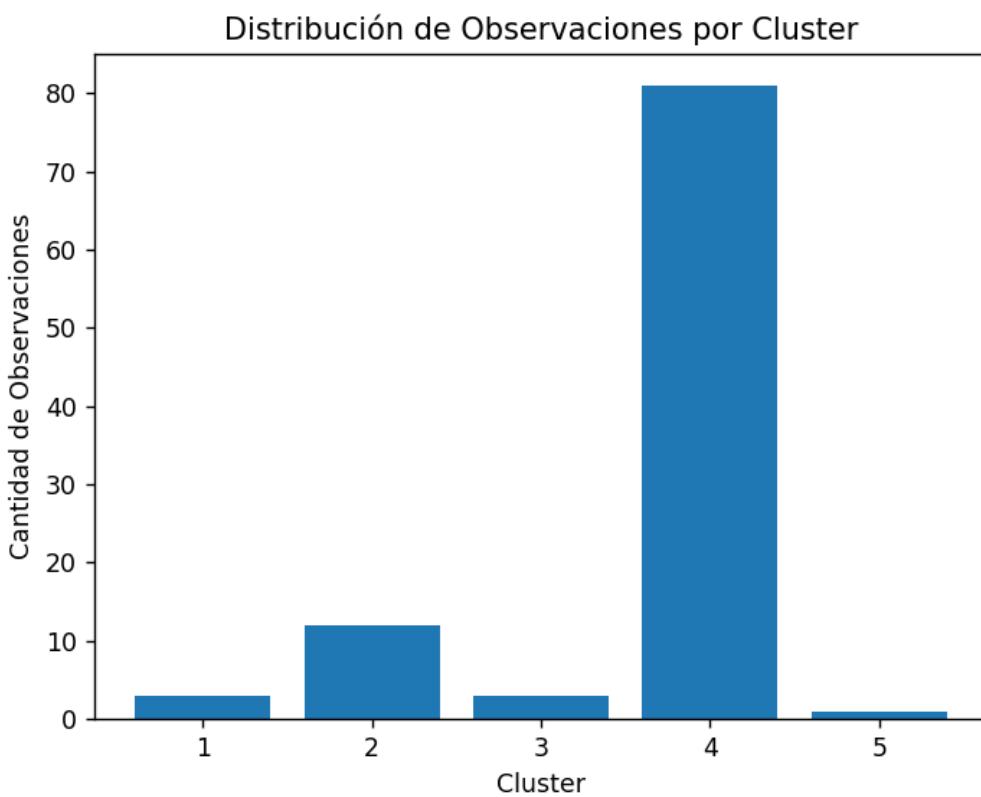
Otra vez se generó un cluster que contiene casi todas las observaciones. Esto no nos sirve.

Muestra de 100 observaciones

Con estandarización

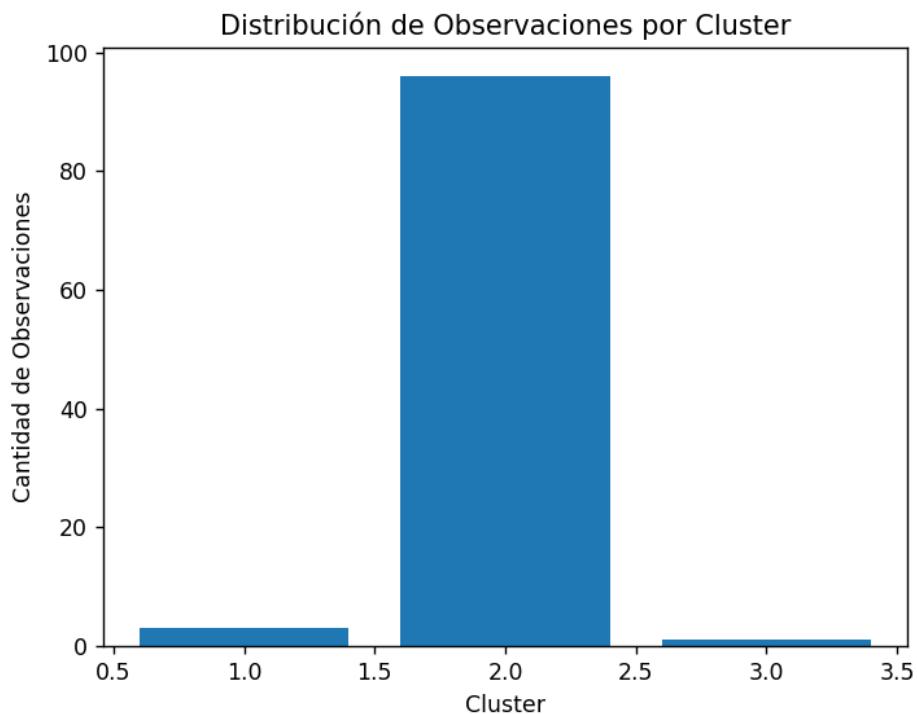


Criterio “distance”

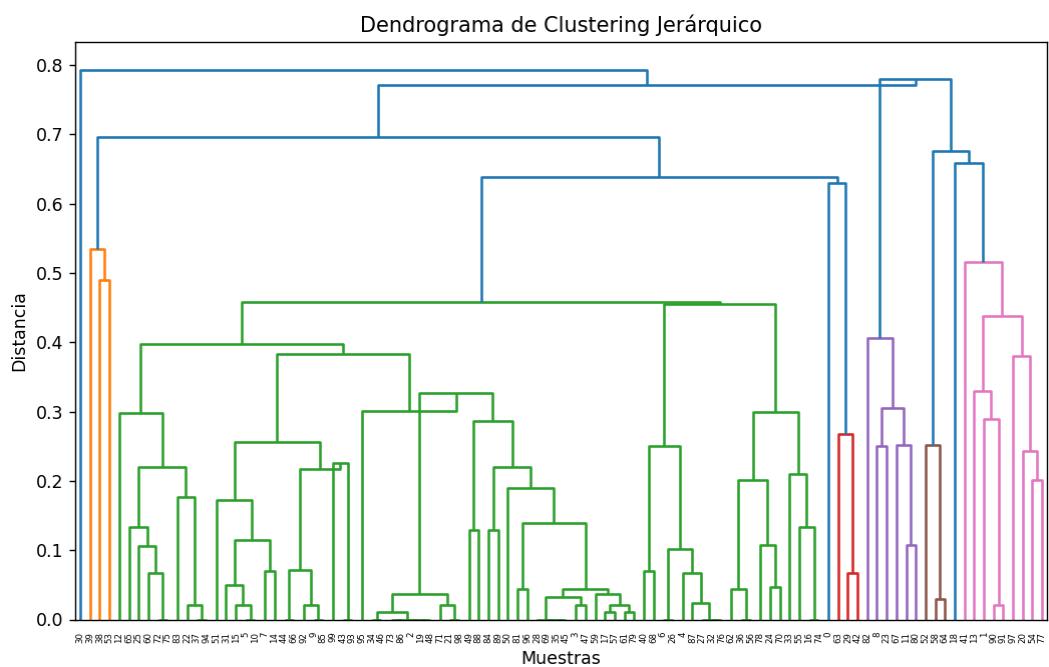


Se generaron 5 clusters, pero el 80% de los datos se encuentran en solo uno.

Criterio “maxclust”

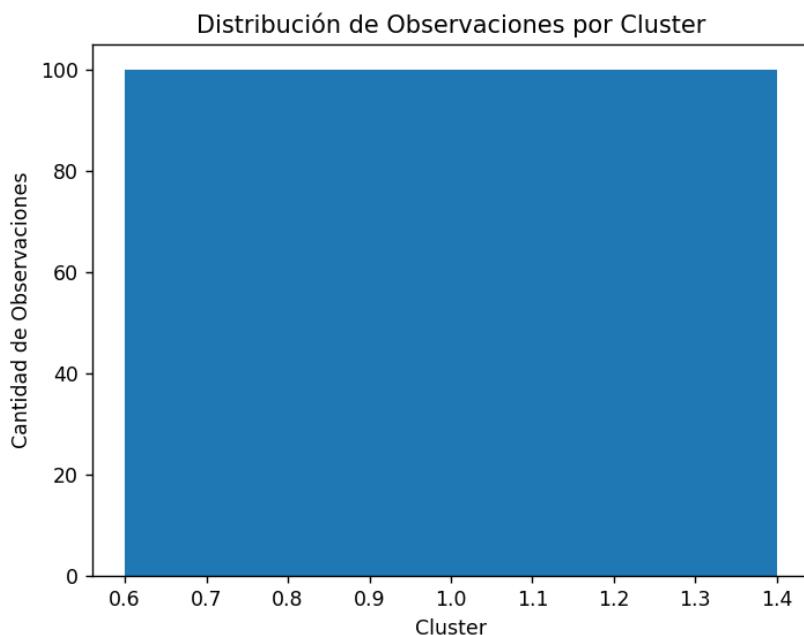


Con normalización

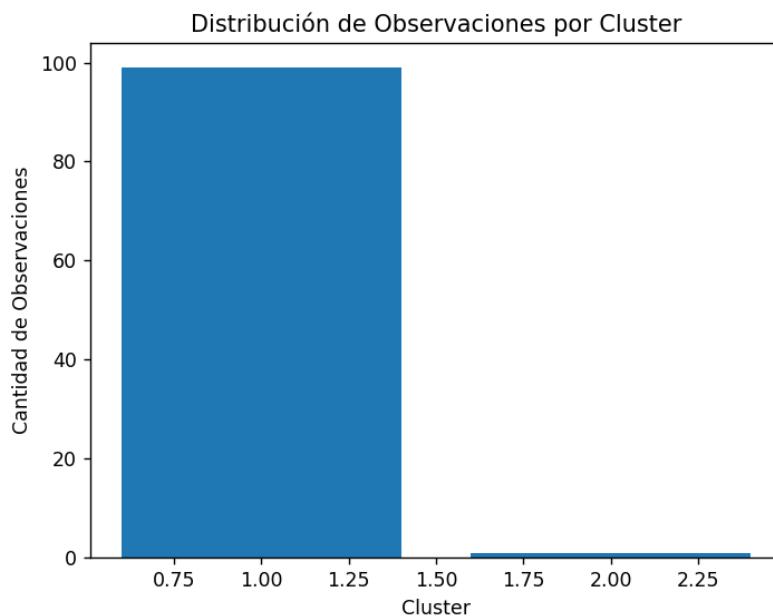


El dendrograma parecería ser bastante decente. Sin embargo, con un corte de 3 clusters, parecería ser que un cluster tendría solo una observación y otro tendría un 80%. Observemos cuantos clusters se generan.

Criterio “distance”



Criterio “maxclust”



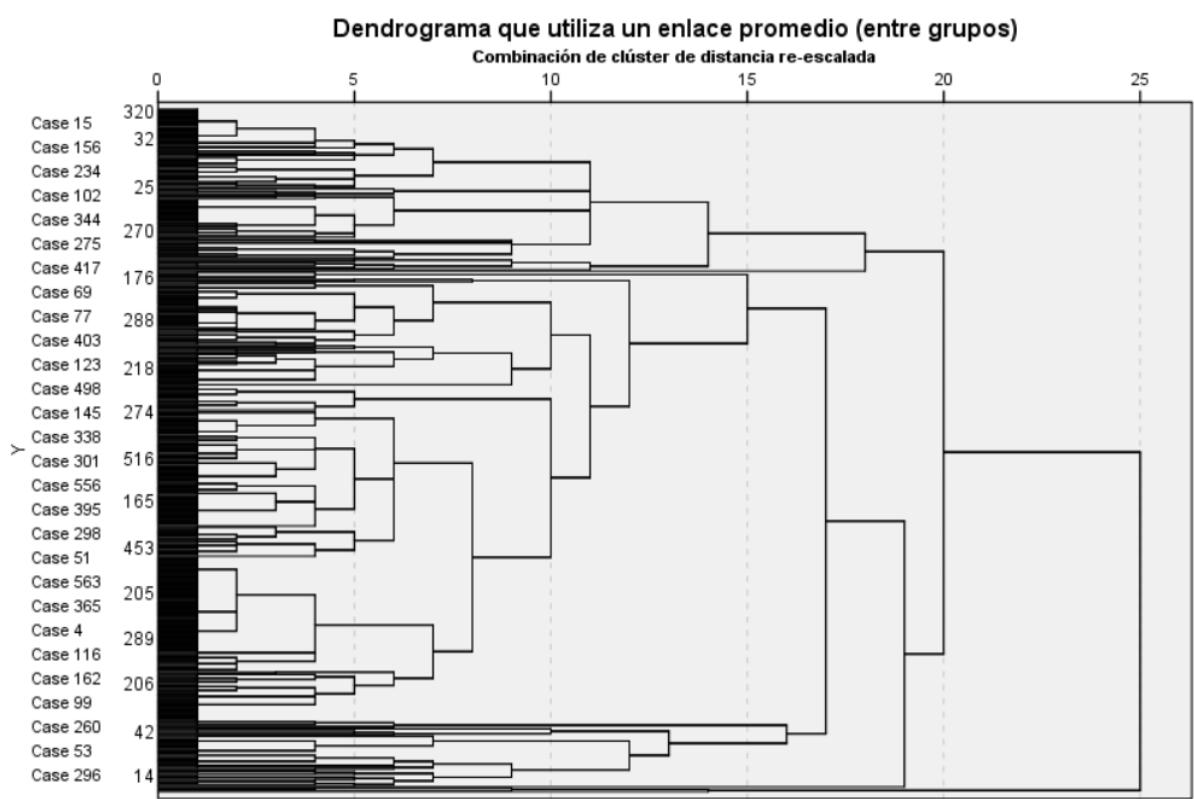
Conclusión

Concluimos que los clusters jerárquicos generados en Python no son de muy buena calidad. Algunos clusters tienen muchísimas observaciones a comparación de los demás, no podemos distinguir los tipos de clientes, algunos parámetros generaron solo un cluster, etc.

Por lo tanto, decidimos emplear el software SPSS Statistics para ver si obtenemos mejores resultados. De más está decir que descartamos los clústeres obtenidos en Python.

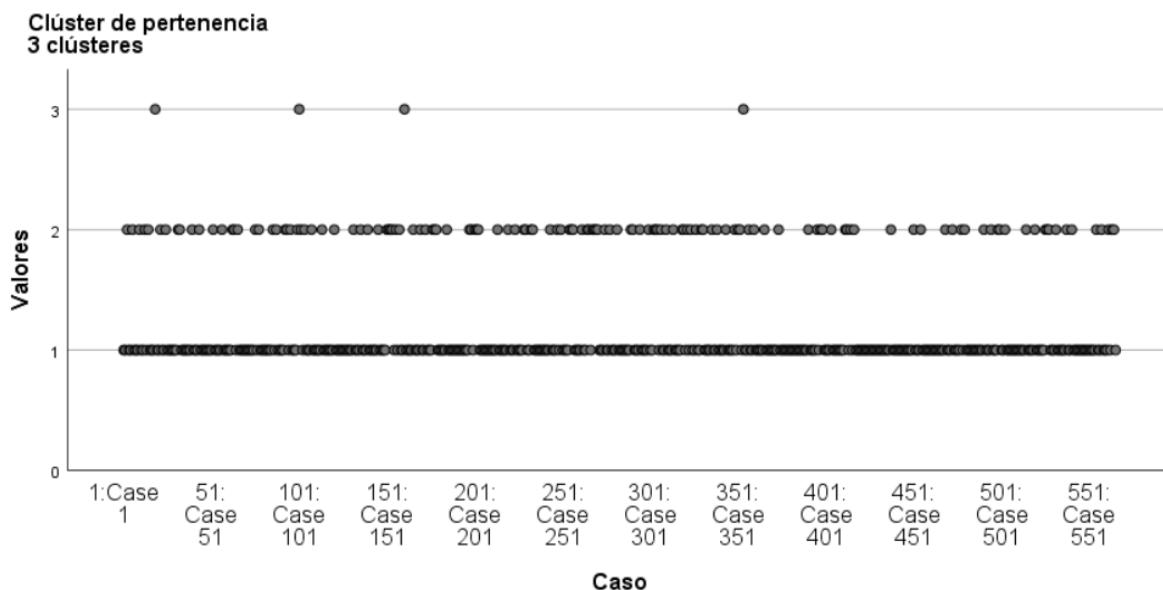
Uso de SPSS Statistics

Utilizamos la distancia euclídea y normalizamos entre 0 y 1. Obtuimos el siguiente dendrograma:



Realizaremos tres, cuatro y cinco cortes y analizaremos los clusters obtenidos.

3 clusters



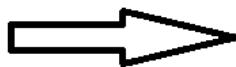
Para analizar mejor, decidimos pasar la tabla generada por el software a Excel. El objeto de esto es poder filtrar la tabla para obtener la cantidad de observaciones de cada cluster (SPSS Statistics no nos proporcionaba ninguna opción de filtrado).

SPSS Statistics

Clúster de pertenencia	
Caso	3 clústeres
1:Case 1	1
2:Case 2	1
3:Case 3	2
4:Case 4	1
5:Case 5	1
6:Case 6	2
7:Case 7	1
8:Case 8	1
9:Case 9	1
10:Case 10	2
11:Case 11	1
12:Case 12	1
13:Case 13	2
14:Case 14	1
15:Case 15	2

Excel

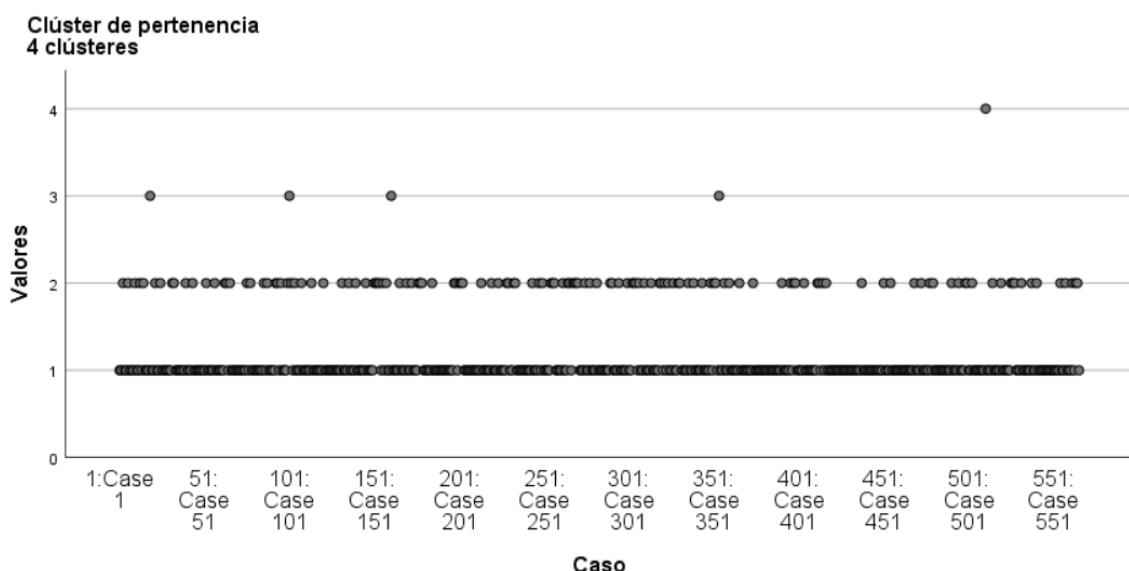
Clúster de pertenencia	
Caso	3 clústeres
1:Case 1	1
2:Case 2	1
3:Case 3	2
4:Case 4	1
5:Case 5	1
6:Case 6	2
7:Case 7	1
8:Case 8	1
9:Case 9	1
10:Case 10	2
11:Case 11	1
12:Case 12	1
13:Case 13	2
14:Case 14	1
15:Case 15	2



Con esto obtuvimos la siguiente información:

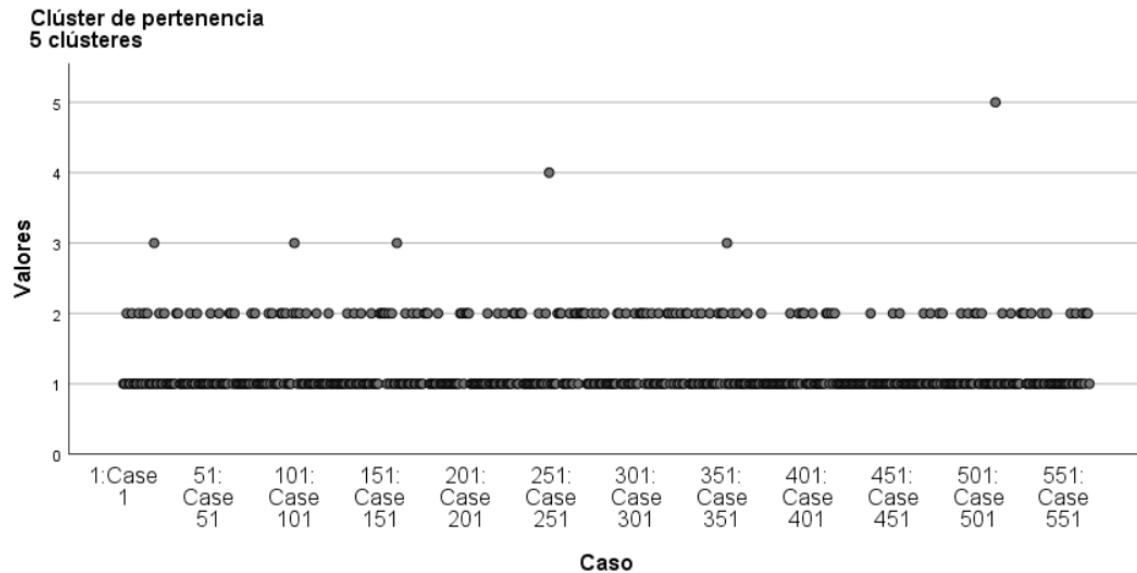
- Cluster 1: 427 observaciones.
- Cluster 2: 135 observaciones.
- Cluster 3: 4 observaciones.

4 clusters



- Cluster 1: 426 observaciones.
- Cluster 2: 135 observaciones.
- Cluster 3: 4 observaciones.
- Cluster 4: 1 observación.

5 clusters



- Cluster 1: 426 observaciones.
- Cluster 2: 134 observaciones.
- Cluster 3: 4 observaciones.
- Cluster 4: 1 observación.
- Cluster 5: 1 observación.

Conclusión

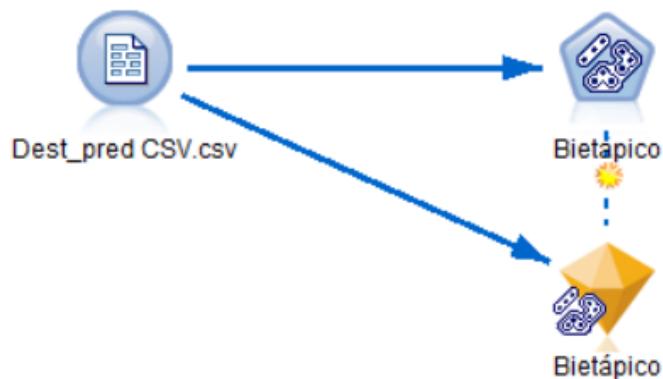
Creemos que los resultados son malos debido a que siempre hay un cluster con muy pocas observaciones. Tanto en Python como en SPSS Statistics se obtuvieron resultados similares, por lo que descartamos clustering jerárquico para realizar la clasificación de clientes.

Por ahora, nos quedamos con clustering k-means que hicimos en RapidMiner y procederemos a analizar clustering bietápico para ver si obtenemos mejores resultados.

Clustering bietápico

El clustering bietápico es un método de agrupación en clústeres que actúa como modelo híbrido, ya que combina el funcionamiento clustering jerárquico y no jerárquico. Este tipo de clustering es útil para la conformación e interpretación de clusters de grandes volúmenes de datos.

Para realizar el clustering bietápico, empleamos SPSS Modeler. El modelo nos quedó de la siguiente forma:



Luego de generar todos los modelos, nos quedó:



Cómo se puede ver en la imagen de arriba, fuimos variando las variables y el número de clusters:

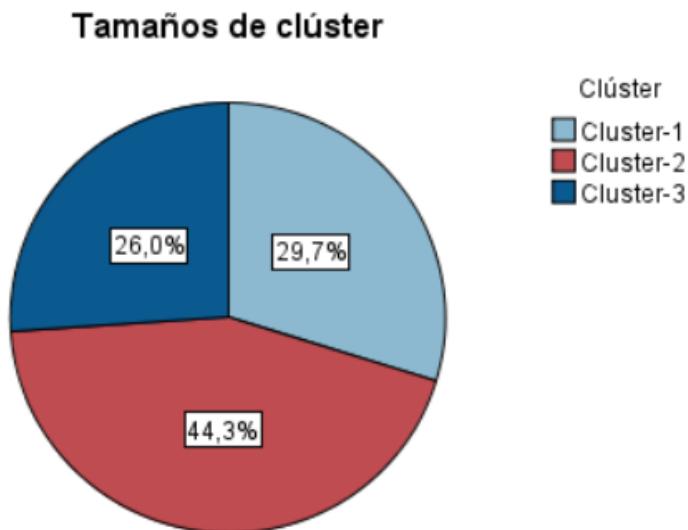
- Número de clusters: fuimos variando de 3 a 5 clusters.
- Variables:
 - Todas las variables.
 - Todas las variables categóricas.
 - Dos variables categóricas.
 - Todas las variables numéricas.
 - Algunas variables numéricas.

Todas las variables

Con 3 clusters:



Lo bueno de SPSS Modeler es que nos genera la calidad del cluster utilizando la medida de silueta. Utilizaremos la calidad proporcionada por el software para decidir con qué cluster bietápico nos quedaremos. Se considera que el modelo es decente con una silueta mayor a 0.5. Con una silueta de 0.7 o más, se considera que el modelo es muy bueno.



También nos genera un gráfico de torta que muestra el porcentaje de registros de cada cluster. A partir de aquí mostraremos los modelos que generamos y solo comentaremos los “decentes”, descartando aquellos cuya calidad es mala.

Con 4 clusters:



Con 5 clusters:



No ponemos la gráfica de tamaño de los clusters debido a que la calidad de los modelos es mala.

Todas las variables categóricas

Con todas las variables categóricas:

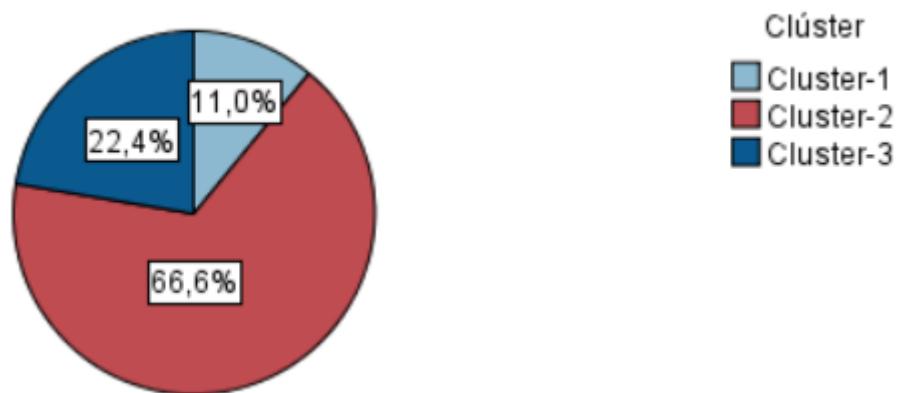


Todas las variables numéricas



Utilizando todas las variables numéricas obtuvimos el primer modelo decente, con una medida de silueta de 0.6.

Tamaños de clúster



Probamos agregando algunas variables categóricas para ver si mejoraba el modelo.

Agregando “EstadoCivil” y “Distancia”:



Disminuyó la calidad del cluster.

Probamos ahora quitando esas dos variables y agregando “Educación” y “Ocupación”:



La calidad disminuyó aún más. No fue buena idea agregar variables categóricas.

Lo mejor es seguir trabajando con todas variables numéricas. Probamos con 4 clusters:



Con 5 clusters:

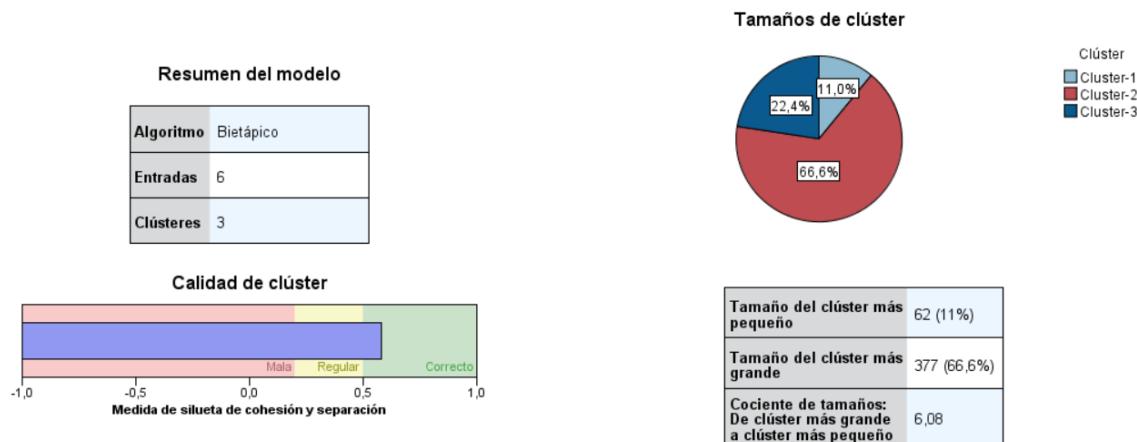


Con 4 y 5 clusters la calidad disminuyó. Es curioso que el modelo de 4 clusters sea peor que el de 5. Claramente el modelo con 3 clusters fue el mejor. Tendremos en cuenta esta información a futuro.

Numéricas sin algunas variables

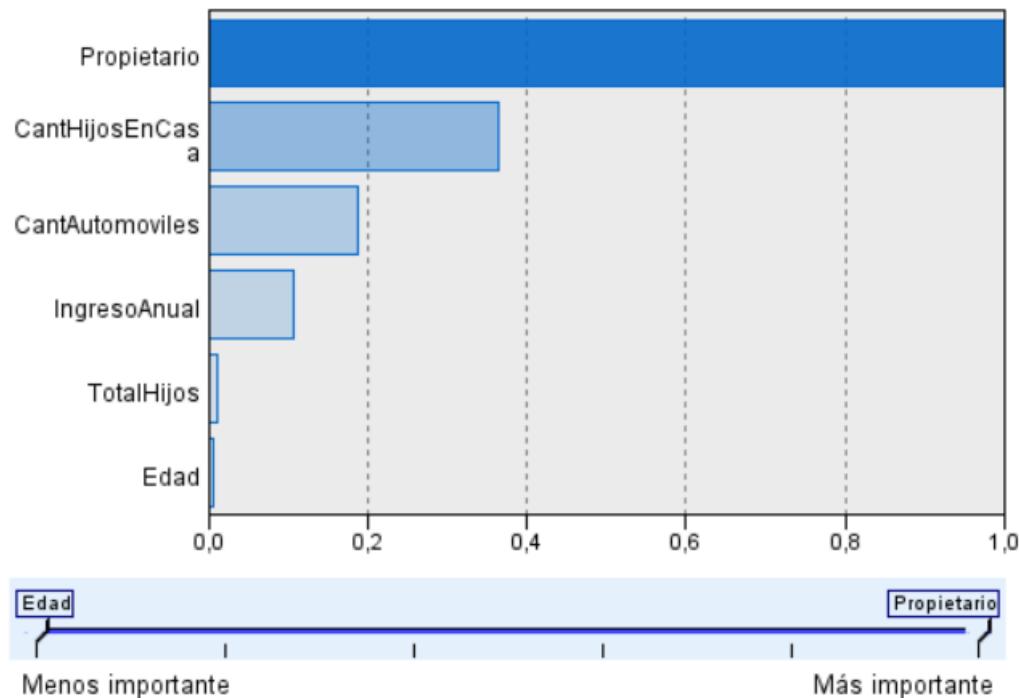
Hasta aquí, el mejor modelo es el generado por todas las variables numéricas con 3 clusters. Por lo tanto, vamos a trabajar a partir de ese. Decidimos quitar algunas variables para probar si aumenta o disminuye la calidad. Para decidir qué variable quitar, nos fijamos la importancia del predictor en el mejor modelo generado hasta ahora: todas numéricas y 3 clusters.

Pero primero, queremos recordar como es el modelo:

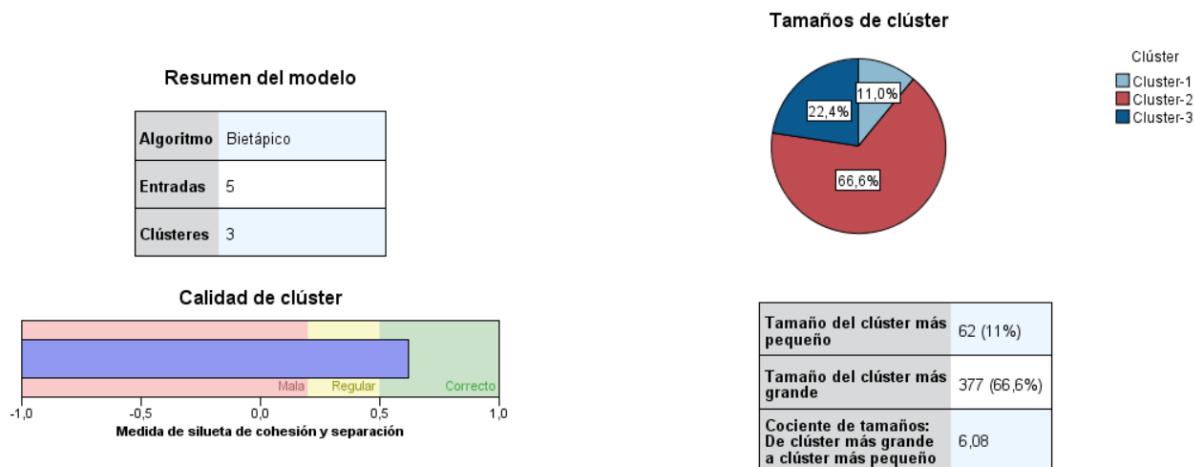


La silueta promedio es de casi 0.6. Procedemos a analizar ahora los predictores:

Importancia del predictor

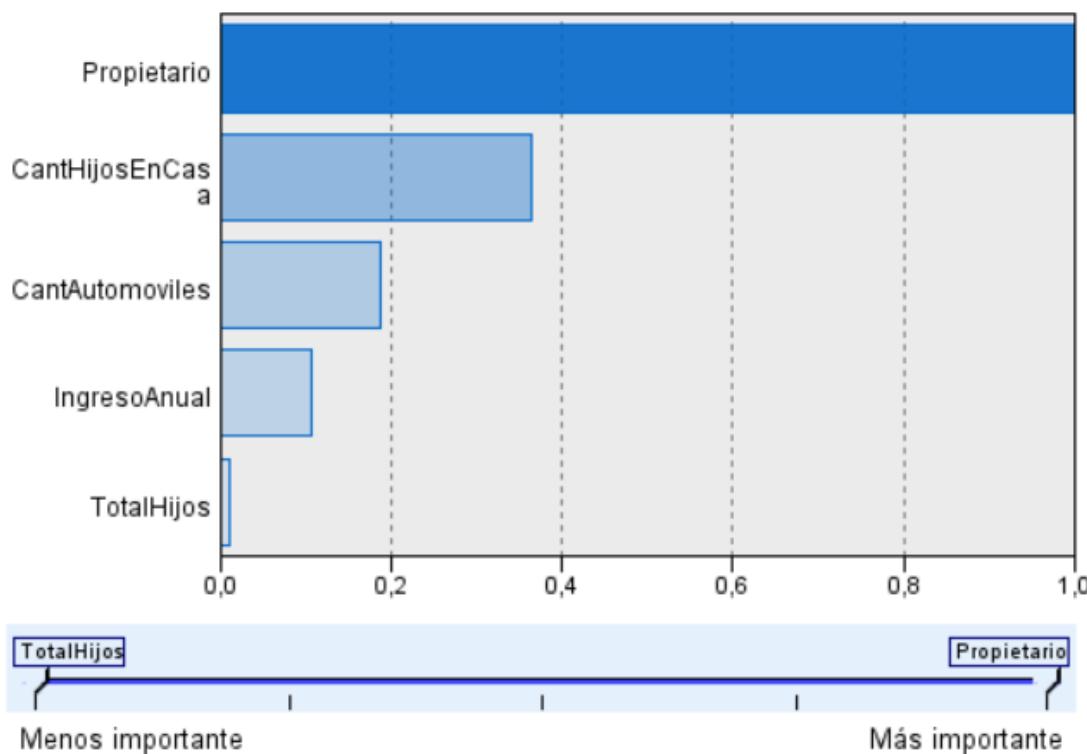


El predictor de menos importancia claramente es “Edad”, seguido por “TotalHijos”. Probamos entonces quitar la primera variable del modelo. Obtuvimos lo siguiente:

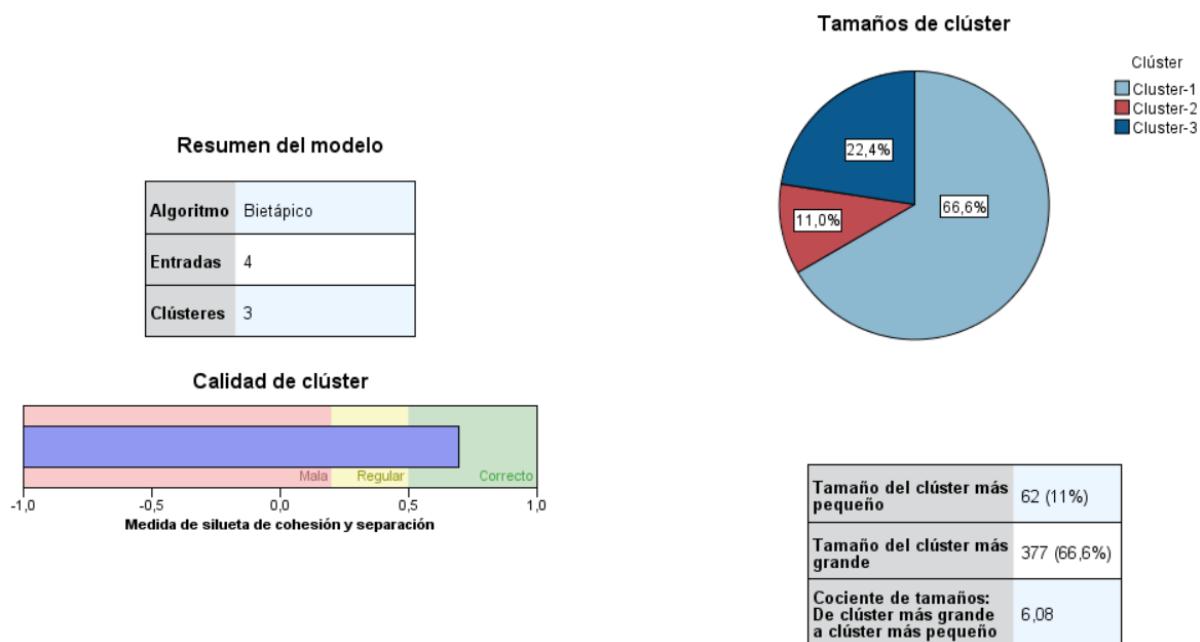


Quizás a simple vista no se vea, pero la calidad del modelo aumentó ligeramente. La silueta promedio es de 0.6.

Importancia del predictor

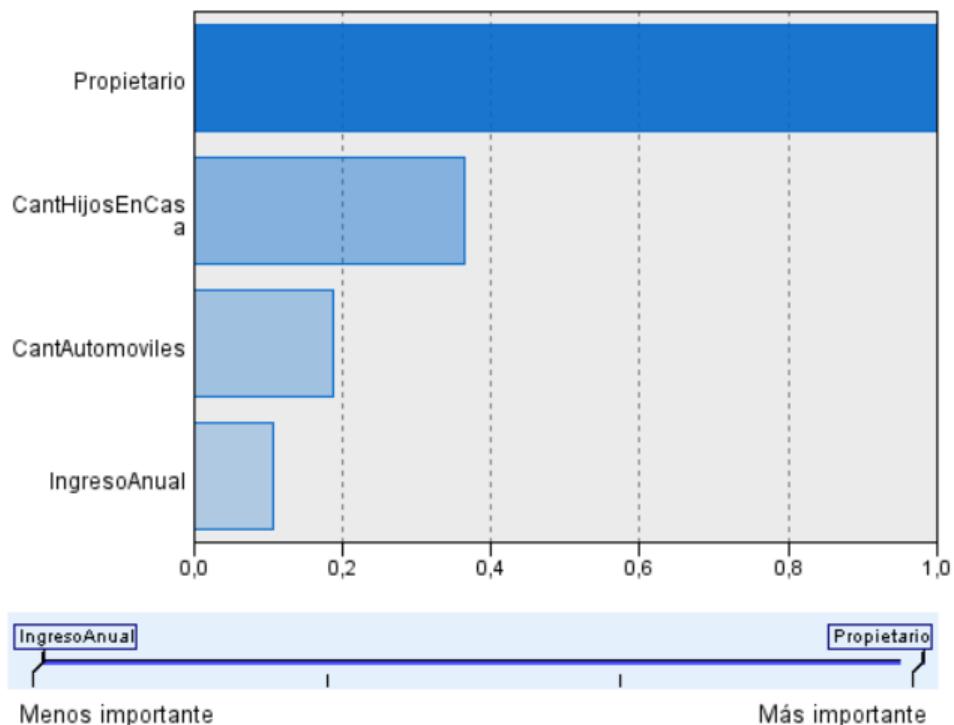


Analizando otra vez los predictores, la variable “TotalHijos” sigue siendo despreciable. Procederemos a quitarla para ver qué conseguimos.



Es el mejor modelo obtenido hasta ahora, con una silueta promedio de 0.7 (un valor que es muy bueno). Veremos la importancia de los predictores:

Importancia del predictor



Si bien la variable “IngresoAnual” no es tan importante como las demás, nosotros no la consideraríamos como despreciable. Por lo tanto, no optaremos por quitarla y decidimos que este es el modelo final de cluster bietápico que generaremos y con el que trabajaremos.

Análisis del mejor modelo

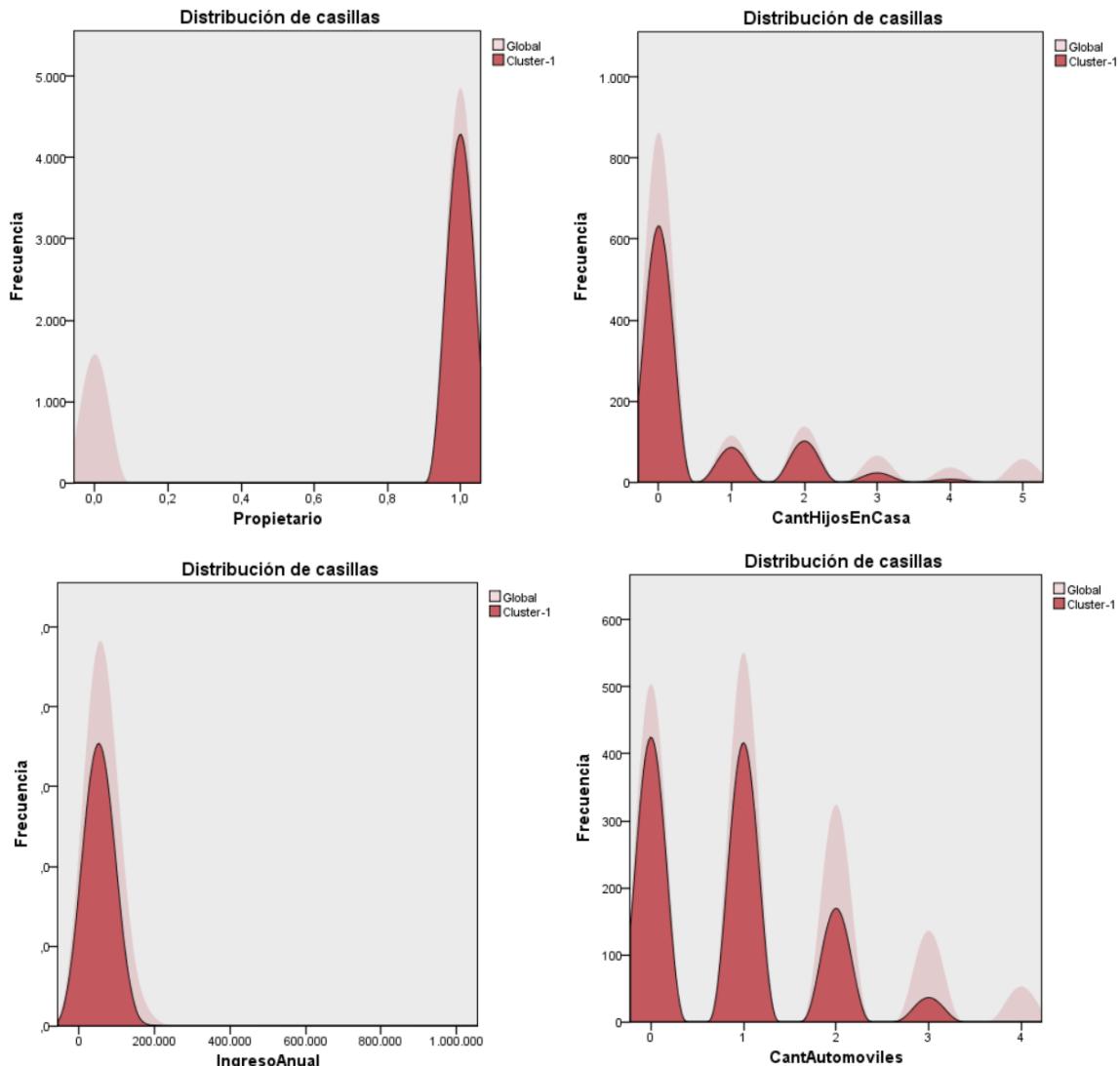
Analizaremos el contenido de los tres clusters creados:

Clústeres		Importancia de entrada (predictor)		
Clúster	Cluster-1	Cluster-3	Cluster-2	
Etiqueta				
Descripción				
Tamaño	66,6% (377)	22,4% (127)	11,0% (62)	
Entradas	Propietario 1,00	Propietario 0,00	Propietario 0,81	
	CantHijosEnCasa 0,45	CantHijosEnCasa 0,35	CantHijosEnCasa 3,79	
	CantAutomoviles 0,82	CantAutomoviles 1,30	CantAutomoviles 2,90	
	IngresoAnual 52.944,30	IngresoAnual 58.110,24	IngresoAnual 152.903,23	

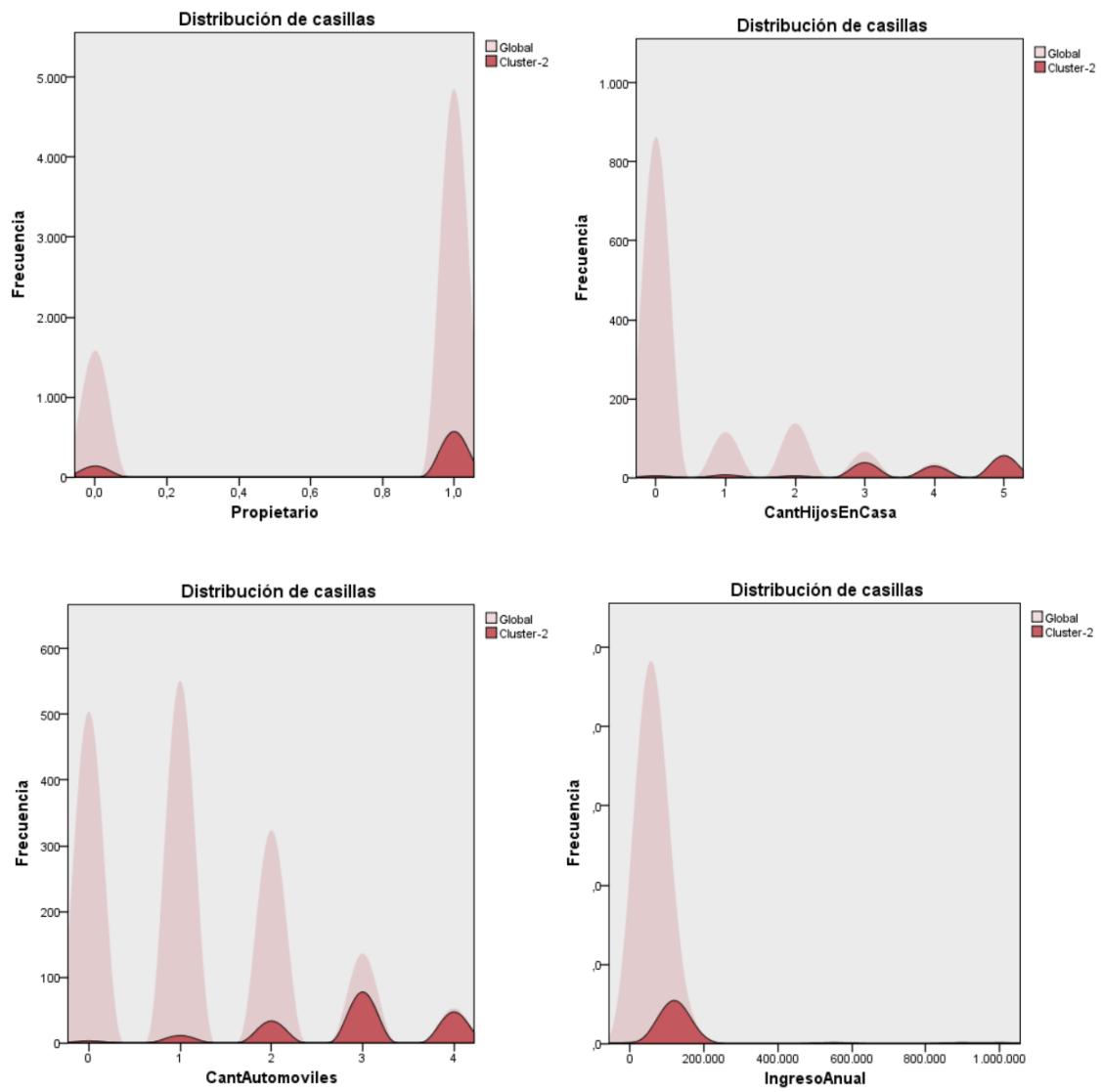
Gracias a esta tabla, podemos decir:

- Cluster 1: contiene el 66.6% de las observaciones (377 registros).
 - Propietario = 1: todos los clientes son propietarios de casas.
 - CantHijosEnCasa = 0.45: la mayoría de los clientes no tienen hijos.
 - CantAutomoviles = 0.82: los clientes tienen un auto o ninguno.
 - IngresoAnual = 52.944,30: el promedio del ingreso anual de los clientes es de aproximadamente \$53.000.
 - Son clientes con bajos ingresos, con máximo un auto y casi sin hijos. Las bicicletas “basic” serían perfectas para estos clientes ya que son versátiles y pueden adaptarse a diferentes necesidades, siendo una opción popular para uso recreativo y desplazamientos urbanos sin gasto de combustible.

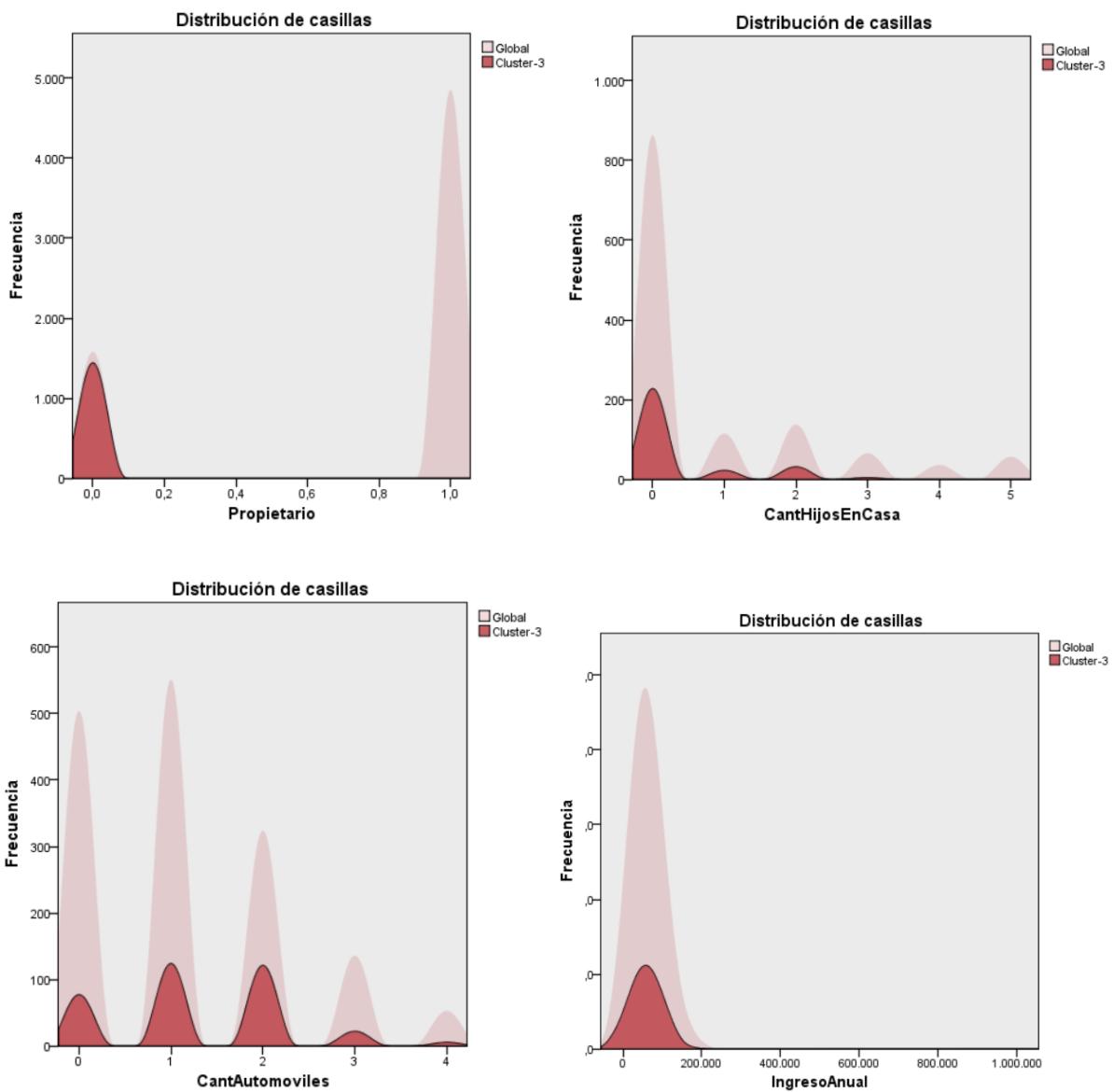
Generamos las gráficas de los predictores:



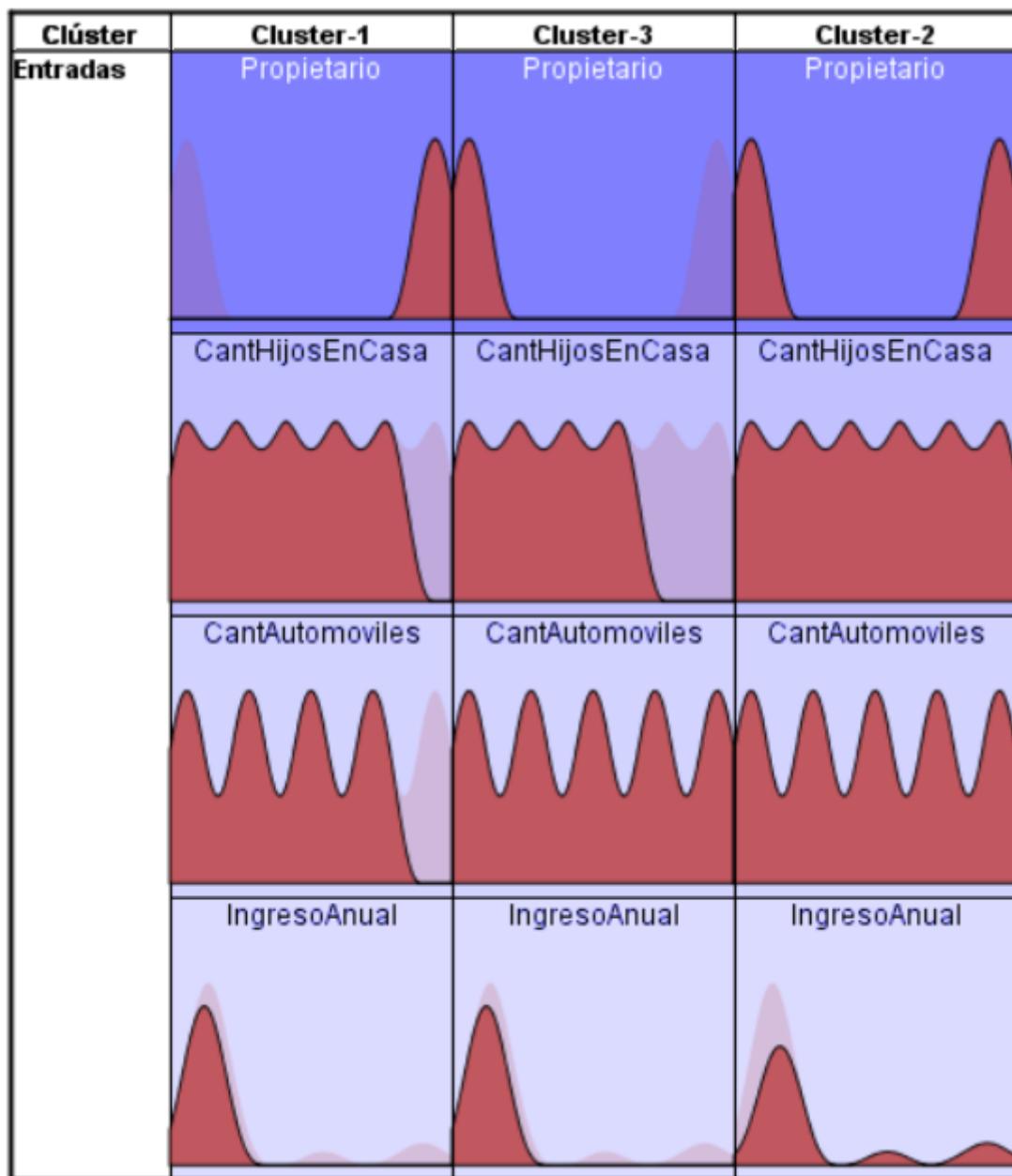
- Cluster 2: contiene el 11% de las observaciones (62 registros).
 - Propietario = 0.81: la mayoría de los clientes tienen casas.
 - CantHijosEnCasa = 3.79: el promedio de hijos por cliente es de 4.
 - CantAutomoviles = 2.90: el promedio de autos por cliente es de 3.
 - IngresoAnual = 152.903,23: el promedio del ingreso anual de los clientes es de aproximadamente \$153.000.
 - Son clientes de grandes recursos económicos, con muchos hijos y muchos autos. Dada la alta cantidad de hijos y automóviles, es posible que las bicicletas de tipo "kinder" (para niños) sean relevantes para ellos. Sin embargo, considerando su alto ingreso anual, también podrían estar interesados en bicicletas de tipo "sports" de mayor calidad y rendimiento.



- Cluster 3: contiene el 22.4% de las observaciones (127 registros).
 - Propietario = 0: los clientes no tienen casas.
 - CantHijosEnCasa = 0.35: la mayoría de los clientes no tienen hijos.
 - CantAutomoviles = 1.30: los clientes tienen un auto o dos.
 - IngresoAnual = 58.110,24: el promedio del ingreso anual de los clientes es de aproximadamente \$58.000.
 - Son clientes sin casas, con ingresos medios, casi sin hijos y en promedio tienen un auto o dos. Para ellos, las bicicletas de tipo "basic" podrían ser una opción adecuada debido a la funcionalidad y rendimiento equilibrados que brindan.



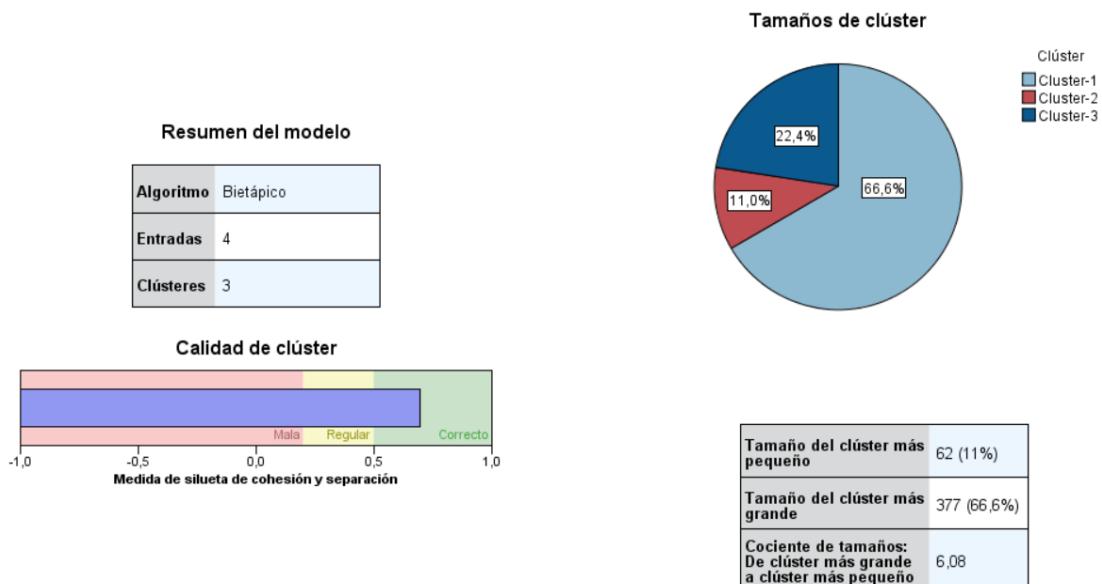
Generamos también los predictores en forma de distribución absoluta:



Mejor técnica de clustering

Detalles

Luego de realizar los tres tipos de clustering (k-medias, jerárquico y bietápico), decidimos que la técnica que nos dió mejores resultados fue la del clustering bietápico. Esto es debido a que nos proporcionó información muy detallada sobre los clusters generados, además de obtenerse un modelo con una silueta promedio de 0.7 (un valor relativamente alto).

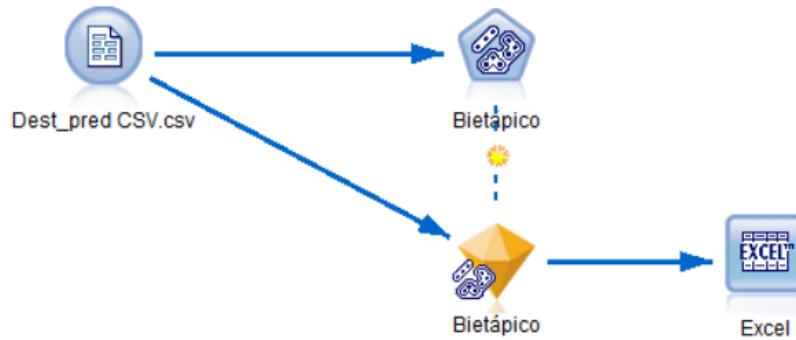


Sin embargo, hemos de aclarar que el modelo generado por el clustering k-means elaborado en RapidMiner era bastante decente. Incluso, los tamaños de los clusters generados eran muy similares al obtenido por medio del clustering bietápico. Creemos que cualquiera de los dos habría sido buena opción.

Volviendo al modelo seleccionado, luego de analizar los clusters detenidamente, determinamos las siguientes recomendaciones:

- Cluster 1: Basic (normales).
- Cluster 2: Kinder (para niños).
- Cluster 3: Basic (normales).

Generamos un Excel con los clusters obtenidos en SPSS Modeler:



Y luego asignamos el tipo de bicicleta según el cluster asignado a cada cliente por medio del uso de fórmulas de Excel. Los resultados parciales son:

- Bicicletas “basic” recomendadas: 504.
- Bicicletas “kinder” recomendadas: 62.
- Bicicletas “sports” recomendadas: 0.

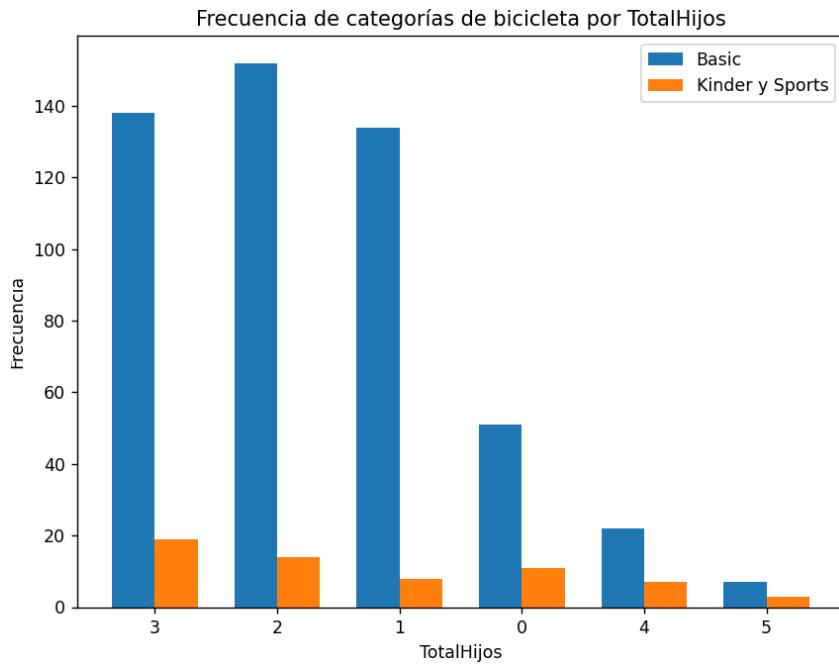
Como observación, a los clientes del cluster 2 se le podrían haber recomendado también las bicicletas “sports” en vez de “kinder”. Luego de pensar y debatir un poco, debido al alto poder adquisitivo de estos clientes creemos que es factible enviarles el doble de publicidad. Con esto, nos referimos a enviarles una publicidad de las bicicletas “kinder” y otra publicidad de las bicicletas “sports”. Creemos que ofrecerles dos alternativas de bicicletas es muy buena idea: pueden decidir si comprar una bicicleta para los hijos o una bicicleta para ellos mismos.

Concluyendo, los resultados finales son:

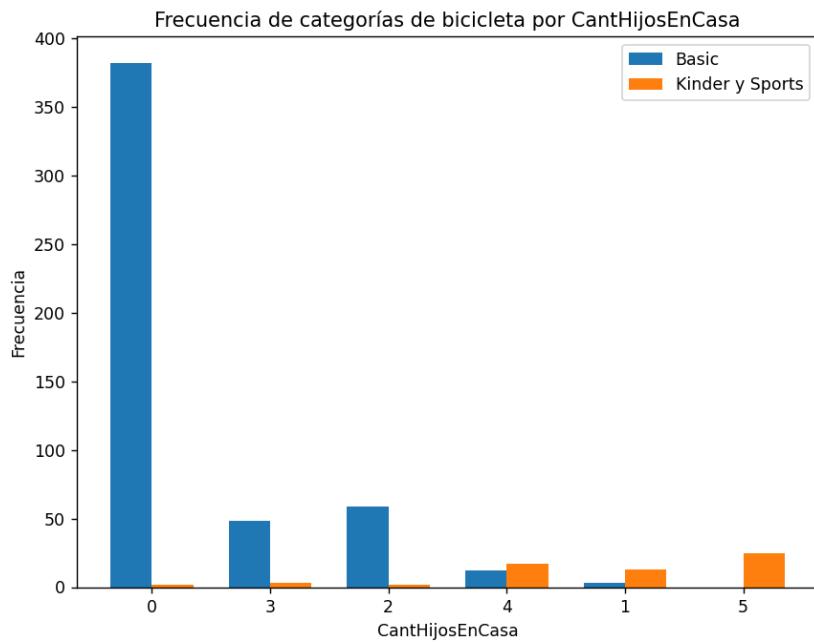
- Bicicletas “basic” recomendadas: 504.
- Bicicletas “kinder” recomendadas: 62.
- Bicicletas “sports” recomendadas: 62 (a los mismos clientes que recibieron publicidad “kinder”).

Gráficas

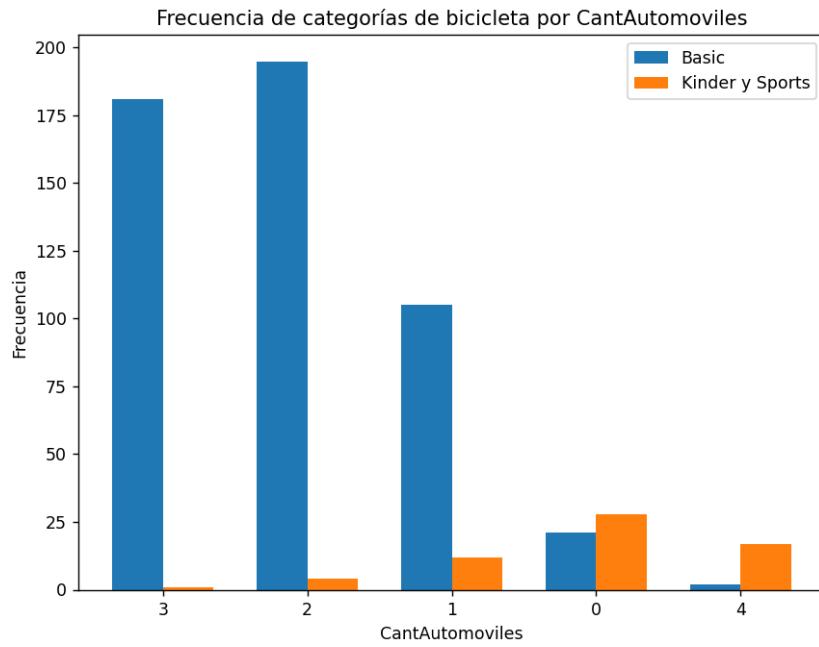
Decidimos generar algunas gráficas para ver qué tal los resultados obtenidos. Primero generamos las gráficas de algunas variables por tipo de bicicleta clasificado:



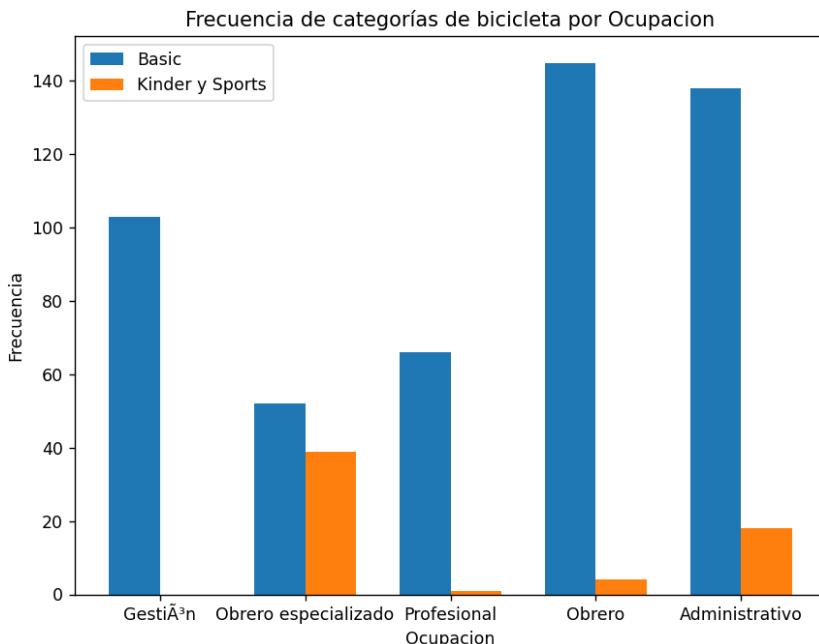
Respecto a la cantidad de hijos, la bicicleta categoría Basic es la más popular en todos los casos.



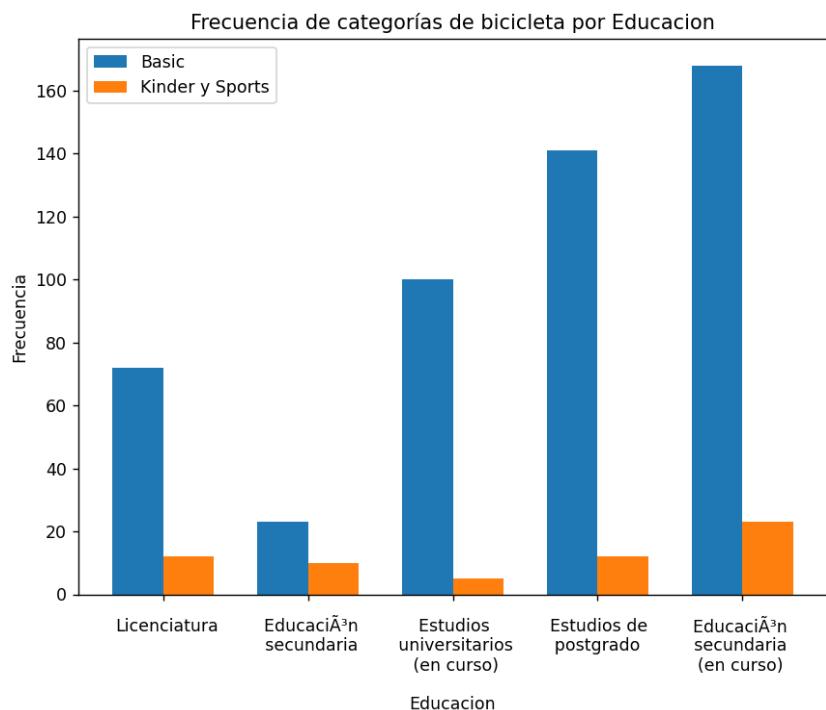
En cuanto a la cantidad de hijos en casa, observamos que la Basic domina en aquellos que tienen dos, tres o ningún hijo, mientras que las Kinder o Sport tienen mayor frecuencia para aquellos clientes con uno, cuatro o cinco hijos. Consideramos que esto está bien, ya que puede que los padres necesiten bicicletas Kinder para sus hijos.



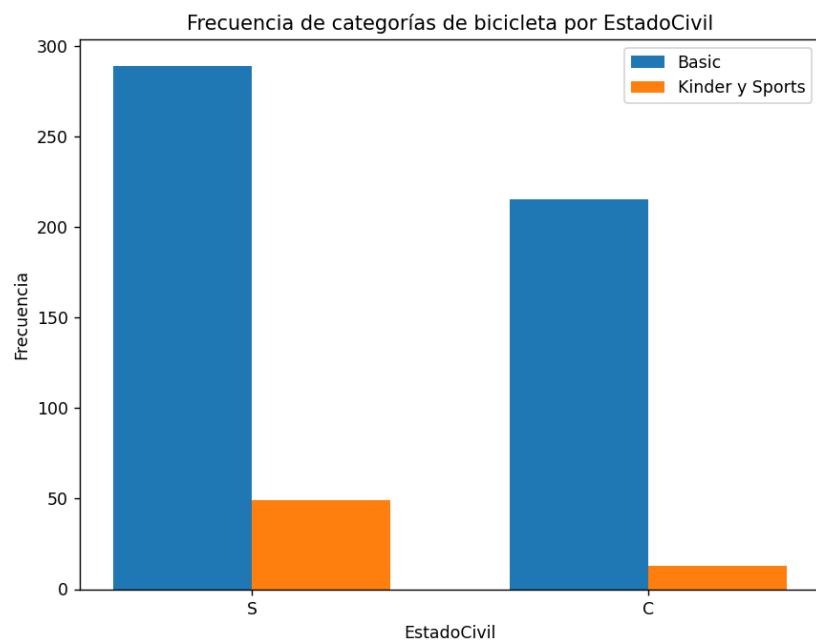
Respecto a la cantidad de automóviles, aquellos clientes que cuentan con uno, dos o tres optarán casi en su totalidad por la Basic, mientras que aquellos que tienen cuatro por la Kinder o Sport. En aquellos clientes con ningún automóvil, encontramos mayor similitud en sus preferencias. Sin embargo, al no tener auto, es probable que se decanten por una bicicleta Sports para poder ir a trabajar con mayor comodidad, por ello es mayor la altura de la barra naranja.



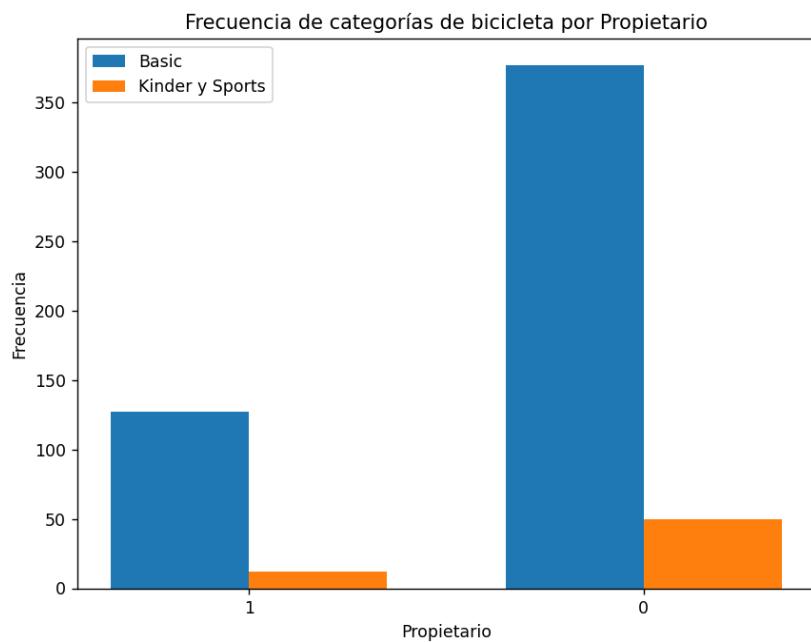
En todas las ocupaciones observamos que la Basic domina por mucho, excepto en aquellos clientes que son obreros especializados, donde la diferencia es más ajustada. Nos es curioso que no hay clientes cuya profesión es “Gestión” que compren bicicletas Kinder o Sports.



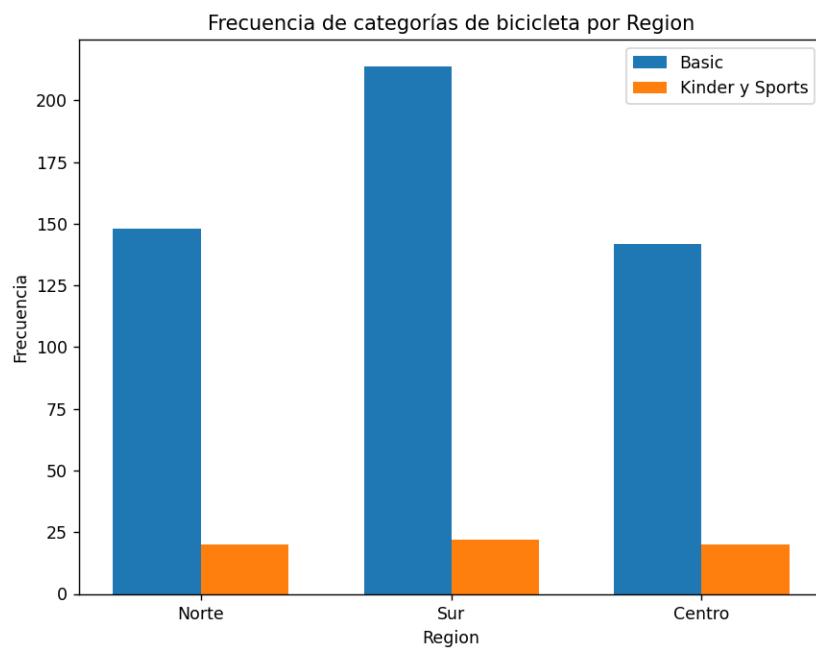
No vemos nada interesante.



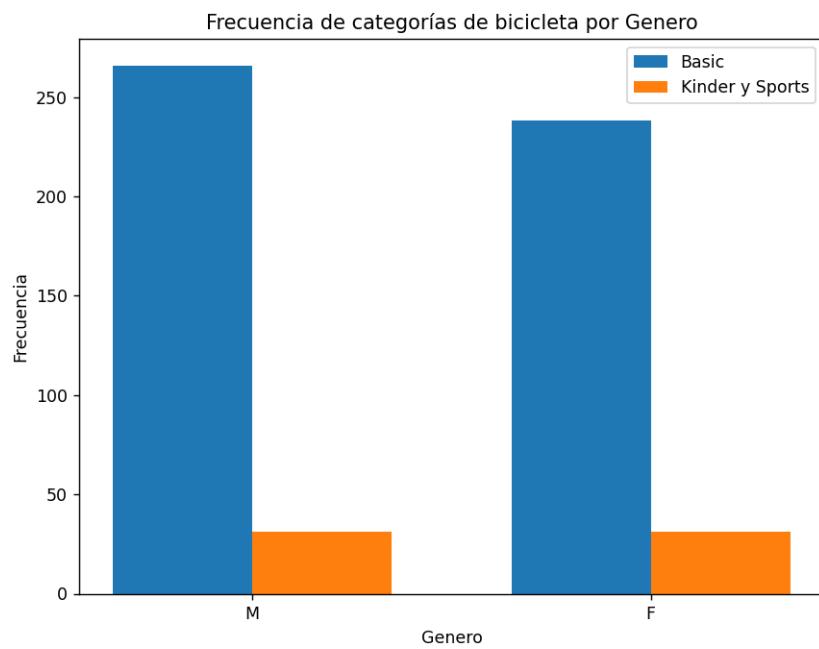
Tampoco vemos nada interesante.



Nuevamente no observamos nada relevante.

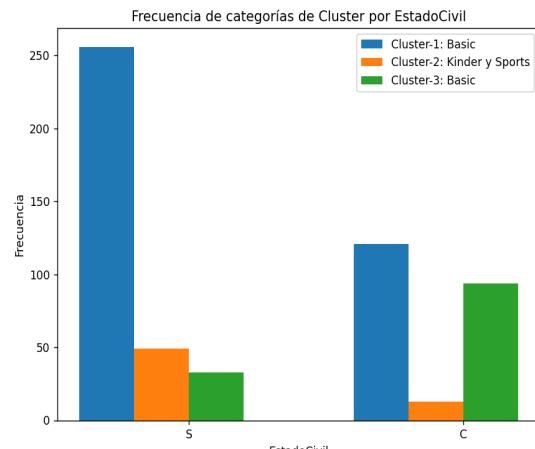
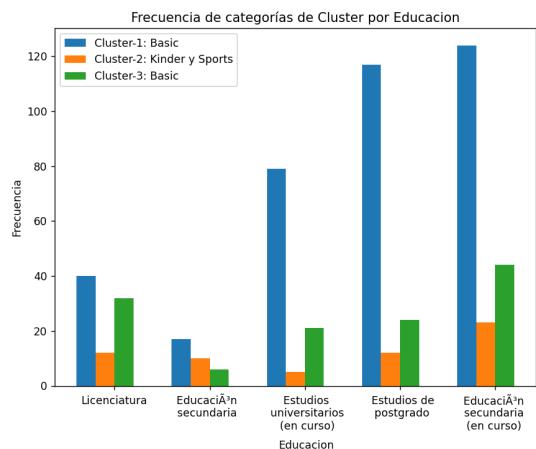
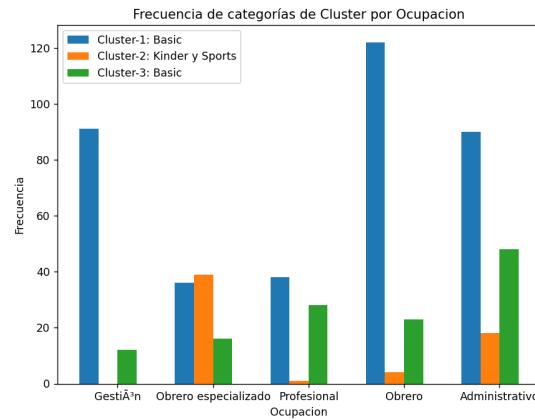
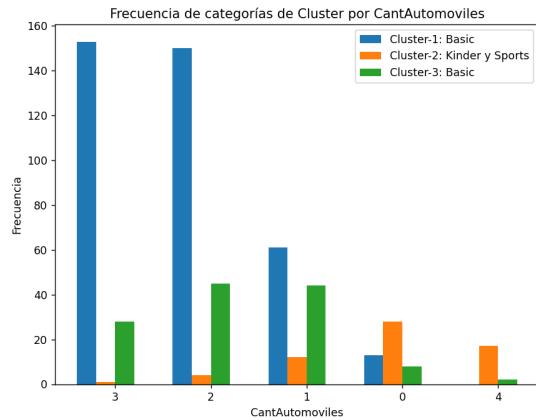
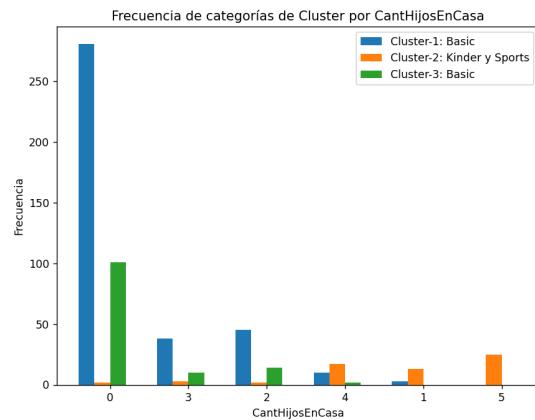
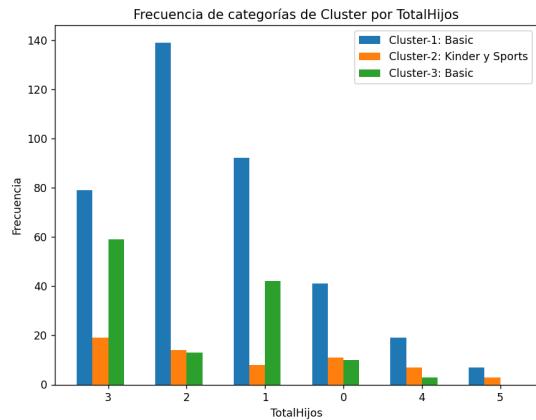


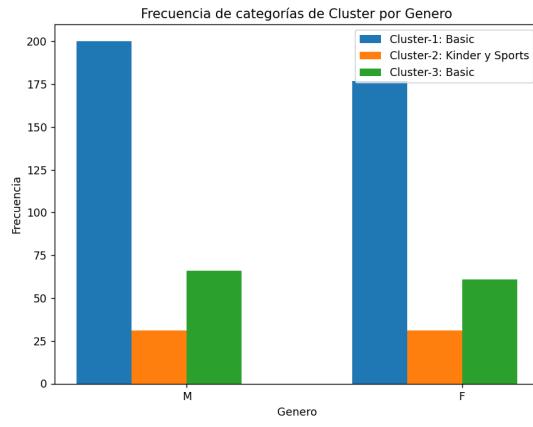
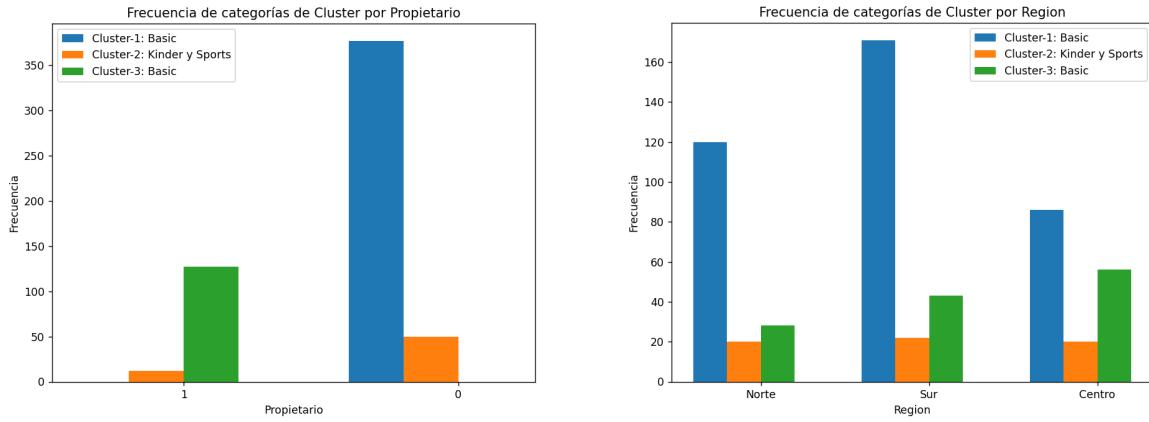
Observamos una similitud en la cantidad de clientes que comprarían Kinder o Sport según la región, pero siendo en todos los casos muchos más quienes optarán por la Basic.



Encontramos una similitud respecto al género, en ambos casos optarán por la Basic casi en su totalidad.

Ahora generamos gráficas pero por tipo de cluster clasificado. Las mostramos de forma general porque las gráficas son muy similares: predomina el cluster 1, seguido por el cluster 3. Por lo tanto, no nos detendremos en analizar cada una.





Como dijimos anteriormente, las tendencias de que el cluster 1 sea dominante, seguido por el cluster 3, se repiten en la mayoría de gráficas. Vemos algunos casos fuera de lo común como por ejemplo, no hay clientes propietarios en el cluster 1 y no hay clientes no propietarios en el cluster 3. Fuera de eso, creemos que estas gráficas son interesantes pero las gráficas por tipo de bicicleta son mejores.

Mercados objetivos

El objetivo de esta etapa consiste en determinar 3 mercados objetivos similares a Argentina puesto que la empresa tiene planeado expandir el alcance de su negocio lanzando la campaña publicitaria en 3 países similares.

Estructura de datos

Para realizar el análisis del mercado, se nos brindó un Excel con los datos de los mercados extranjeros. Las variables con las que trabajaremos son:

- Horas de trabajo promedio [hs/año].
- Días de vacaciones promedio (por año).
- Inflación 2006.
- Inflación 2007.
- Inflación 2008.
- Inflación 2009.
- Inflación 2010.
- Inflación 2011.
- Alquiler departamento 3 ambientes [USD por mes].
- Contribución al seguro social (%).
- Sueldo promedio maestro de escuela primaria [USD por año].
- Sueldo promedio chofer colectivo [USD por año].
- Sueldo promedio mecánico de automóviles [USD por año].
- Sueldo promedio arquitecto [USD por año].
- Sueldo promedio cocinero [USD por año].
- Sueldo promedio ingeniero [USD por año].
- Sueldo promedio secretaria [USD por año].
- Sueldo promedio vendedor [USD por año].
- Sueldo promedio analista financiero [USD por año].
- Ciudad.

Hay 74 registros y 20 variables. Dentro de la variable “Ciudad” tenemos a Buenos Aires. Usaremos el correspondiente registro para realizar las comparaciones de los demás mercados.

Sin embargo, debido a la gran cantidad de variables que tenemos, es complicado ver qué ciudades son similares a Buenos Aires. Podríamos intentar realizar un análisis exploratorio univariante o multivariante pero aún así será difícil determinar una relación. Además,

Ante esta situación, nos vemos obligados a tener que reducir la dimensión de los datos. Buscamos reducir la cantidad de variables, intentando mantener la mayor variabilidad posible.

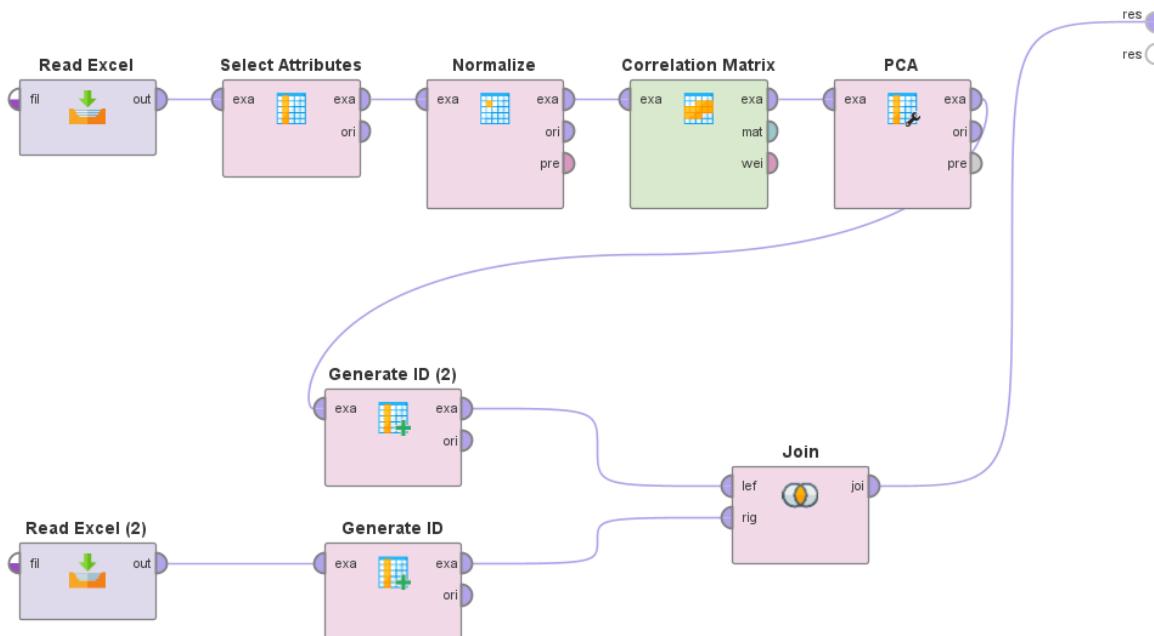
Utilizaremos la técnica ACP (análisis de componentes principales). El objetivo es obtener un conjunto de nuevas variables independientes que sean combinación lineal de las variables originales.

Análisis de componentes principales

Esta técnica, además de reducir la dimensión del conjunto de datos, transforma las variables originales en una representación con variables incorreladas (las variables originales generalmente están correladas).

Observando brevemente los datos, notamos que hay bastante diferencia entre las escalas de las diferentes variables. Por ejemplo, la escala de los salarios es muy grande comparada a la de los días de vacaciones. Entonces, es necesario que normalicemos las variables numéricas.

Utilizaremos el software RapidMiner para aplicar esta técnica. El modelo nos quedó de la siguiente forma:



- Camino 1:
 - Arranca con el módulo “Read Excel”, el cual lee el excel de mercados objetivos.
 - Quitamos la variable categórica “Ciudad” con el módulo “Select Attributes”.
 - Normalizamos las variables numéricas luego de analizar el Excel leído por medio del módulo “Normalize”.
 - Generamos una matriz de correlación con el módulo “Correlation Matrix” para tener una vista previa de las variables y cómo se relacionan entre sí.
 - Con el módulo “PCA” aplicamos el análisis de componentes principales.
 - Generamos un ID para luego poder unir los datos filtrados con los originales.
- Camino 2:
 - Leemos nuevamente el Excel con el módulo “Read Excel”.
 - Generamos un ID para este conjunto de datos, con el objetivo de unirlo con los datos filtrados.
 - Aplicamos un Join para unir ambos Excel. Ahora, tenemos la variable “Ciudad” con el análisis de componentes principales aplicado.

Como ya explicamos, comenzamos leyendo el Excel y de forma rápida analizamos un poco los datos:

▼ Inflación 2011	Real	0	Min -0.283	Max 26.090	Average 4.248
▼ Alquiler departamento 3 ambientes	Real	0	Min 226.500	Max 2149.500	Average 783.007
▼ Contribución al seguro social (%)	Integer	0	Min 0	Max 42	Average 20.808
▼ Sueldo promedio maestro de escuela primaria [USD por año]	Integer	0	Min 1600	Max 113300	Average 33106.849

Viendo solamente estas cuatro variables nos damos cuenta que los valores mínimos y máximos son muy distintos. Confirmamos entonces que es necesario estandarizar. Para ello, utilizamos el módulo “Normalize” con el parámetro “Z-transformation”.

Sin embargo, antes de la normalización, filtramos las variables numéricas. Esto lo hicimos para poder aplicar la técnica de ACP. Tuvimos que quitar la variable “Ciudad” para que la técnica funcione (es la única variable categórica).

Decidimos generar la matriz de correlación para ver que obtenemos:

Attributes	Horas d...	Días de ...	Inflació...	Inflació...	Inflació...	Inflació...	Inflació...	Alquiler ...	Contrib...	Sueldo ...									
Horas de trabajo promedio [hs/año]	1	-0.637	0.290	0.271	0.223	0.246	0.225	0.222	0.042	-0.575	-0.323	-0.401	-0.489	-0.423	-0.201	-0.326	-0.411	-0.437	-0.320
Días de vacaciones promedio (por año)	-0.637	1	-0.057	-0.022	-0.027	-0.059	-0.159	-0.134	-0.061	0.306	0.118	0.169	0.213	0.131	0.084	0.112	0.189	0.212	0.151
Inflación 2006	0.290	-0.057	1	0.828	0.781	0.591	0.534	0.645	-0.079	-0.418	-0.512	-0.492	-0.490	-0.494	-0.438	-0.465	-0.518	-0.512	-0.532
Inflación 2007	0.271	-0.022	0.828	1	0.917	0.702	0.622	0.693	-0.046	-0.408	-0.503	-0.513	-0.518	-0.532	-0.376	-0.465	-0.529	-0.525	-0.533
Inflación 2008	0.223	-0.027	0.781	0.917	1	0.767	0.654	0.730	-0.079	-0.370	-0.528	-0.520	-0.494	-0.513	-0.432	-0.497	-0.541	-0.531	-0.521
Inflación 2009	0.246	-0.059	0.591	0.702	0.767	1	0.894	0.865	-0.149	-0.253	-0.465	-0.443	-0.423	-0.415	-0.312	-0.431	-0.466	-0.429	-0.447
Inflación 2010	0.225	-0.159	0.534	0.622	0.654	0.894	1	0.893	-0.101	-0.290	-0.388	-0.378	-0.365	-0.355	-0.263	-0.378	-0.405	-0.362	-0.415
Inflación 2011	0.222	-0.134	0.645	0.693	0.730	0.865	0.893	1	-0.069	-0.259	-0.446	-0.445	-0.438	-0.422	-0.333	-0.430	-0.469	-0.445	-0.477
Alquiler departamento 3 ambientes [USD por año]	0.042	-0.061	-0.079	-0.046	-0.079	-0.149	-0.101	-0.069	1	-0.039	0.556	0.483	0.461	0.495	0.598	0.597	0.526	0.509	0.556
Contribución al seguro social (%)	-0.575	0.306	-0.418	-0.408	-0.370	-0.253	-0.290	-0.259	-0.039	1	0.383	0.455	0.543	0.529	0.391	0.416	0.512	0.509	0.386
Sueldo promedio maestro de escuela primaria [USD por año]	-0.323	0.118	-0.512	-0.503	-0.528	-0.465	-0.388	-0.446	0.556	0.383	1	0.930	0.861	0.845	0.774	0.877	0.883	0.879	0.839
Sueldo promedio chofer colectivo [USD por año]	-0.401	0.169	-0.492	-0.513	-0.520	-0.443	-0.378	-0.445	0.483	0.455	0.930	1	0.894	0.878	0.765	0.845	0.905	0.891	0.831
Sueldo promedio mecánico de automóvil [USD por año]	-0.489	0.213	-0.490	-0.518	-0.494	-0.423	-0.365	-0.438	0.461	0.543	0.861	0.894	1	0.952	0.752	0.860	0.945	0.936	0.833
Sueldo promedio arquitecto [USD por año]	-0.423	0.131	-0.494	-0.532	-0.513	-0.415	-0.355	-0.422	0.495	0.529	0.845	0.878	0.952	1	0.780	0.872	0.917	0.929	0.816
Sueldo promedio cocinero [USD por año]	-0.201	0.084	-0.438	-0.376	-0.432	-0.312	-0.263	-0.333	0.598	0.391	0.774	0.765	0.752	0.780	1	0.795	0.790	0.792	0.846
Sueldo promedio ingeniero [USD por año]	-0.326	0.112	-0.465	-0.465	-0.497	-0.431	-0.378	-0.430	0.597	0.416	0.877	0.845	0.860	0.872	0.795	1	0.900	0.886	0.853
Sueldo promedio secretaria [USD por año]	-0.411	0.189	-0.518	-0.529	-0.541	-0.466	-0.405	-0.469	0.526	0.512	0.883	0.905	0.945	0.917	0.790	0.900	1	0.964	0.862
Sueldo promedio vendedor [USD por año]	-0.437	0.212	-0.512	-0.525	-0.531	-0.429	-0.362	-0.445	0.509	0.509	0.879	0.891	0.936	0.929	0.792	0.886	0.964	1	0.842
Sueldo promedio analista financiero [USD por año]	-0.320	0.151	-0.532	-0.533	-0.521	-0.447	-0.415	-0.477	0.556	0.386	0.839	0.831	0.833	0.816	0.846	0.853	0.862	1	

Mientras más oscuro el morado, más correlación hay entre las variables. A simple vista, lo que detectamos es:

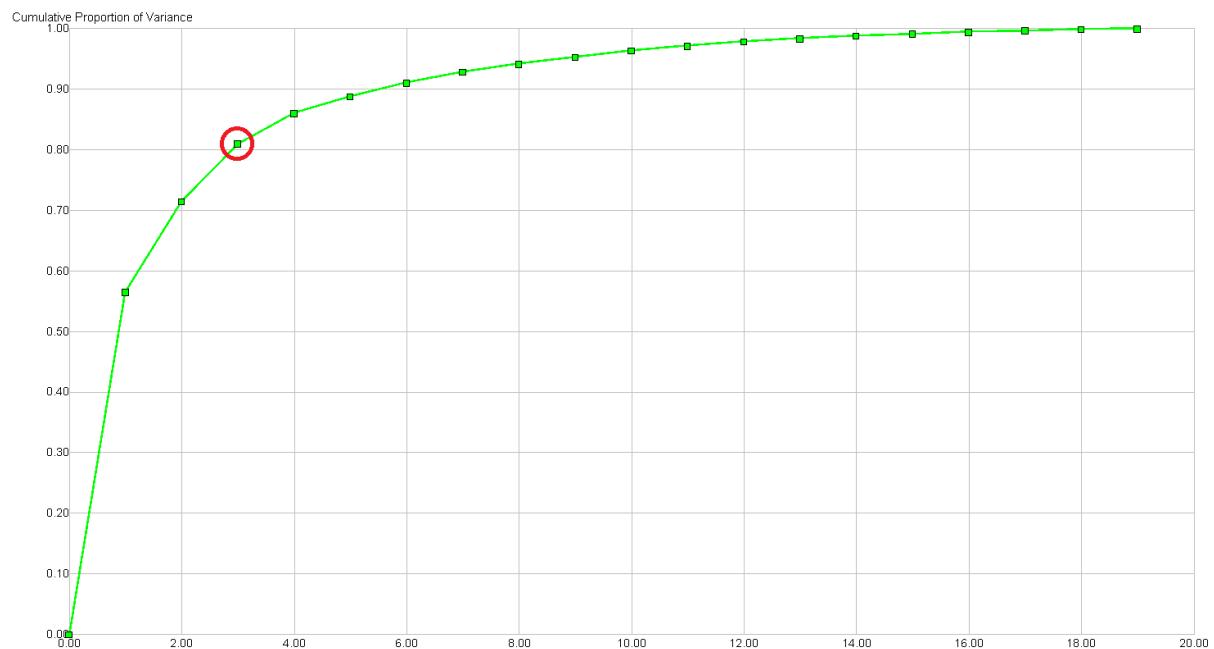
- La variable “Días de vacaciones promedio (por año)” es quizás la que menos correlación tiene con las demás.
- Las variables de sueldos están muy correlacionadas entre sí.
- Las variables de inflación también están muy correlacionadas entre sí.

Utilizamos el nodo del análisis de componentes principales y obtuvimos:

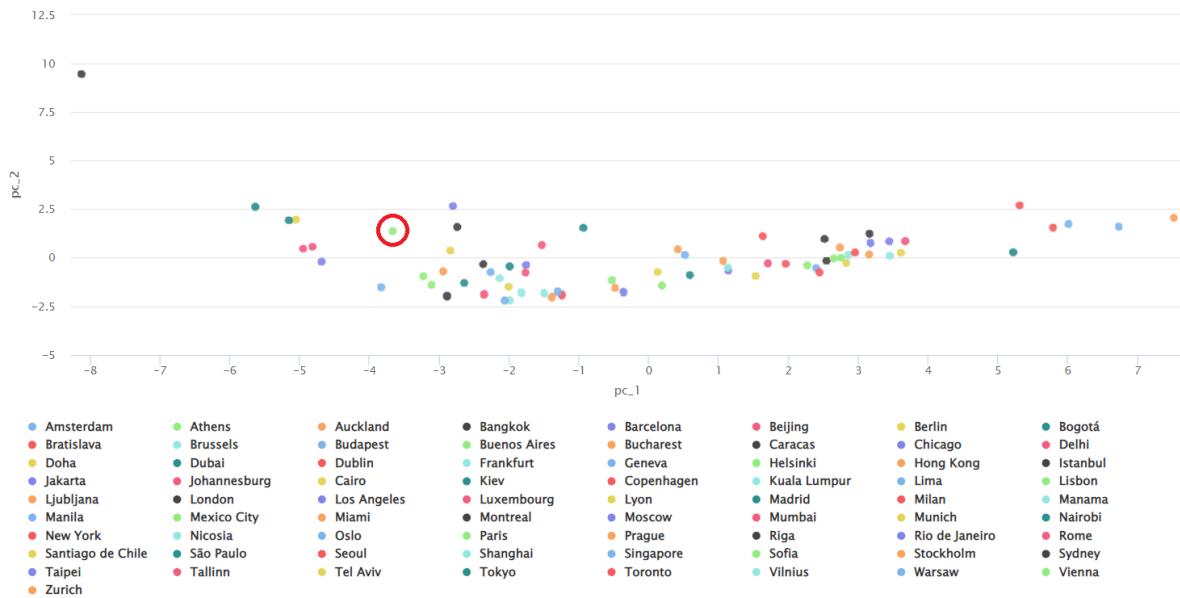
Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	3.277	0.565	0.565
PC 2	1.685	0.150	0.715
PC 3	1.347	0.096	0.810
PC 4	0.978	0.050	0.860
PC 5	0.727	0.028	0.888
PC 6	0.651	0.022	0.911
PC 7	0.592	0.018	0.929
PC 8	0.496	0.013	0.942
PC 9	0.460	0.011	0.953
PC 10	0.448	0.011	0.964
PC 11	0.384	0.008	0.971
PC 12	0.374	0.007	0.979
PC 13	0.310	0.005	0.984
PC 14	0.292	0.004	0.988
PC 15	0.247	0.003	0.992
PC 16	0.235	0.003	0.994
PC 17	0.213	0.002	0.997
PC 18	0.200	0.002	0.999
PC 19	0.143	0.001	1.000

Vemos que la componente principal n° 3 acumula el 81% de los datos. Consideramos que es la más apropiada para utilizar.

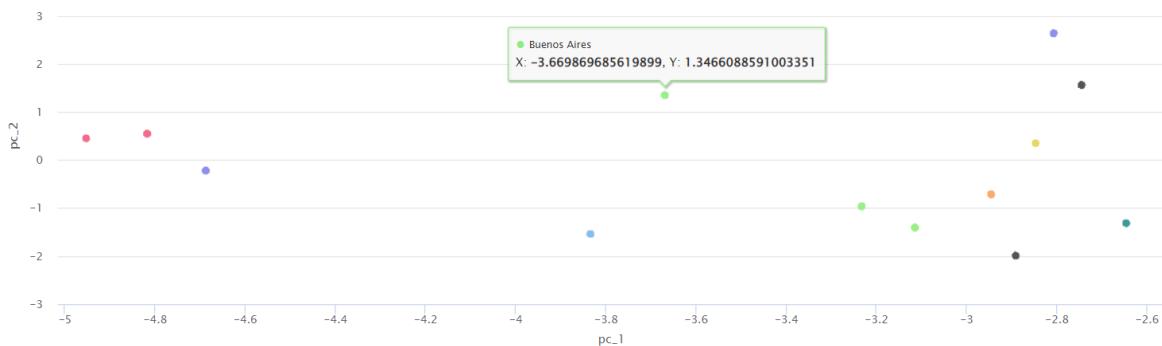
A continuación presentamos la gráfica de la acumulación de la varianza. En rojo está marcada la componente n° 3.



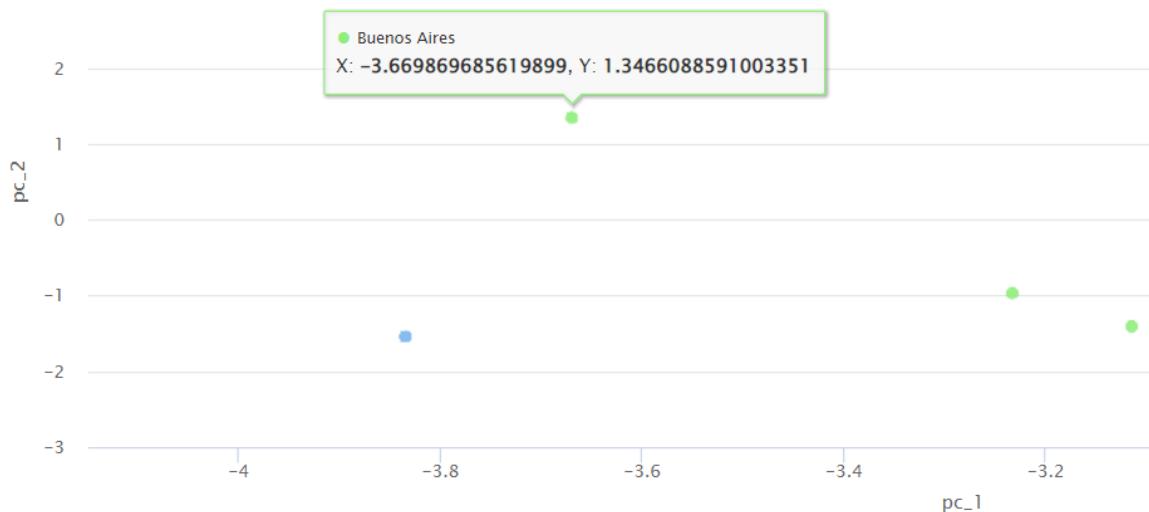
Utilizamos un diagrama de dispersión para graficar la componente 1 vs la componente 2. Coloreamos por la variable Ciudad:



Si hacemos un poco más de zoom:



Y un poco más:



Ahora estamos listos para analizar la cercanía de los puntos. En esta última gráfica vemos tres puntos, los cuales son los más próximos a nuestro mercado.

Las ciudades más cercanas (y por lo tanto, más parecidas) a Buenos Aires son:

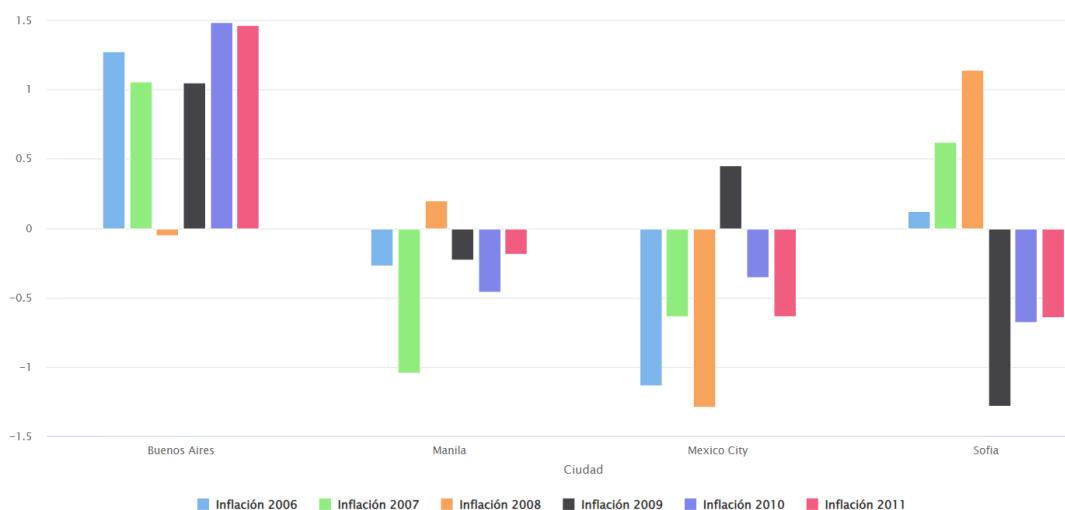
- Manila (Filipinas).
- Sofia (Bulgaria).
- Ciudad de México (Méjico).

Concluyendo, estos serían los mercados candidatos que le recomendaríamos al gerente de ventas.

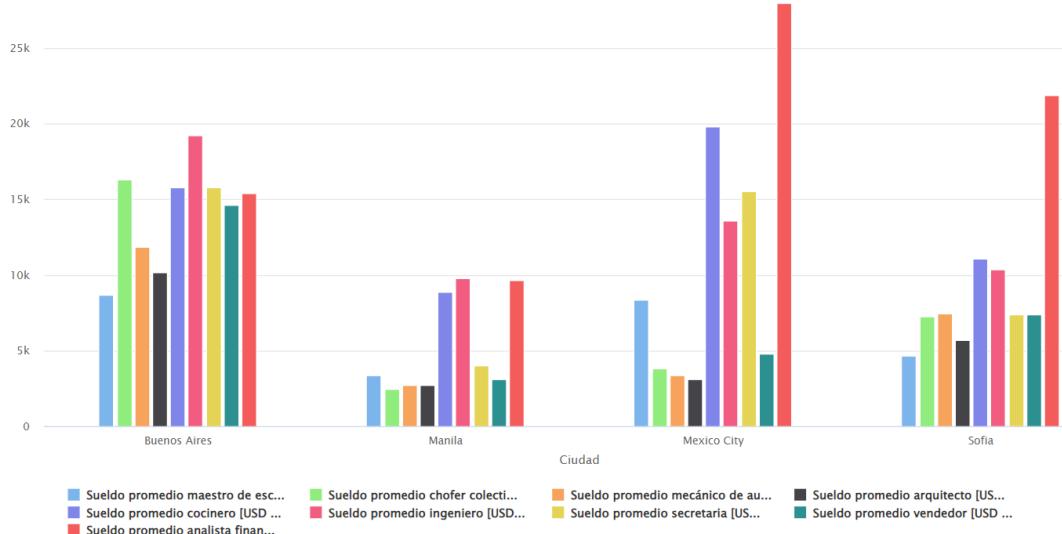
Validación de hipótesis

Para confirmar que el análisis de componentes principales fue exitoso y que las tres ciudades elegidas son las más parecidas a Buenos Aires, decidimos comparar los valores de las variables originales.

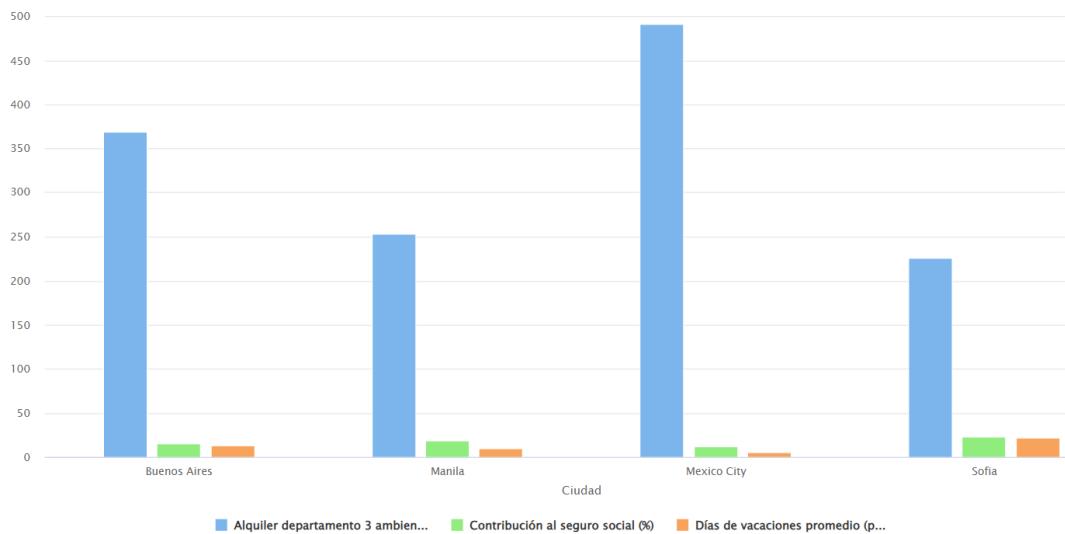
Para ello, generamos gráficas para cada ciudad para varias variables.



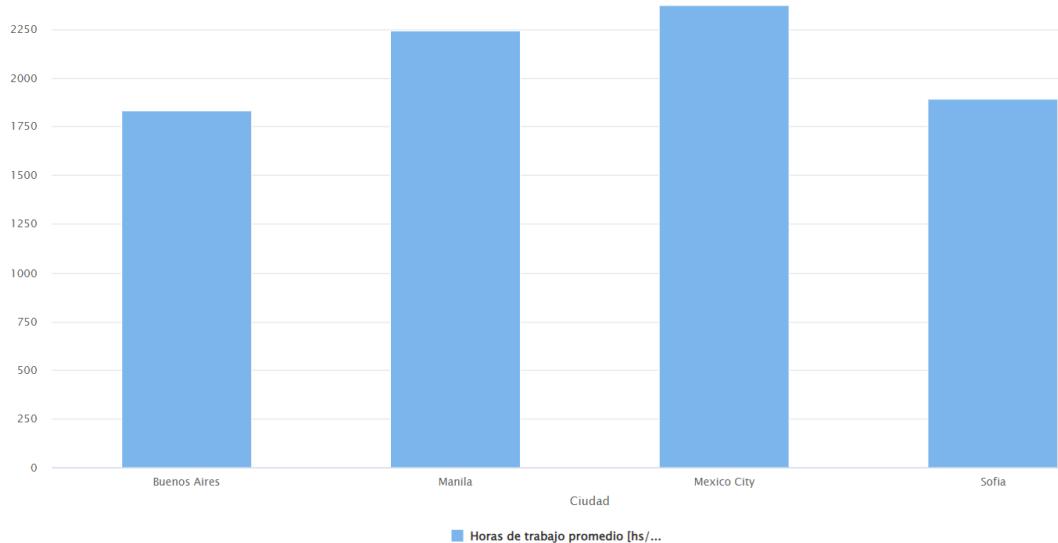
Si analizáramos sólo la inflación, parecería ser que las ciudades son bastante diferentes entre sí. Sin embargo, hay que tener en cuenta que la inflación en Argentina es muy cambiante y extraña, por lo que es una información a agarrar con pinzas.



En cuanto a los sueldos promedios según las ocupaciones, encontramos que los menores se encuentran en la ciudad de Manila. Observamos mayor regularidad en los sueldos de Buenos Aires pero picos más altos en Ciudad de México. Los sueldos no parecerían estar muy relacionados, por lo menos los de Manila. Los de Ciudad de México son raros, algunos muy bajos y otros muy grandes. Los sueldos de Sofia son los más parecidos a los de Buenos Aires.



Con respecto a los alquileres, los valores de Manila y Sofia son menores que los de Buenos Aires, mientras que los valores de Ciudad de México son más altos. Sin embargo, la diferencia con respecto a nuestra capital no es muy grande. Viendo la contribución social y los días de vacaciones, los valores de las cuatro ciudades son muy similares y están relacionados.



Observamos que la ciudad con menos horas de trabajo promedio es Buenos Aires mientras que la que más tiene es Ciudad de México. La que más se asemeja a Buenos Aires es Sofía, teniendo una pequeña diferencia en el promedio. No obstante, las cuatro ciudades tienen horas de trabajo muy similares.

Nos resulta interesante como Buenos Aires es la ciudad con menor horas promedio trabajadas pero con mayor uniformidad en cuanto a los salarios de las distintas ocupaciones.

Luego de analizar estas gráficas, consideramos que los mercados seleccionados son buenos. Creemos que las ciudades elegidas son similares a Buenos Aires, y por lo tanto, la comercialización de la nueva línea de bicicletas será un éxito.