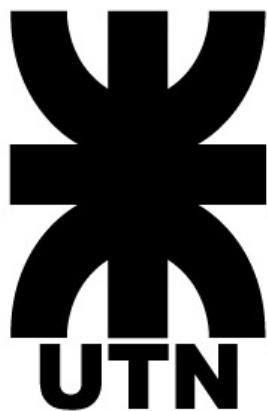


Universidad Tecnológica Nacional

FRRO



**Trabajo Práctico Integrador
Minería de Datos
Año 2023**

Docentes

- Cristian Bigatti.
- Martina Di Carlo.

Alumnos

Nombre y Apellido	Legajo	Email
Antonella Bologna	46906	antonellabologna.21@gmail.com
Sofía Buljubasich	47380	sofiabuljuba@gmail.com
Manuel Cantarini	46759	manu100cantarini@gmail.com
Fadua Dora	47019	faduadora77@gmail.com

Índice

Enunciado	4
Fase de Análisis del problema	5
Definición del Problema y Objetivos	5
Técnicas a Utilizar	5
Fase de pre-procesamiento y análisis de los datos	6
Análisis Exploratorio de Datos Univariante	6
Estado Civil	6
Género	7
Educación	8
Ocupación	8
Distancia	9
Región	11
ComproBicicleta	11
Propietario	12
IngresoAnual	13
TotalHijos	13
CantAutomoviles	14
Edad	16
Análisis Exploratorio de Datos Multivariante	16
Edad vs ComproBicicleta	17
Edad vs IngresoAnual	18
Edad vs Ocupación	18
Edad vs Distancia	19
IngresoAnual vs ComproBicicleta	20
IngresoAnual vs Ocupación	21
TotalHijos vs ComproBicicleta	22
TotalHijos vs EstadoCivil	23
CantAutomoviles vs ComproBicicleta	24
Region vs ComproBicicleta	25
Distancia vs ComproBicicleta	26
Distancia vs CantAutomoviles	27
CantAutomoviles vs IngresoAnual vs ComproBicicleta	27
CantAutomoviles vs TotalHijos vs ComproBicicleta	28
Matriz R	28
Matriz S	29
Proceso de Limpieza de Datos	29
Vista Minable	31
Fase de Modelado	31
Árboles de Decisión	33
Funcionamiento	33

Árbol CHAID	33
Importancia del Predictor	34
Matriz de confusión entrenamiento	34
Matriz de confusión validación	35
Árbol QUEST	36
Importancia del Predictor	36
Matriz de confusión de entrenamiento	36
Matriz de confusión de validación	37
Árbol C5.0 Sin Poda	39
Importancia del Predictor	39
Matriz de confusión de entrenamiento	39
Matriz de confusión de validación	39
Árbol C5.0 con Poda	41
Importancia del Predictor	41
Matriz de confusión de entrenamiento	41
Matriz de confusión de validación	42
KNN	43
Análisis discriminante	46
Análisis de Supuestos	46
Función Discriminante	47
¿Existe Modelo?	49
Análisis de Matriz de Confusión	49
Fase de evaluación	50
Análisis de Costos para los Modelos de Árboles	51
Fase de Implementación	51
Análisis descriptivo	52
KMedias	52
K = 2	53
Caracterización de los cluster con K = 2	59
K = 3	59
Caracterización de los cluster con K = 3	64
K = 4	65
Caracterización de los cluster con K = 4	70
Conclusión	71
Cluster Jerárquico	72
Cluster Bietápico	90
Criterio del tipo de bicicleta a promocionar	93
Análisis de Mercado	93

Enunciado

La empresa AllHome se dedica a la venta de una amplia gama de productos para el hogar. Fue fundada en el año 2005 y desde entonces ha incorporado sucursales en numerosas provincias de nuestro país.

Su evolución en el tiempo le permitió incorporar distintos rubros de productos: electrodomésticos, TV, audio, video, computación, telefonía celular, bazar, muebles, bicicletas y motos. A su vez, en el año 2007 ha logrado crear su propia marca: AllMyHome.

Con el objetivo de mantener el buen posicionamiento de la empresa en el mercado, el gerente general ha elaborado una serie de estrategias comerciales para el próximo año.

Una de las estrategias consiste en impulsar la comercialización de bicicletas mediante un convenio exclusivo con una reconocida marca nacional, la cual fabricará una línea exclusiva formada por tres tipos de productos:

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

El sector de marketing está trabajando intensivamente en el diseño de campañas de publicidad por correo electrónico. El gerente general le solicita a usted que con sus conocimientos de minería de datos colabore en esta tarea.

La jefa de marketing cuenta con un archivo de 1500 potenciales clientes (“destinatarios.txt”) sobre los cuales habrá que decidir si se le envía o no la publicidad y el contenido del correo.

El gerente de ventas le ha proporcionado un archivo histórico “clientes.csv” donde se detalla cada cliente de la empresa y si alguna vez ha comprado una bicicleta.

El archivo está compuesto por los siguientes campos:

- IdCliente
- IdCiudad
- Nombre
- Apellido
- FechaNacimiento
- EstadoCivil
- Genero
- Email
- IngresoAnual
- TotalHijos
- Educación
- Ocupación
- Propietario
- CantAutomoviles
- Dirección
- Teléfono

- Distancia al trabajo (km)
- Región
- Edad
- ComproBicicleta

La jefa de marketing le solicita que considere el hecho de que es preferible enviarle innecesariamente un correo a una persona que no resulte comprador, y no perder un potencial cliente porque no se le mandó la publicidad.

El jefe de publicidad ha detectado 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.

El gerente de ventas le comenta 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. Le solicita a usted que recomiende cuáles serían los mercados candidatos, teniendo en cuenta toda la información del archivo “mercados.xlsx”.

Fase de Análisis del problema

Definición del Problema y Objetivos

El gerente de “AllHome” nos solicitó que trabajemos junto con el sector de marketing para llevar a cabo la campaña de comercialización de bicicletas para niños, estándares y deportivas, con el fin de buscar un buen posicionamiento de la empresa en el mercado. La publicidad será distribuida por correo electrónico.

El primer objetivo a alcanzar, es predecir a qué potencial cliente enviar dicho correo electrónico. Para esto, la jefa de marketing cuenta con un archivo de 1500 potenciales clientes, el cual se usará para decidir si se le envía o no la publicidad.

El segundo objetivo es caracterizar a los clientes para determinar qué tipo de bicicleta ofrecer en la publicidad.

El tercer objetivo es determinar cuáles son los países alternativos más convenientes para comenzar a expandirse y hacer campañas en ellos.

Se prefiere enviar innecesariamente un correo a una persona que no resulte comprador, y no perder un potencial cliente porque no se le mandó la publicidad.

Técnicas a Utilizar

Utilizando Python, se realizará un análisis exploratorio de los datos contenidos en el archivo “Clientes”, al cual se le aplicará análisis univariante y multivariante, a fin de comprender los datos, sus relaciones y su comportamiento.

También se realizará una limpieza de los datos que se consideren irrelevantes, como por ejemplo, aquellos que sean únicos para cada registro (id, dni, domicilio) y a la vez se determinara qué estrategia se va a utilizar ante los datos erróneos (eliminarlos o corregirlos) y los outliers.

Posteriormente, se van a crear diferentes modelos con Modeler y RapidMiner para encontrar aquel modelo que permita clasificar a los clientes de manera confiable y al costo más bajo.

- Python para el análisis exploratorio y limpieza de datos.
- RapidMiner para la creación de modelos KNN (vecino más próximo).
- SPSS Modeler para la creación de árboles de decisión.
- SPSS Statistics para realizar el Análisis Discriminante.

Fase de preprocessamiento y análisis de los datos

En primer lugar, descartamos aquellas variables que consideramos que no son necesarias para resolver los objetivos y no nos aportan valor:

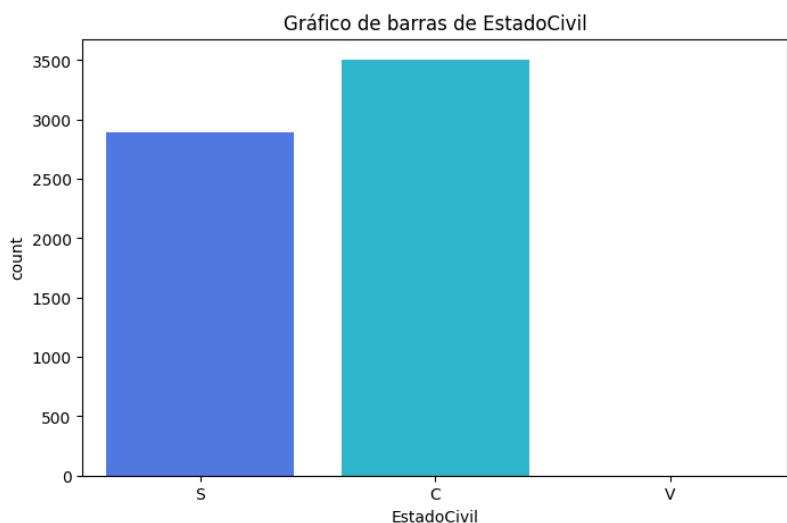
- IdCliente e IdCiudad: Son datos únicos para cada observación y no aportan información. Para el caso de IdCiudad tampoco disponemos de un dataset con la información de qué ciudad corresponde al ID.
- Direccion, Telefono, Nombre, Apellido, Email: No aportan información adicional, solo sirven como identificación del cliente.
- FechaNacimiento: Es redundante ya que se tiene el campo Edad.
- FechaPrimeraCompra: Dato único que solo identifica una fecha en el cliente.

Análisis Exploratorio de Datos Univariante

Estado Civil

Variable cualitativa categórica

Al ser una variable cualitativa, utilizaremos para esta y las demás un gráfico de barras para representar la cantidad de valores para cada categoría de la variable y a partir de ello hacer su respectivo análisis

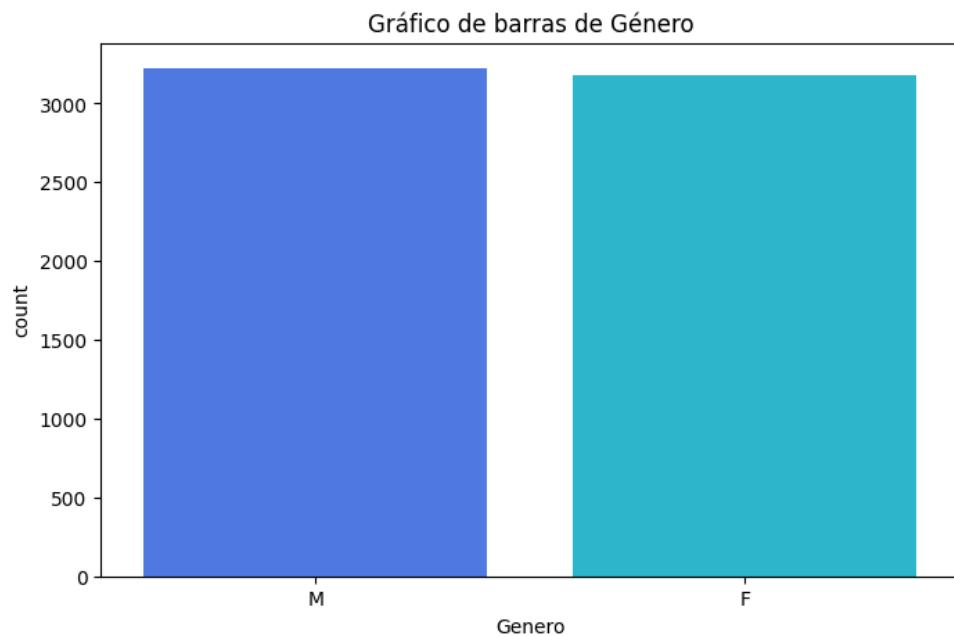


C	3504
S	2894
V	2

En este gráfico podemos concluir que la distribución de los datos entre los solteros y los casados es similar, sin embargo, los casados son más. En el gráfico de barras no se puede visualizar la cantidad de viudos porque es muy inferior con respecto a los otros dos valores posibles vistos.

Género

Variable cualitativa categórica

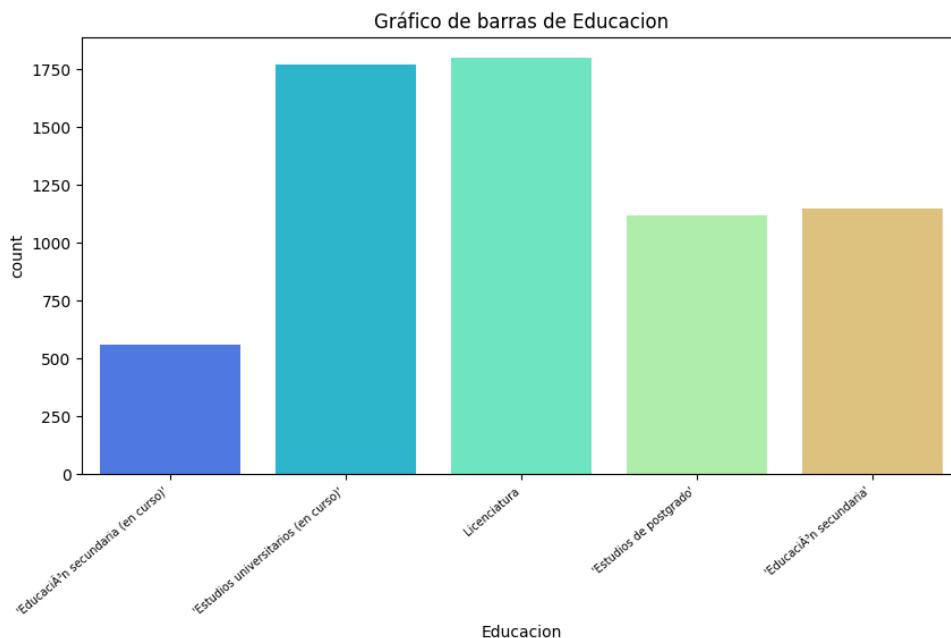


M	3223
F	3177

Se puede observar que la cantidad de registros de ambos géneros son similares.

Educación

Variable Cualitativa Categórica

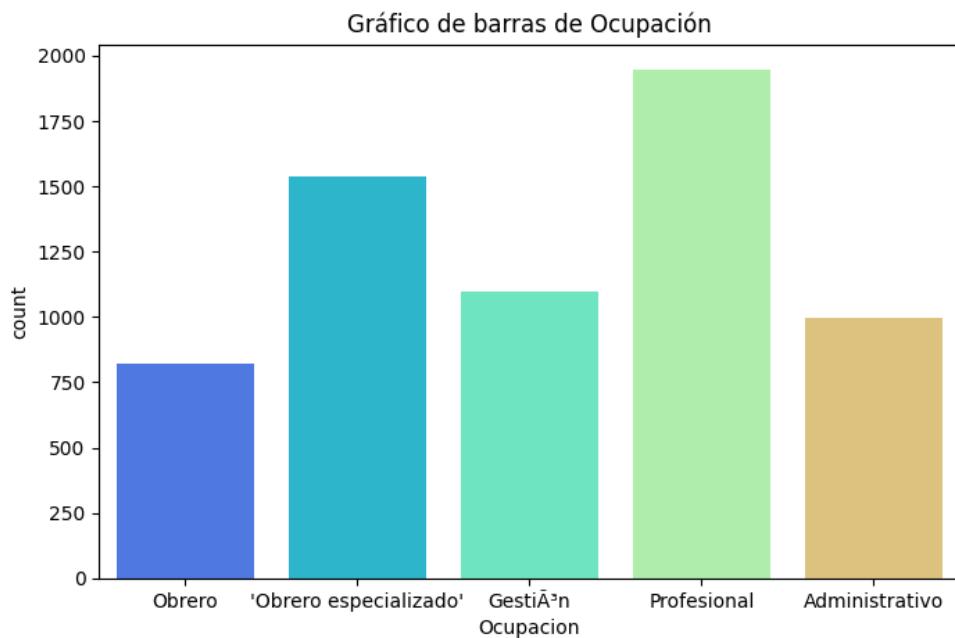


Educación secundaria (en curso)	557
Estudios Universitarios (en curso)	1774
Licenciatura	1800
Estudios de postgrado	1119
Educación secundaria	1150

Se puede observar que es pequeño el porcentaje de observaciones que aún no finalizaron sus estudios secundarios. Por otro lado, más del 50% de las observaciones cuentan con estudios académicos ya finalizados o en curso (Licenciatura, Estudios universitarios en curso).

Ocupación

Variable cualitativa categórica

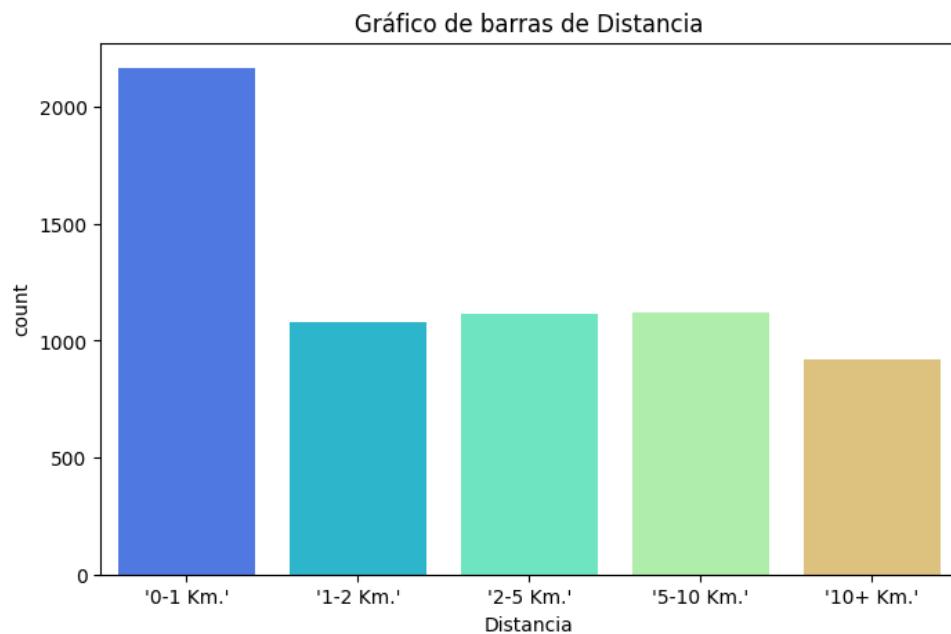


Obrero	821
Obrero especializado	1537
Gestión	1098
Profesional	1946
Administrativo	998

Este gráfico muestra que la mayoría de las personas son profesionales, seguido de una gran cantidad de obreros especializados. Entre los dos forman un poco más del 50% de las observaciones. Las demás ocupaciones tienen cantidades similares y no superan el 18% cada una.

Distancia

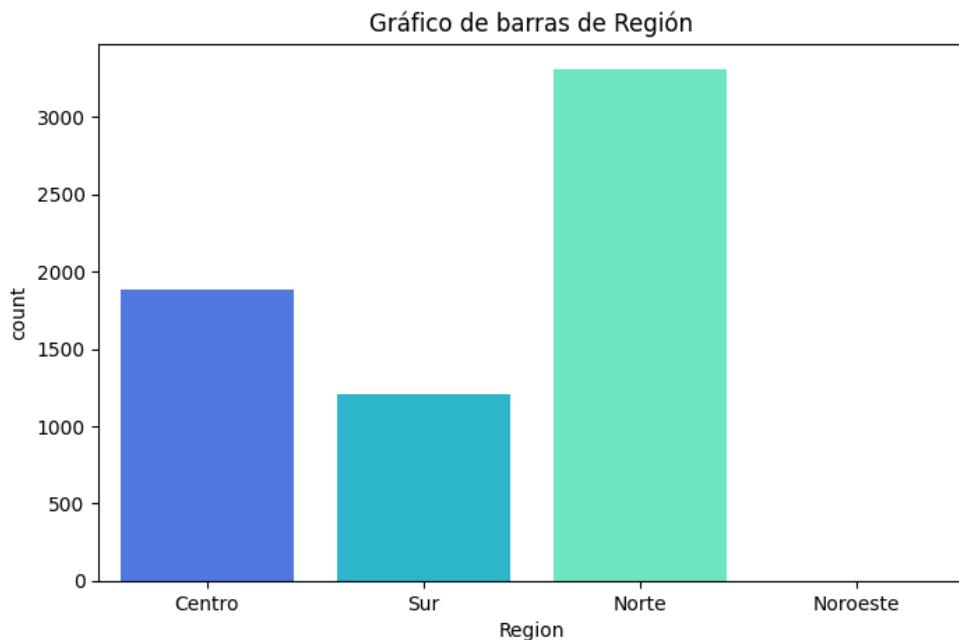
Variable cualitativa categórica



0 - 1 km	33.84%
1 - 2 km	16.88%
2 - 5 km	17.41%
5 - 10 km	17.53%
10+ km	14.34%

Este gráfico describe la proporción de la población en función de la distancia a su respectivo trabajo. Se puede observar que la población de 1-2 km, 2-5 km, 5-10 km son relativamente homogéneas. La población que está más cerca del trabajo es la mayor.

Región

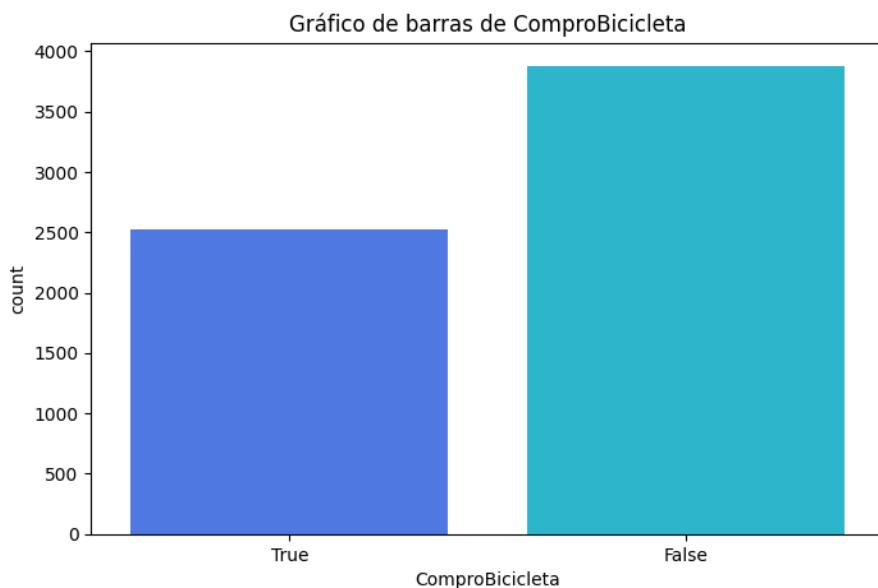


Centro	29.38%
Sur	18.89%
Norte	51.72%
Noroeste	0.01%

Este gráfico representa la proporción de la población en función de la zona en la que residen. A partir de este, se puede observar que aproximadamente más del 50% de los clientes se alojan en el Norte. En el otro 50% predomina la población de clientes del centro y solo uno de los clientes se encuentra en el Noroeste.

ComproBicicleta

Variable cualitativa categórica



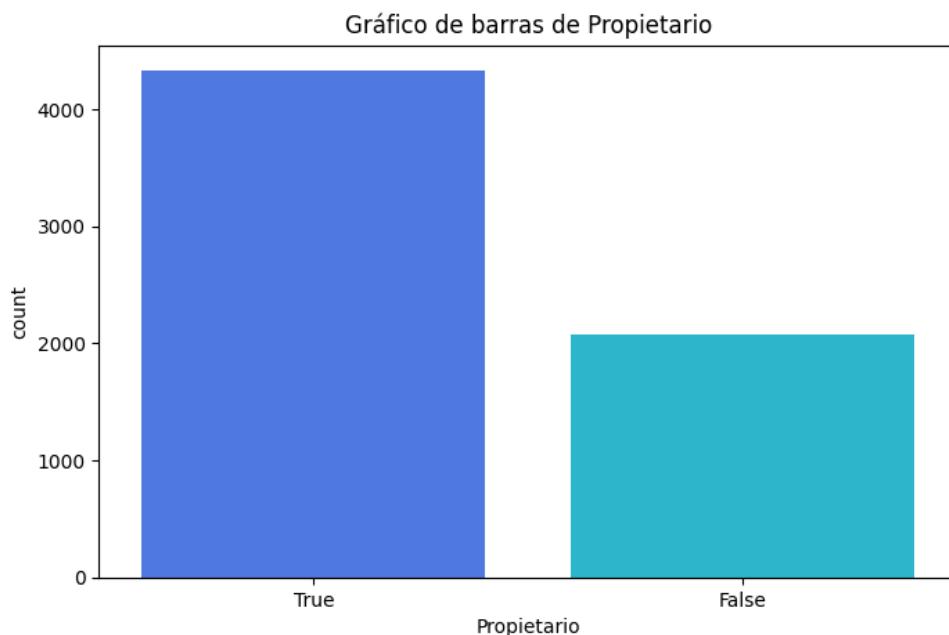
False	60.56%
True	39.44%

- False: Indica que no compro bicicleta.
- True: Indica que compraron alguna vez bicicleta.

A partir del gráfico y análisis se puede demostrar que aproximadamente el 60% de los clientes no compraron alguna vez una bicicleta.

Propietario

Variable cualitativa categórica



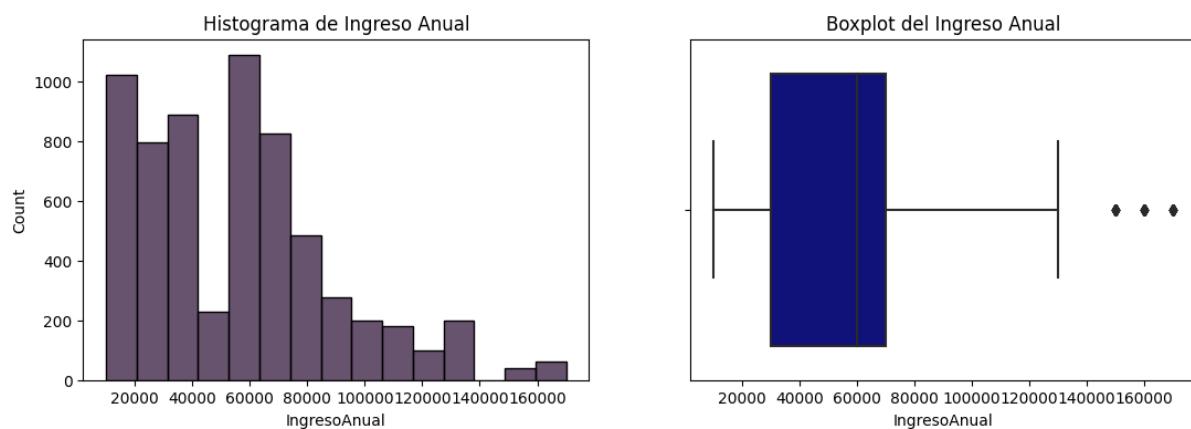
True	67.66%
False	32.34%

- False: Indica que no es propietario.
- True: Indica que es propietario.

A partir del gráfico y análisis se puede demostrar que aproximadamente 67% de los clientes son propietarios.

IngresoAnual

Variable cuantitativa continua

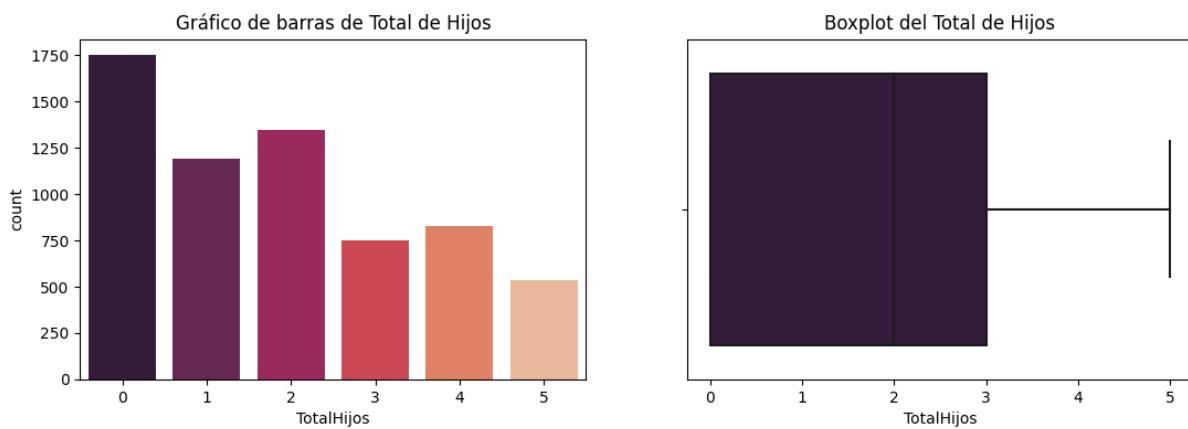


cantidad observaciones	6390
promedio	57532.08
desvío estándar	32331.97
min	10000
1er cuartil (Q1 = 25%)	30000
mediana (Q2 = 50%)	60000
3er cuartil (Q3 = 75%)	70000
max	170000

En base al histograma y al boxplot podemos observar que el 50% de los clientes tienen un ingreso anual entre 30000 y 70000. Se visualizan 3 outliers que consideraremos más adelante.

TotalHijos

Variable cuantitativa discreta



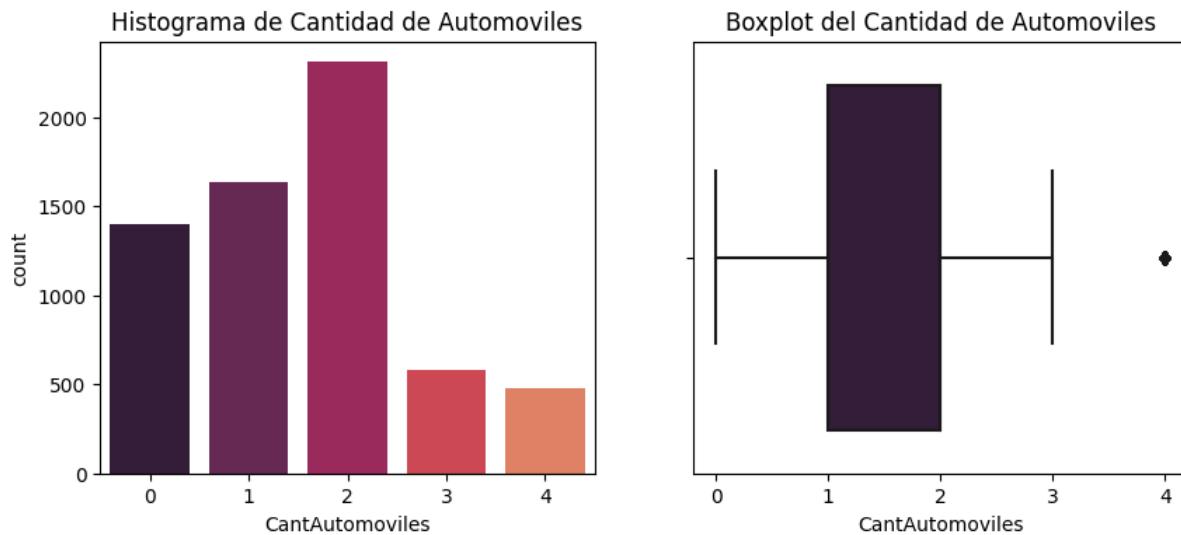
0	27.34%
1	18.60%
2	21.02%
3	11.69%
4	12.94%
5	8.41%

cantidad observaciones	6400
promedio	1.894844
desvío estándar	1.630993
min	0
1er cuartil (Q1 = 25%)	0
mediana (Q2 = 50%)	2
3er cuartil (Q3 = 75%)	3
max	5

Aproximadamente el 30% de los clientes no tienen hijos, y el 50% de las observaciones tienen entre 0 y 3 hijos (Q1 y Q3). Estos datos serán relevantes más adelante para determinar a quién mandar promociones de las bicicletas Kinder.

CantAutomoviles

Variable cuantitativa discreta



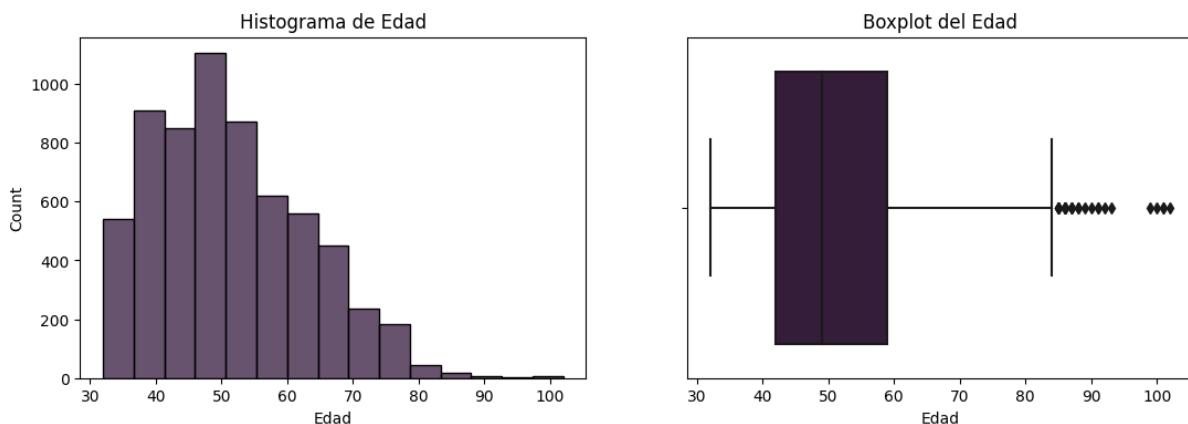
0	21.86%
1	25.55%
2	36.06%
3	9.03%
4	7.50%

Cantidad observaciones	6400
Promedio	1.547656
Desvío estándar	1.147060
Min	0
1er cuartil (Q1 = 25%)	1
Mediana (Q2 = 50%)	2
3er cuartil (Q3 = 75%)	2
Max	4

Se observa que aproximadamente el 36% de las observaciones tienen 2 vehículos. Más del 50% de los clientes tienen entre 1 y 2 automóviles, aunque hay más del 20% que no tienen ninguno. En el gráfico de barras y en el Box Plot se observa que la mediana es de 2 autos.

Edad

Variable cuantitativa discreta



Cantidad Observaciones	6400
Promedio	51.195469
Desvío Estándar	11.517698
Min	32
1er Cuartil (Q1 = 25%)	42
Mediana (Q2 = 50%)	49
3er Cuartil (Q3 = 75%)	59
Max	102

El histograma es asimétrico a la derecha. El 50% de las observaciones tienen entre 42 y 59 años (en base a los cuartiles Q1 y Q3). En el Boxplot se observan varios valores anómalos. Desde los 70 años de edad en adelante decrece la cantidad de clientes. No hay clientes menores de 32 años.

Análisis Exploratorio de Datos Multivariante

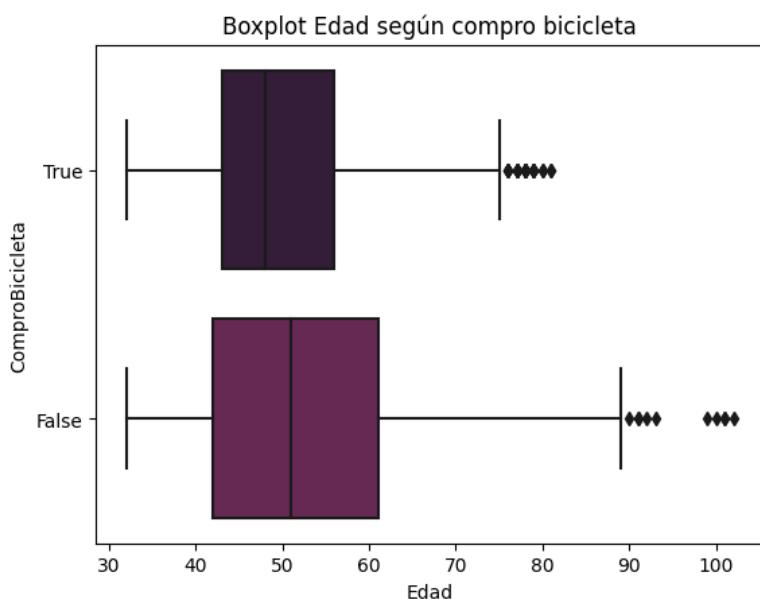
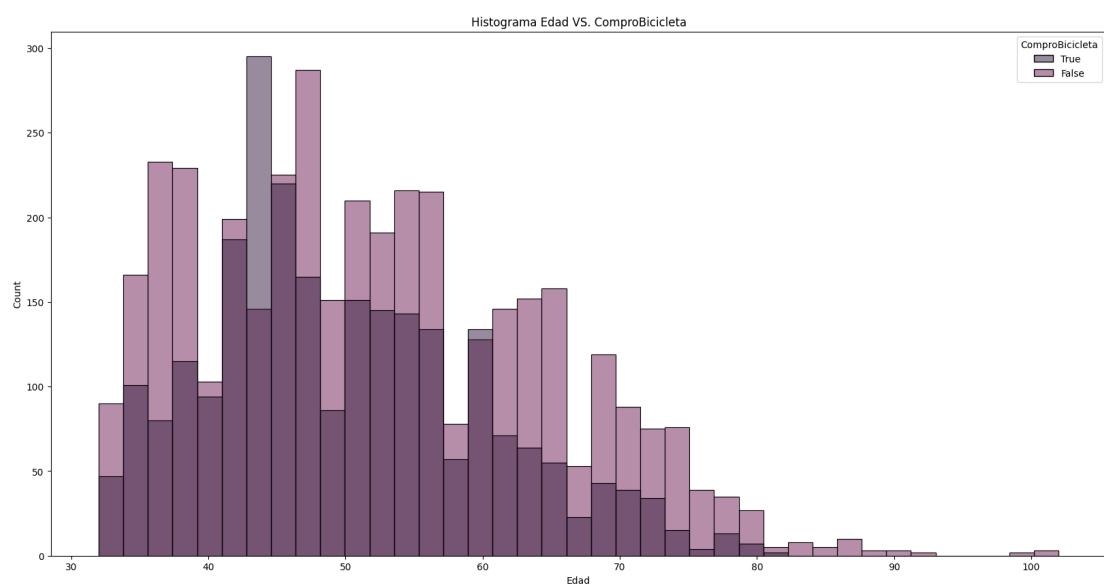
Vamos a analizar en conjunto las siguientes variables:

- Edad vs Compro bicicleta
- Edad vs Ingreso Anual
- Edad vs Ocupación
- Edad vs Distancia
- Ingreso Anual vs Compro bicicleta
- Ingreso Anual vs Ocupación
- Total Hijos vs Compro bicicleta
- Total Hijos vs Estado Civil

- Cantidad Automóviles vs Compro bicicleta
- Región vs Compro bicicleta
- Distancia vs Compro bicicleta
- Distancia vs Cantidad de Automóviles
- Cantidad de autos vs Ingreso Anual vs Compro bicicleta
- Cantidad de autos vs Total hijos vs Compro bicicleta
- Distancia vs Ocupación vs Compro bicicleta

Edad vs ComproBicicleta

Al ser una variable numérica y una cualitativa, utilizaremos un histograma de color junto a boxplots para representarlas.



Donde:

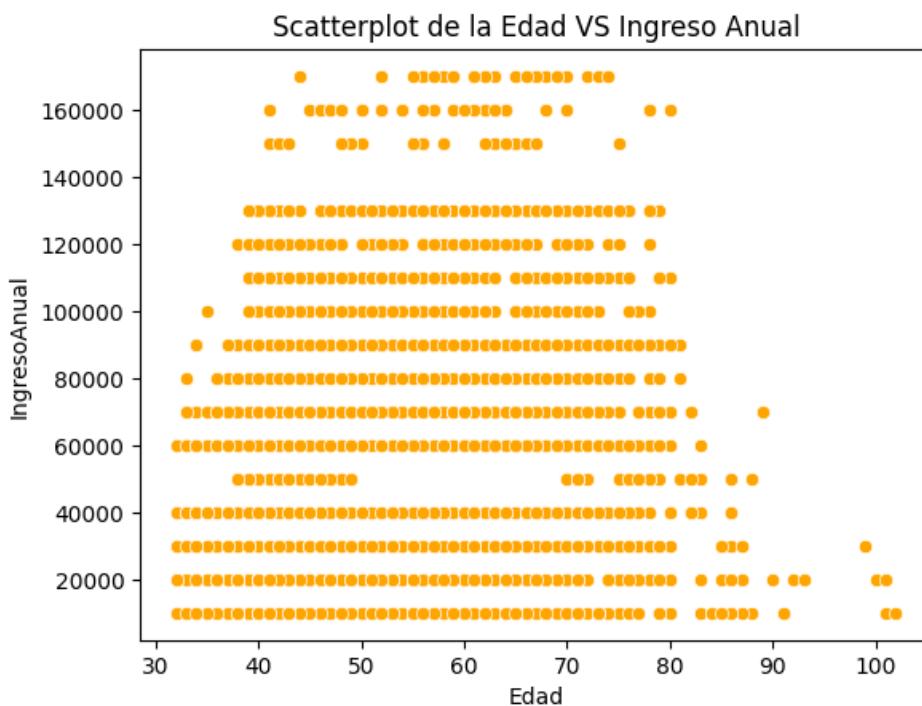
- True: Indica que si compro bicicleta
- False: Indica que no compro bicicleta

El 50% de los compradores de bicicletas tienen entre 40 y 55 años. Los que no compran bicicletas tienen un comportamiento similar, el 50% se ubica entre 43 y 60 años. A partir de los 70 años, comienza a disminuir la cantidad de compradores hasta los 80 años, donde no hay compradores.

En base a este análisis, consideramos evaluar la variable Edad con la variable Ingreso Anual, ya que podría ocurrir que las edades más propensas a comprar bicicletas correspondan a personas con sueldos altos.

Edad vs IngresoAnual

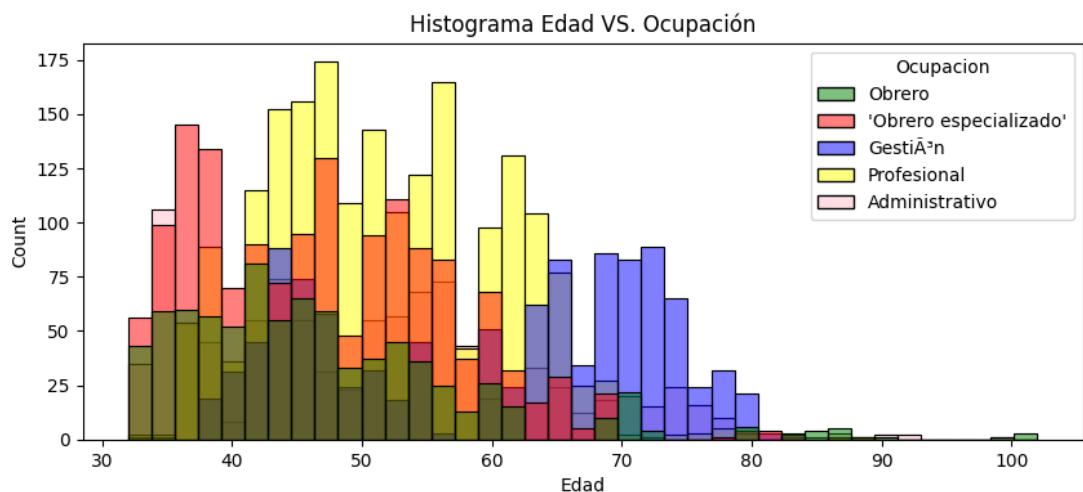
Al ser ambas variables numéricas, utilizamos un diagrama de dispersión para representarlas



Aproximadamente a partir de los 85 años, la mayoría de las observaciones no tiene ingresos mayores a 40000. Esto puede deberse a que las personas mayores son jubiladas. Las personas más jóvenes no tienen ingresos superiores a 100000. Los sueldos más altos están comprendidos en aquellos clientes que tienen entre 40 y 80 años de edad. Esto puede deberse a su ocupación.

Edad vs Ocupación

Al ser una variable numérica y una cualitativa, utilizaremos un histograma de color para representarlas



En el gráfico observamos que aquellas personas jóvenes que tienen sueldos bajos se debe a que su ocupación es Obrero. Quienes tienen los sueldos más altos comprendidos entre los 42 y 65 años aproximadamente se debe a que en su mayoría son Profesionales.

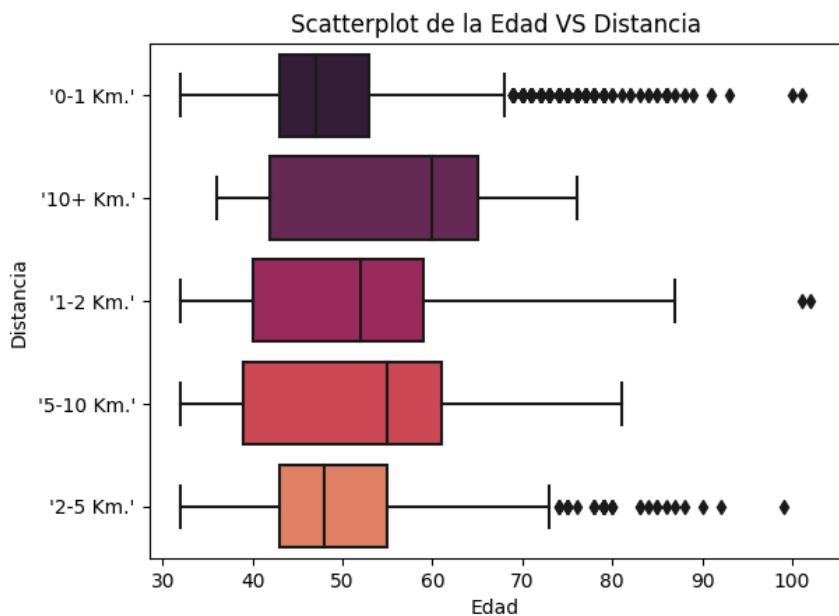
Quienes tienen sueldos medianamente altos y son de mayor edad, tienen ocupaciones de Gestión, que puede deberse a sus años de experiencia tener ese tipo de ocupaciones.

Quienes tienen sueldos medianamente altos y son de menor edad, tienen ocupaciones de Obrero Especializado. Puede deberse a sus años de experiencia como Obrero.

En base a este análisis, consideramos que será muy interesante relacionar la Ocupación con el Ingreso anual.

Edad vs Distancia

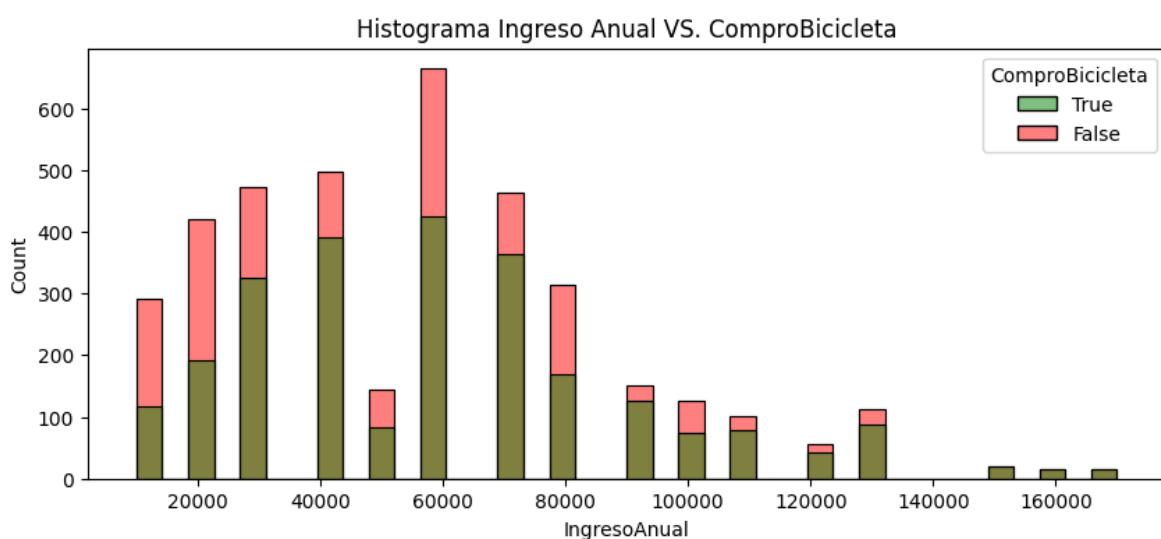
Al ser una variable numérica y una cualitativa, utilizaremos un boxplot para representarlas.

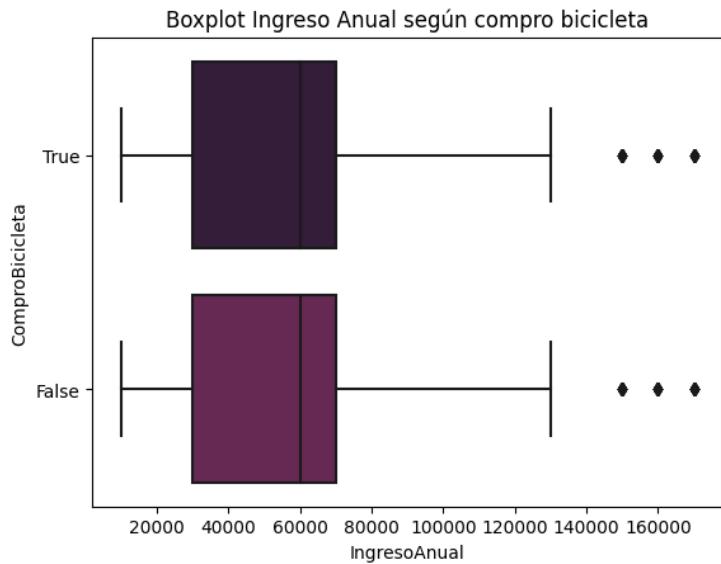


La distribución de los datos es bastante similar. Podemos observar que hay varios outliers en aquellos rangos de menor y mayor distancia.

De todas formas, vemos que la edad no es una variable significativa en cuanto a si las personas mayores o menores prefieren alejarse o no de sus trabajos.

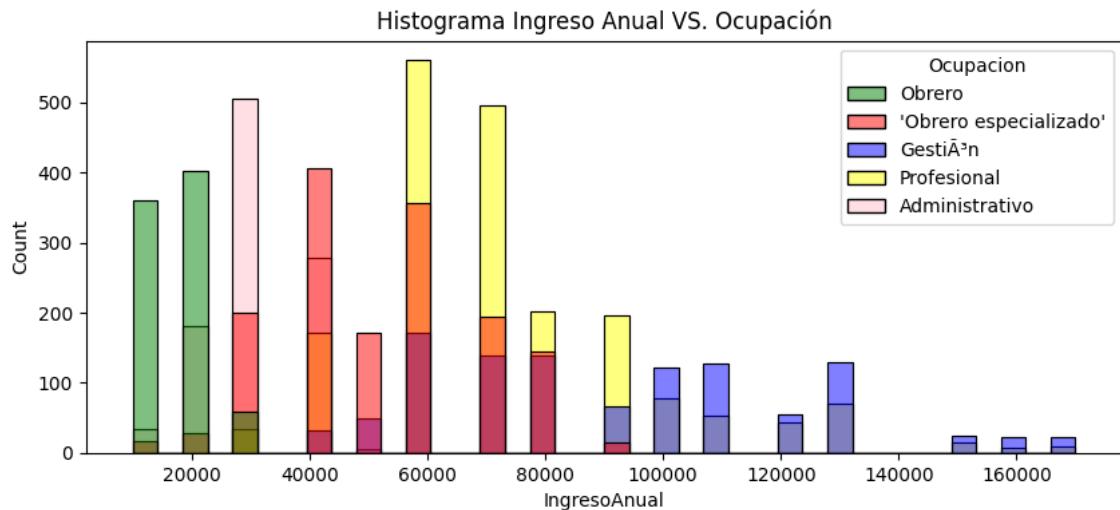
IngresoAnual vs ComproBicicleta

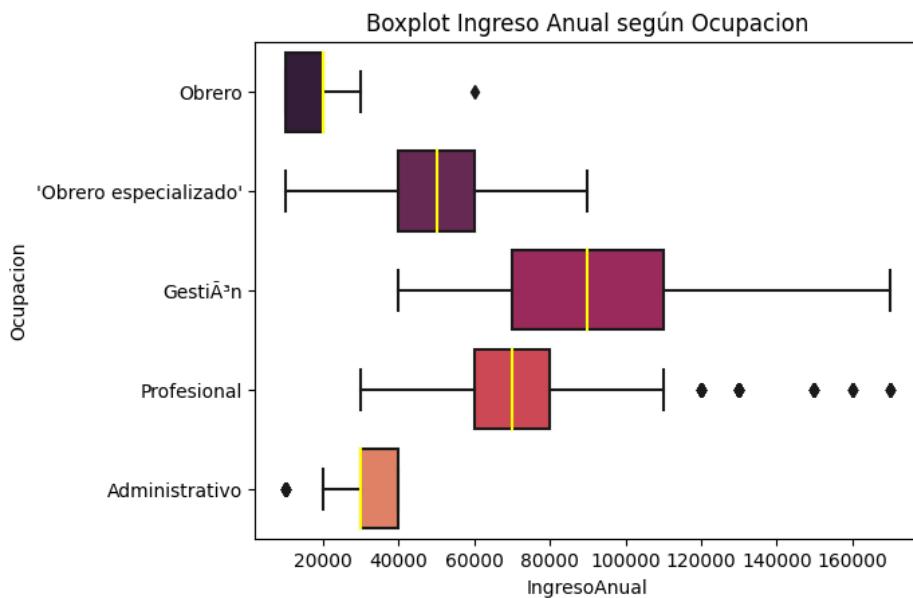




Podemos observar que mientras más aumenta el ingreso anual de las personas, mayor es la probabilidad de que se compren una bicicleta. Los boxplots se ven similares porque la distribución tanto de la compra o no de las bicicletas es similar. Se observan outliers.

IngresoAnual vs Ocupación



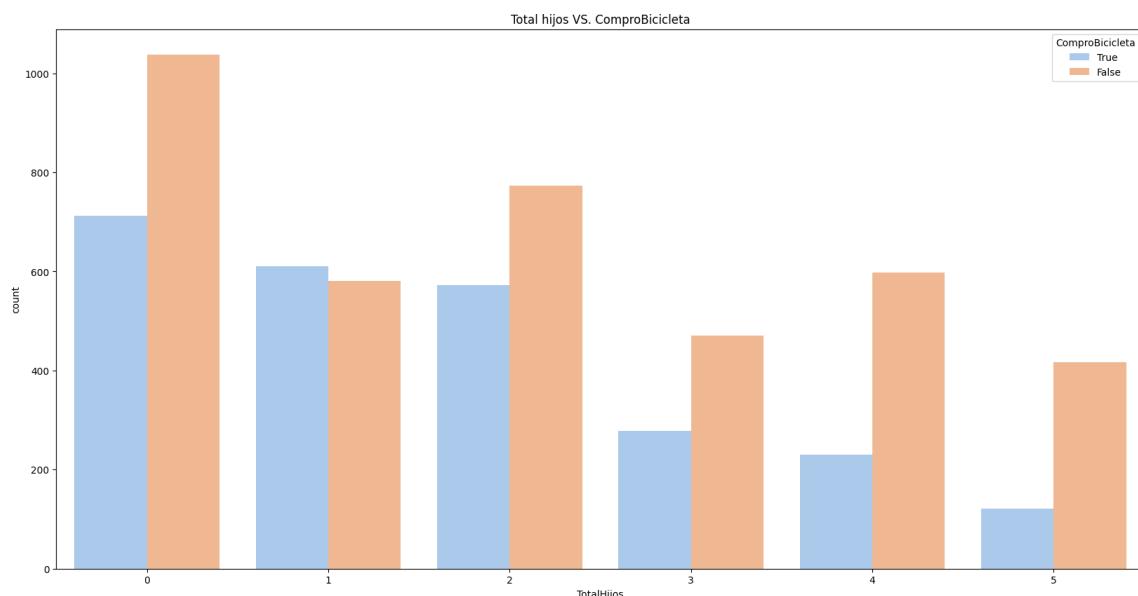


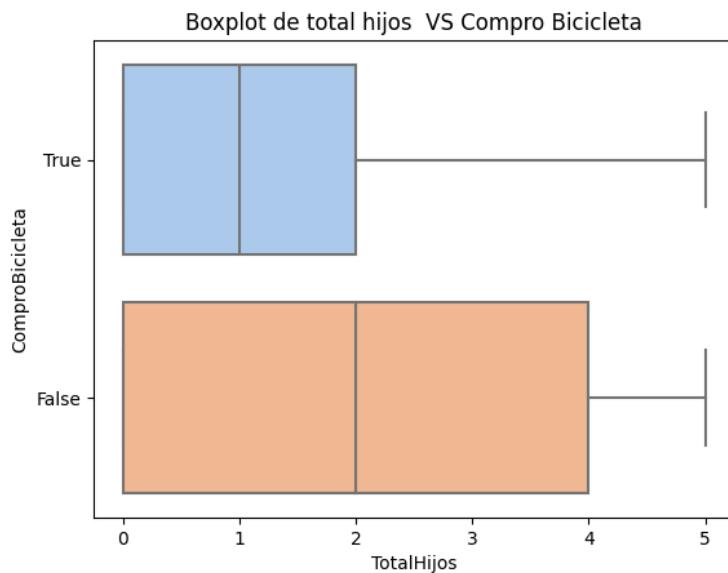
Observamos en la gráfica de Edad VS Ocupación que las personas mayores eran las que desarrollan ocupaciones de Gestión, y a la vez, en la gráfica de Edad VS ComproBicicleta, que estas personas no eran tan propensas a comprar bicicletas.

Las personas más propensas a comprar bicicletas, tenían entre 40 y 55 años, y estos en su mayoría eran profesionales, obreros u obreros especializados.

Esto nos hace ver que los más jóvenes, y con sueldos medianamente buenos (ni muy bajos ni muy altos), son personas más propensas a comprar bicicletas.

TotalHijos vs ComproBicicleta





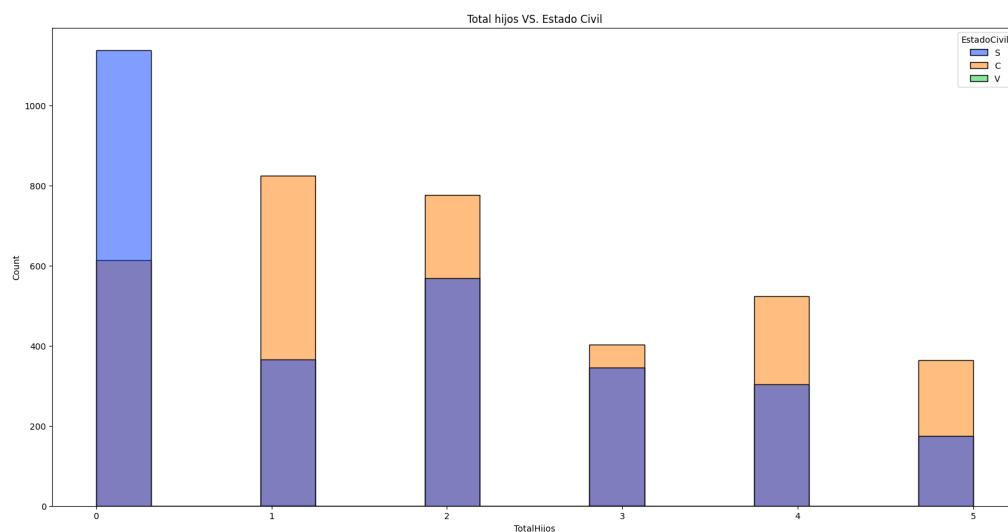
Donde:

- True: Indica que si compro bicicleta
- False: Indica que no compro bicicleta

El 50% de los que compraron bicicletas tienen entre 0 y 2 hijos. El 50% de los que no compraron bicicletas tienen entre 0 y 4 hijos. Esto también puede deberse a que aquellos que tienen más hijos, pueden compartir la bicicleta de sus hermanos.

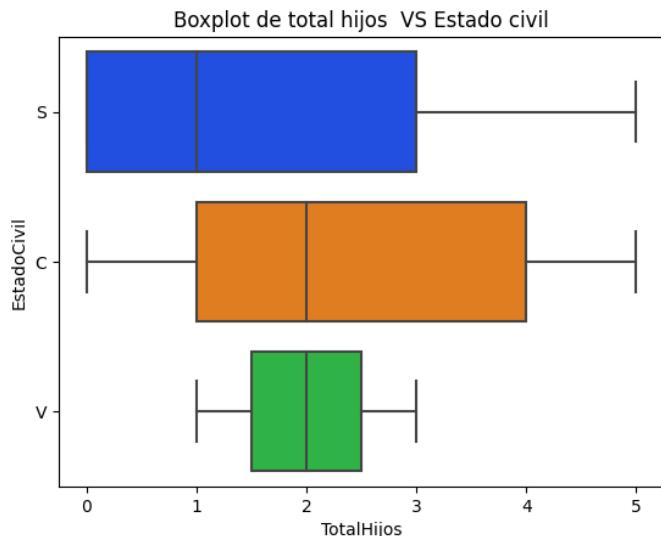
Sería interesante analizar quiénes tienen menos hijos, si los clientes casados o solteros.

TotalHijos vs EstadoCivil



Aquellas personas casadas (Estado Civil=Casado), suelen tener más hijos, y en el gráfico anterior vimos que las personas con menos cantidad de hijos eran más

propensas a comprar bicicletas, por lo que esta información es de gran valor. Hay una observación correspondiente a estado civil viudo para 1 y 3 hijos.

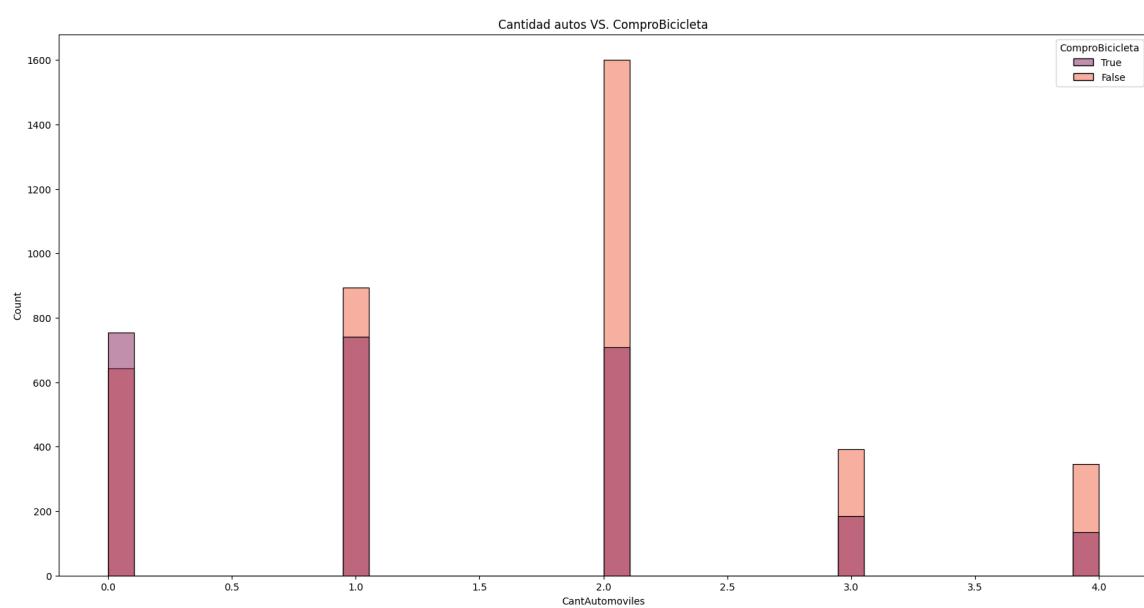


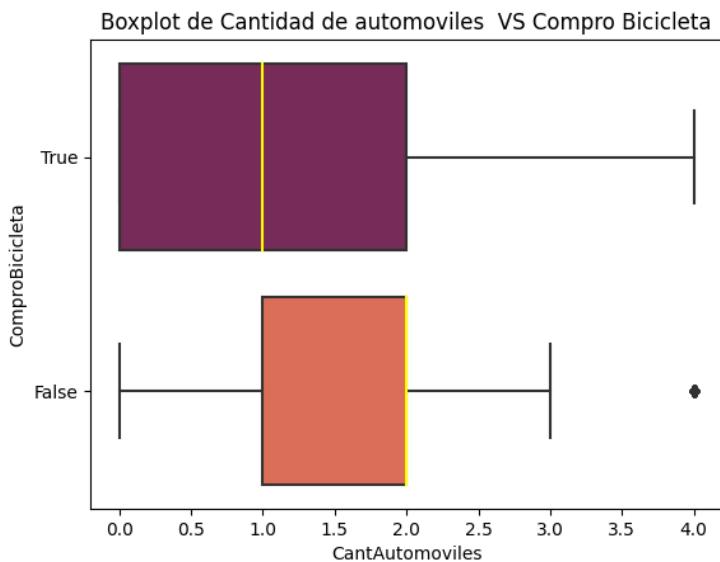
El Boxplot nos muestra que el 50% de las observaciones de los solteros tienen entre 0 y 3 hijos, donde la mediana es 1 hijo.

El 50% de las observaciones de los casados tienen entre 1 y 4 hijos.

Hay 2 observaciones de viudos, de los cuales uno tiene 1 hijo y el otro 3 hijos.

CantAutomoviles vs ComproBicicleta





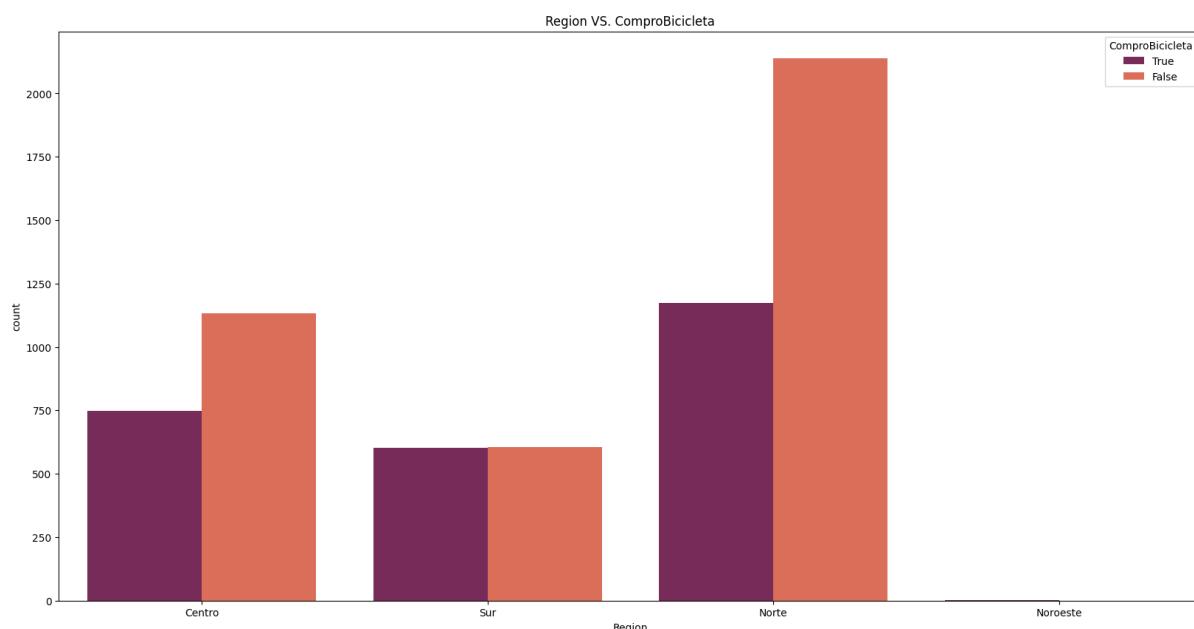
Donde:

- True: Indica que si compro bicicleta
- False: Indica que no compro bicicleta

Se observa en el gráfico de barras que es más propenso a comprar una bicicleta una persona que no tiene automóvil. Desde que tienen 2 automóviles ya son menos propensos a comprarla.

Los boxplot nos muestran a la vez que 50% de los que compraron bicicletas tienen entre 0 y 2 autos y que al no tener auto, un cliente es mucho más propenso a comprar una bicicleta.

Region vs ComproBicicleta

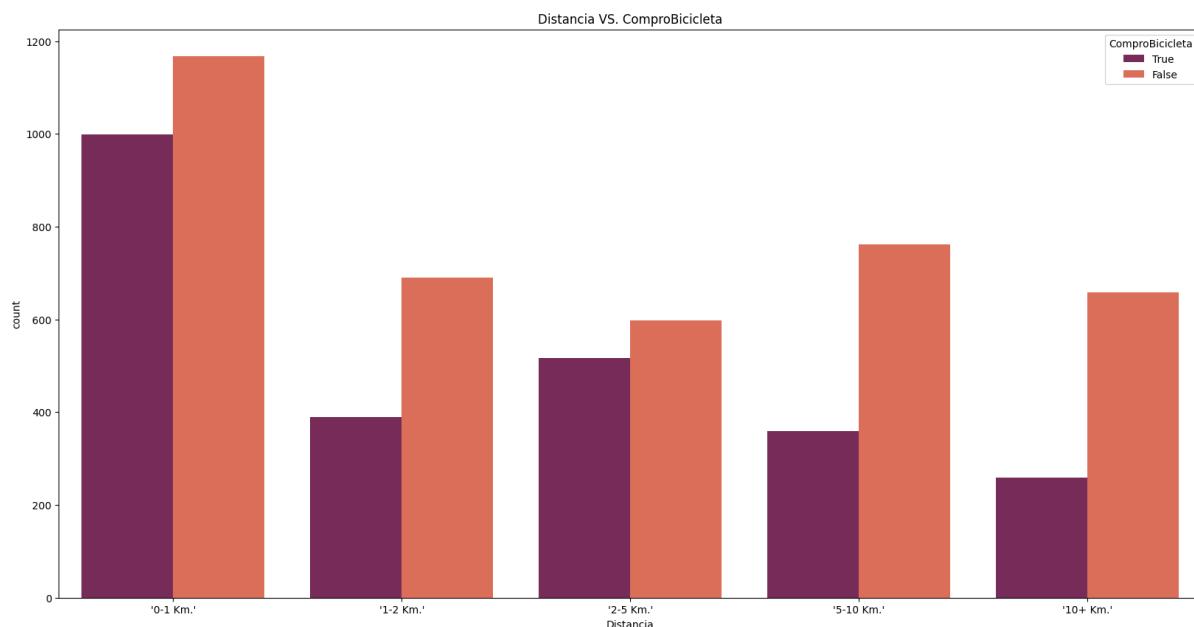


Donde:

- True: Indica que si compro bicicleta
- False: Indica que no compro bicicleta

Se observa que en el norte del país las personas no son propensas comprar bicicletas, caso contrario se da en el Sur.

Distancia vs ComproBicicleta

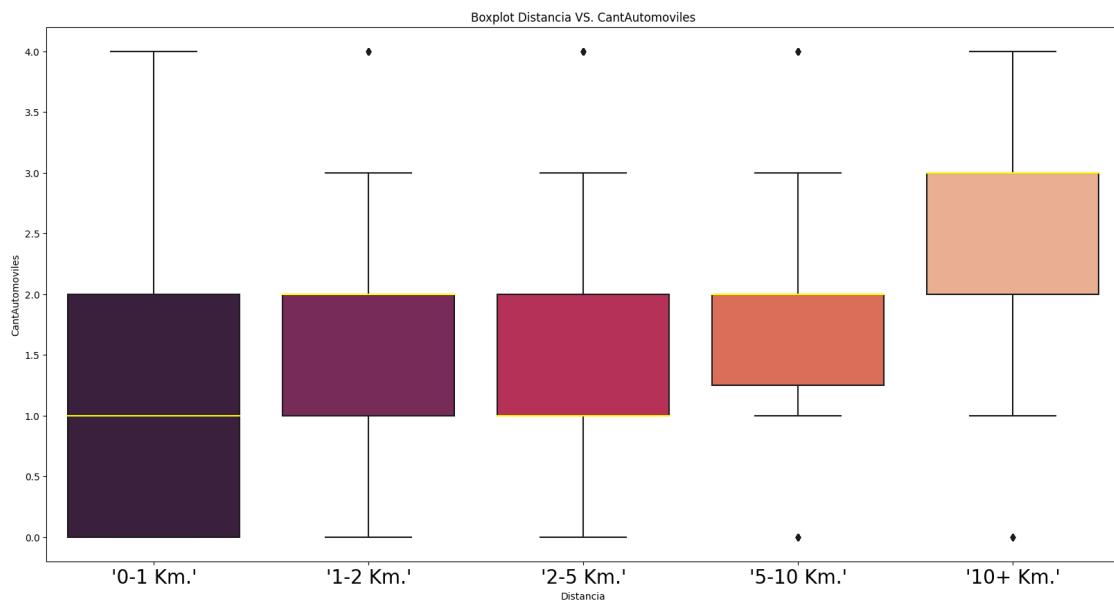


Donde:

- True: Indica que si compro bicicleta
- False: Indica que no compro bicicleta

Se puede observar que mientras más lejos es la distancia a la que viven las personas, disminuye la cantidad de personas que compraron bicicletas. Es probable que estas personas tengan automóviles o usen otro medio de transporte para llegar al trabajo debido a la distancia. Haremos el análisis en la siguiente gráfica.

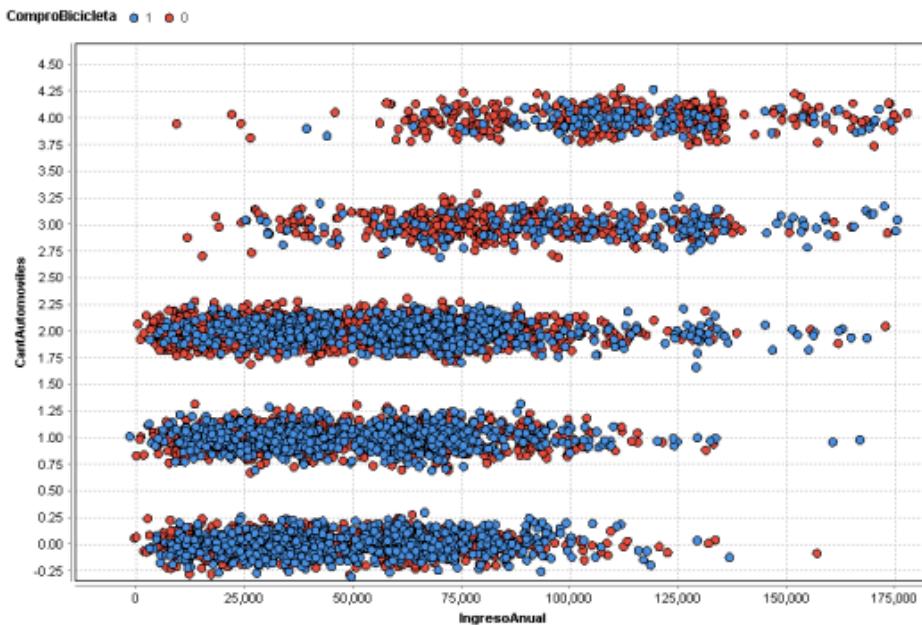
Distancia vs CantAutomoviles



Se puede observar que quienes viven a distancias mayores a 5Km tienen un vehículo o más, y la mayoría de quienes están más cerca no tienen, o como mucho tienen 2 vehículos.

CantAutomoviles vs IngresoAnual vs ComproBicicleta

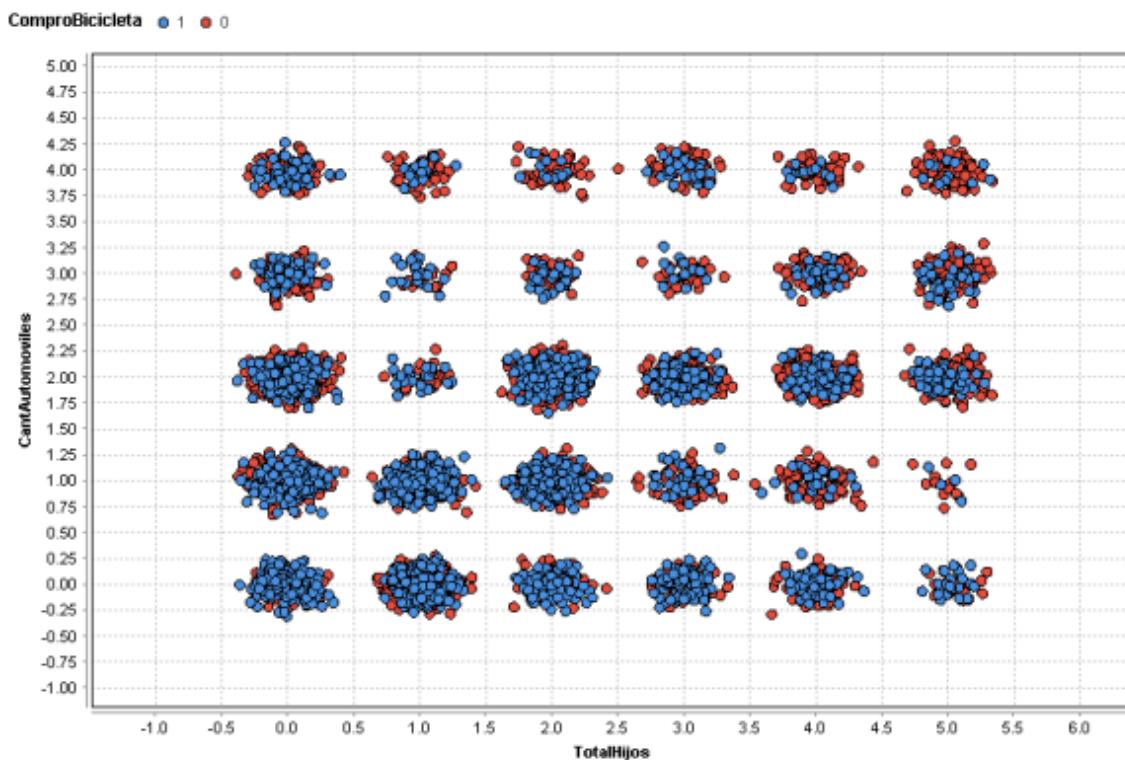
Al ser el análisis de 3 variables, una numérica y dos cualitativas, utilizaremos un gráfico de dispersión para visualizar la relación.



Se puede observar que quienes menos vehículos tienen poseen ingresos en general menores a los 120000, y son mucho más propensos a comprar bicicletas.

En cambio, quienes tienen entre 3 y 4 vehículos, poseen ingresos anuales mayores a 30000 y son propensos a no comprar bicicletas.

CantAutomoviles vs TotalHijos vs ComproBicicleta



Se puede observar que si se sigue la diagonal desde la esquina inferior izquierda (sin vehículos y sin hijos) hasta la esquina superior derecha (muchos vehículos y muchos hijos) decrece la probabilidad de comprar bicicleta.

Este gráfico es muy útil, nos hace pensar también que quienes tienen más hijos necesitan un transporte más grande que una bicicleta para poder transportarlos a todos juntos. En cambio, quienes tengan un solo hijo o ninguno, con una bicicleta podrían movilizarse.

Matriz R

Matriz de Correlación: muestra la relación lineal entre cada par de variables, donde un valor cercano a 1 indica una correlación positiva fuerte (es decir, que ambas variables tienden a aumentar o disminuir juntas), un valor cercano a -1 indica una correlación negativa fuerte (es decir, que cuando una variable aumenta, la otra tiende a disminuir) y un valor cercano a 0 indica que no hay una correlación lineal significativa entre las variables.

	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 este caso se puede observar que no hay ninguna relación evidente entre las variables.

Matriz S

Matriz de Covarianza: En esta matriz, lo que podemos ver son las direcciones de los signos, pero no podemos ver la magnitud, y no nos permite afirmar nada.

A diferencia de la correlación, la covarianza no está normalizada y su magnitud depende de las unidades de las variables involucradas. Esto significa que la covarianza no puede ser utilizada para comparar la fuerza de la relación lineal entre dos pares de variables que tienen unidades diferentes.

	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

Proceso de Limpieza de Datos

En esta sección vamos a analizar aquellas observaciones con valores anómalos o valores nulos en alguna columna y evaluaremos si quitamos esos registros o los reemplazamos por algún valor (por ejemplo la mediana).

Análisis de la columna Estado Civil

Se decide ver si las únicas dos observaciones de EstadoCivil = V (viudo) tienen valores atípicos o si le faltan valores, para decidir si se las deja o no.

EstadoCivil	Genero	IngresoAnual	TotalHijos	Educacion	Ocupacion	Propietario	CantAutomoviles	Distancia	Region	Edad	ComproBicicleta
5494	V	F	70000.0	3	'EducaciÃ³n secundaria'	Profesional	True	0	'1-2 Km.'	58	True
6038	V	M	60000.0	1	Licenciatura	Profesional	False	1	'0-1 Km.'	50	True

Decidimos dejar estas dos observaciones en lugar de eliminarlas, ya que creemos que no son datos erróneos, además la edad de las personas no es un indicador de que no sean posibles compradores de bicicletas.

Análisis de la Columna Edad

Mostramos los valores de los datos cuyo valor en edad sea mayor a 85 para determinar si puede llegar a haber algún dato erróneo o inesperado.

EstadoCivil	Genero	IngresoAnual	TotalHijos	Educacion	Ocupacion	Propietario	CantAutomoviles	Distancia	Region	Edad	ComproBicicleta
336	C	F	10000.0	4 'EducaciÃ³n secundaria'	Obrero	False	2	'0-1 Km.'	Centro	101	False
516	C	M	10000.0	4 'Estudios universitarios (en curso)'	Obrero	False	2	'0-1 Km.'	Centro	88	False
571	C	F	10000.0	4 'EducaciÃ³n secundaria (en curso)'	Obrero	False	2	'1-2 Km.'	Centro	102	False
855	C	F	40000.0	1 Licenciatura	Administrativo	True	0	'0-1 Km.'	Centro	86	False
1182	S	M	40000.0	2 'EducaciÃ³n secundaria'	'Obrero especializado'	False	1	'0-1 Km.'	Norte	86	False
1407	C	M	30000.0	1 Licenciatura	Administrativo	True	1	'0-1 Km.'	Centro	86	False
1460	S	M	20000.0	2 'EducaciÃ³n secundaria (en curso)'	Obrero	False	2	'0-1 Km.'	Centro	100	False
1499	S	M	50000.0	1 Licenciatura	Profesional	False	1	'0-1 Km.'	Sur	86	False
1778	C	M	20000.0	2 'EducaciÃ³n secundaria'	Obrero	True	1	'1-2 Km.'	Centro	101	False
1873	C	F	20000.0	2 Licenciatura	Administrativo	True	1	'2-5 Km.'	Centro	92	False
2230	S	F	30000.0	1 'Estudios universitarios (en curso)'	Administrativo	True	1	'2-5 Km.'	Centro	99	False
2491	C	M	30000.0	1 Licenciatura	Administrativo	True	1	'2-5 Km.'	Centro	87	False
2658	C	M	20000.0	2 'EducaciÃ³n secundaria'	Obrero	False	1	'1-2 Km.'	Centro	87	False
3420	C	M	10000.0	3 Licenciatura	Administrativo	True	2	'0-1 Km.'	Centro	91	False
3502	S	F	50000.0	1 'Estudios de postgrado'	Profesional	True	1	'2-5 Km.'	Norte	88	False
3587	C	M	20000.0	3 'EducaciÃ³n secundaria (en curso)'	Administrativo	False	2	'0-1 Km.'	Norte	93	False
3800	S	F	10000.0	4 'Estudios universitarios (en curso)'	Obrero	False	2	'0-1 Km.'	Centro	87	False
3825	C	M	10000.0	4 'EducaciÃ³n secundaria'	Obrero	False	2	'0-1 Km.'	Centro	86	False
4582	C	F	70000.0	1 Licenciatura	'Obrero especializado'	True	1	'0-1 Km.'	Norte	89	False
4604	S	F	10000.0	4 'EducaciÃ³n secundaria'	Obrero	True	2	'0-1 Km.'	Centro	86	False
6133	C	M	20000.0	2 'Estudios universitarios (en curso)'	Obrero	True	1	'2-5 Km.'	Centro	90	False

Decidimos conservar los registros, porque podrían ser útiles a la hora de predecir a quién no enviar los correos publicitarios.

Análisis de la Columna IngresoAnual

Mostramos los datos cuya variable de IngresoAnual era nula para determinar el mejor método para imputar los campos

	EstadoCivil	Genero	IngresoAnual	TotalHijos	Educacion	Ocupacion	Propietario	CantAutomoviles	Distancia	Region	Edad	ComproBicicleta
61	C	F	NaN	1	Licenciatura	'Obrero especializado'	True	1	'0-1 Km.'	Norte	44	False
725	S	F	NaN	0	'Estudios de postgrado'	'Obrero especializado'	False	0	'0-1 Km.'	Norte	39	False
1205	S	F	NaN	1	Licenciatura	Profesional	True	1	'5-10 Km.'	Sur	50	True
2523	S	F	NaN	4	'Educación secundaria'	Profesional	False	2	'1-2 Km.'	Norte	72	False
3082	C	M	NaN	2	'Estudios universitarios (en curso)'	Profesional	True	1	'2-5 Km.'	Norte	62	False
4531	S	M	NaN	0	Licenciatura	Profesional	True	3	'10+ Km.'	Sur	37	False
4871	C	M	NaN	1	'Estudios de postgrado'	'Obrero especializado'	True	0	'1-2 Km.'	Norte	41	False
5264	S	F	NaN	3	'Educación secundaria'	Profesional	True	4	'5-10 Km.'	Centro	57	True
6018	C	M	NaN	3	'Estudios universitarios (en curso)'	Profesional	True	4	'10+ Km.'	Centro	58	False
6214	S	F	NaN	2	'Educación secundaria'	Obrero	True	0	'0-1 Km.'	Centro	43	True

En este caso decidimos imputar la mediana, en lugar de eliminar las observaciones, tomamos esta decisión, ya que las filas no están completamente vacías, sino que por alguna razón falta el valor del Ingreso Anual, y si eliminamos la fila completa solo porque falta un valor, estaríamos perdiendo información.

Vista Minable

En la vista minable resultante tendremos la variable a predecir "ComproBicicleta" y el resto serán las variables predictoras.

	EstadoCivil	Genero	IngresoAnual	TotalHijos	Educacion	Ocupacion	Propietario	CantAutomoviles	Distancia	Region	Edad	ComproBicicleta
0	S	M	10000.0	4	'Educación secundaria (en curso)'	Obrero	True	1	'0-1 Km.'	Centro	46	True
1	S	M	70000.0	1	'Estudios universitarios (en curso)'	'Obrero especializado'	False	1	'0-1 Km.'	Sur	55	True
2	C	F	40000.0	1	Licenciatura	'Obrero especializado'	True	1	'0-1 Km.'	Centro	39	True
3	C	M	60000.0	3	'Estudios de postgrado'	Gestión	True	2	'10+ Km.'	Norte	74	False
4	S	F	60000.0	0	'Estudios universitarios (en curso)'	'Obrero especializado'	False	2	'1-2 Km.'	Norte	36	False

Fase de Modelado

Para los modelos utilizaremos los árboles CHAID, QUEST y C5.0, KNN y análisis discriminante. En todos los casos el dataset lo separaremos en un dataset de "Entrenamiento" y "Validación" en una proporción 70-30.

Para analizar cada modelo utilizaremos la matriz de confusión. Recordando que significa cada valor en la matriz de confusión:

Matriz de confusión	Actual True	Actual False
predicted True	TP	FP
predicted False	FN	TN

- **TP(True Positive)**: Predijo verdadero y es cierto.
- **FP(False Positive)**: Predijo verdadero y es falso.
- **FN(False Negative)**: Predijo falso y es verdadero.
- **TN(True Negative)**: Predijo falso y es falso.

Además, podemos obtener las siguientes métricas:

- **Accuracy**: Es el porcentaje total de elementos clasificados correctamente.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall**: Es el número de elementos identificados correctamente como positivos del total de positivos verdaderos.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: Es el número de elementos identificados correctamente como positivos del total de positivos verdaderos.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Teniendo en cuenta lo anterior y que el objetivo es enviar la publicidad a potenciales clientes y se prefiere enviar un correo a una persona que no resulte comprador, y no perder un potencial cliente porque no se le mandó la publicidad, se busca que:

- 1) El valor de TP(True Positive) sea lo más alto posible ya que representa la cantidad de clientes compradores que fueron predichos como compradores.
- 2) El valor de FN(False Negative) sea lo más bajo posible ya que representa la cantidad de potenciales clientes perdidos (Clientes que habían comprado pero se predijeron como **no** compradores).
- 3) El valor de FP(False Positive) y TN(True Negative) no es de importancia porque representaría enviar un correo a un cliente no comprador.

Por lo tanto, a partir de aquí, para cada modelo analizaremos y daremos mayor importancia al modelo que dé un Recall más elevado.

Árboles de Decisión

Es un algoritmo de aprendizaje supervisado (Predictivo), que se utiliza para tareas de clasificación y de regresión. Tiene una estructura de árbol jerárquico, que consta de un nodo raíz, que se extiende por sus ramas, y estas pueden terminar en nodos hoja o en un nuevo nodo de decisión.

Es un conjunto de reglas jerarquizadas, de tal modo que cada camino me lleva a una decisión. Por esto, todo árbol puede ser escrito como un conjunto de reglas.

El aprendizaje del árbol emplea una estrategia de división de variables predictoras de la matriz X, con estas divisiones se utilizan medidas como la entropía que busca entender qué tan homogéneas o heterogéneas son estas divisiones, para así poder identificar los puntos de división óptimos dentro de un árbol. Este proceso de división se repite de forma recursiva de arriba hacia abajo hasta que todos o la mayoría de los registros se hayan clasificado bajo etiquetas de clase específicas. Que todos los puntos de datos se clasifiquen o no como conjuntos homogéneos depende en gran medida de la complejidad del árbol de decisión.

Los árboles más pequeños son más fáciles de obtener nodos hoja puros, es decir, puntos de datos en una sola clase. Sin embargo, a medida que crece en tamaño, se vuelve cada vez más difícil mantener esta pureza.

Toda rama finaliza en una hoja que puede ser pura o impura.

No permite más de una clase/etiqueta para una instancia, es decir, una instancia solo puede ser de una clase.

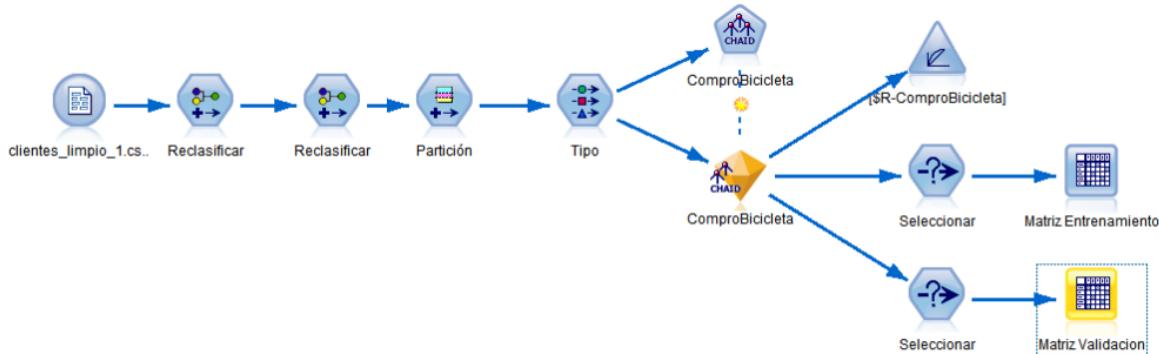
Funcionamiento

El algoritmo va construyendo el árbol añadiendo particiones y los hijos resultantes de cada partición. En cada partición, los ejemplos se van dividiendo entre los hijos. Finalmente, se llega a la situación en la que todos los ejemplos que caen en los nodos inferiores son de la misma clase y esa rama ya no sigue creciendo.

La única condición que hay que exigir es que las particiones al menos separen ejemplos en distintos hijos, con lo que la cardinalidad de los nodos irá disminuyendo a medida que se desciende en el árbol.

Árbol CHAID

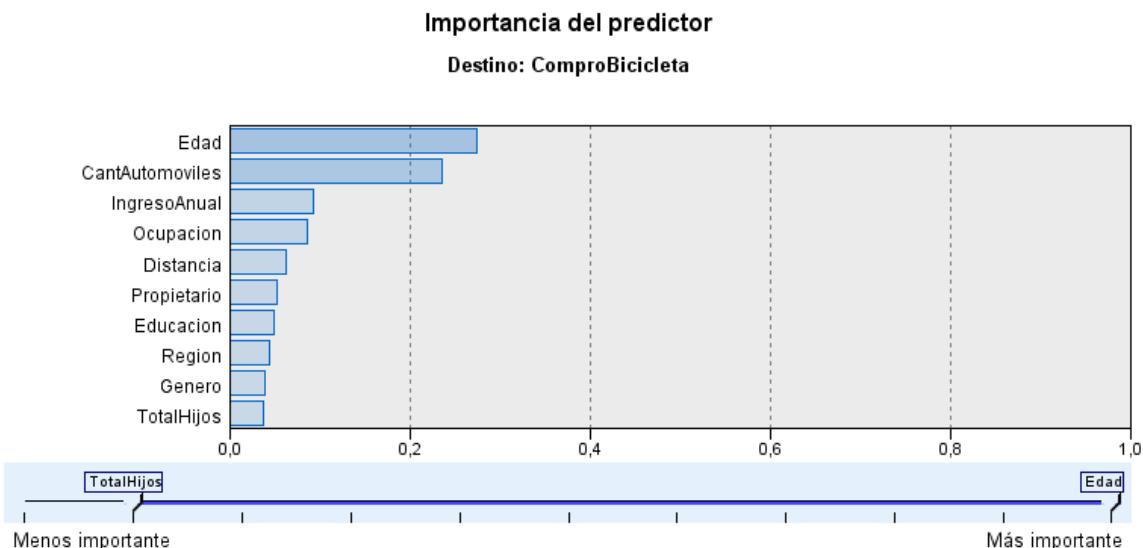
Utilizando SPSS Modeler se generó el siguiente modelo:



Utilizamos los nodos “Reclasificar” para corregir los términos “EducaciÃ³n secundaria (en curso)” y “EducaciÃ³n secundaria” de la variable Educación y el término “GestiÃ³n” de la variable Ocupación.

Además, usamos “Partición” para especificar los porcentajes 70% para Entrenamiento y 30% para Validación.

Importancia del Predictor



Como puede observarse en este árbol, el nodo raíz es la variable predictora “Edad”, con una ganancia cercana al 0,3.

Matriz de confusión entrenamiento

Accuracy = 71.3%

Recall = 54.2%

		Predicho
ComproBicicleta		

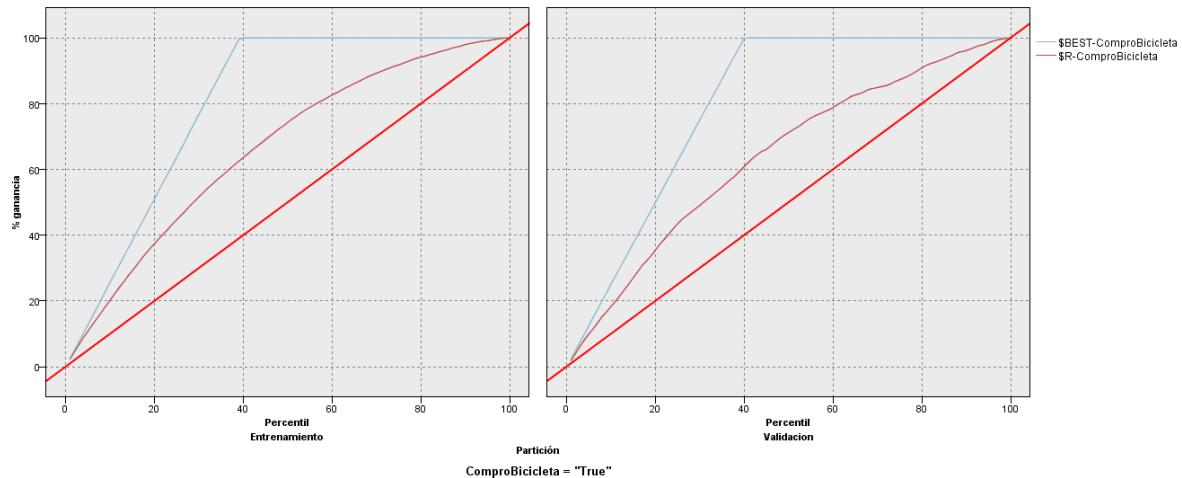
		False	True
False	Recuento	2233	478
	% de Filas	82.368	17.632
True	Recuento	800	947
	% de Filas	45.793	54.207

Matriz de confusión validación

Accuracy= 68.9%

Recall = 51.6%

ComproBicicleta		Predicho	
		False	True
False	Recuento	938	227
	% de Filas	80.515	19.485
True	Recuento	376	401
	% de Filas	48.391	51.609



Para determinar que el modelo es bueno, los valores de la diagonal de la matriz de confusión deben ser similares y a su vez, la de entrenamiento y validación entre sí.

En este caso, comparando las diagonales de entrenamiento y validación son similares, por lo que el modelo aprende bien de los datos y funciona con datos nuevos. Sin embargo, en los porcentajes de las diagonales se observa que el resultado obtenido no es óptimo debido a que hay una gran diferencia, casi el 30 %, entre el porcentaje obtenido por el valor False de la variable comprada con el valor

True.

Conclusión:

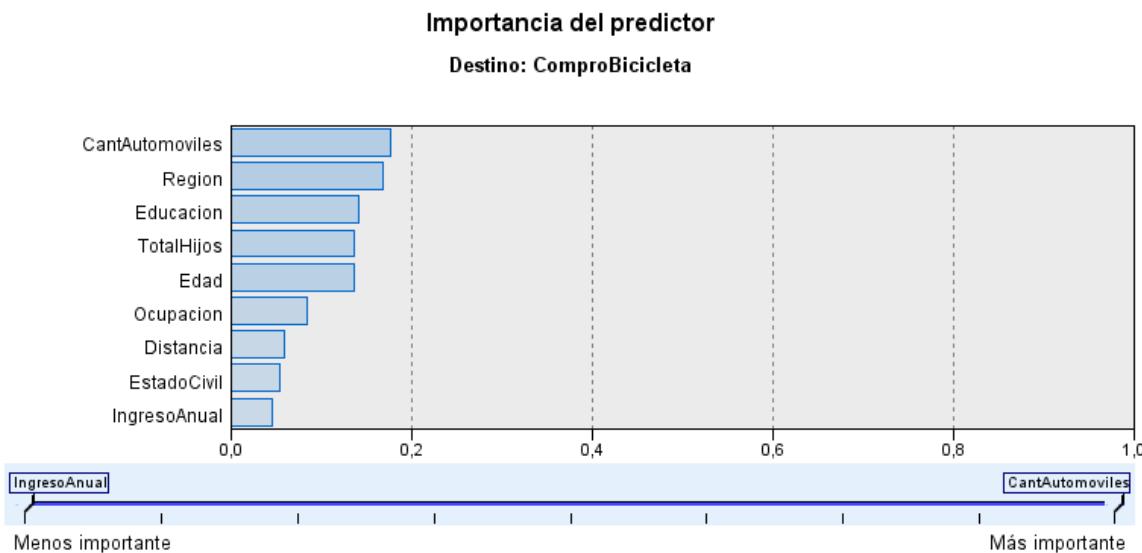
Análisis de la matriz de confusión de la partición de validación:

- El 80,515% de los clientes que se predijeron que no comprarían bicicletas, realmente no compraron.
- El 51,60% de los clientes que se predijeron que compraron, realmente compraron bicicletas.
- El 19.485% de los clientes que no compraron bicicletas, en realidad fueron predichos como que compraron bicicletas
- El 48.391% de los clientes que compraron bicicletas fueron predichos incorrectamente como no compradores de bicicletas

Este modelo no es óptimo porque no provee buenos resultados en base a los objetivos de estudio, ya que tiene una baja tasa de acierto general (68,9%), y sólo el 51,60% de los clientes que compraron bicicletas se les enviaría el mail.

Árbol QUEST

Importancia del Predictor



Como puede observarse en este árbol, el nodo raíz es la variable predictora “Cantidad de Automóviles” con una ganancia de 0,19. Sin embargo, la variable “Región” posee una ganancia similar de 0,18.

Matriz de confusión de entrenamiento

Accuracy= 66.57%

Recall= 43.8%

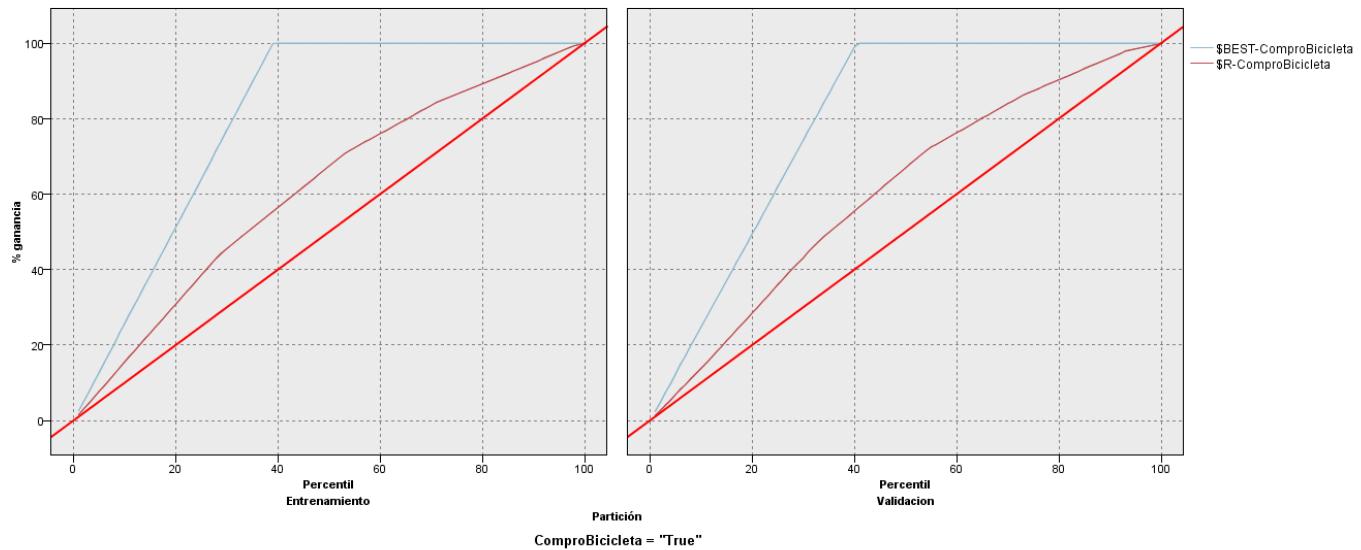
ComproBicicleta		Predicho	
		False	True
False	Recuento	2206	512
	% de Filas	81.163	18.837
True	Recuento	978	762
	% de Filas	56.207	43.793

Matriz de confusión de validación**Accuracy= 64.26%****Recall= 42.21%**

ComproBicicleta		Predicho	
		False	True
False	Recuento	917	241
	% de Filas	79.188	20.812
True	Recuento	453	331
	% de Filas	57.781	42.219

Para determinar que el modelo es bueno, los valores de la diagonal de la matriz de confusión deben ser similares y a su vez, la de entrenamiento y validación entre sí.

En este caso, comparando las diagonales de entrenamiento y validación son similares, por lo que el modelo aprende bien de los datos y funciona con datos nuevos. Sin embargo, en los porcentajes de las diagonales se observa que el resultado obtenido no es óptimo debido a que hay una gran diferencia de casi el 40 %, entre el porcentaje obtenido por el valor False de la variable comprado con el valor True.



Conclusión:

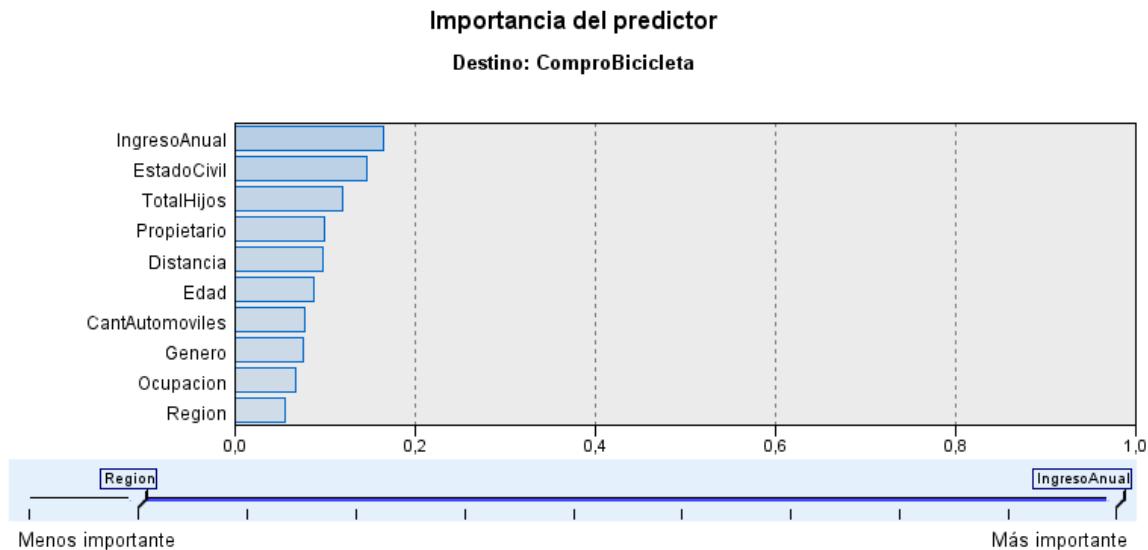
Análisis de la matriz de confusión de la partición de validación:

- El 79.188% de los clientes que se predijeron que no comprarían bicicletas, realmente no compraron.
- El 42.219% de los clientes que se predijeron que compraron, realmente compraron bicicletas.
- El 20.812% de los clientes que no compraron bicicletas, en realidad fueron predichos como que compraron bicicletas
- El 57.781% de los clientes que compraron bicicletas fueron predichos incorrectamente como no compradores de bicicletas

Este modelo no provee buenos resultados en base a los objetivos de estudio, ya que tiene una baja tasa de acierto general (64,26%), y sólo el 42.219% de los clientes que compraron bicicletas se les enviaría el mail.

Árbol C5.0 Sin Poda

Importancia del Predictor



Como puede observarse en este árbol, el nodo raíz es la variable predictora “Ingreso Anual” con una ganancia de 0,18.

Matriz de confusión de entrenamiento

Accuracy= 85.08%

Recall = 78,56 %

ComproBicicleta		Predicho	
		False	True
False	Recuento	2426	292
	% de Filas	89.257	10.743
True	Recuento	373	1367
	% de Filas	21.437	78.563

Matriz de confusión de validación

Accuracy= 77.39%

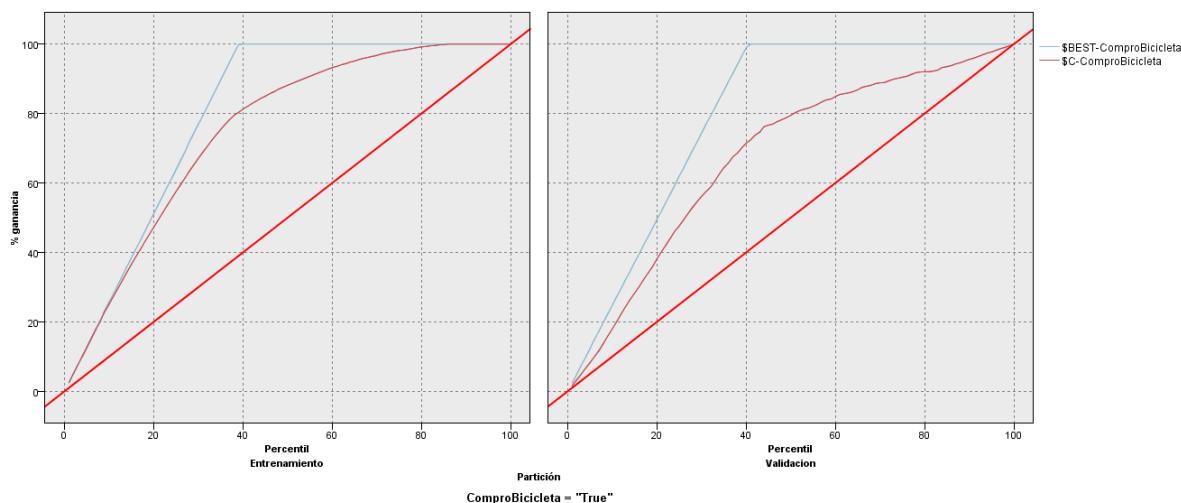
Recall= 71,30%

		Predicho	
ComproBicicleta		False	True

False	Recuento	944	214
	% de Filas	81.520	18.480
True	Recuento	225	559
	% de Filas	28.699	71.301

Para determinar que el modelo es bueno, los valores de la diagonal de la matriz de confusión deben ser similares y a su vez, la de entrenamiento y validación entre sí.

En este caso, comparando las diagonales de entrenamiento y validación son similares, por lo que el modelo aprende bien de los datos y funciona con datos nuevos. En los porcentajes de las diagonales se observa que el resultado obtenido es óptimo, pues solo hay una diferencia del 10% entre el porcentaje obtenido por el valor False de la variable comparada con el valor True.



Conclusión:

Análisis de la matriz de confusión de la partición de validación:

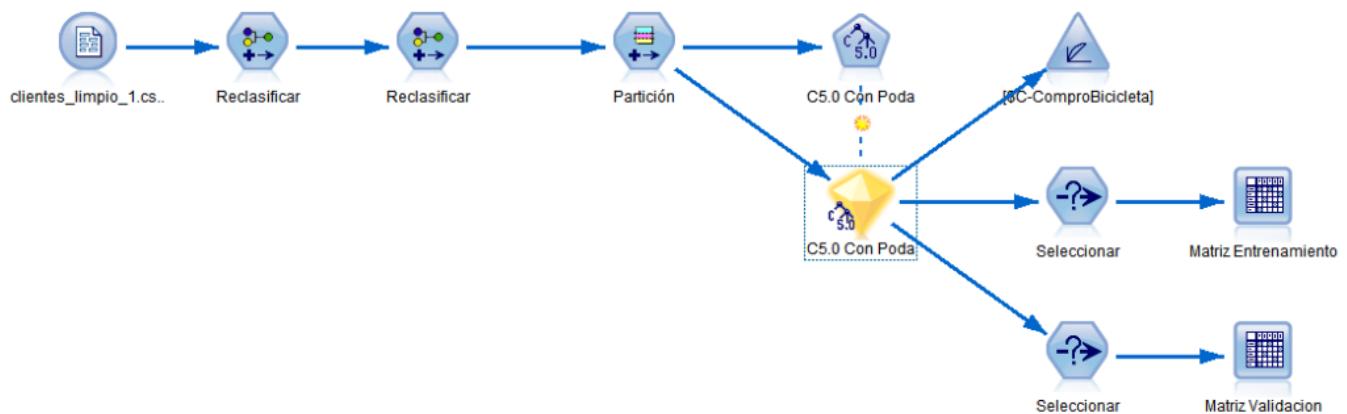
- El 81.520% de los clientes que se predijeron que no comprarian bicicletas, realmente no compraron.
- El 71.301% de los clientes que se predijeron que compraron, realmente compraron bicicletas.
- El 18.480% de los clientes que no compraron bicicletas, en realidad fueron predichos como que compraron bicicletas
- El 28.699% de los clientes que compraron bicicletas fueron predichos incorrectamente como no compradores de bicicletas

Este modelo no provee resultados tan malos comparados a los modelos anteriores en base a los objetivos de estudio, ya que tiene una tasa de acierto general (77,39%), y el 71.301% de los clientes que compraron bicicletas se les enviaría el mail.

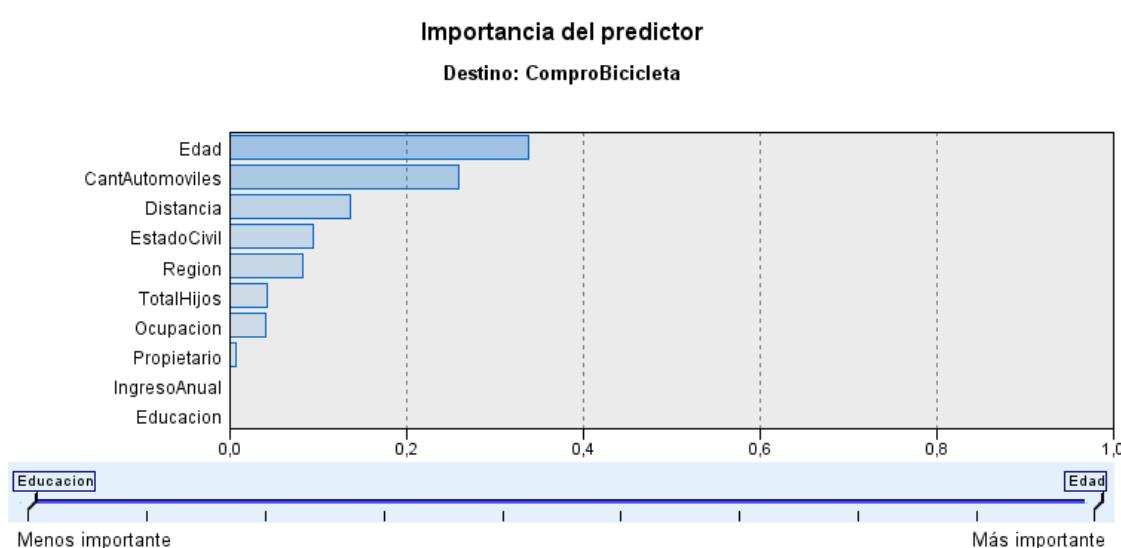
Árbol C5.0 con Poda

Construimos el árbol utilizando los siguientes parámetros:

- Gravedad de la poda: 75
- Número mínimo de registros por rama hija: 5
- Poda global



Importancia del Predictor



Como puede observarse en este árbol, el nodo raíz es la variable predictora “Edad” con una ganancia de 0,32.

Matriz de confusión de entrenamiento

Accuracy= 80.66%

Recall = 66.667%

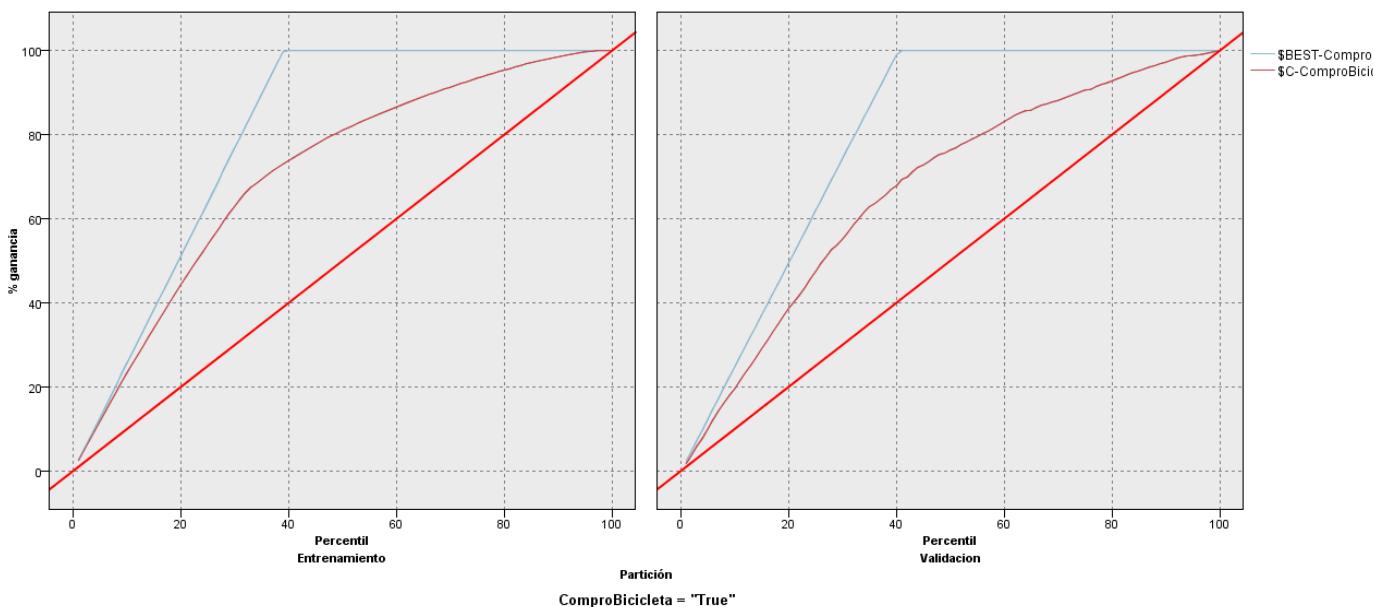
ComproBicicleta		Predicho	
		False	True
False	Recuento	2436	282
	% de Filas	89.625	10.375
True	Recuento	580	1160
	% de Filas	33.333	66.667

Matriz de confusión de validación

Accuracy= 75.43%

Recall = 61.862%

ComproBicicleta		Predicho	
		False	True
False	Recuento	980	178
	% de Filas	84.629	15.371
True	Recuento	299	485
	% de Filas	38.138	61.862



Conclusión:

Análisis de la matriz de confusión de la partición de validación:

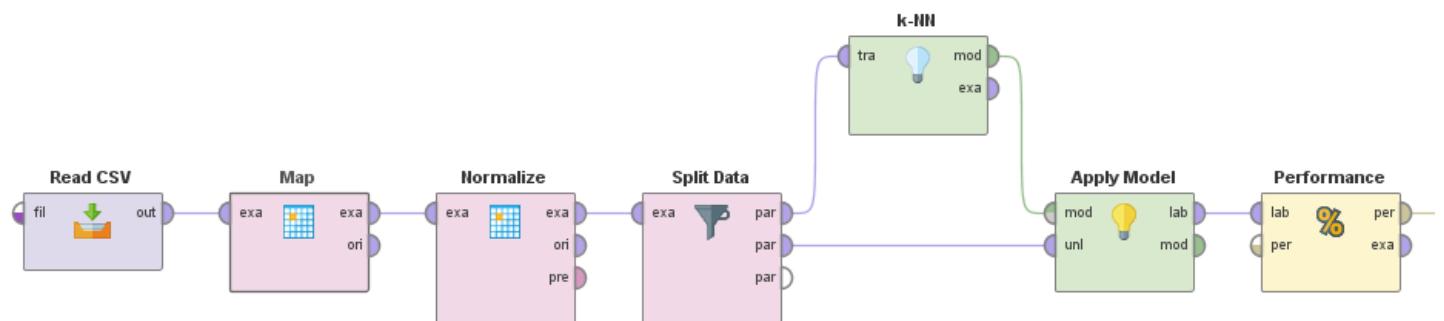
- El 84.629% de los clientes que se predijeron que no comprarían bicicletas, realmente no compraron.
- El 61.862% de los clientes que se predijeron que compraron, realmente compraron bicicletas.

- El 15.371% de los clientes que no compraron bicicletas, en realidad fueron predichos como que compraron bicicletas
- El 38.138% de los clientes que compraron bicicletas fueron predichos incorrectamente como no compradores de bicicletas

Este modelo provee mejores resultados que los árboles CHAID y QUEST basándonos en los objetivos de estudio, pero peor que el árbol Sin Poda, ya que tiene una tasa de acierto general del 75.43%, y a un 61.862% de los clientes que compraron bicicletas se les enviaría el mail.

KNN

Con RapidMiner se construyó el siguiente modelo:



Primero con el nodo Map, renombramos valores de las columnas ocupación y educación que aparecían con errores por el tilde. Luego, normalizamos las variables y dividimos el dataset en “Entrenamiento” y “Validación”.

Utilizando como medida la distancia euclídea mixta para los distintos K se obtuvieron las siguientes matrices de confusión:

K=1

Accuracy = 72.40%

Recall = 64,33%

	real True	real False	class Precision
pred. True	487	260	65.19%
pred. False	270	903	76.98%
class recall	64.33%	77.64%	

K=2

Accuracy = 71.15%

Recall = 68,56%

	real True	real False	class Precision

pred. True	519	316	62.16%
pred. False	238	847	78.06%
class recall	68.56%	72.83%	

K=3

Accuracy = 73.80%

Recall = 66,05%

	real True	real False	class Precision
pred. True	500	246	67.02%
pred. False	257	917	78.11%
class recall	66.05%	78.85%	

K=4

Accuracy = 74.58%

Recall = 62,88%

	real True	real False	class Precision
pred. True	476	207	69.69%
pred. False	281	956	77.28%
class recall	62.88%	82.20%	

K=5

Accuracy = 73.75%

Recall = 60,24%

	real True	real False	class Precision
pred. True	456	203	69.20%
pred. False	301	960	76.13%
class recall	60.24%	82.55%	

K=6

Accuracy = 75.10%

Recall = 61.29%

	real True	real False	class Precision
pred. True	464	185	71.49%
pred. False	293	978	76.95%
class recall	61.29%	84.09%	

K=7

Accuracy = 73.02%

Recall = 56.41%

	real True	real False	class Precision
pred. True	427	188	69.43%
pred. False	330	975	74.71%
class recall	56.41%	83.83%	

K=8

Accuracy = 75.47%

Recall = 60.37%

	real True	real False	class Precision
pred. True	457	171	72.77%
pred. False	300	992	76.78%
class recall	60.37%	85.30%	

K=9

Accuracy = 73.85%

Recall = 57.99%

	true True	true False	class Precision
pred. True	439	184	70.47%
pred. False	318	979	75.48%
class recall	57.99%	84.18%	

K= 10

Accuracy = 75.42%

Recall = 59.7%

	true True	true False	class Precision
pred. True	452	167	73.02%
pred. False	305	996	76.56%
class recall	59.7%	85.64%	

Teniendo en cuenta lo anterior el valor de K que da el mejor rendimiento del modelo es K =2 con un accuracy de 71.15% y un recall de 68.56%.

Análisis discriminante

El objetivo del análisis discriminante es encontrar la combinación lineal de las variables independientes que mejor permite diferenciar (discriminar) a los grupos. Una vez encontrada, esa combinación podrá ser utilizada para clasificar nuevos casos.

Se trata de una técnica de análisis multivariante que es capaz de aprovechar las relaciones existentes entre una gran cantidad de variables independientes para maximizar la capacidad de discriminación.

Las observaciones para poder utilizar este modelo tienen que cumplir los siguientes supuestos:

- No multicolinealidad.
- Distribución normal de las variables.
- Matrices de varianza covarianza iguales.

Análisis de Supuestos

Analizaremos los supuestos del análisis discriminante, para determinar si el modelo es válido o no. Recordamos que solo se trabaja con variables cuantitativas y continuas, por lo que de acuerdo a las variables de nuestro dataset, las únicas que cumplen esta condición son el Ingreso Anual y la Edad

1) Distribución normal de las variables

Las poblaciones deben tener una distribución normal. Para analizar esto, planteamos las siguientes hipótesis:

H_0 : Las variables tienen distribución normal.

H_1 : Las variables tienen distribución distinta a la normal.

Luego de realizar el análisis en el SPSS Statistics se obtuvo el siguiente cuadro de prueba de normalidad:

Pruebas de normalidad

Kolmogorov-Smirnov ^a			
	Estadístico	gl	Sig.
IngresoAnual	,129	6400	,000
Edad	,079	6400	,000

a. Corrección de significación de Lilliefors

Para que se cumpla la hipótesis nula H_0 , **p-value** (*Columna Sig.*) cuando lo comparamos con **$\alpha=0.05$** (alfa), debe ser **p-value > α** . para que se cumpla la hipótesis nula.

Al ser **p-value** cero en ambas variables no se cumple la hipótesis nula, es decir, las variables no tienen distribución normal. Por lo tanto, no se cumple el supuesto.

2) No multicolinealidad

Supone que las variables explicativas no están correlacionadas, es decir, que son independientes.

Diagnósticos de colinealidad^a

Modelo	Dimensión	Autovalor	Índice de condición	Proporciones de varianza		
				(Constante)	IngresoAnual	Edad
1	1	2,811	1,000	,01	,03	,01
	2	,165	4,128	,04	,97	,05
	3	,024	10,750	,95	,00	,95

a. Variable dependiente: ID

Observamos el índice de condición (IC), mirando el valor de la última fila. Si el IC es mayor que 30 hay multicolinealidad, si está entre 20 y 30 es moderada y en caso de ser menor que 20 no hay multicolinealidad. Como el IC = 10,750 el supuesto se cumple, y no hay multicolinealidad.

3) Matrices de Covarianza Iguales

Utilizando la M de Box, para pasar el supuesto el *p value* tiene que ser mayor al α .

Resultados de prueba

M de Box	130,797
F	Aprox. 43,583
gl1	3
gl2	1461698477
Sig.	,000

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

Apreciamos que el valor del *p value* es 0, por lo que no es mayor a α , por lo tanto no se cumple el supuesto.

Función Discriminante

Más allá de que no todos los supuestos son válidos, continuaremos el estudio de este modelo.

Coeficientes de la Función

Para determinar la cantidad de funciones discriminantes necesarias debemos realizar el siguiente cálculo:

$$r = \min(G - 1, p)$$

Donde:

- G: Cantidad de poblaciones. En este caso tenemos 2 poblaciones.
- p: Cantidad de variables independientes. En este caso tenemos 2 variables independientes.

$$r = \min(2 - 1, 2) = 1$$

Obtenemos los coeficientes de la función discriminante:

**Coeficientes de la
función
discriminante
canónica**

	Función
	1
IngresoAnual	,000
Edad	,080
(Constante)	-3,043

Coefficientes no estandarizados

La función discriminante será:

$$D = -3.043 + 0.080 * Edad$$

Centroides

En la siguiente imagen podemos identificar los centroides:

**Funciones en centroides
de grupo**

	Función
	1
ComproBicicleta_int	1
,00	,101
1,00	-,154

Las funciones discriminantes canónicas sin estandarizar se han evaluado en medias de grupos

Matriz de Estructura

En esta matriz lo que observamos es el cálculo de la correlación entre cada una de las variables y la función discriminante. Este valor sirve para evaluar qué variables tienen mayor importancia, mientras mayor sea la correlación más influencia tiene esa variable a la hora de hacer la discriminación.

Matriz de estructuras

	Función
	1
Edad	,820
IngresoAnual	-,434

En este caso la variable con mayor importancia es Edad.

¿Existe Modelo?

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,985	98,556	2	,000

Aquí podemos observar que debido a que el p-value = 0,00 es menor al valor de alpha = 0,05 no hay modelo. Sin embargo, como mencionamos anteriormente, este modelo no va a ser bueno, ya que no se cumple el supuesto de Normalidad y el supuesto de Matrices de Covarianzas Varianzas Iguales.

Análisis de Matriz de Confusión

En la siguiente figura podemos ver la matriz de confusión obtenida para el análisis discriminante, tanto para el conjunto de prueba como el de entrenamiento.

Resultados de clasificación^{a,b}

			Pertenencia a grupos pronosticada		Total
			ComproBicicleta_int	,00	
Casos seleccionados	Original	Recuento	,00	1269	1429
			1,00	736	1055
	%		,00	47,0	53,0
			1,00	41,1	58,9
Casos no seleccionados	Original	Recuento	,00	567	611
			1,00	275	458
	%		,00	48,1	51,9
			1,00	37,5	62,5
			100,0		

a. 51,8% de casos agrupados originales seleccionados clasificados correctamente.

b. 53,6% de casos agrupados originales sin seleccionar clasificados correctamente.

Recall entrenamiento = 58.90%

Recall validación= 62.48%

Analizando los valores tanto de los datos de prueba como los de entrenamiento notamos que existe cierta variación entre ellos. Por lo tanto, no creemos que el modelo sea muy bueno.

Fase de evaluación

Para la evaluación decidimos trabajar con una matriz de costos, donde se evalúa el costo entre los valores predichos y los reales del modelo, penalizando aquellos errores del mismo.

Como nuestro caso considera más costoso no enviar mail a posibles compradores, es decir, que haya comprado bicicleta y el modelo me prediga que no lo hizo, le pondremos para ese caso un costo de 3. A su vez, como no es tan importante si se envía mail a personas que no compraron, le ponemos un costo de 1.

Esto último se establece de esta manera ya que es preferible enviarle innecesariamente un correo a una persona que no resulte comprador, y no perder un potencial cliente porque no se le envió el correo con la publicidad.

		PREDICHO	
		0	1
TRUE	0	0	1
	1	3	0

Luego, para cada modelo realizado de:

- Árboles
- KNN
- Análisis Discriminante

donde se ha obtenido las matrices de confusión correspondientes en cada caso, se multiplicarán las mismas por la matriz de costos mencionada, llegando a un valor al cual se lo denominará **Costo Total** de ese modelo. El cálculo del Costo Total será: $3*FN + 1*FP$ ($FN =$ Falsos Negativos, $FP =$ Falsos Positivos).

Esto nos permitirá obtener el modelo donde el costo total sea menor y a su vez tenga una alta tasa de aciertos o al menos que la misma sea aceptable.

Análisis de Costos para los Modelos de Árboles

Árbol	Tasa de Aciertos	Recall	Costo Total
CHAID	68.9%	51.6%	1355
QUEST	64.26%	42.21%	1600
C5 sin Poda	77.39%	71.30%	889
C5 con Poda	75.43%	66,667%	1048
KNN 1	72.40%	64.33%	1070
KNN 2	71.15%	68.56%	1030
KNN 3	73.80%	66.05%	1017
KNN 4	74.58%	62.88%	1050
KNN 5	73.75%	60.24%	1106
KNN 6	75.10%	61.29%	1064
KNN 7	73.02%	56.41%	1178
KNN 8	75.47%	60.37%	1071
KNN 9	73.85%	57.99%	1138
KNN 10	75.42%	59.7%	1082
Análisis discriminante	53.6%	62.48%	1436

A partir del Análisis de Costos decidimos seleccionar como modelo el Árbol C5 sin Poda, ya que tiene el mayor Recall y Accuracy logrando como consecuencia el menor costo.

Fase de Implementación

En esta fase aplicamos el archivo de destinatarios al modelo del árbol C5.0 escogido. La salida correspondiente a dicho modelo se muestra resumida en la siguiente imagen:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	IdCiudad	Nombre	Apellido	EstadoCivil	Genero	Email	IngresoAnui	TotalHijos	Educacion	Ocupacion	Propie	CantAu	Direccion	Telefono	Distancia	Region	Edad	Compra Bicic	Probabilidad	Enviar Mail
2	184	Ashlee	Xie	S	F	ashlee10@mineriadedatos.com.a	10000	1	Educación secu	Obrero	0	2	92 rue Ste-Honoré	500 555-0159	0-1 Km.	Centro	43	False	0.849	0
3	383	Marcus	Campbell	S	M	marcus44@mineriadedatos.com.	100000	3	Licenciatura	Gestión	1	3	8819 Camino Norte	911-555-0178	2-5 Km.	Norte	32	True	0.725	1
4	220	Monique	Vazquez	C	F	monique11@mineriadedatos.cor	20000	4	Educación secu	Obrero	1	2	2 rue Lafayette	500 555-0129	0-1 Km.	Centro	34	False	0.849	0
5	311	Adriana	Malhotra	C	F	adriana5@mineriadedatos.com.e	80000	3	Licenciatura	Obrero especia	1	3	1944 Serene Court	911-555-0122	10+ Km.	Norte	48	False	0.808	0
6	258	Johnny	Goel	C	M	johnny21@mineriadedatos.com..	30000	1	Estudios de po:	Administrativo	1	0	5869 Clayton Road	500 555-0155	0-1 Km.	Centro	75	False	0.727	0
7	10	Corey	Lal	C	M	corey8@mineriadedatos.com.ar	20000	4	Educación secu	Obrero especia	1	2	9161 Viking Drive	500 555-0194	5-10 Km.	Sur	65	False	0.870	0
8	203	Omar	Goel	C	M	omar40@mineriadedatos.com.ar	80000	4	Educación secu	Profesional	1	3	1 avenue des Cham	500 555-0134	10+ Km.	Centro	61	False	1.000	0
9	155	Christian	Long	S	M	christian24@mineriadedatos.con	20000	0	Educación secu	Obrero	1	2	Heidestieg Straße 84	500 555-0111	1-2 Km.	Centro	33	False	0.898	0
10	205	Ronnie	Wagner	S	M	ronnie1@mineriadedatos.com.ar	10000	3	Estudios univer	Obrero	1	1	49 rue Royale	500 555-0152	1-2 Km.	Centro	70	True	0.583	1
11	32	Daisy	Navarro	S	F	daisy6@mineriadedatos.com.ar	70000	2	Educación secu	Obrero especia	0	2	7427 Grove Way	500 555-0120	1-2 Km.	Sur	56	True	0.838	1
12	553	Hunter	Li	S	M	hunter21@mineriadedatos.com.i	60000	1	Licenciatura	Profesional	1	1	2612 Berry Dr	402-555-0113	2-5 Km.	Norte	54	True	0.780	1
13	177	Ebony	Martinez	S	F	ebony18@mineriadedatos.com.e	20000	2	Educación secu	Obrero	1	0	Heideweg 1442	500 555-0165	0-1 Km.	Centro	47	True	0.667	1
14	244	Derrick	Moreno	S	M	derrick6@mineriadedatos.com.a	30000	0	Educación secu	Obrero	0	1	8942 Sierra Road	500 555-0195	1-2 Km.	Centro	40	True	0.960	1

El modelo determinó que 638 de los 1500 destinatarios son posibles compradores y se les debería enviar mail.

	Enviar mail	No enviar mail
Cantidad	638	862
Porcentaje	42.53333333	57.46667

Análisis descriptivo

En cuanto al objetivo de seleccionar el tipo de bicicleta más adecuado para cada cliente, se llevarán a cabo diversos análisis de clasificación con el propósito de identificar grupos de clientes que comparten características similares. Este enfoque nos permitirá personalizar la oferta de bicicletas y brindar una experiencia más satisfactoria a los clientes.

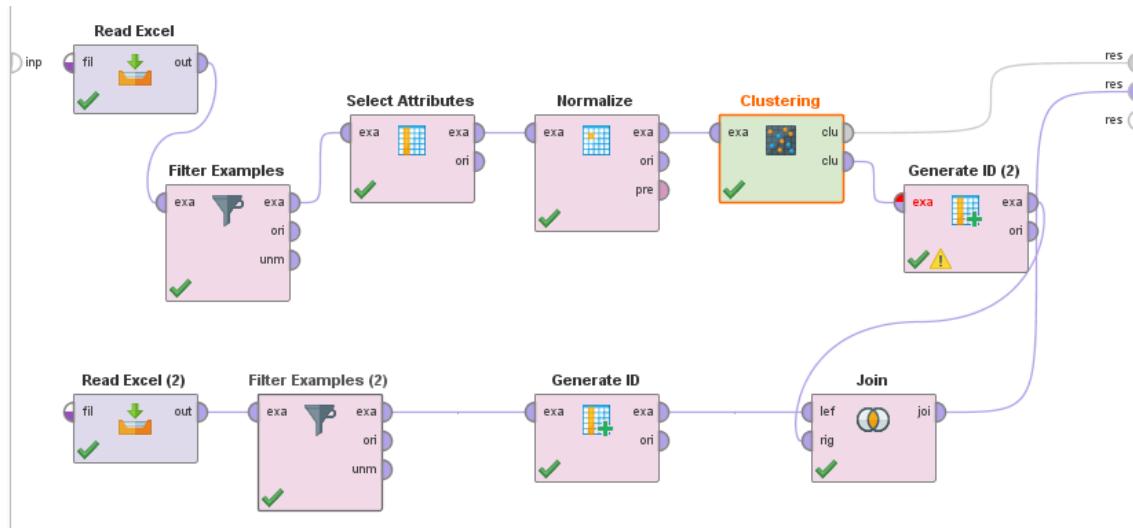
Para lograr este objetivo, utilizaremos tres técnicas específicas: K-Medias, Clustering Jerárquico y Clustering Bietápico. Estas técnicas nos ayudarán a agrupar a nuestros clientes en categorías distintas en función de sus preferencias, necesidades y características demográficas.

KMedias

K-Medias es un algoritmo descriptivo no supervisado. El parámetro K se lo define de forma manual, y si se quisiera, podrían encontrarse muchos grupos. Sin embargo, se debe encontrar la cantidad de grupos más óptima.

El objetivo es que los grupos sean diferentes entre sí, pero homogéneos dentro del mismo grupo.

Utilizando el Software RapidMiner construimos el siguiente modelo:



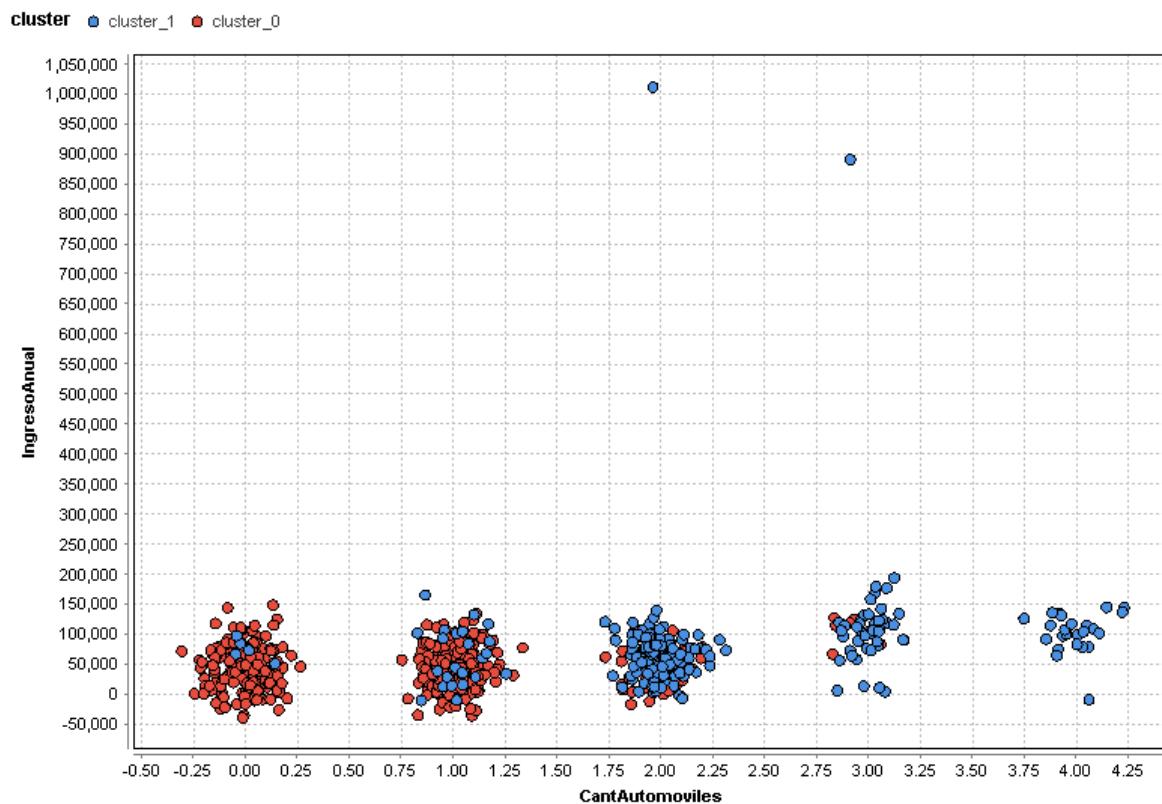
A partir de la población estandarizada se seleccionan los atributos numéricos para aplicar clustering variando la cantidad de K y luego uniendo los datos originales para analizar las variables categóricas de los clusters.

A continuación analizamos los resultados para los distintos valores de K.

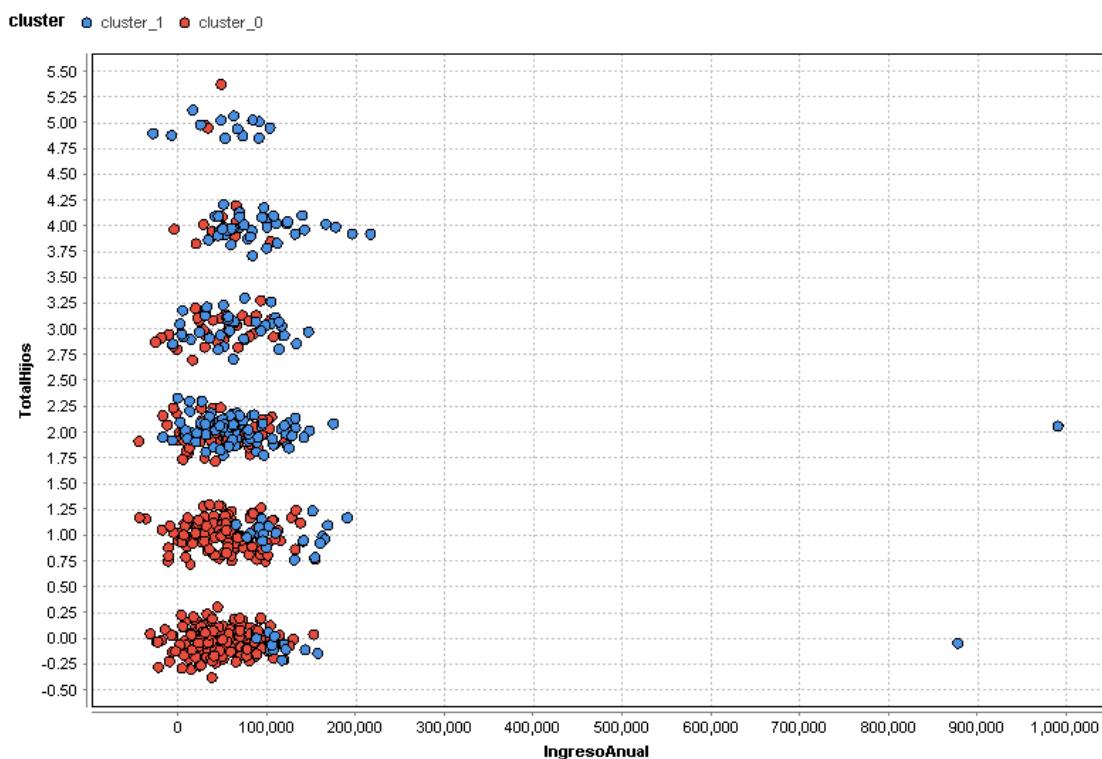
K = 2

El modelo agrupó 432 observaciones en el grupo 0 y 206 observaciones en el grupo 1 en una población de 638 observaciones.

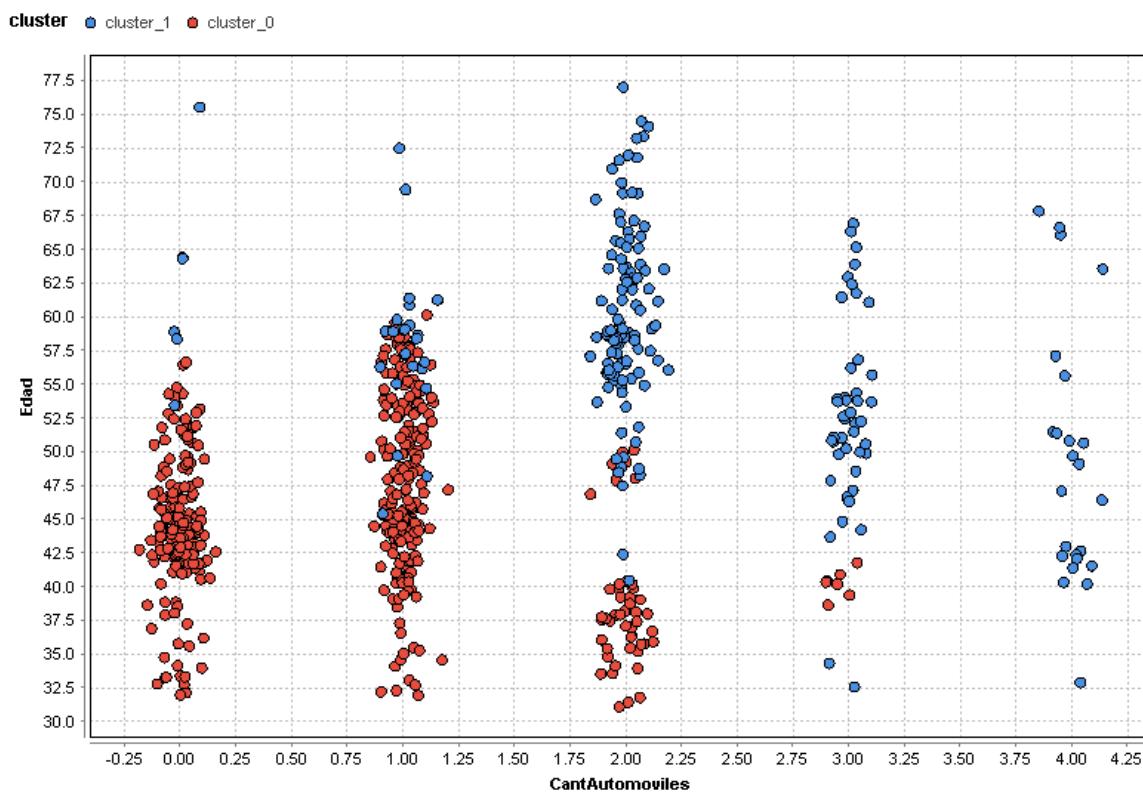
CantAutomoviles vs IngresoAnual



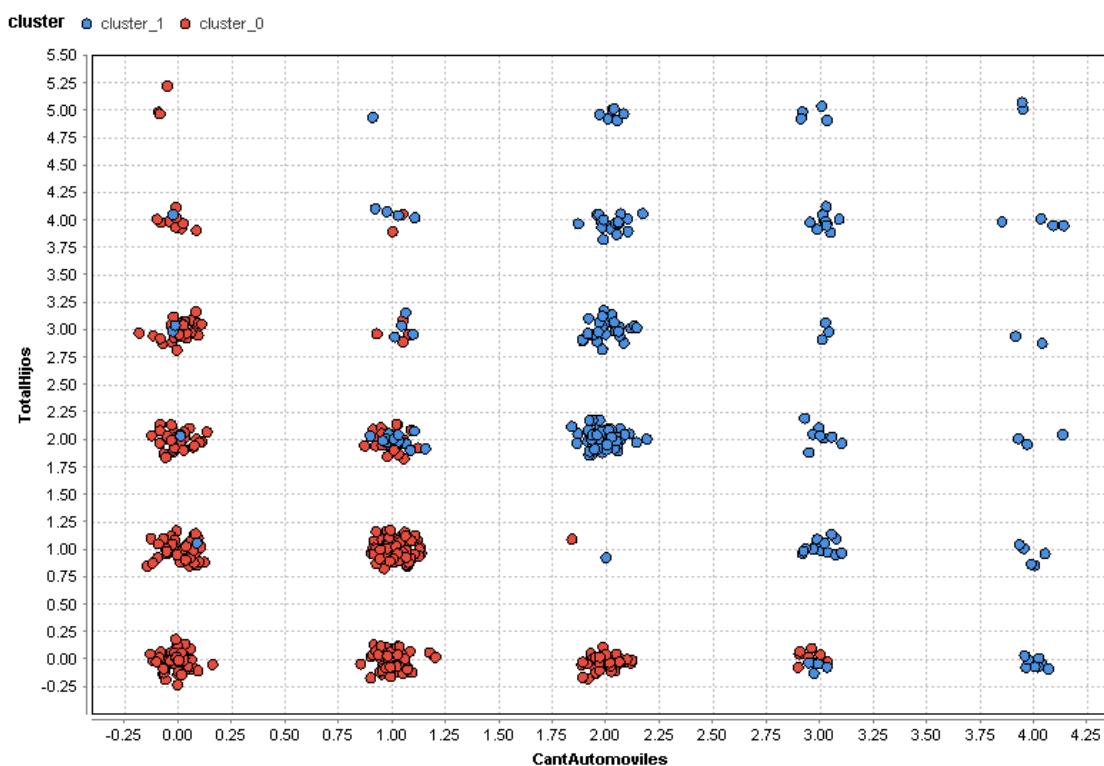
IngresoAnual vs TotalHijos



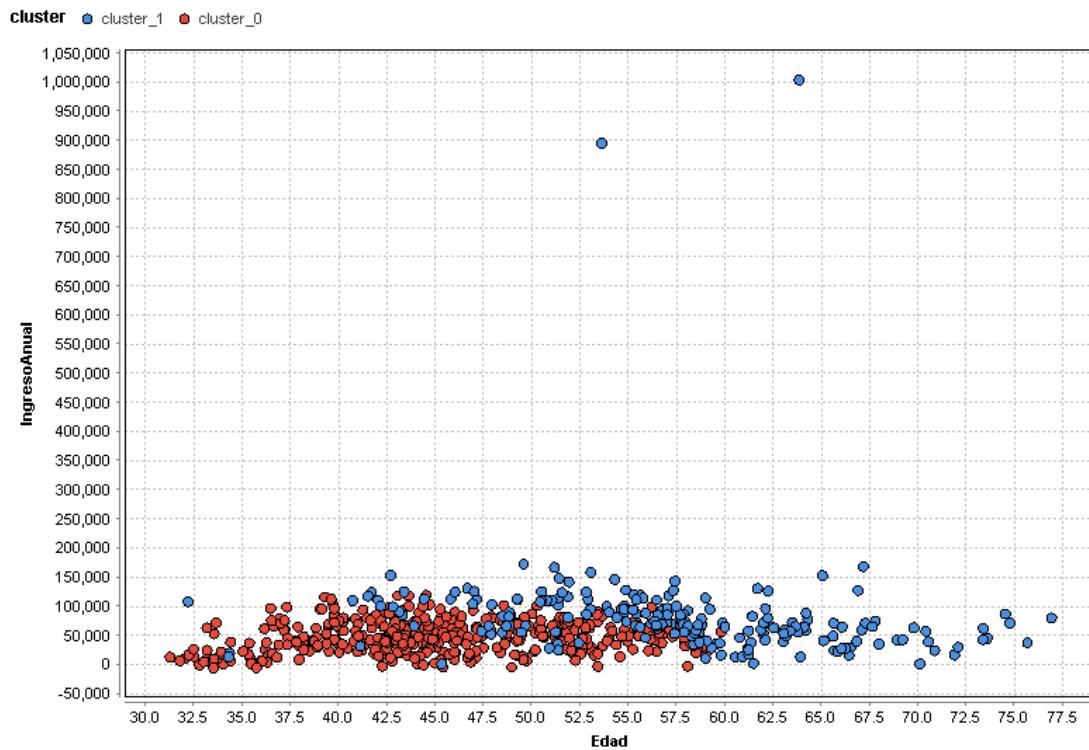
CantAutomoviles vs Edad



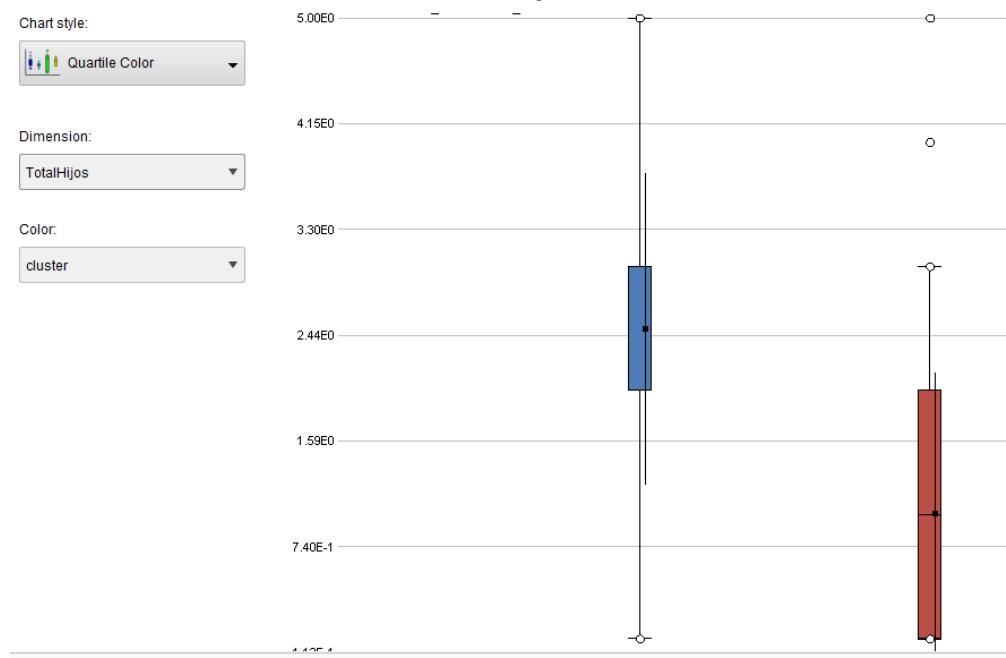
CantAutomoviles vs TotalHijos



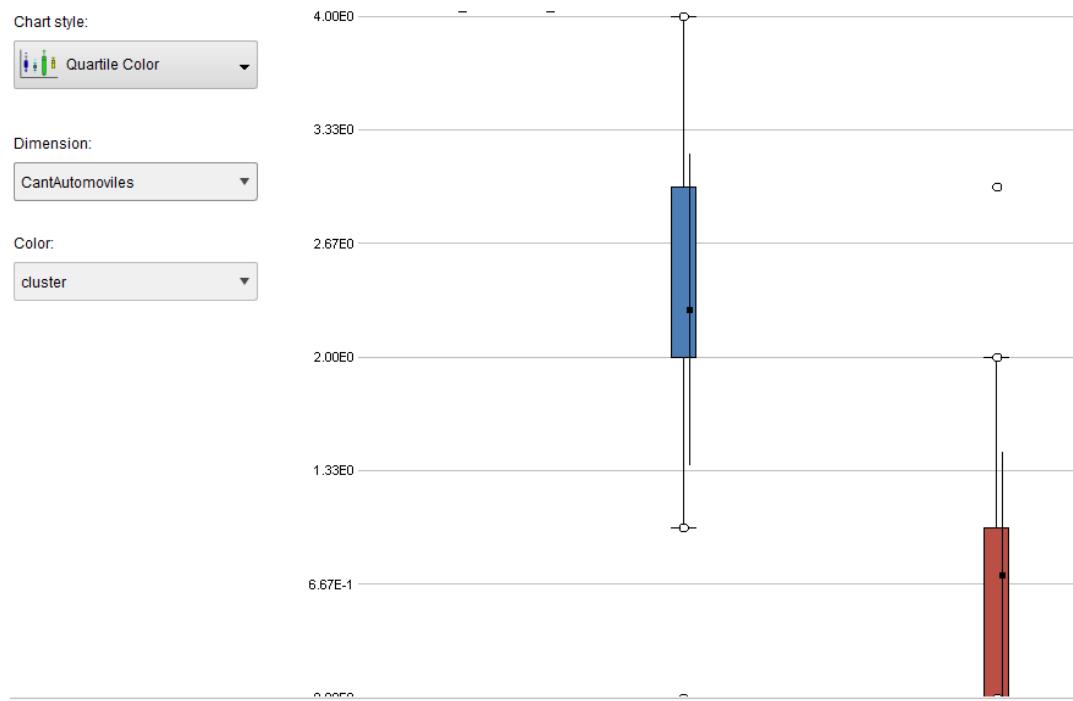
Edad vs IngresoAnual



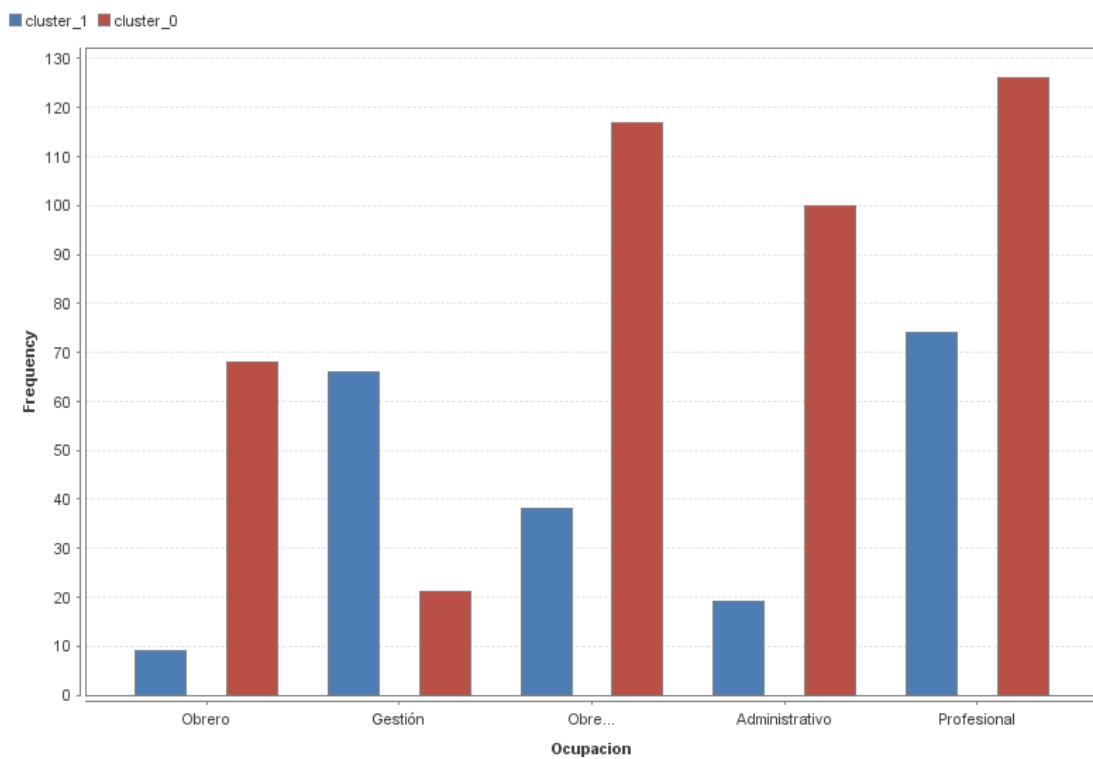
Total de Hijos en cada cluster



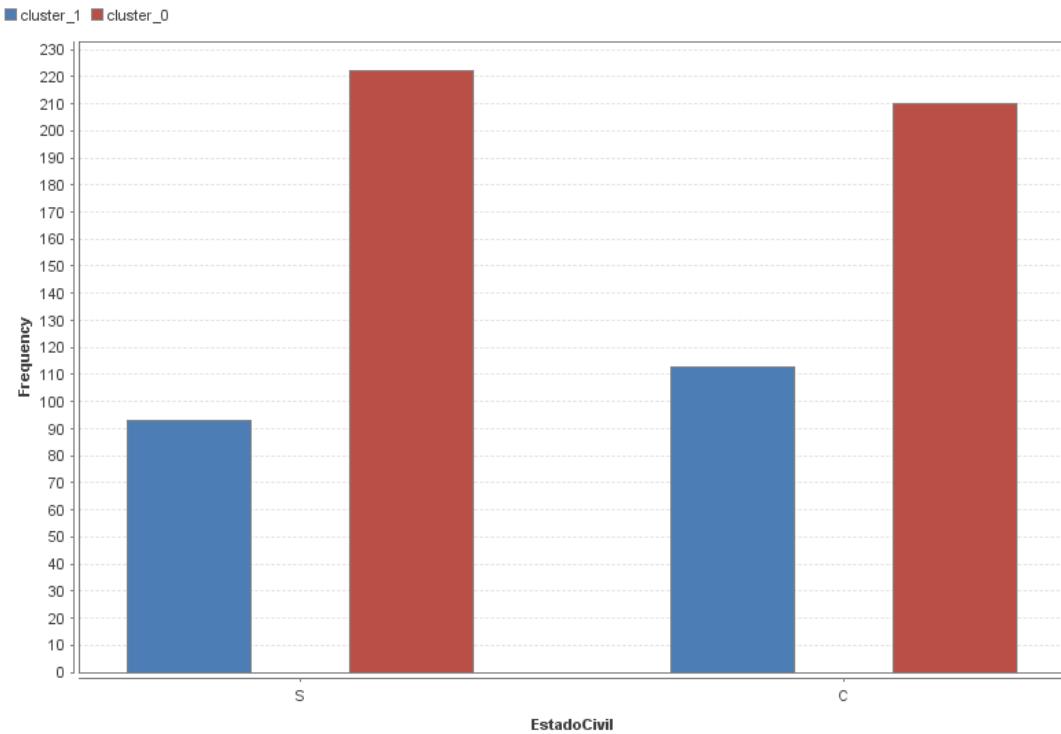
Cantidad de Automóviles en cada cluster



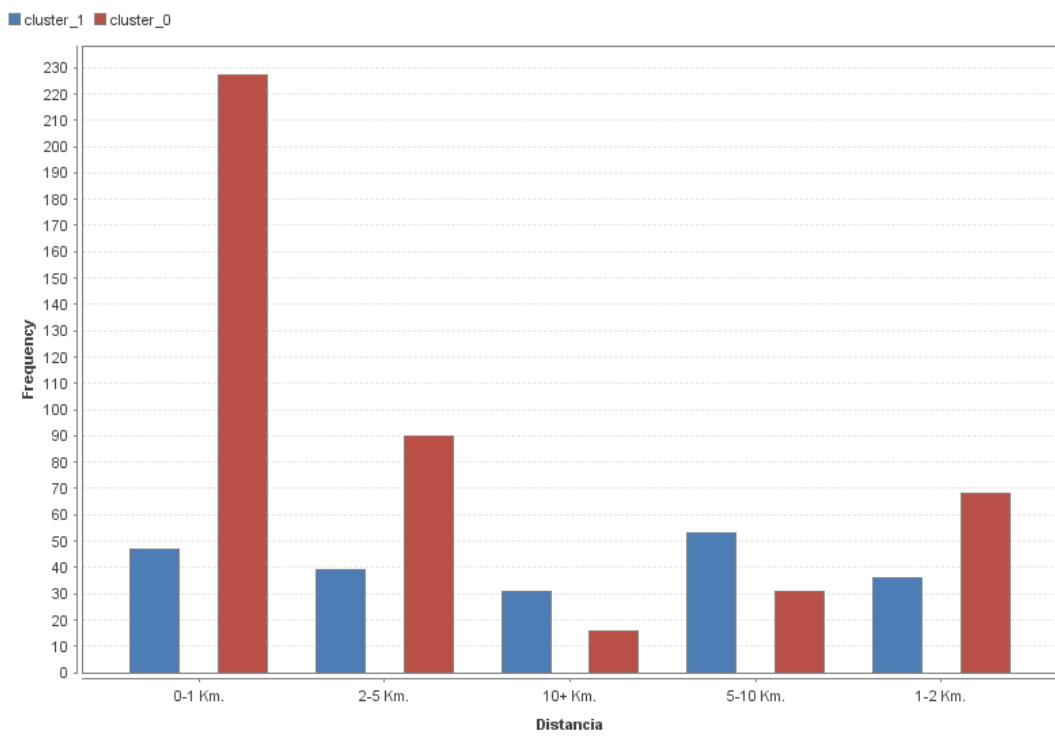
Ocupación más *frecuente* en cada cluster



Estado Civil más *frecuente* en cada cluster



Distancia al trabajo más *frecuente* en cada cluster



Caracterización de los clusters con K = 2

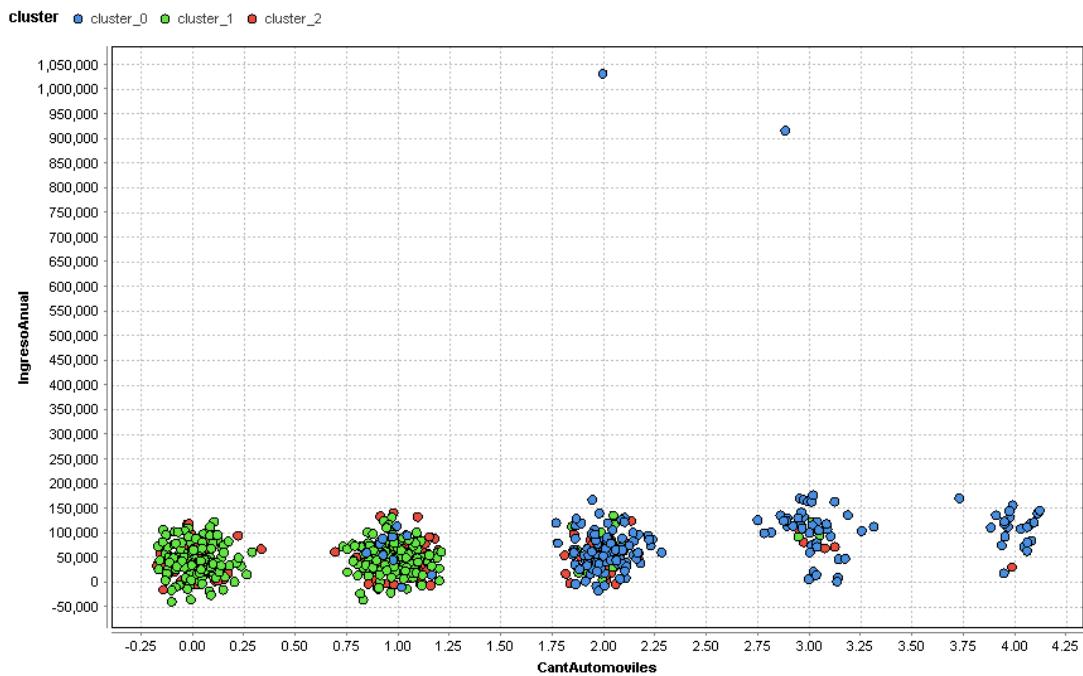
ATRIBUTO	CLUSTER 0	CLUSTER 1
Ingreso Anual	Menor Valor	Mayor Valor
Edad	Mayormente jóvenes y adultos	Mayormente adultos y ancianos
Cant Automóviles	Menor Cantidad	Mayor Cantidad
Ocupación	En mayor proporción existen ocupaciones de Profesionales, Obreros Especializados y Administrativos.	En mayor proporción existen ocupaciones de Gestión y Profesionales.
TotalHijos	Menor Cantidad	Mayor Cantidad
Estado Civil	-	-
Distancia	La mayoría viven cerca o a media distancia (0-1 km, 1-2 km, 2-5 km)	-

K = 3

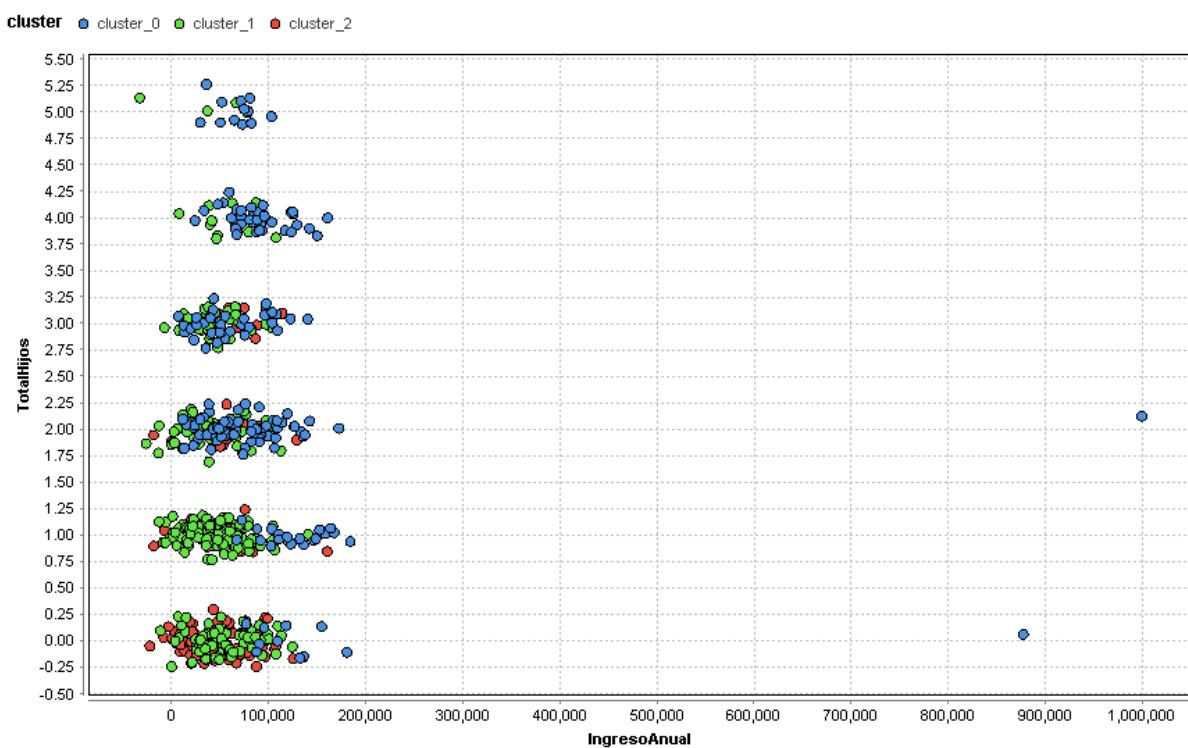
De un total de 638 observaciones el modelo agrupó:

- 185 observaciones en el cluster 0.
- 299 observaciones en el cluster 1.
- 154 observaciones en el cluster 2.

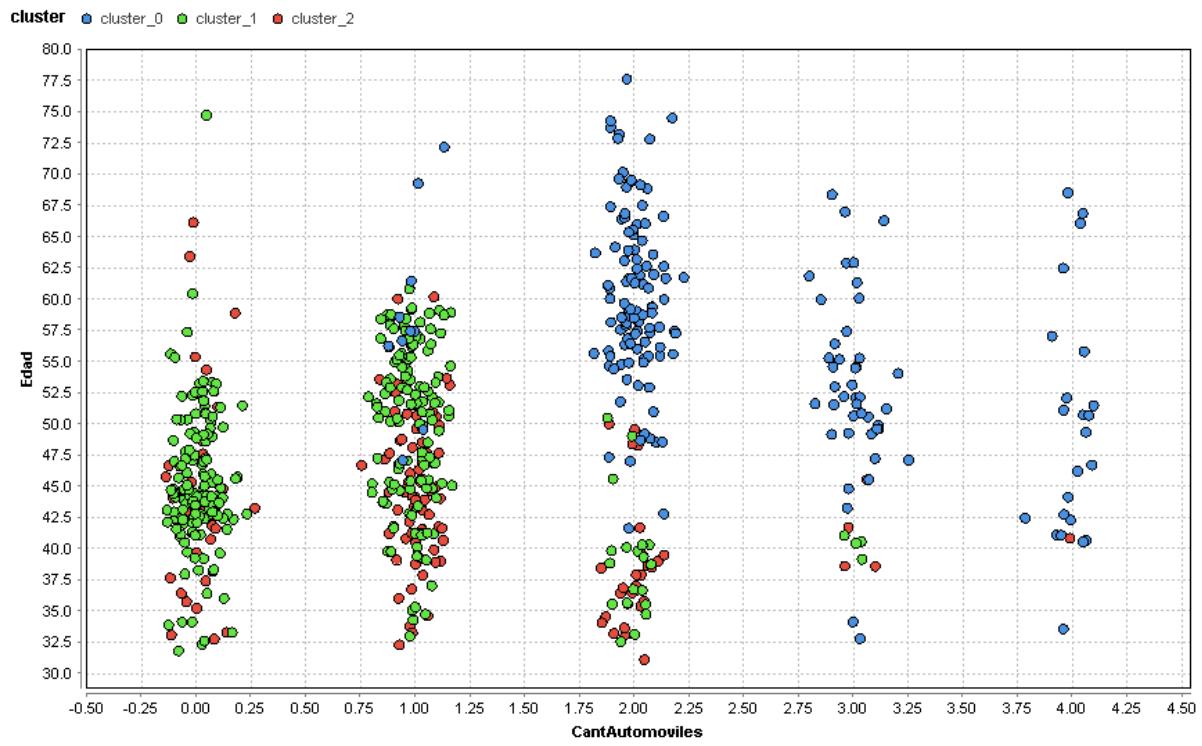
CantAutomoviles vs IngresoAnual



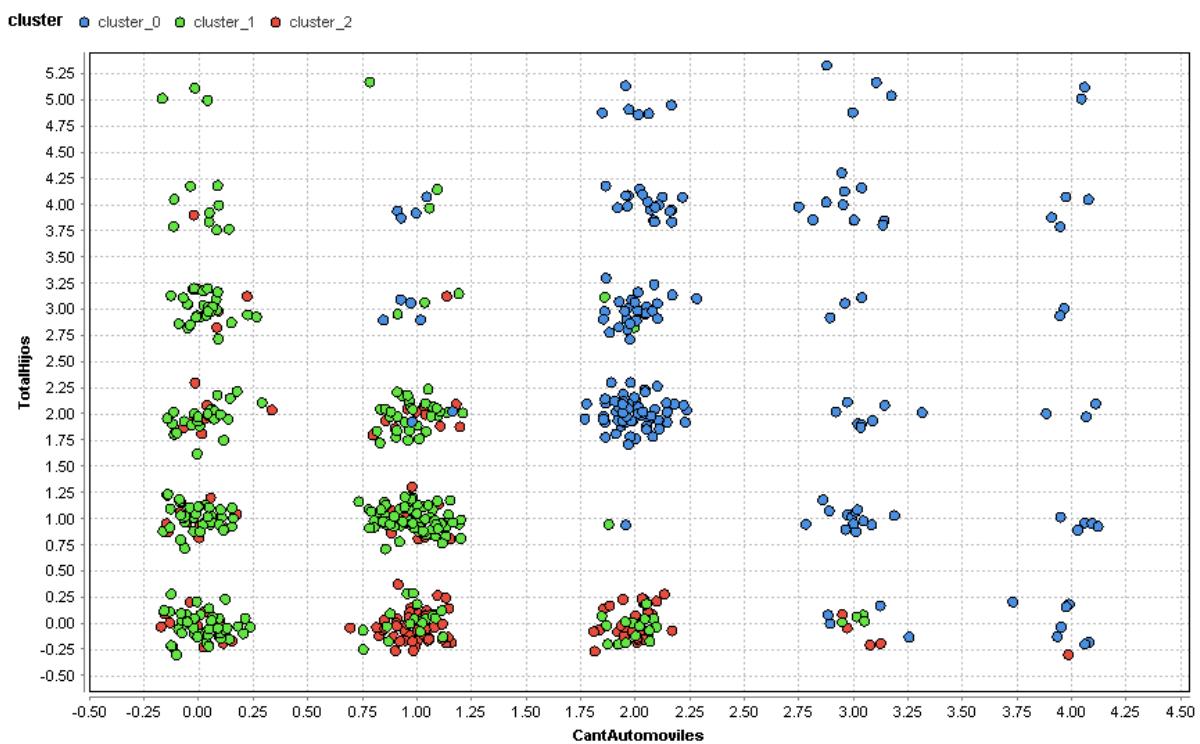
IngresoAnual vs TotalHijos



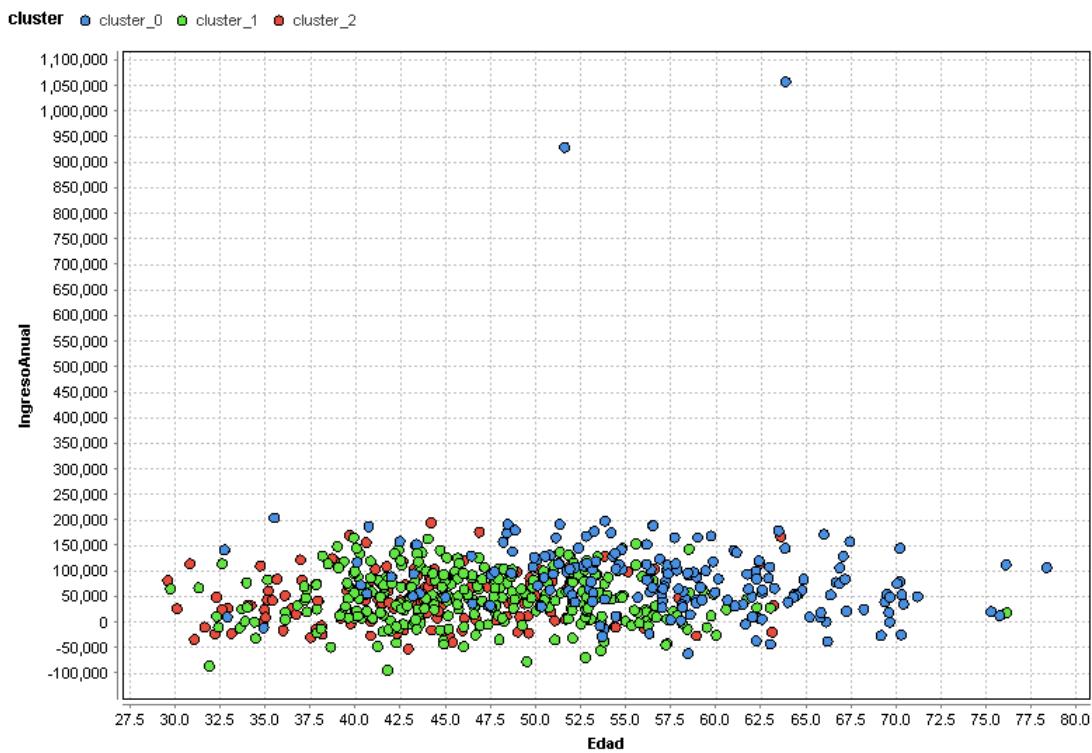
CantAutomoviles vs Edad



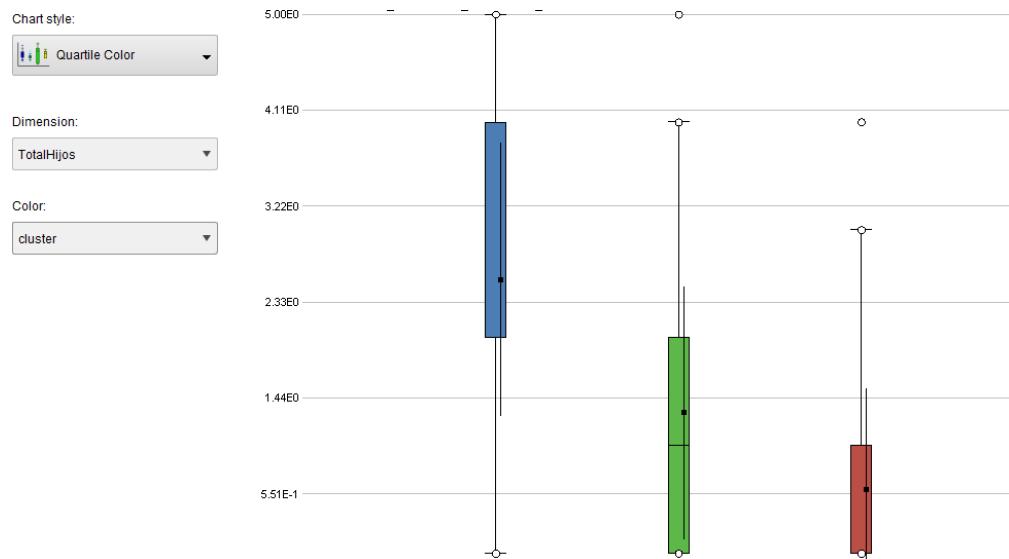
CantAutomoviles vs TotalHijos



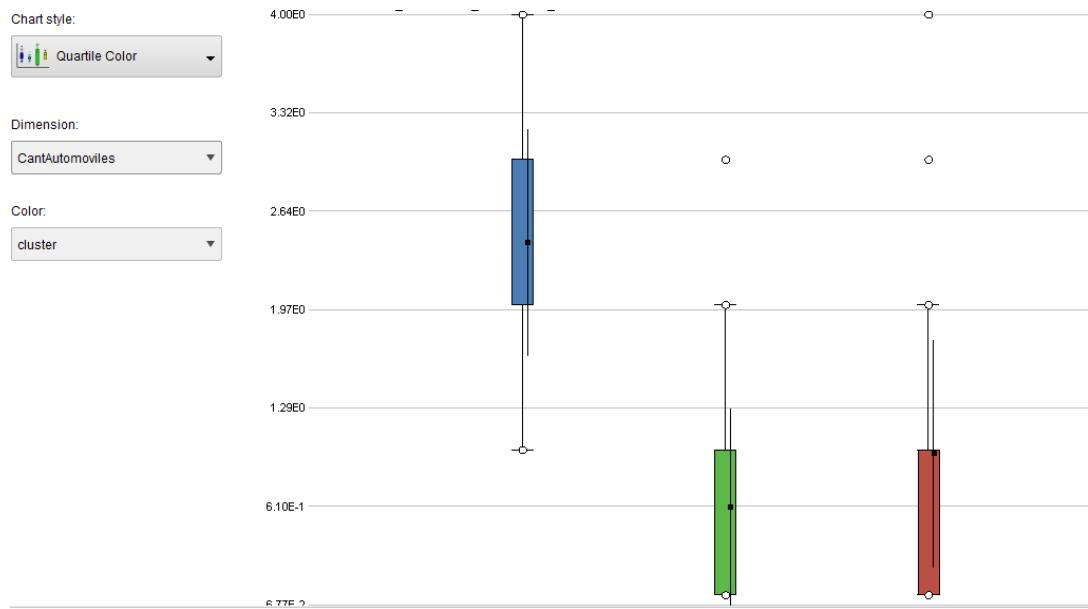
Edad vs IngresoAnual



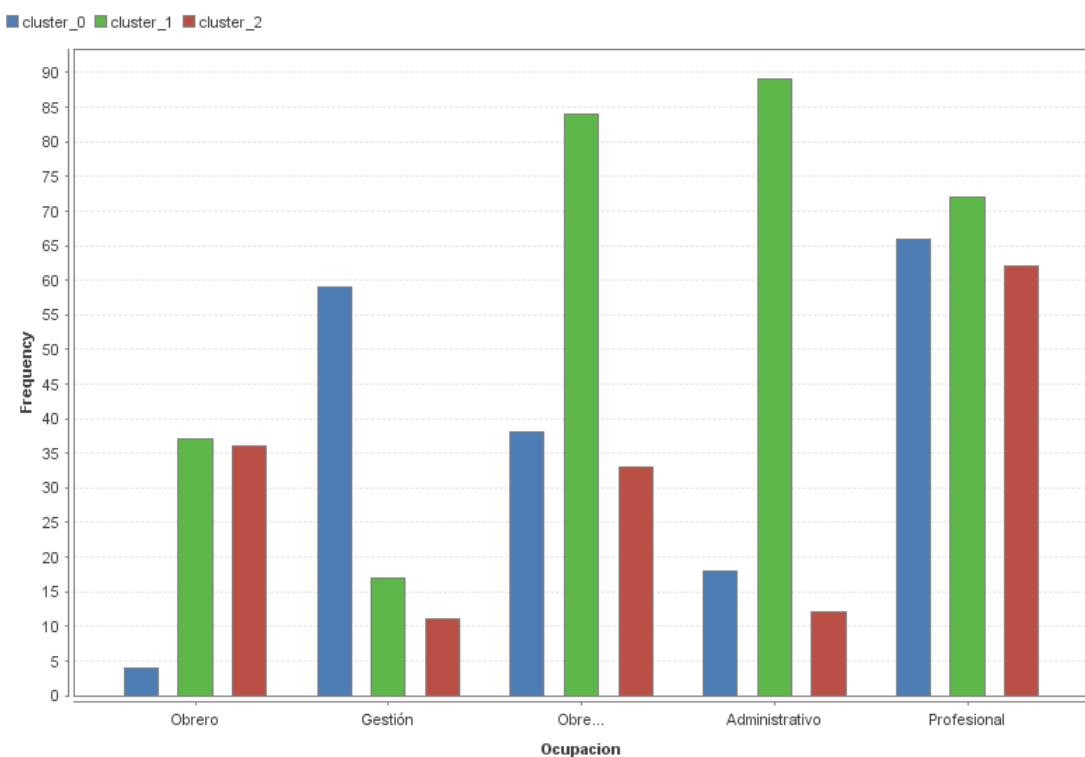
Total de Hijos en cada cluster



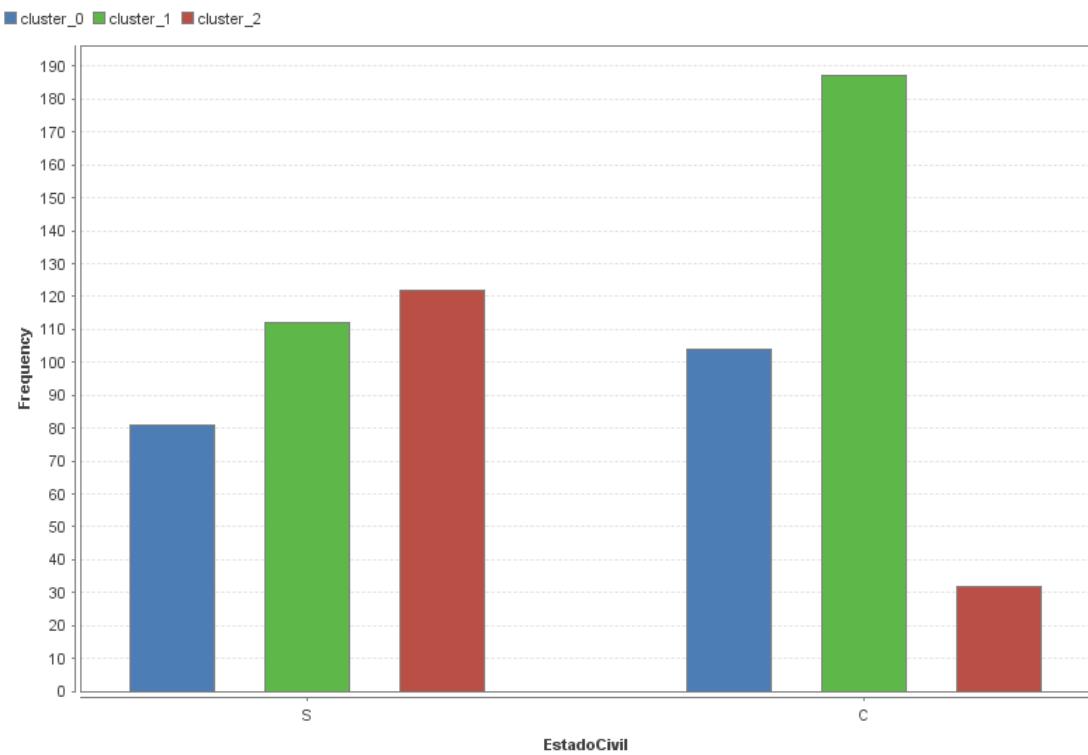
Cantidad de Automóviles en cada cluster



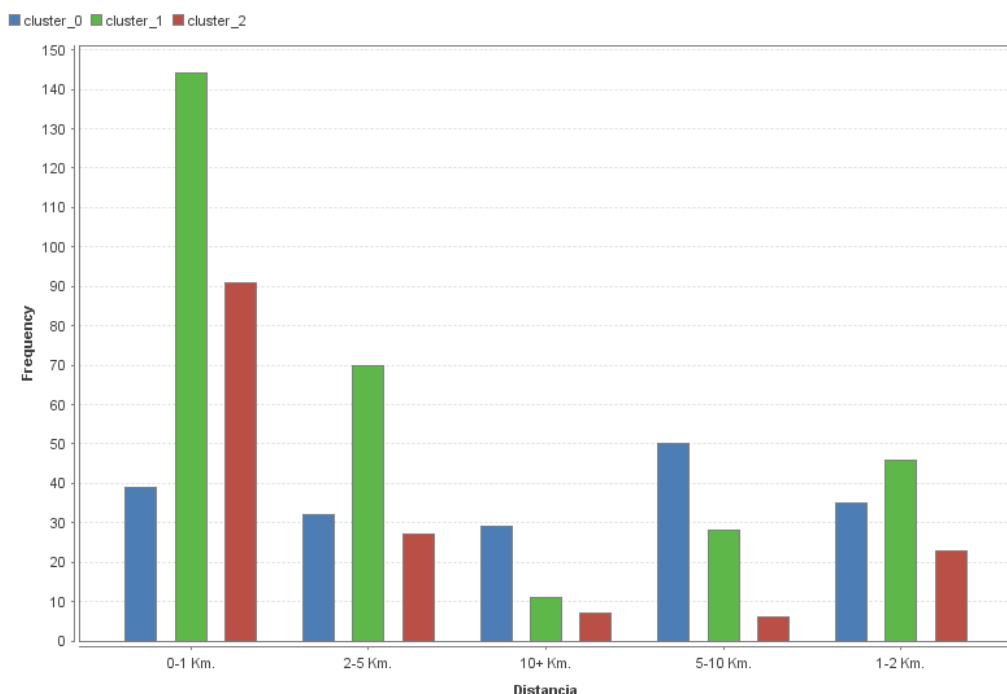
Ocupación más frecuente en cada cluster



Estado Civil más frecuente en cada cluster



Distancia al trabajo más frecuente en cada cluster



Caracterización de los cluster con K = 3

ATRIBUTO	CLUSTER 0	CLUSTER 1	CLUSTER 2
Ingreso Anual	Mayor Valor	En su mayoría de Menor	En su mayoría de Menor

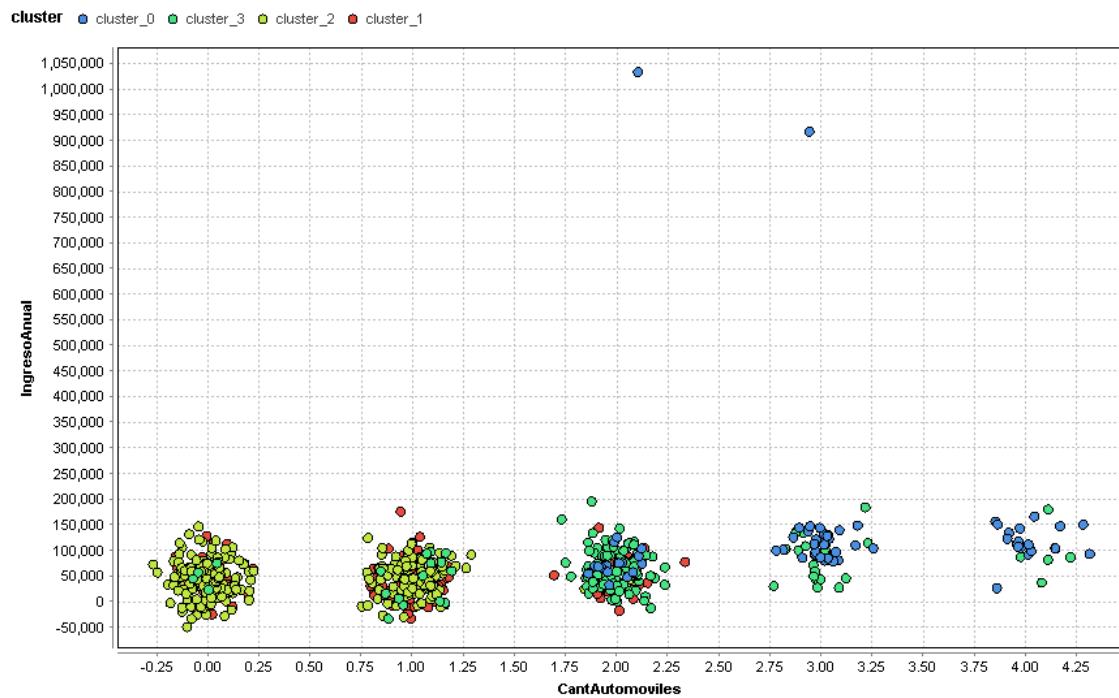
		Valor	Valor
Edad	Mayormente adultos y ancianos	Mayormente jóvenes y adultos	Mayormente adultos y ancianos
Cant Automóviles	Mayor Cantidad	Menor Cantidad	Menor Cantidad
Ocupación	En mayor proporción existen ocupaciones de Gestión y Profesionales.	En mayor proporción existen ocupaciones de Profesionales, Administrativos y Obreros especializados.	En mayor proporción existen ocupaciones de Profesionales.
TotalHijos	-	La mayoría tiene como máximo 1 hijo	No tienen hijos en su mayoría
Estado Civil	-	Mayormente Casados	Mayormente Solteros
Distancia	-	La mayoría viven cerca o a media distancia (0-1 km, 1-2 km, 2-5 km)	La mayoría viven cerca o a media distancia (0-1 km, 1-2 km, 2-5 km)

K = 4

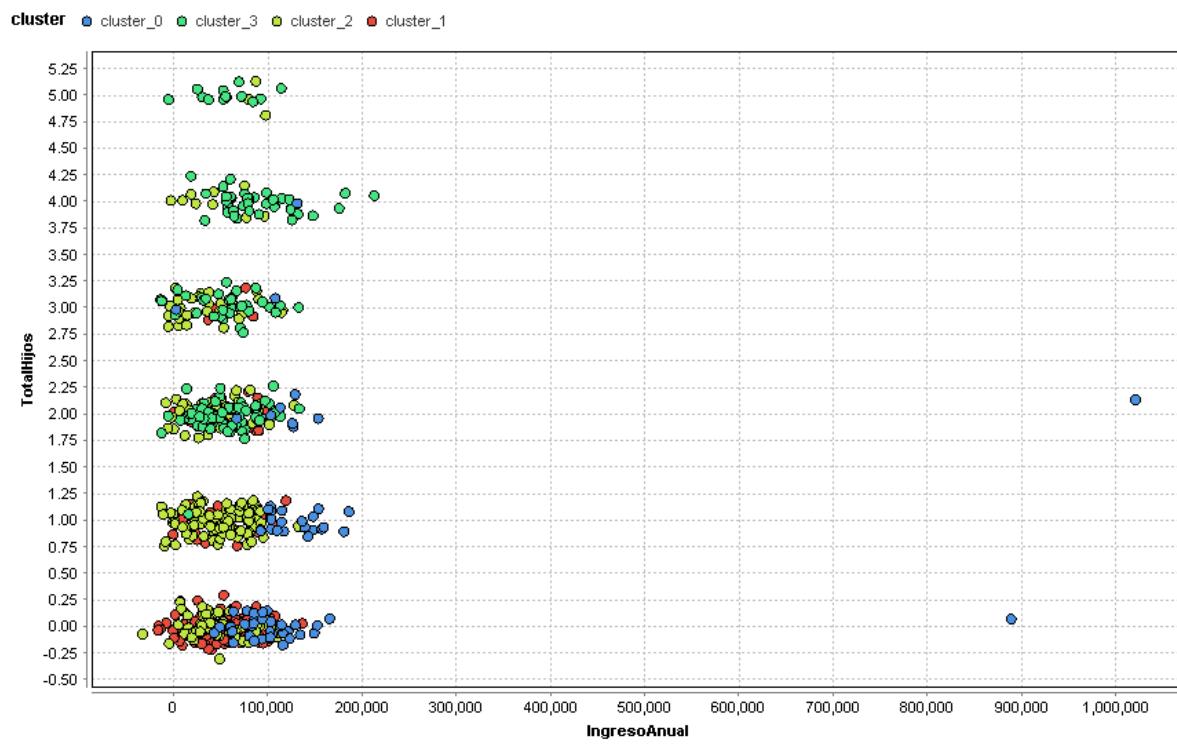
De un total de 638 observaciones, el modelo agrupó:

- 67 observaciones en el cluster 0.
- 143 observaciones en el cluster 1.
- 277 observaciones en el cluster 2.
- 151 observaciones en el cluster 3.

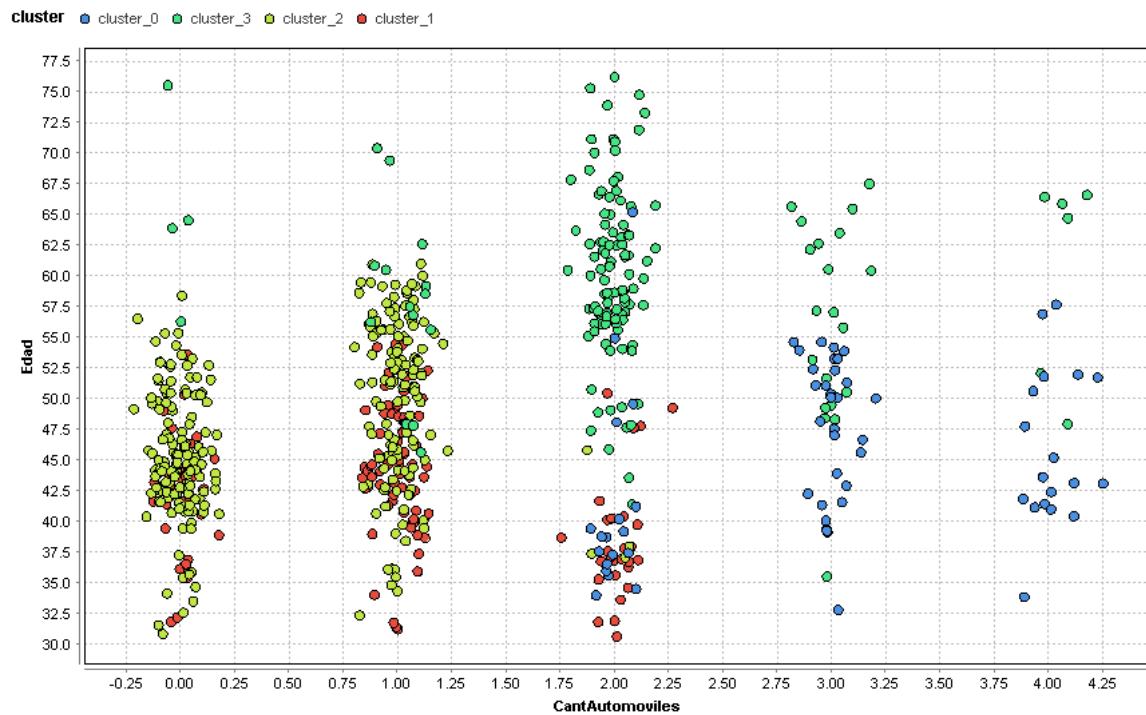
CantAutomoviles vs IngresoAnual



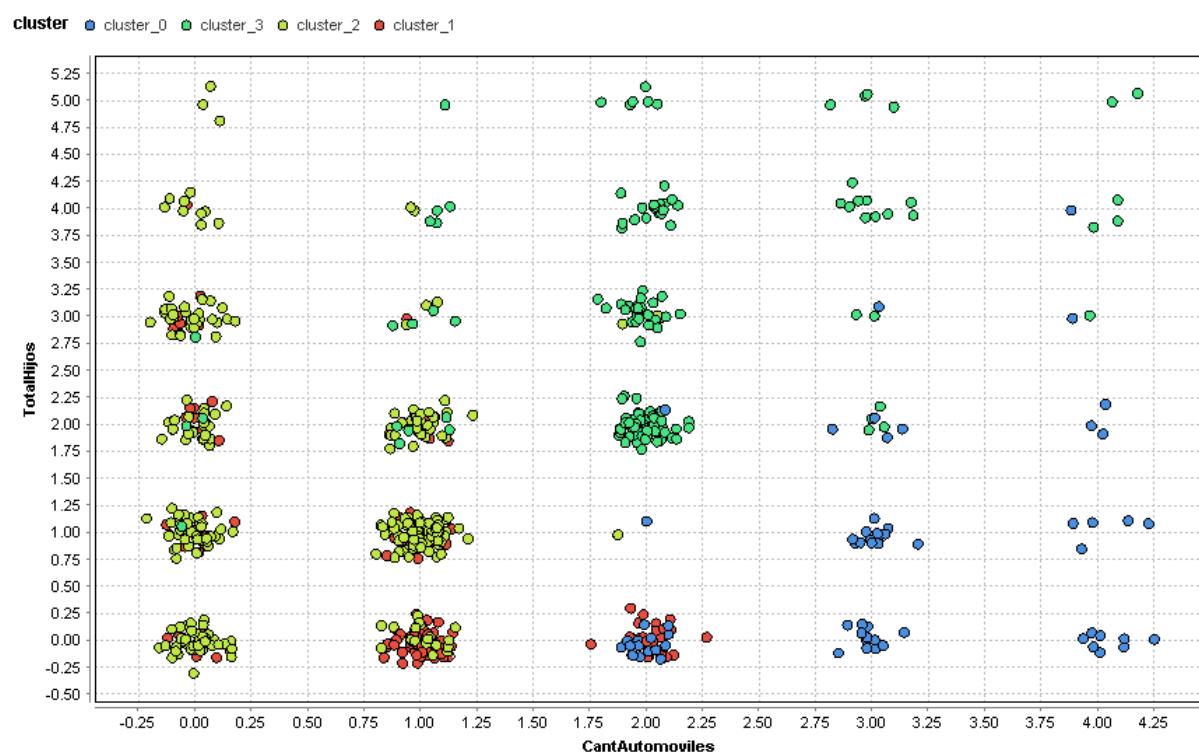
IngresoAnual vs TotalHijos



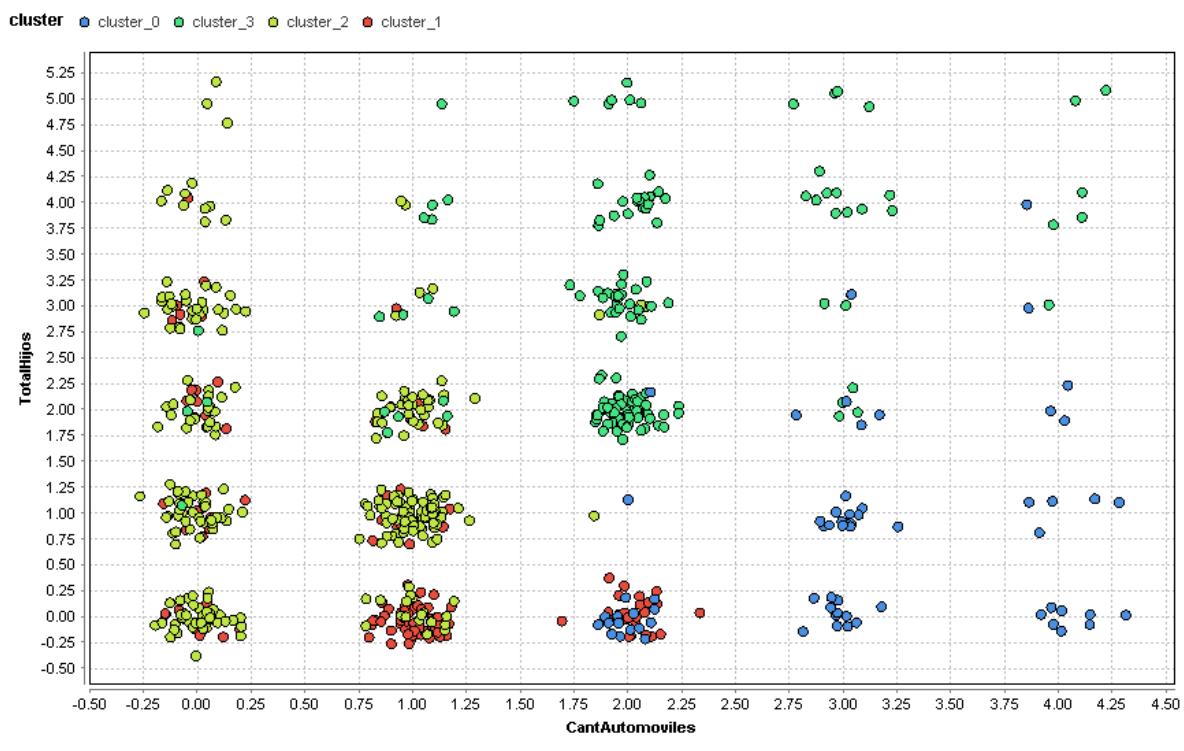
CantAutomoviles vs Edad



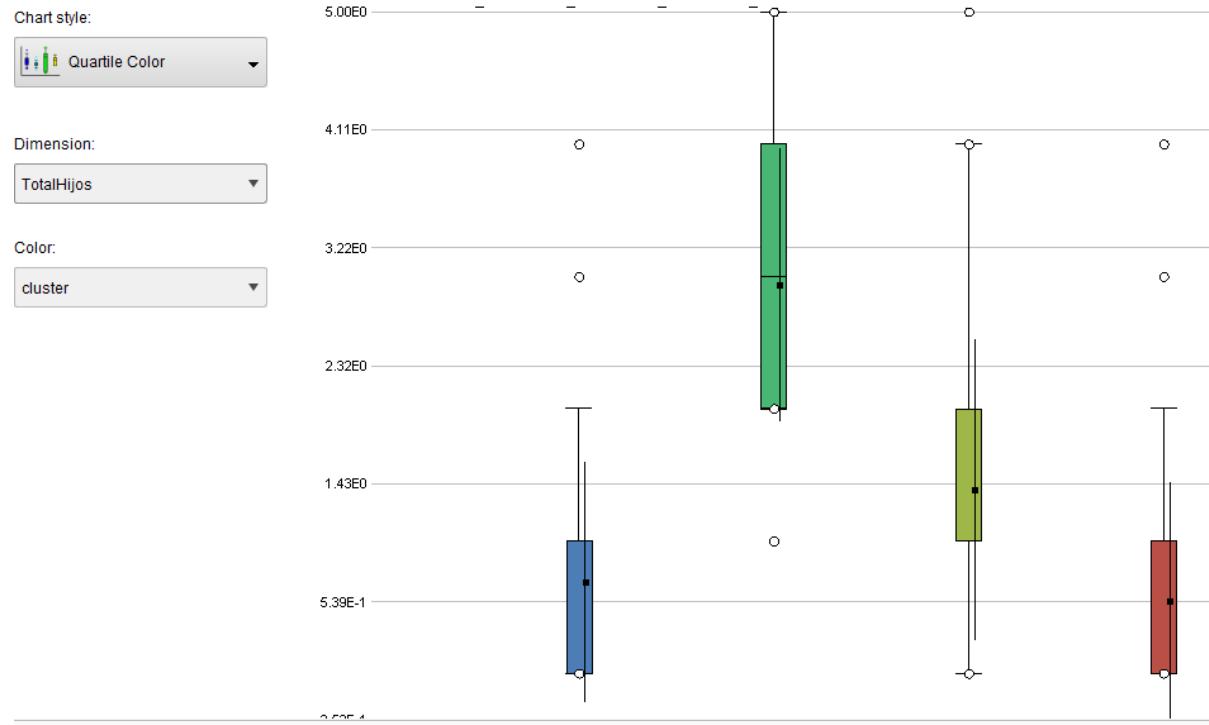
CantAutomoviles vs TotalHijos



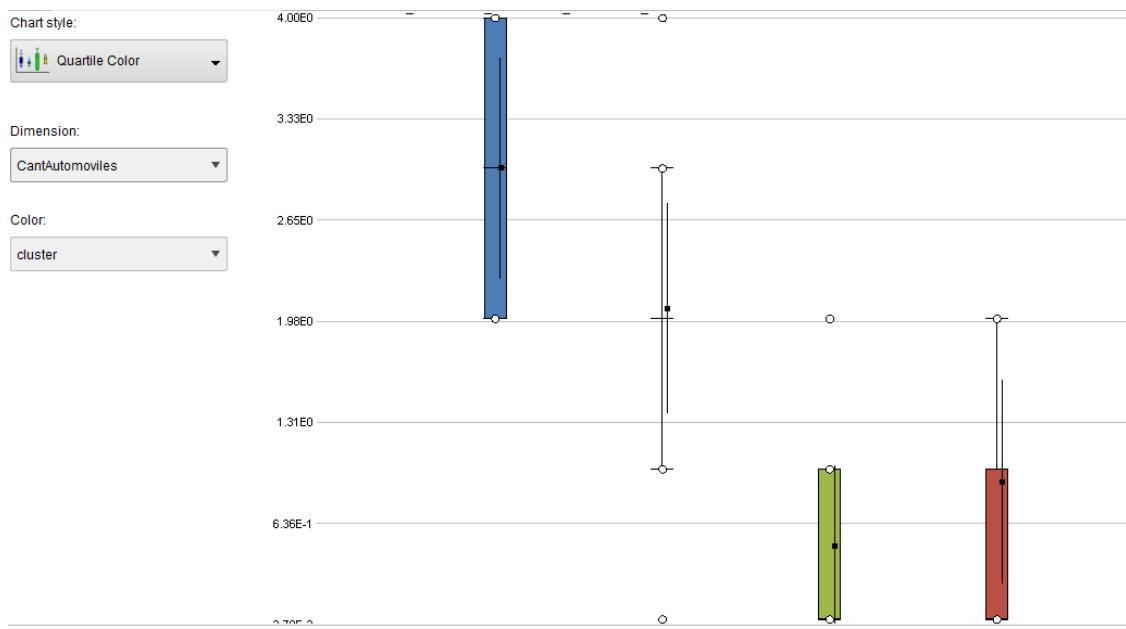
Edad vs IngresoAnual



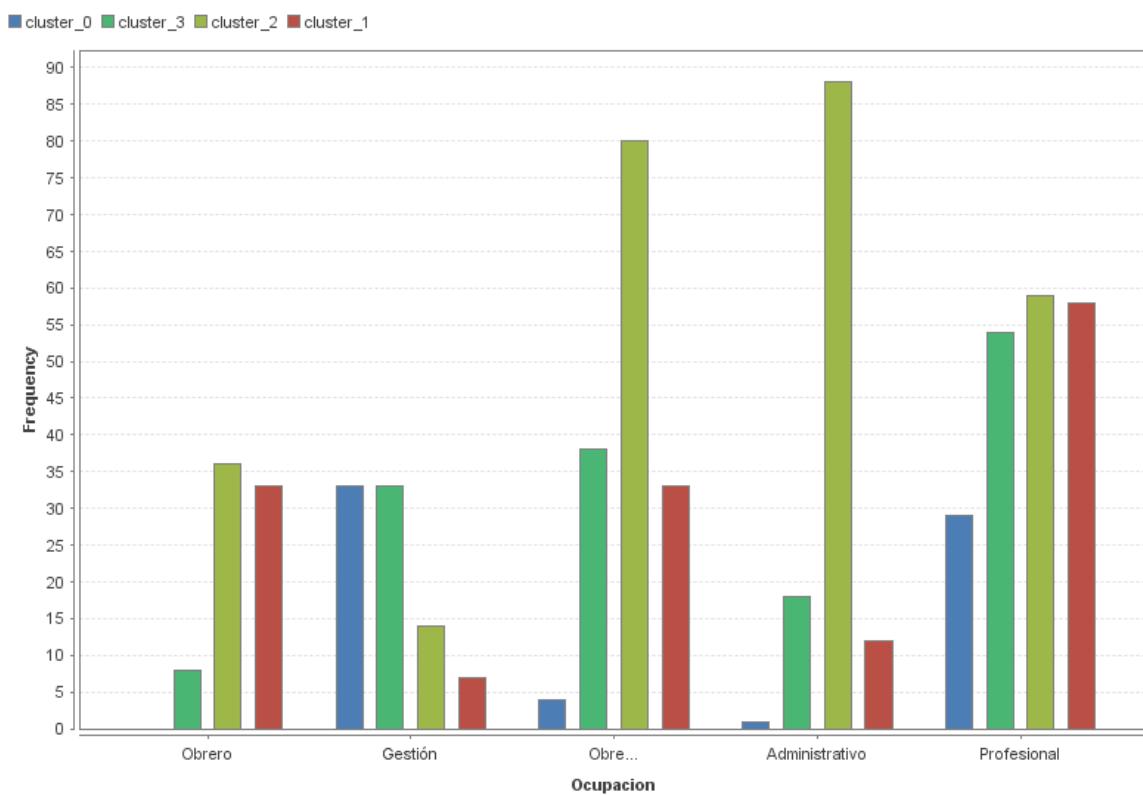
Total de Hijos en cada cluster



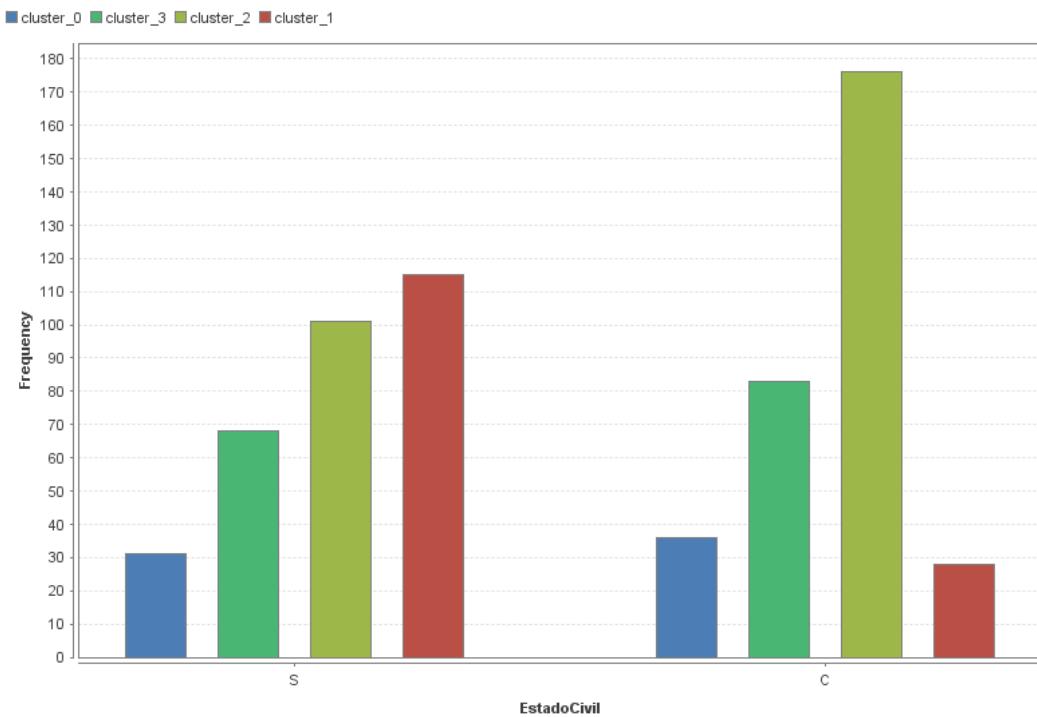
Cantidad de Automóviles en cada cluster



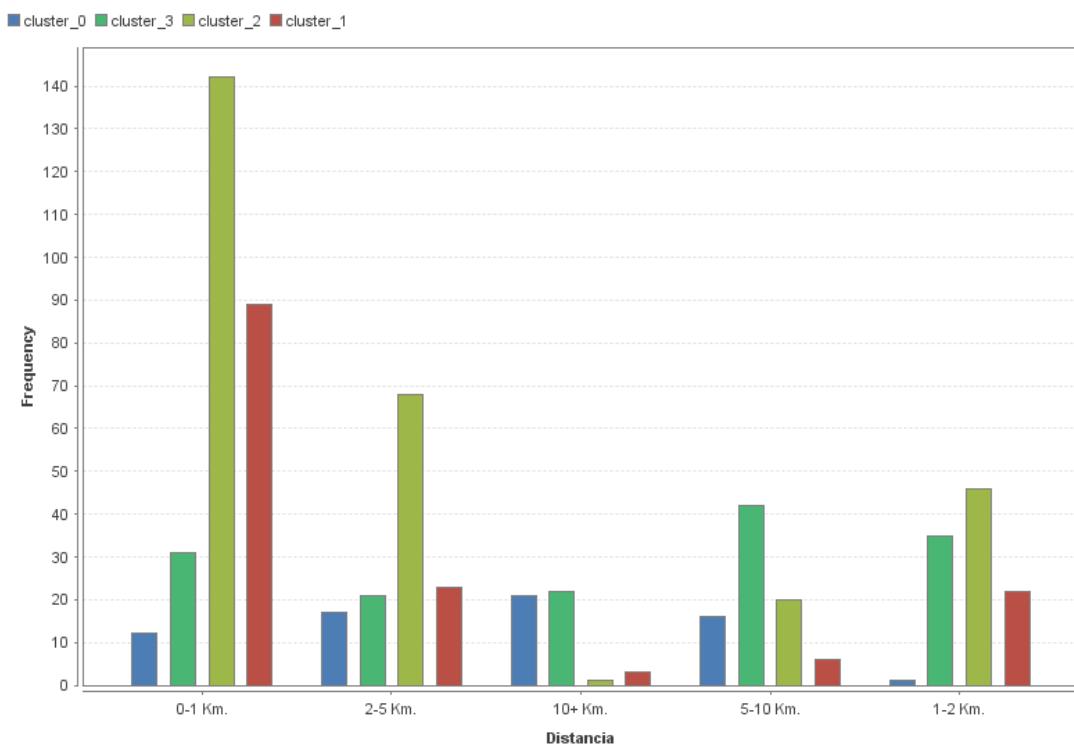
Ocupación más frecuente en cada cluster



Estado Civil más frecuente en cada cluster



Distancia al trabajo más frecuente en cada cluster



Caracterización de los cluster con K = 4

ATRIBUTO	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3
Ingreso Anual	Mayor Valor	Menor Valor	Menor Valor	Mayor Valor
Edad	Mayormente jóvenes y adultos	Mayormente jóvenes y adultos	Mayormente jóvenes y adultos	Mayormente adultos y ancianos
Cant Automóviles	2 - 3 - 4	0 - 1 - 2	0 - 1 - 2 (aunque la mayoría tiene como máximo 1 vehículo)	-
Ocupación	En su mayoría tienen ocupaciones de Gestión y Profesionales.	En su mayoría tienen ocupaciones de Profesionales, Obreros y Obreros Especializados.	En su mayoría tienen ocupaciones de Profesionales, Obreros Especializados y Administrativo.	En su mayoría tienen ocupaciones de Profesionales, Obreros Especializados y Gestión.
TotalHijos	0 - 1 - 2 - 3 - 4 (aunque la mayoría tiene como máximo 2 hijos)	0 - 1 - 2 - 3	-	1 - 2 - 3 - 4 - 5 (La mayoría tiene al menos 2 hijos)
Estado Civil	-	Mayormente Solteros	Mayormente Casados	-
Distancia	-	Mayormente viven cerca (0-1 km)	Mayormente viven cerca o a media distancia (0-1 km, 1-2 km, 2-5 km)	-

Conclusión

Dentro de los valores de K analizados anteriormente y las caracterizaciones de los cluster obtenidas para cada uno, consideramos que el valor de K que diferencia y agrupa de mejor forma es K = 2, siendo también el que divide en cantidades de individuos más homogéneas por cluster.

Cluster Jerárquico

En minería de datos, el clustering jerárquico es un método de análisis de grupos organizándose en forma jerárquica los mismos, es decir, los clusters están anidados.

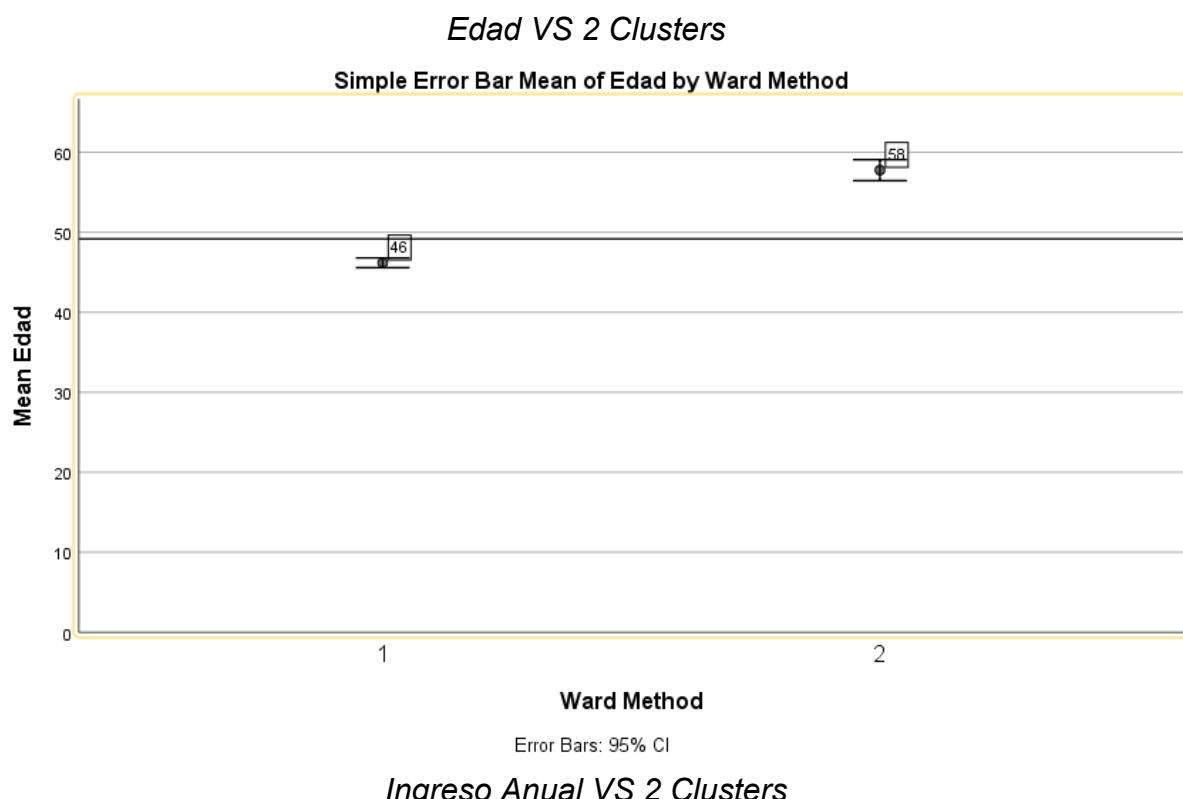
Utilizaremos el Software SPSS Statistics haciendo un análisis de las variables cuantitativas separadas de las cualitativas por las limitaciones del SW.

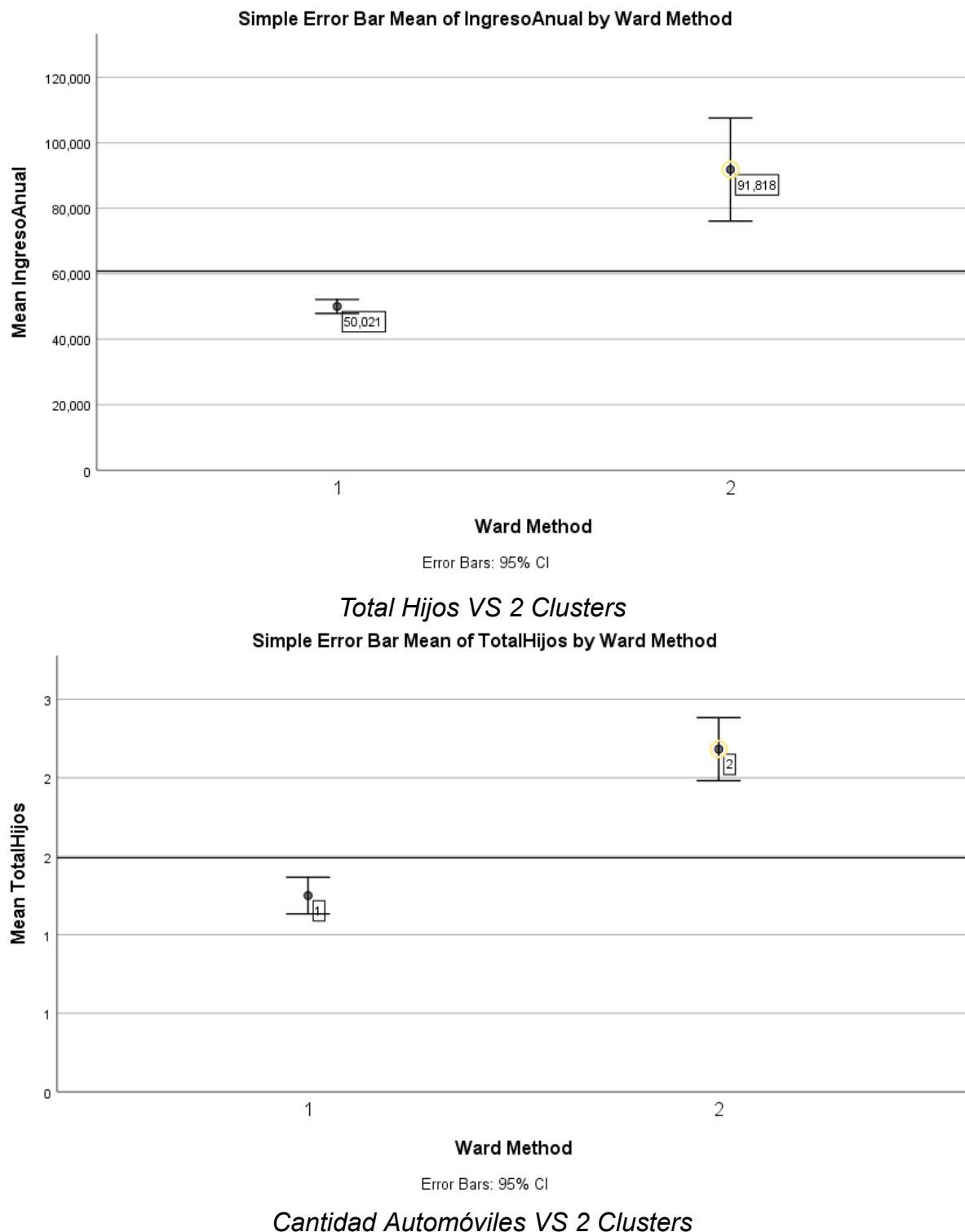
Clustering con variables cuantitativas

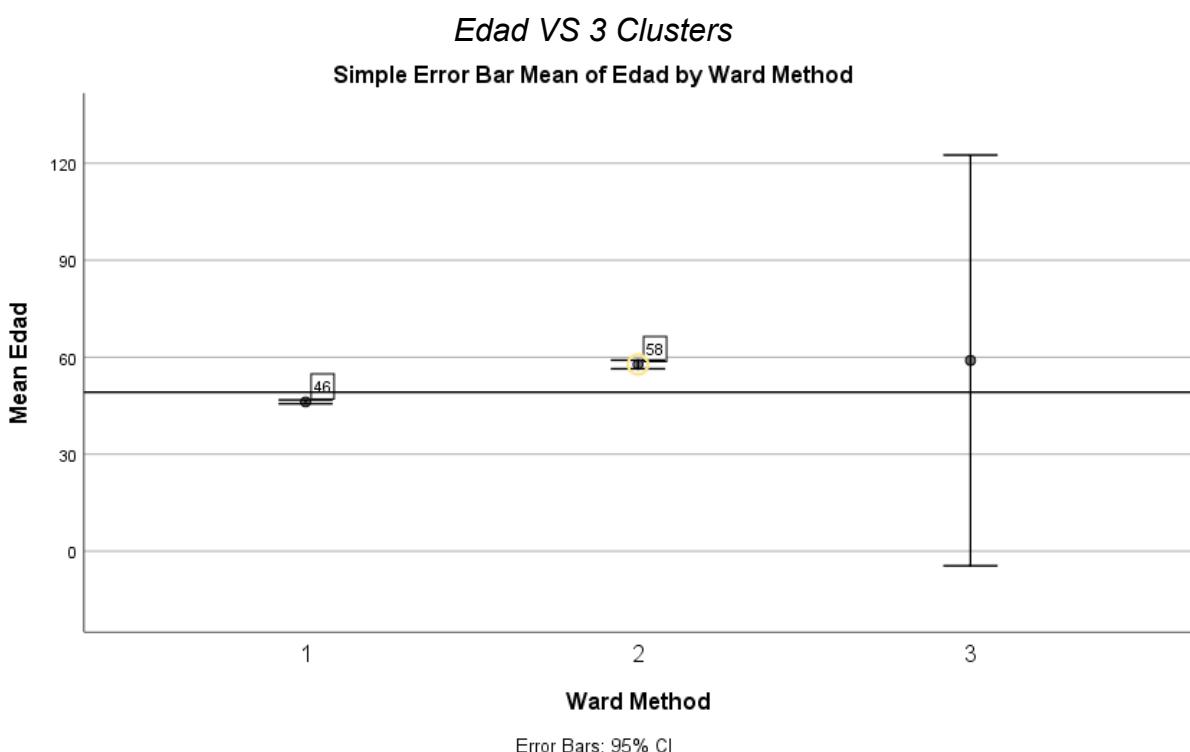
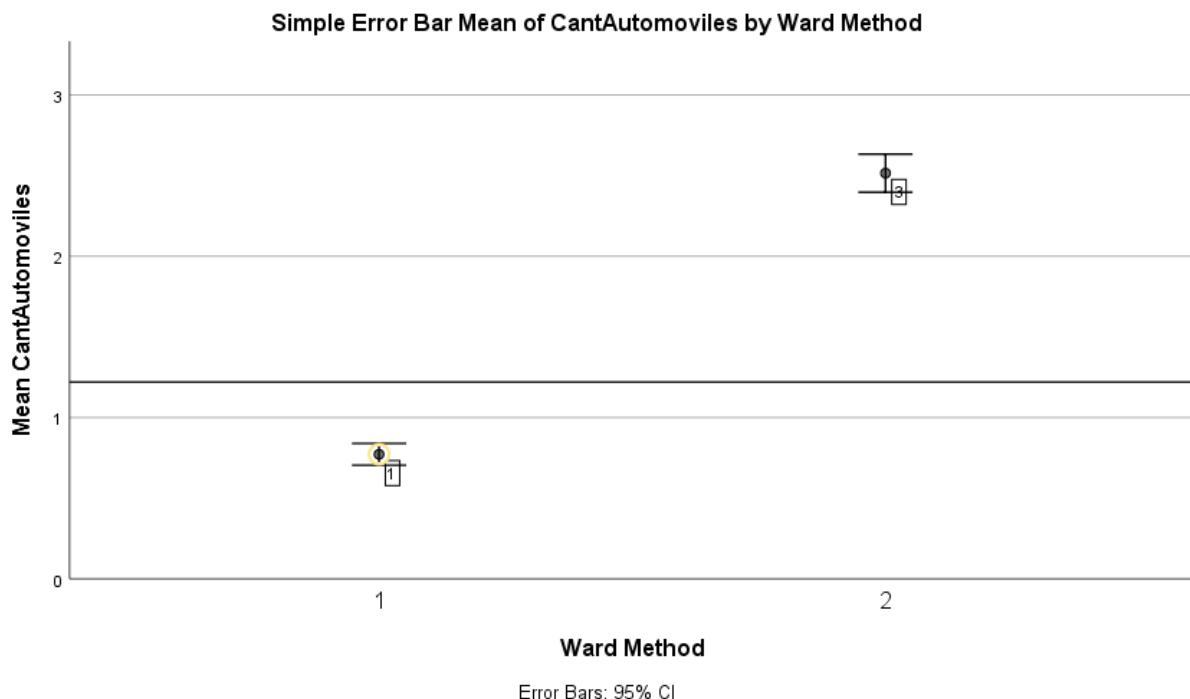
Para las variables estandarizadas, edad, ingreso anual, total hijos y cantidad de automóviles se obtuvo el siguiente dendrograma:

 [Cluster_Jerarquico_Cuantitativas.png](#)

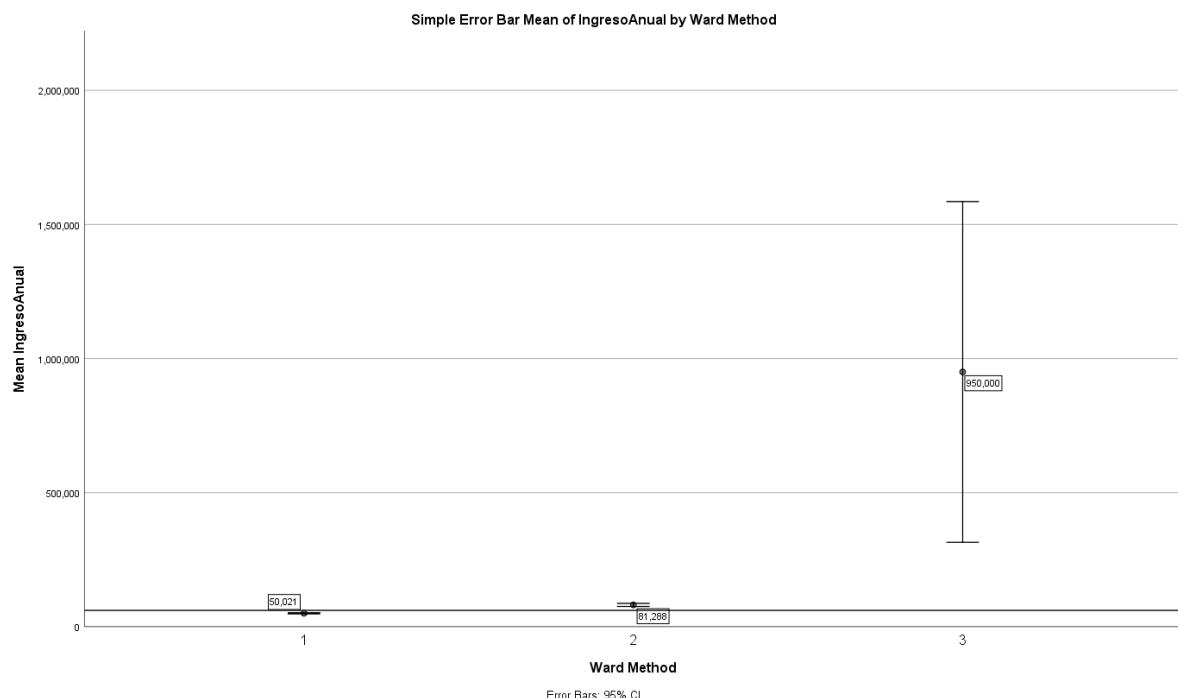
A partir de las siguientes salidas compararemos los clusters formados.



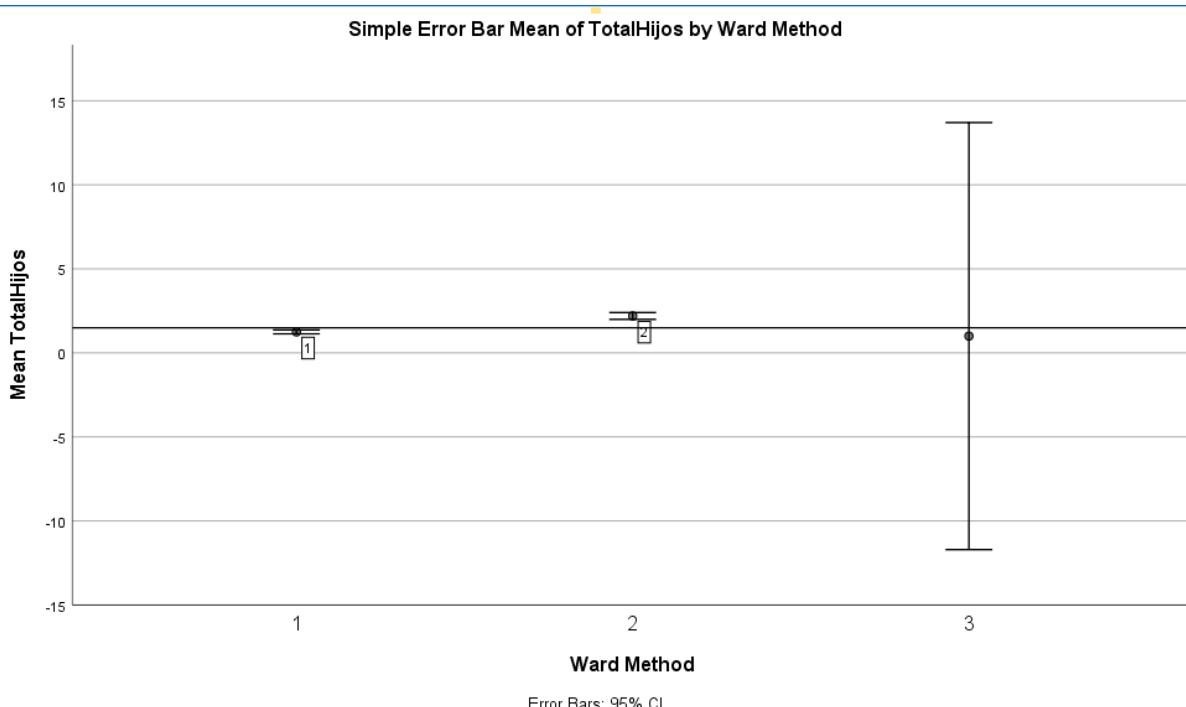




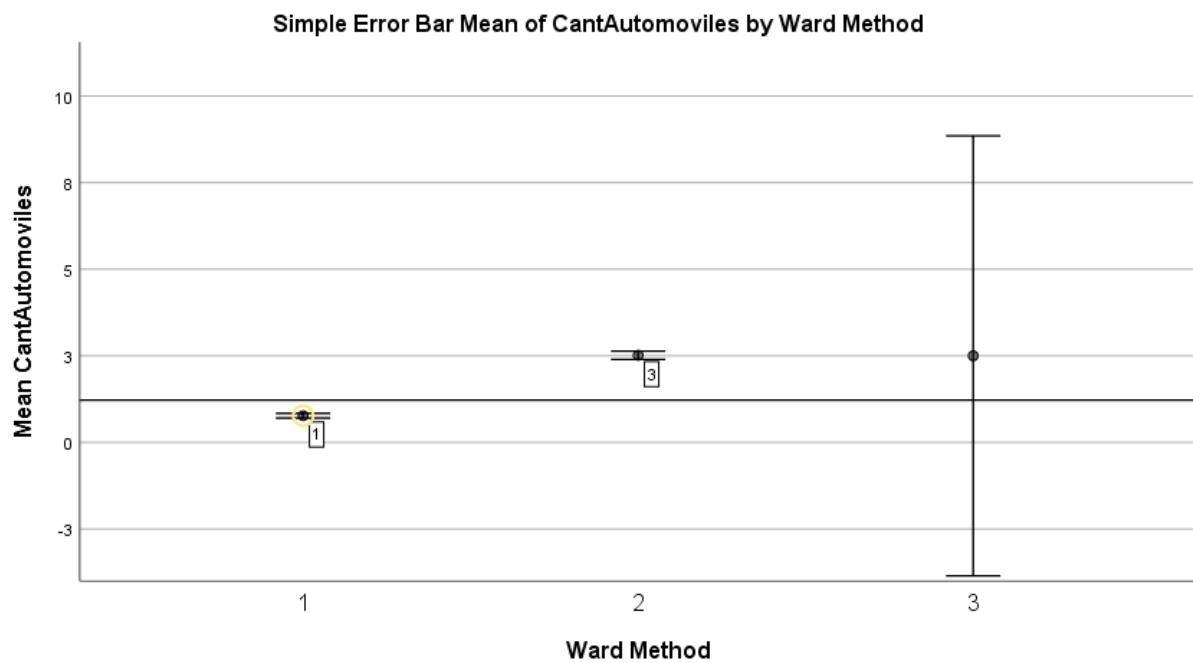
Ingreso Anual VS 3 Clusters



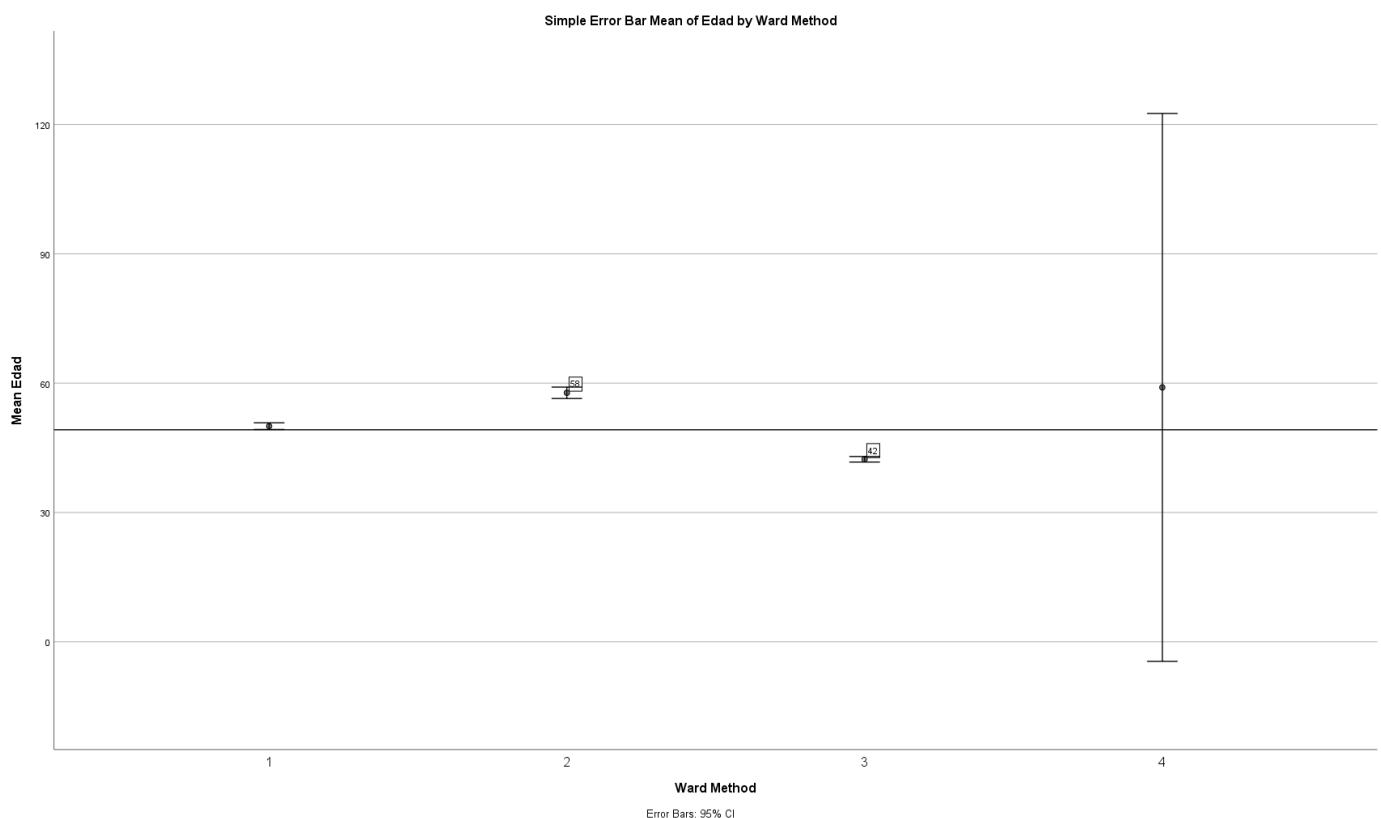
Total Hijos VS 3 Clusters



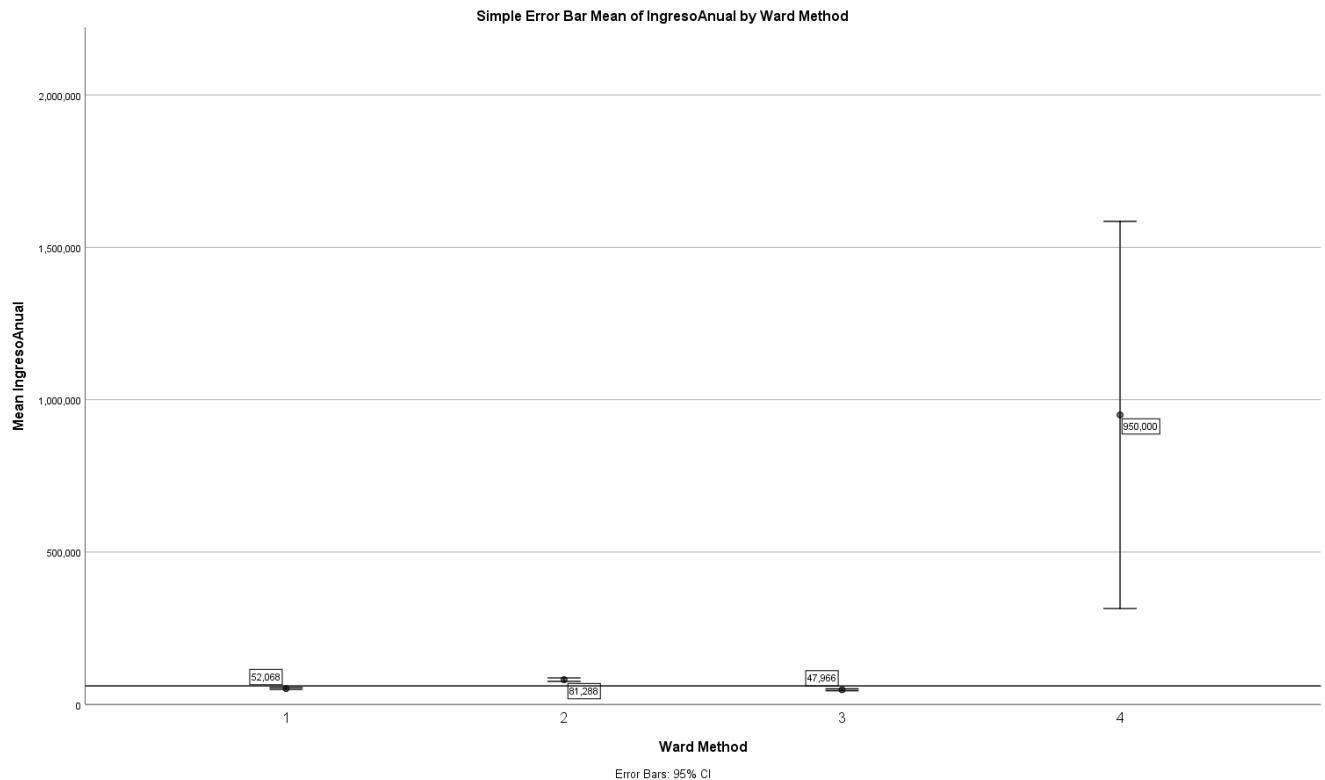
Cantidad Automóviles VS 3 Clusters



Edad VS 4 Clusters

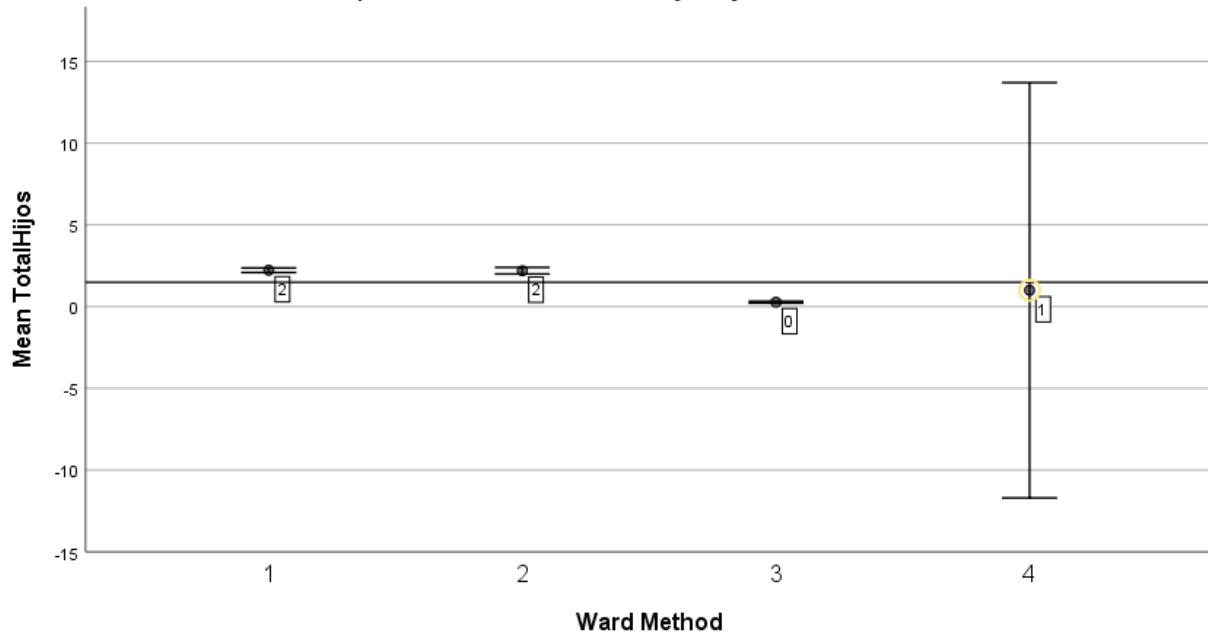


Ingreso Anual VS 4 Clusters

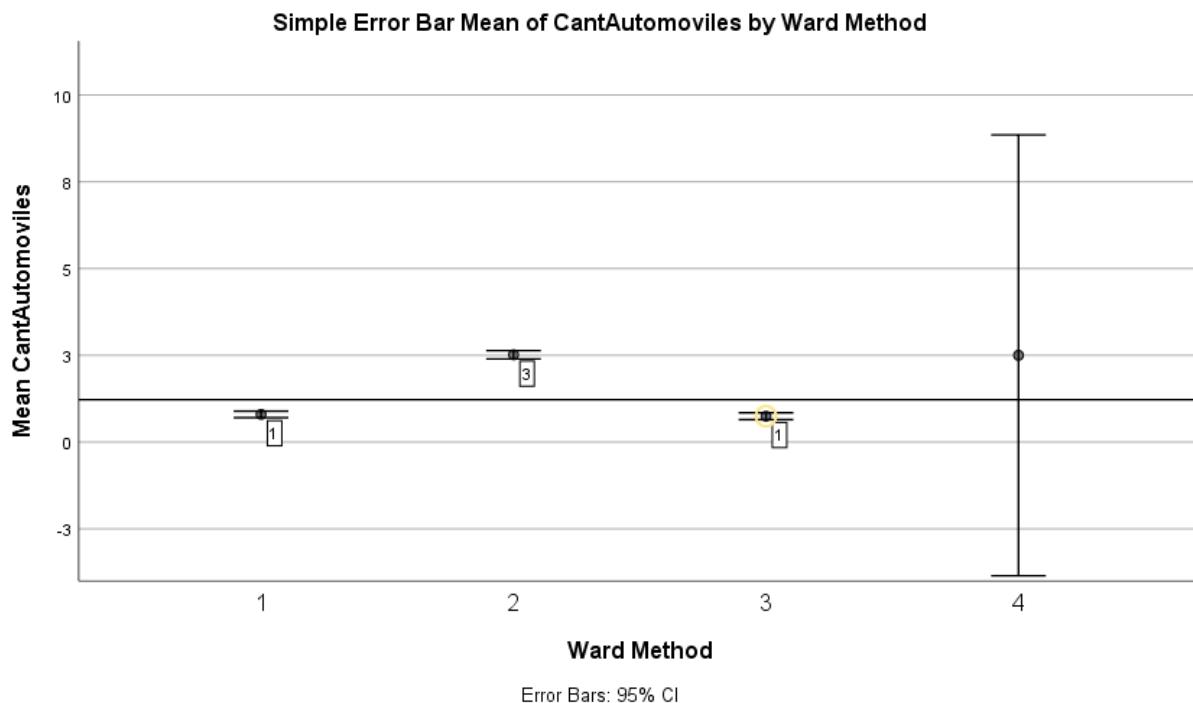


Total Hijos VS 4 Clusters

Simple Error Bar Mean of TotalHijos by Ward Method



Cantidad automóviles VS 4 Clusters



Con las gráficas anteriores se completó la siguiente tabla indicando:

- Un valor si los intervalos no se solapan.
- Si los intervalos se solapan con la media (Línea horizontal) colocamos un - debido a que la variable dentro del grupo no se diferencia con la población.
- Si los intervalos entre clusters se solapan, se ingresa el valor promedio entre los clusters.

		2		3			4			
Media Poblacional	Variable	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
49,18	Edad	46	58	46	58	-	-	58	42	-
60830,72	Ingreso Anual	50021	91818	50021	81288	950000	50017	81288	50017	950000
1,49	Total Hijos	1	2	1	2	-	2	2	0	-
1,22	CantAutomobiles	1	3	1	3	-	1	3	1	-

Se puede observar que tanto para 3 clusters, las variables edad, total hijos y cantidad de automóviles no aportan información en el tercer cluster.

Para 4 clusters encontramos:

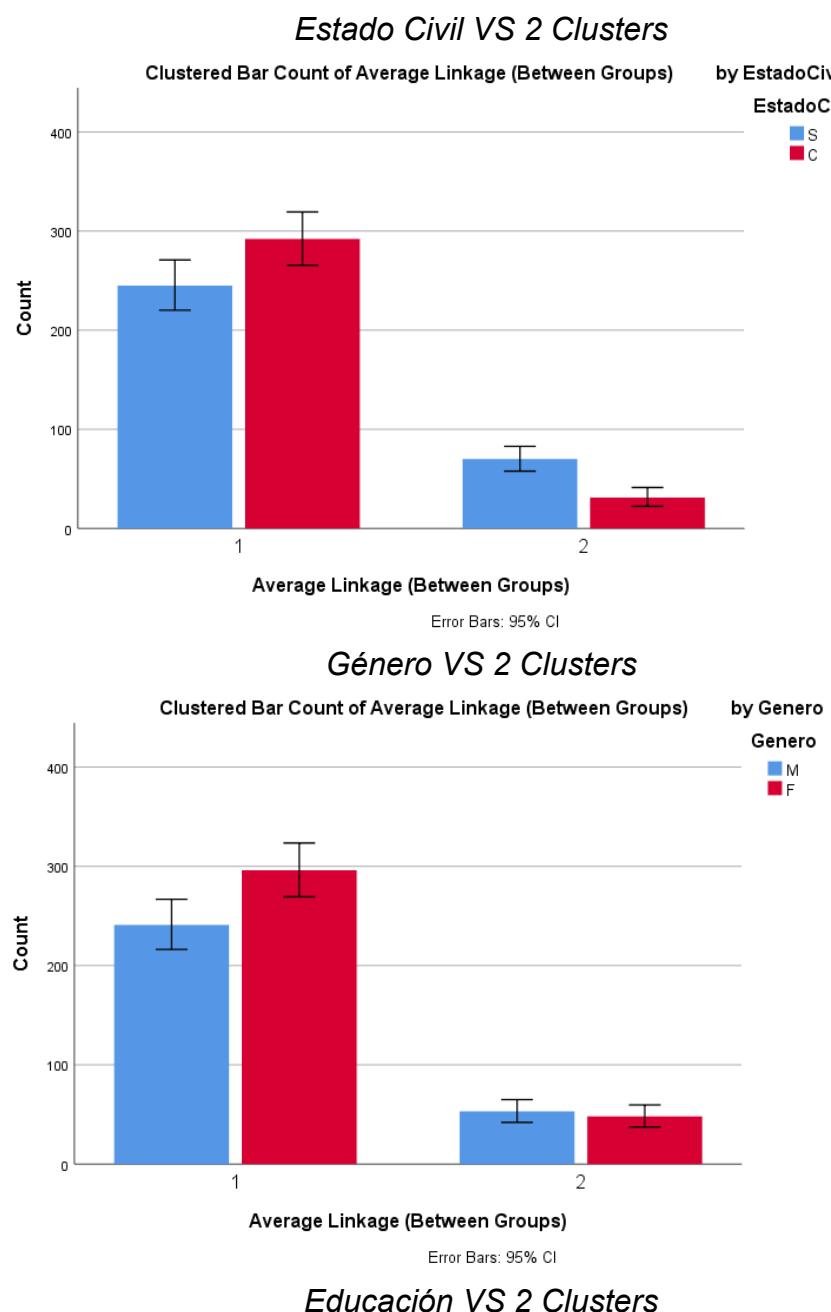
- La variable edad no aporta información en el grupo 1.
- La variable edad, total hijos y cantidad de automóviles no aporta información en el grupo 4.

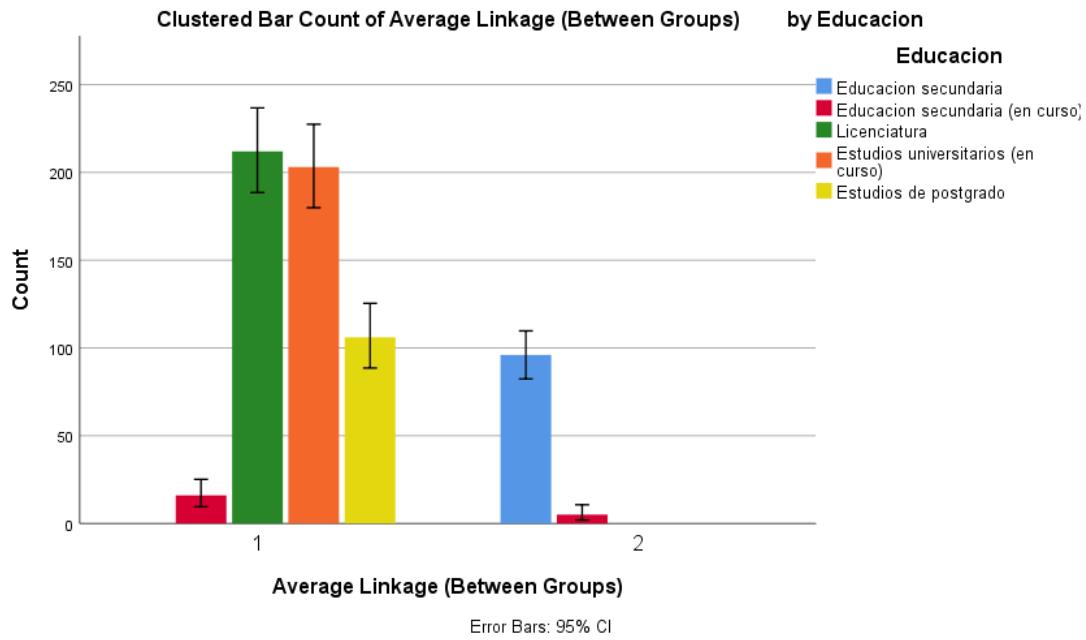
Por lo tanto, elegiremos 2 clusters, ya que todas las variables aportan información y son diferenciadas entre sí.

Clustering con variables cualitativas

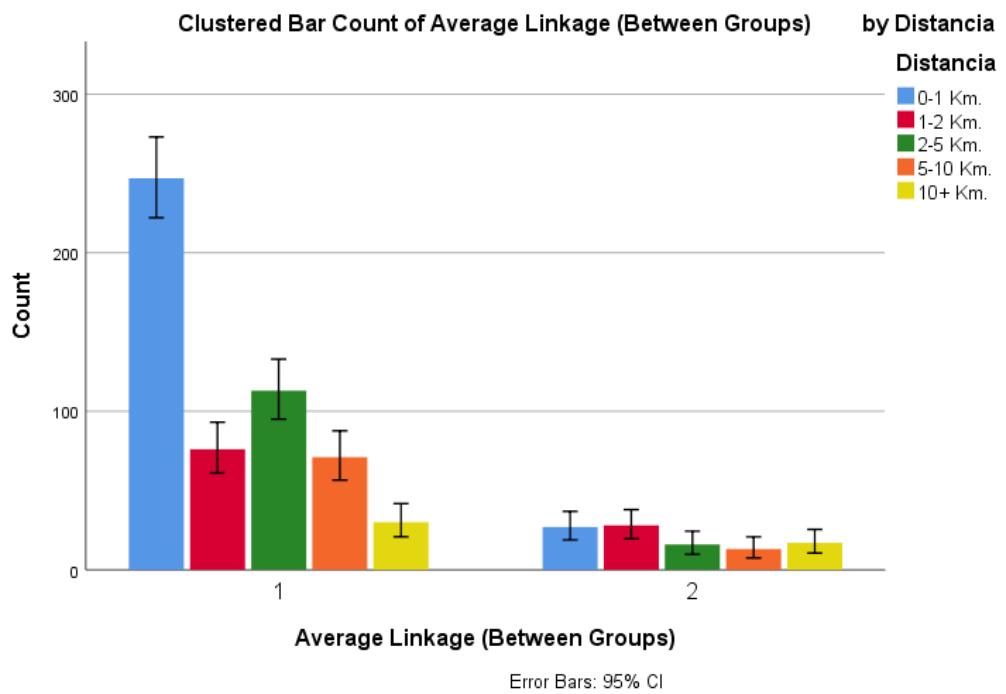
Aplicamos clustering a las variables Estado Civil, Género, Educación, Distancia, Región, Ocupación y Propietario, obtuvimos el siguiente dendrograma:

Cluster_jerárquico_cualitativas.png

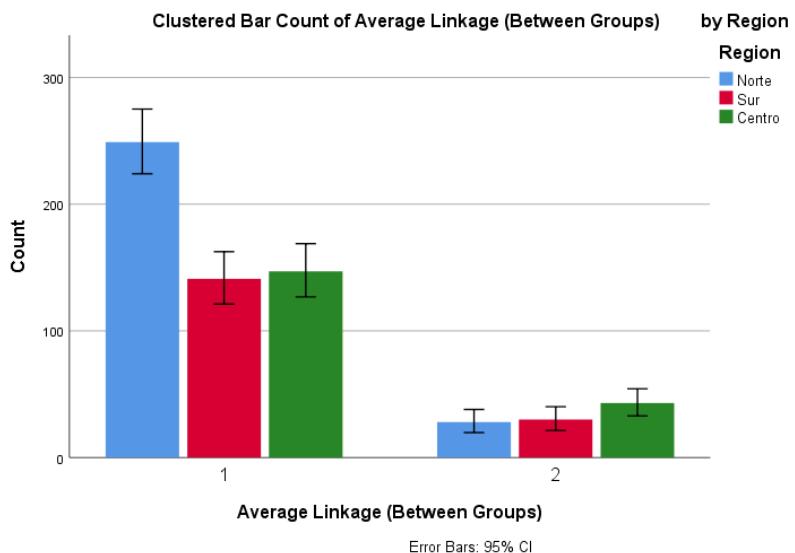




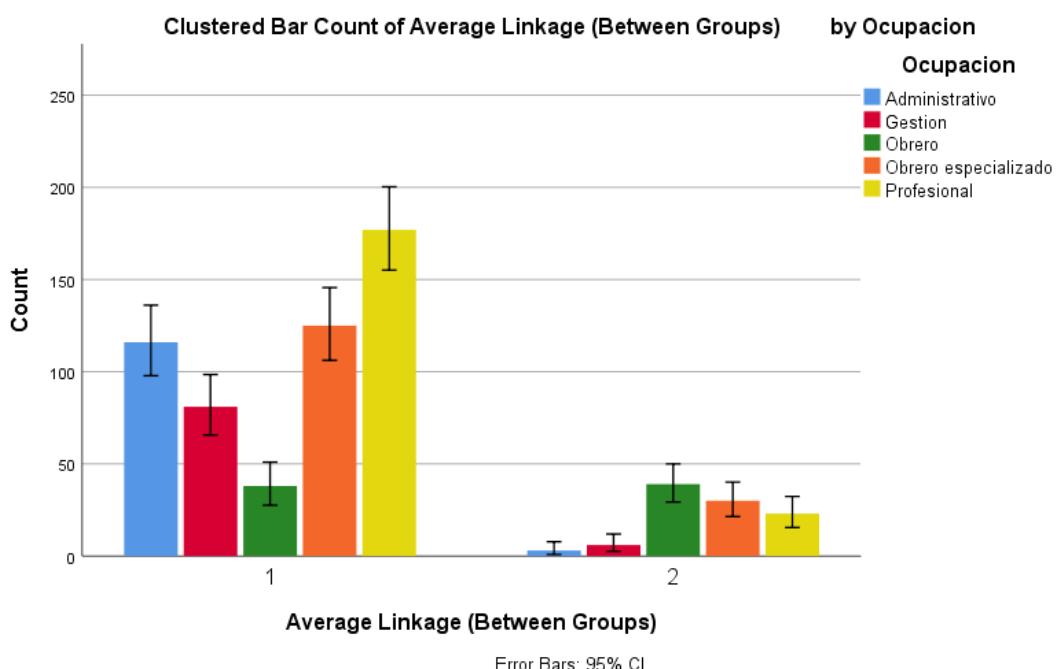
Distancia VS 2 Clusters



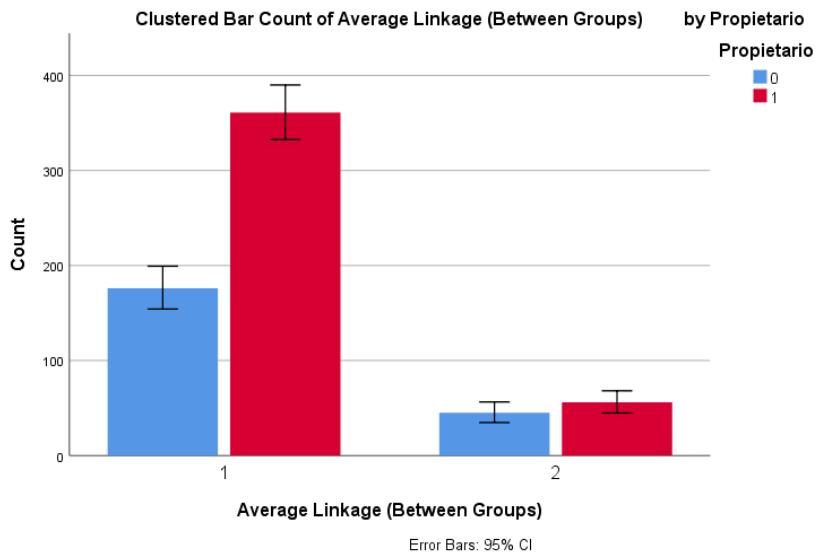
Región VS 2 Clusters



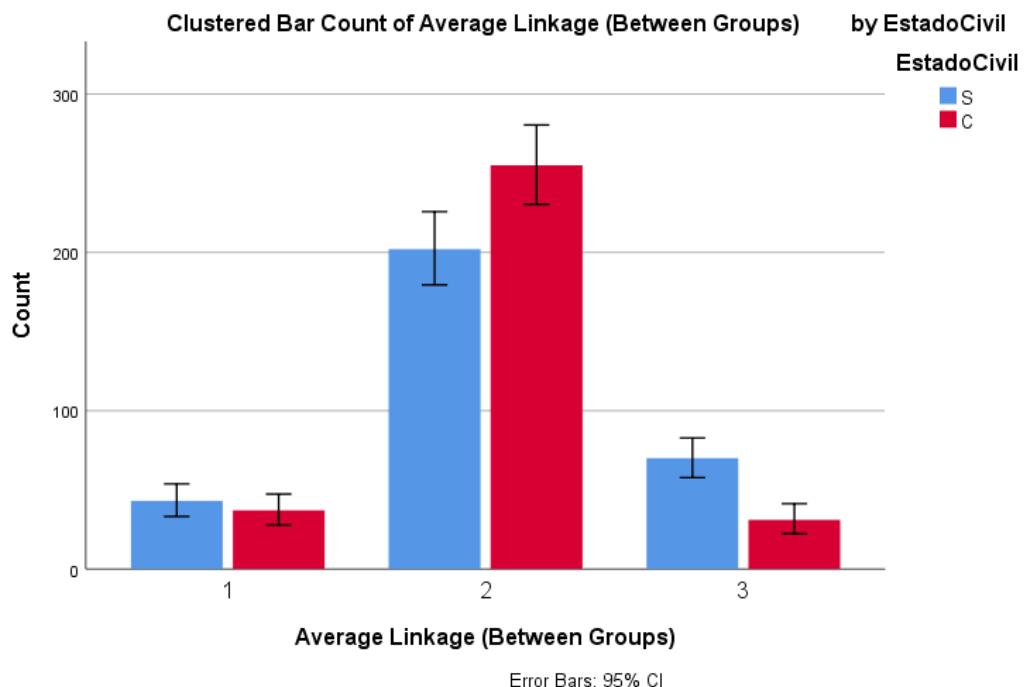
Ocupación VS 2 Clusters



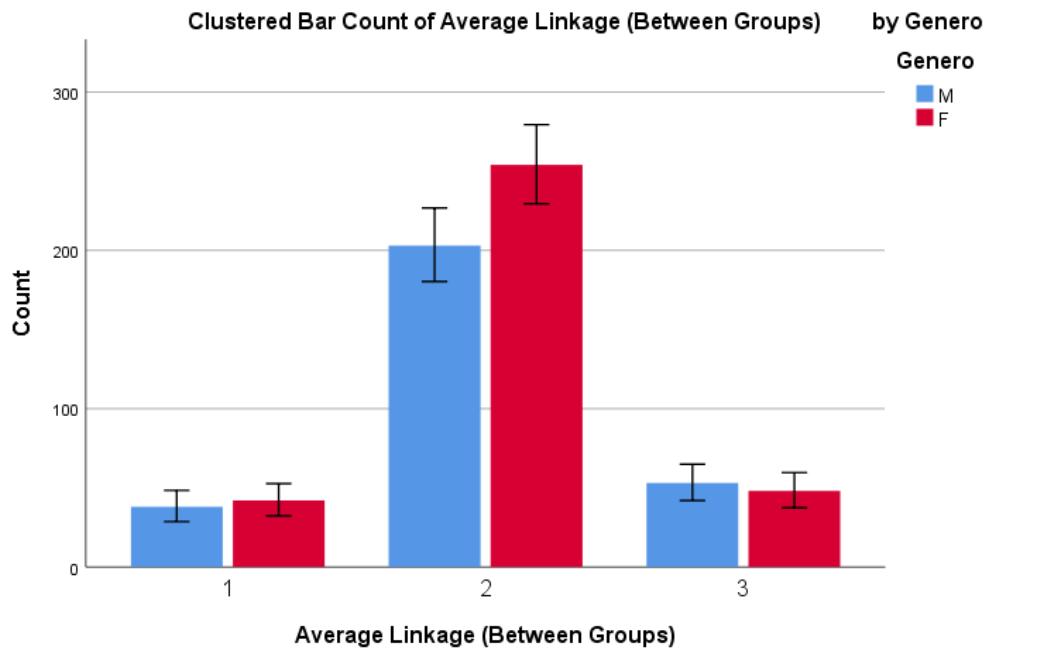
Propietario VS 2 Clusters



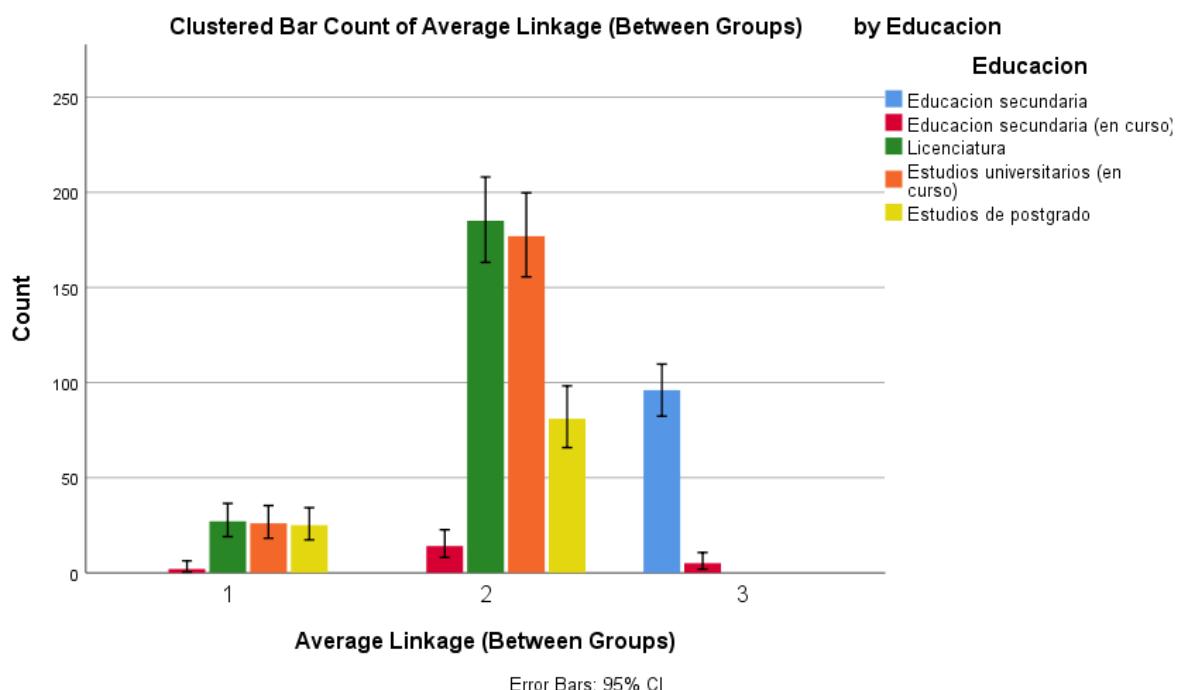
Estado civil VS 3 Clusters



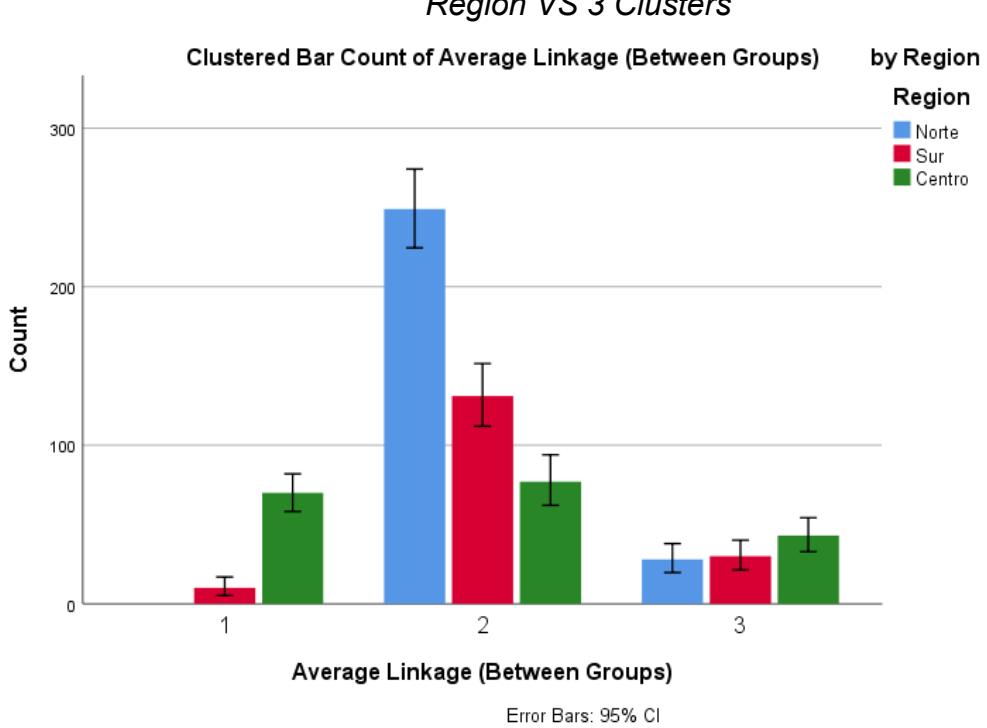
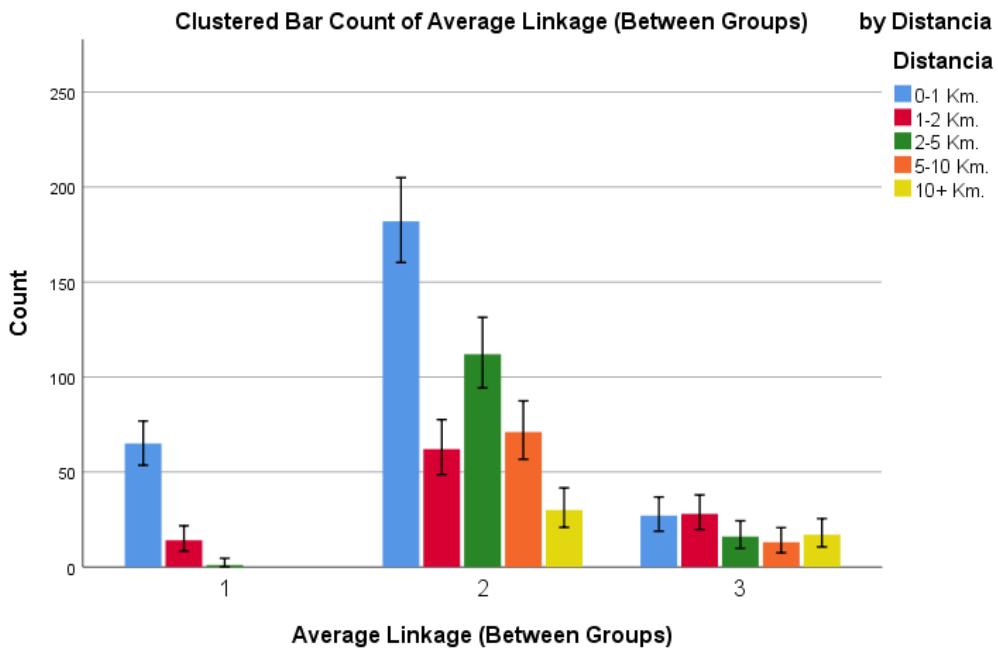
Género VS 3 Clusters



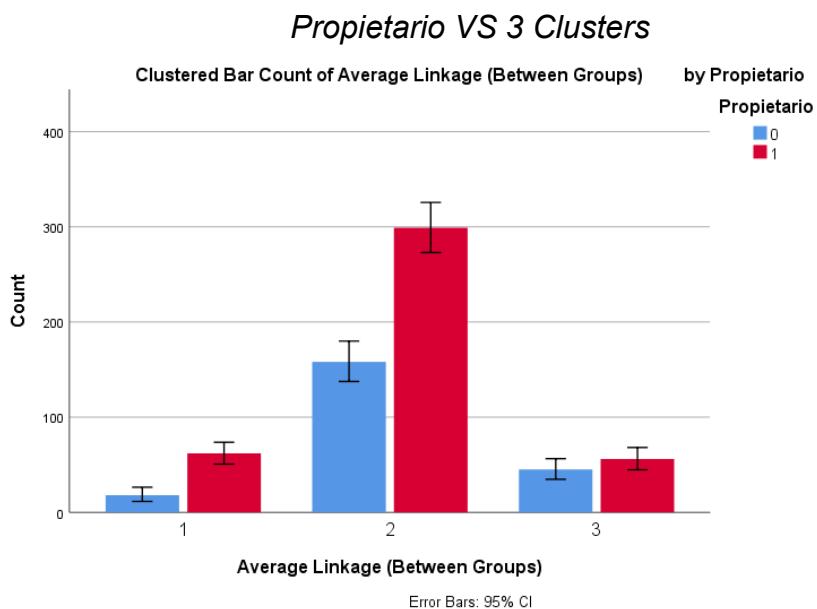
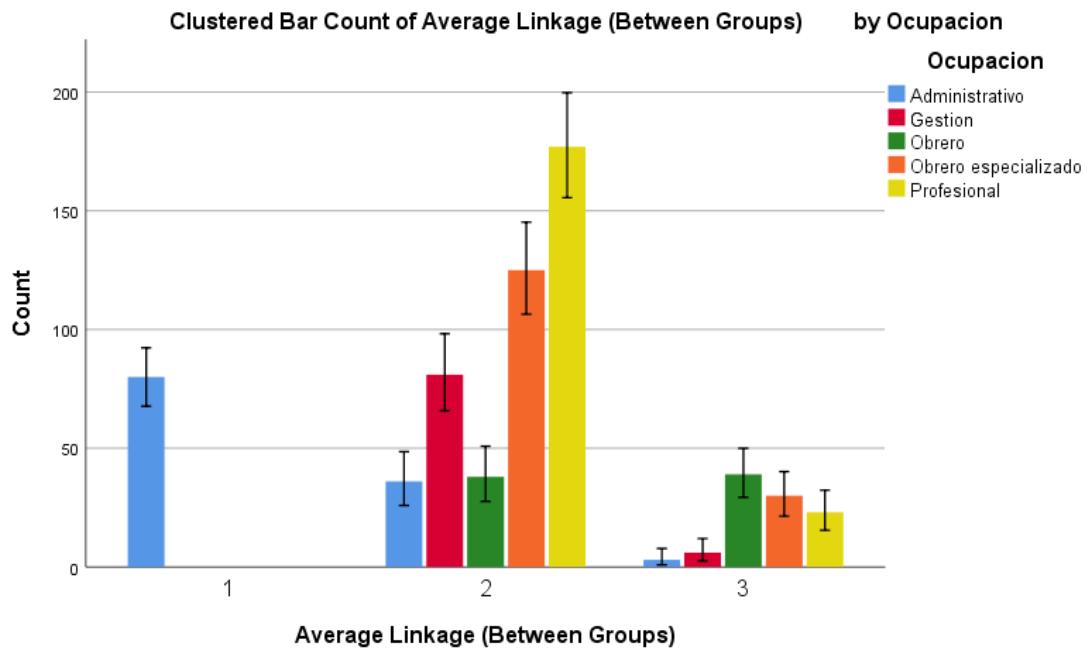
Educación VS 3 Clusters



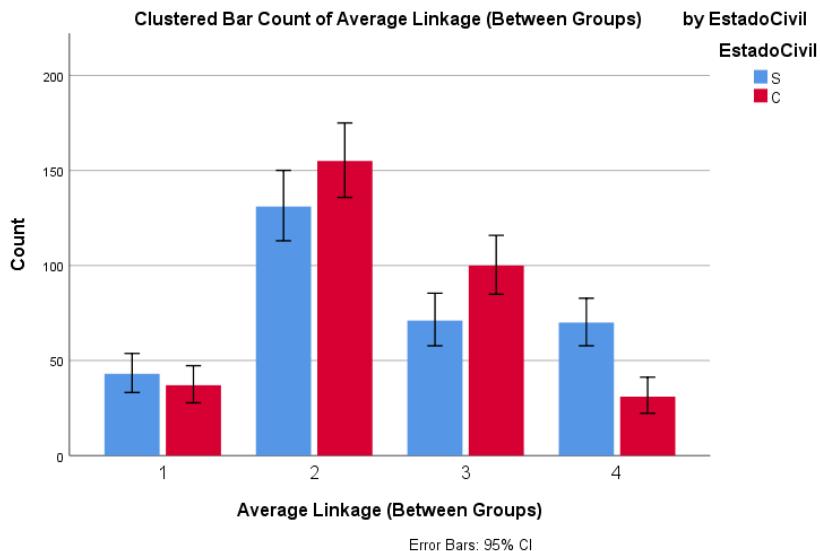
Distancia VS 3 Clusters



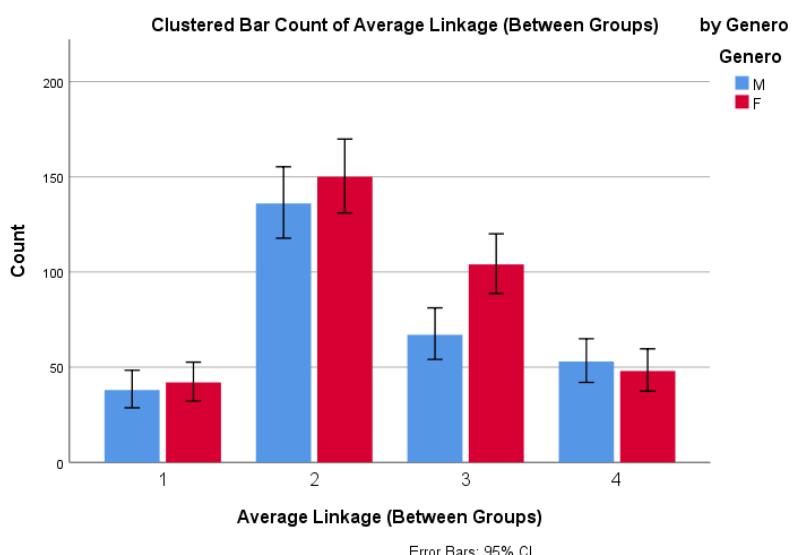
Ocupación VS 3 Clusters



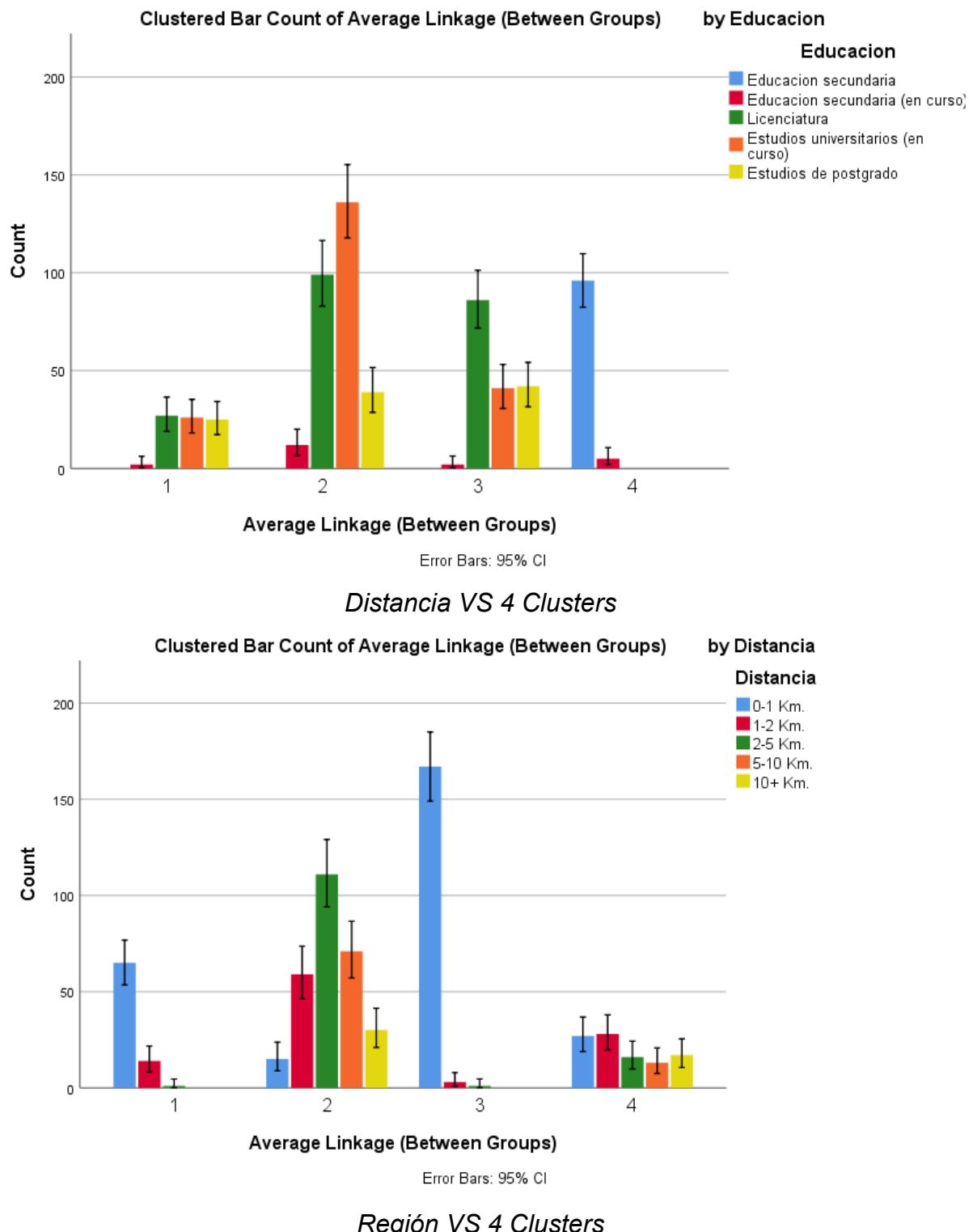
Estado civil VS 4 Clusters

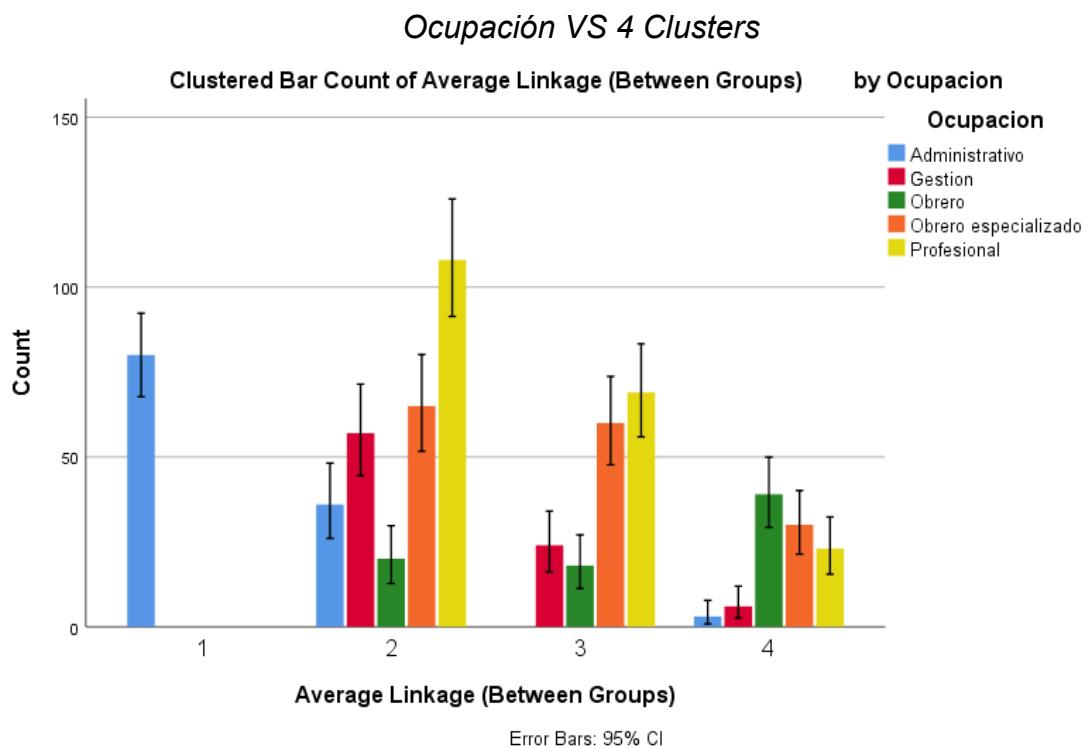
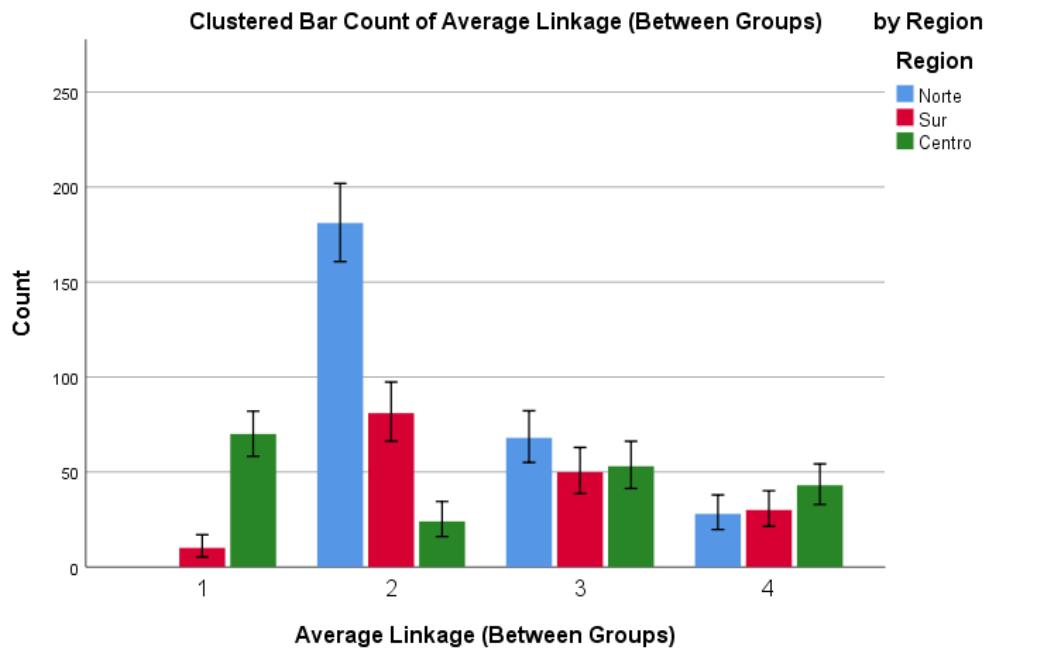


Género VS 4 Clusters

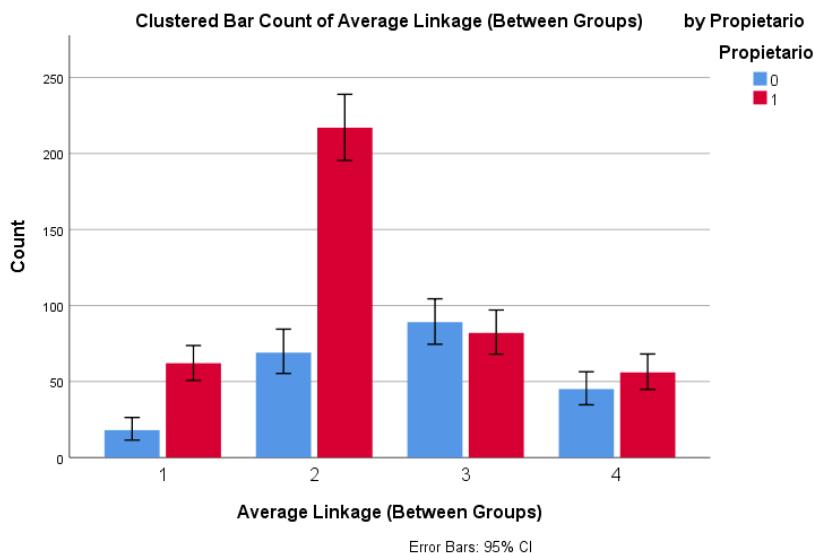


Educación VS 4 Clusters





Propietario VS 4 Clusters



Con las gráficas anteriores se completó la siguiente tabla indicando:

- Un valor si su media superaba y no se solapa con las medias de los otros valores.
- Si los intervalos de dos o más valores se solapan entre sí, colocamos un - debido a que la variable dentro del grupo no se diferencia con la población.
- Si los intervalos en un cluster son similares, pero predominan sobre la población, ingresamos todos los valores de los intervalos que predominan.

	2		3			4			
Variable	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Estado civil	-	Soltero	-	Casado	Soltero	-	-	-	Soltero
Género	Femenino	-	-	Femenino	-	-	-	Femenino	-
Educación	-	Educación secundaria	Lic-UniversitarioCurso-Postgrado	-	Ed. Secundaria	Lic-UniversitarioCurso-Postgrado	Universitario en curso	Licenciatura	Educación secundaria
Distancia	0-1Km	-	0-1 Km	0-1 Km	-	0-1Km	2-5 Km	0-1 Km	-
Región	Norte	-	Centro	Norte	-	Centro	Norte	-	-
Ocupación	Profesional	-	Administrativo	Profesional-Gestión-Obrero Especializado	-	Administrativo	Profesional	-	-
Propietario	Es propietario	-	Es propietario	Es propietario	-	Es propietario	Es propietario	-	-

Encontramos que:

- La variable región, ocupación, y propietario utilizando 4 clusters aporta la misma información que si utilizamos 3 clusters.
- Para cualquier cluster, Género casi ni aporta información y no sirve para agrupar, por lo que se debería eliminar.
- El cluster 2 del grupo 2, coincide con el cluster 3 del grupo 3 y el 4 del grupo 4.
- Utilizando 4 clusters muchas variables no aportan información del grupo.

Por lo tanto, elegiremos 2 clusters, ya que todas las variables aportan información y son diferenciadas entre sí.

Conclusión

Basándonos en los análisis realizados tanto para las variables cuantitativas como cualitativas, consideramos que lo mejor es trabajar con 2 Clusters, puesto que es el que menos variables solapadas tiene, y es en el que más se diferencian los clusters.

Cluster Bietápico

Utilizando el Software SPSS Modeler creamos el siguiente modelo. Antes de realizar el clustering, transformamos las variables Total Hijos, Ingreso Anual, Edad, Distancia y Cantidad de automóviles a categóricas.

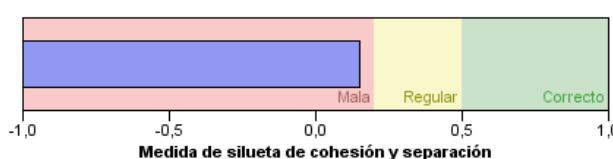


La primera ejecución generó 3 grupos pero el modelo tiene un mal desempeño.

Resumen del modelo

Algoritmo	Bietápico
Entradas	11
Clústeres	3

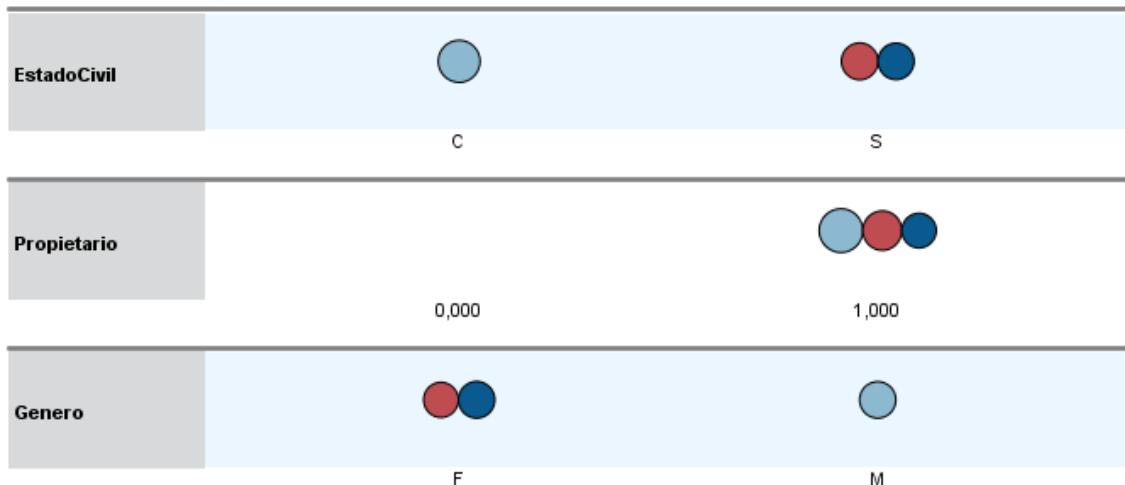
Calidad de clúster



Tamaños de clúster



Tamaño del clúster más pequeño	197 (30,9%)
Tamaño del clúster más grande	233 (36,5%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	1,18



La imagen anterior muestra la comparación de clusters por variables. Se puede observar que la variable propietario no aporta información porque los 3 clusters toman la característica de “Son propietarios”. Por lo tanto, eliminaremos la variable propietario y ejecutaremos de nuevo el modelo para evaluar si mejora.

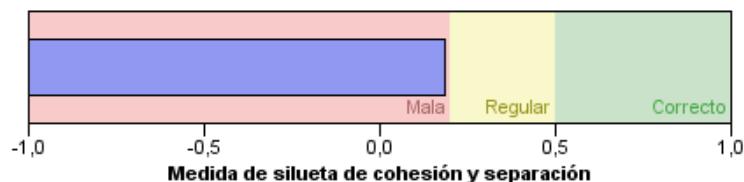
El proceso anterior se repitió 2 veces más eliminando las variables:

1. Género
2. Distancia

Resumen del modelo

Algoritmo	Bielápico
Entradas	8
Clústeres	3

Calidad de clúster



Tamaños de clúster



Tamaño del clúster más pequeño	186 (29,2%)
Tamaño del clúster más grande	251 (39,3%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	1,35

Podemos observar que obtenemos 3 Clústeres cuyo indicador BIC es malo, por lo cual no utilizaremos este modelo.

Criterio del tipo de bicicleta a promocionar

De acuerdo a los dos cluster caracterizados en Cluster Jerárquico, determinamos que:

- Ambos grupos tienen al menos un hijo, por lo que se les enviará publicidad de bicicletas Kinder.
- El Cluster 2, cuenta con personas que tienen ingresos más altos, tienen más automóviles, viven a diferentes distancias de sus trabajos y están solteros, por lo que consideramos una buena opción enviarles publicidad de una bicicleta Sport.
- El Cluster 1, cuenta con personas que tienen ingresos más bajos, son más jóvenes, tienen pocos automóviles y viven cerca de su trabajo, por lo que consideramos que sería correcto enviarles publicidad de las bicicletas Basic.

IdCiudad	Nombre	Apellido	FechaNaci	EstadoCiv	Genero	Email	IngresoAn	TotalHijos	Educacion	Ocupacion	Propietari	CantAutor	Direcc	Tele	Distancia	Region	Edad	EnviarMai	CLU_1	TipoBici
2	Bryant	Perez	#NULL!	0	0	bryant20@e	20000	0	3	2	0	0	1502 N 500 S	0	1	36	1	1	Kinder y Basic	
2	Warren	Zhang	#NULL!	0	0	warren17@	80000	1	3	3	0	1	3905 H 500 S	0	1	57	1	1	Kinder y Basic	
2	Carlos	Edwards	#####	1	0	carlos27@	70000	0	2	4	0	1	5576 V 500 S	0	1	50	1	1	Kinder y Basic	
2	Ariana	Stewart	#####	1	1	ariana21@	40000	2	2	1	1	2	3726 N 500 S	0	1	71	1	1	Kinder y Basic	
2	Renee	Navarro	#NULL!	0	1	renee9@r	70000	0	2	4	0	1	6512 E 500 S	0	1	48	1	1	Kinder y Basic	
2	Jill	Ortega	#NULL!	1	1	jill30@mii	90000	2	2	4	1	0	2824 N 500 S	2	1	45	1	1	Kinder y Basic	
2	Gabriel	Mitchell	#####	0	0	gabriel42@	40000	3	1	0	0	2	6510 N 500 S	1	1	59	1	1	Kinder y Basic	
3	Meredith	Romero	#####	0	1	meredith@	10000	0	3	2	1	1	8335 V 500 S	2	1	33	1	1	Kinder y Basic	
3	Alan	Zhu	#####	0	0	alan17@n	70000	0	2	4	0	1	8995 S 500 S	0	1	48	1	1	Kinder y Basic	
3	Lori	Blanco	2-May-70	1	1	lori16@m	100000	0	1	4	1	4	6966 E 500 S	4	1	43	1	1	Kinder y Basic	
3	Dustin	Sharma	#NULL!	0	0	dustin9@i	80000	2	0	3	0	2	4262 N 500 S	1	1	57	1	2	Kinder y Sport	
3	Michele	Gonzalez	#####	0	1	michele3@	70000	2	0	4	1	2	2748 A 500 S	3	1	55	1	2	Kinder y Sport	
4	Lance	Sanz	#####	0	0	lance20@	10000	0	3	2	0	1	8327 R 500 S	0	1	33	1	1	Kinder y Basic	
4	Ebony	Hernandez	#####	1	1	ebony26@	40000	2	2	1	1	2	7614 I 500 S	3	1	69	1	1	Kinder y Basic	
5	Lori	Moreno	5-Nov-63	0	1	lori9@mir	70000	0	2	4	0	2	792 M 500 S	0	1	49	1	1	Kinder y Basic	
5	Katie	Kumar	9-Sep-76	1	1	katie10@r	80000	0	2	4	1	2	1463 L 500 S	4	1	36	1	1	Kinder y Basic	
5	Kellie	Navarro	#NULL!	1	1	kellie8@n	70000	0	2	4	1	2	1164 A 500 S	4	1	37	1	1	Kinder y Basic	
5	Shawna	Shan	9-Jun-76	0	1	shawna11	10000	1	0	2	1	0	4480 N 500 S	2	1	36	1	2	Kinder y Sport	
6	Kari	Sanz	#####	0	1	kari41@m	80000	0	2	4	0	2	1884 S 500 S	4	1	36	1	1	Kinder y Basic	
6	Levi	Sanchez	#####	0	0	levi18@m	40000	2	2	1	1	2	7226 C 500 S	3	1	70	1	1	Kinder y Basic	
6	Carly	Nath	3-Sep-70	1	1	carly16@r	100000	0	1	4	1	4	5884 N 500 S	4	1	42	1	1	Kinder y Basic	
7	Manuel	Fernandez	#NULL!	0	0	manuel14	10000	0	3	2	1	1	251 R 500 S	1	1	33	1	1	Kinder y Basic	
7	Kaitlin	McDonald	3-Sep-58	1	1	kaitlin14@	100000	1	2	1	1	3	3543 L 500 S	2	1	54	1	1	Kinder y Basic	
8	Clayton	Sharma	#NULL!	0	0	clayton27	10000	0	3	2	0	1	9429 C 500 S	2	1	32	1	1	Kinder y Basic	
9	Gloria	Alvarez	#NULL!	0	1	gloria6@n	70000	0	2	4	0	2	8373 E 500 S	3	1	49	1	1	Kinder y Basic	
9	Lacey	Raje	#####	1	1	lacey4@m	100000	0	1	4	1	4	8986 F 500 S	4	1	43	1	1	Kinder y Basic	

Análisis de Mercado

Realizaremos un análisis de potenciales mercados para la expansión de la marca mediante el método de Análisis de Componentes Principales. El objetivo es encontrar 3 países con características sociales y económicas similares al nuestro para esto disponemos de un dataset con datos sobre los potenciales mercados de distintas ciudades del mundo, incluida la ciudad Buenos Aires.

El ACP es un método que permite simplificar la complejidad de espacios muestrales con muchas dimensiones, mediante la reducción de estas dimensiones de datos.

Permite convertir un conjunto de variables en otro conjunto mucho menor, pero manteniendo la máxima información que proveía el conjunto inicial.

Primero realizamos y analizamos la matriz de correlación para ver si PCA es factible, y si hay variables fuertemente correlacionadas que se puedan reemplazar por una componente.

La matriz de correlación completa se puede observar en el siguiente link:

 Matriz_corr_pca.png . Vista previa:

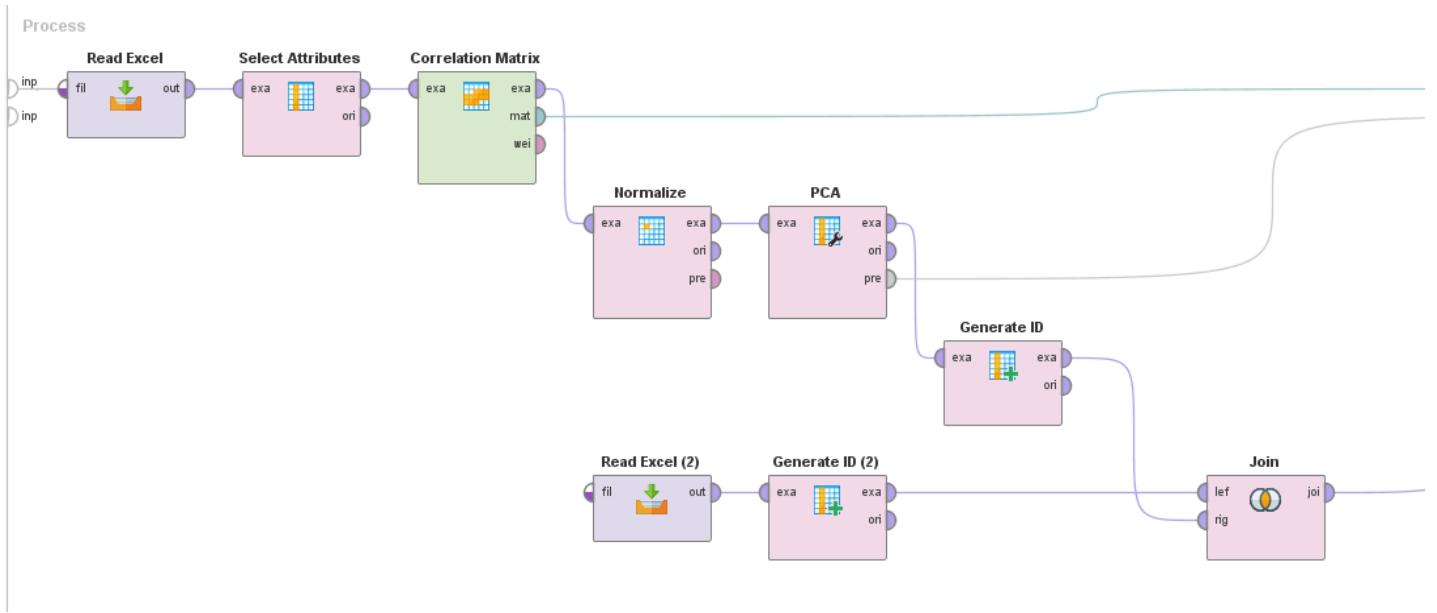
Attributes	Horas ...	Días de...	Inflación...	Inflación...	Inflació...	Inflaci...	Inflació...	Inflaci...	Alquiler ...	Contrib...	Sueldo ...	Sueldo ...
Horas de trabajo promedio [hs/año]	1	-0.637	0.290	0.271	0.223	0.246	0.225	0.222	0.042	-0.575	-0.323	-0.401
Días de vacaciones promedio (por año)	-0.637	1	-0.057	-0.022	-0.027	-0.059	-0.159	-0.134	-0.061	0.306	0.118	0.169
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
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
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
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
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
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
Alquiler departamento 3 ambientes [USD por mes]	0.042	-0.061	-0.079	-0.046	-0.079	-0.149	-0.101	-0.069	1	-0.039	0.556	0.483
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
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
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
Sueldo promedio mecánico de automóviles [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
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
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
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
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
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
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

En este caso podemos apreciar que hay varias variables fuertemente correlacionadas, como por ejemplo:

- Inflación 2007 e inflación 2008 con 0.917.
- Inflación 2009 e inflación 2010 con 0.894.
- Sueldo promedio arquitecto y Sueldo promedio mecánico de automóviles con 0.952, etc.)

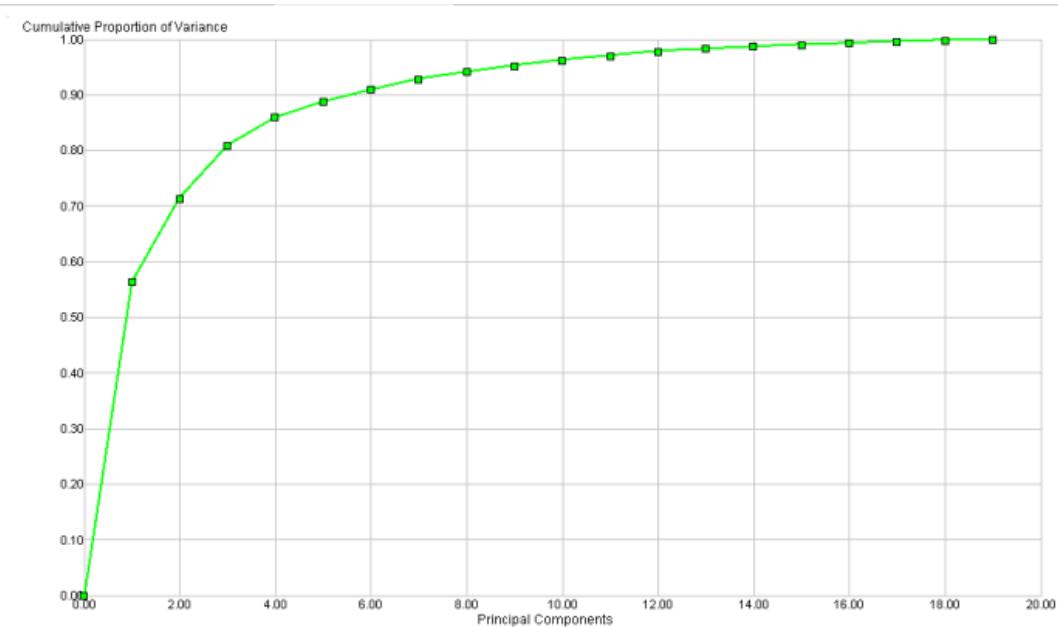
Por lo tanto, tiene sentido continuar realizando PCA.

En RapidMiner realizamos el siguiente modelo:



Utilizando un porcentaje de variabilidad explicativa del 85% obtuvimos 3 Componentes Principales:

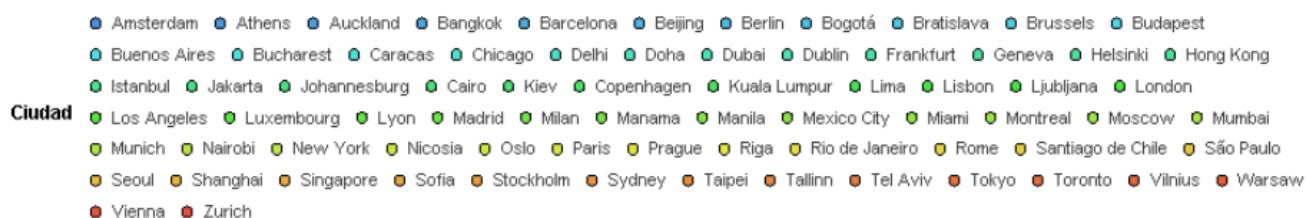
Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	3.277	0.565	0.565
PC 2	1.685	0.150	0.715
PC 3	1.347	0.096	0.810



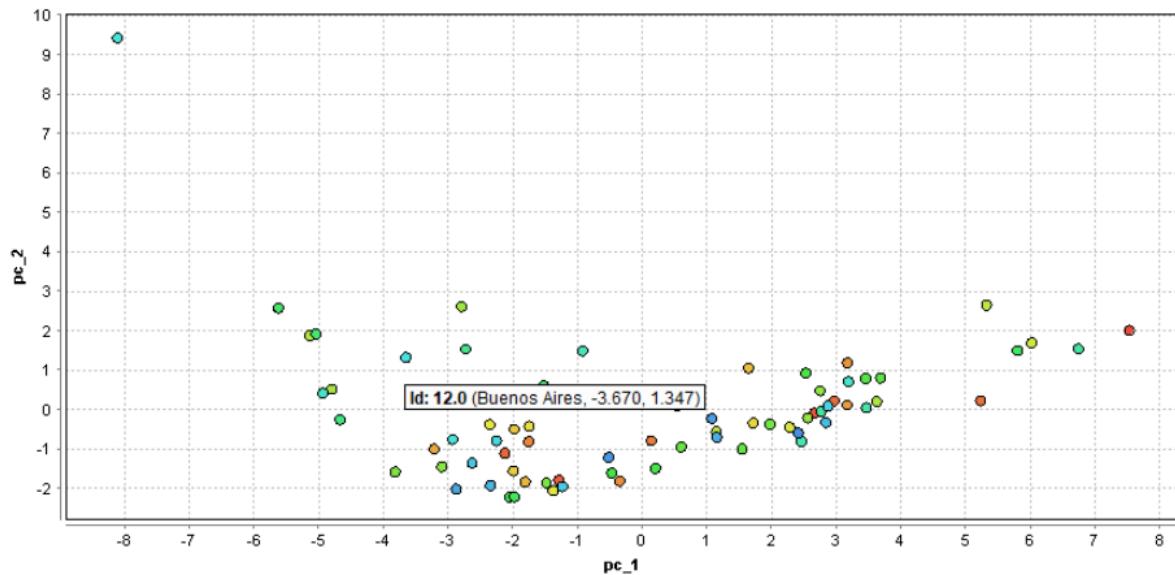
Matriz de componentes para cada componente

Attribute	PC 1	PC 2	PC 3
Horas de trabajo promedio [hs/año]	-0.139	0.010	0.582
Días de vacaciones promedio (por año)	0.062	0.006	-0.576
Inflación 2006	-0.210	0.267	-0.024
Inflación 2007	-0.217	0.320	-0.053
Inflación 2008	-0.220	0.324	-0.095
Inflación 2009	-0.199	0.367	-0.111
Inflación 2010	-0.181	0.378	-0.052
Inflación 2011	-0.200	0.372	-0.076
Alquiler departamento 3 ambientes [USD por mes]	0.143	0.301	0.310
Contribución al seguro social (%)	0.168	-0.026	-0.400
Sueldo promedio maestro de escuela primaria [USD por año]	0.274	0.141	0.081
Sueldo promedio chofer colectivo [USD por año]	0.276	0.141	0.002
Sueldo promedio mecánico de automóviles [USD por año]	0.280	0.151	-0.076
Sueldo promedio arquitecto [USD por año]	0.278	0.153	-0.017
Sueldo promedio cocinero [USD por año]	0.244	0.203	0.117
Sueldo promedio ingeniero [USD por año]	0.272	0.167	0.076
Sueldo promedio secretaria [USD por año]	0.286	0.141	-0.009
Sueldo promedio vendedor [USD por año]	0.283	0.153	-0.034
Sueldo promedio analista financiero [USD por año]	0.271	0.123	0.078

Graficamos las componentes en función de las ciudades y calculamos la distancia euclídea (longitud de la recta que une dos puntos en el espacio) de aquellas ciudades cercanas a Buenos Aires.

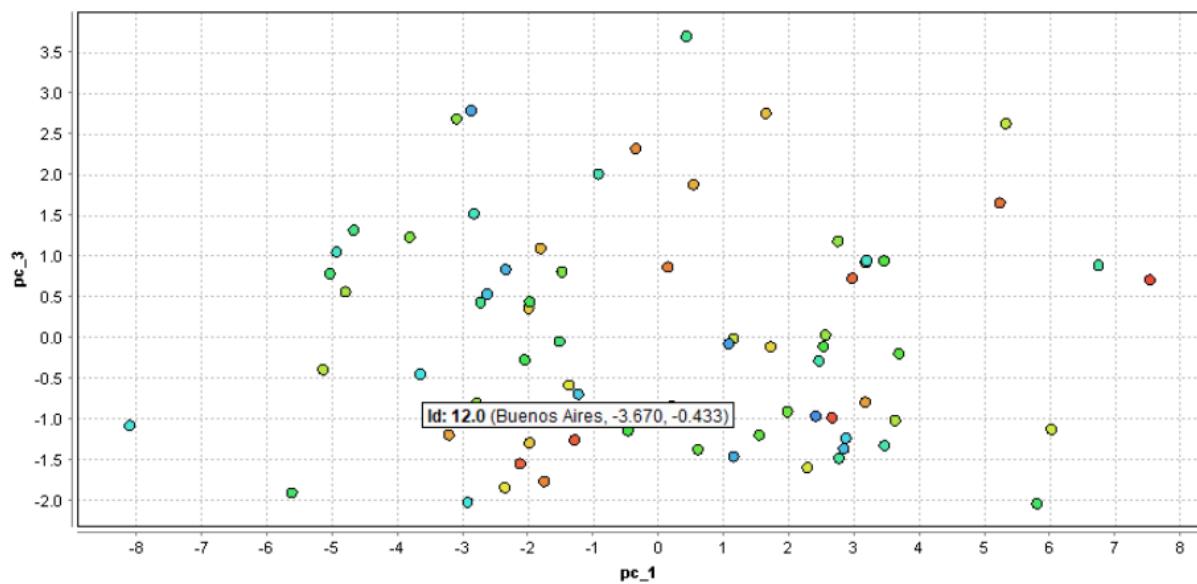


Componente 1 VS Componente 2



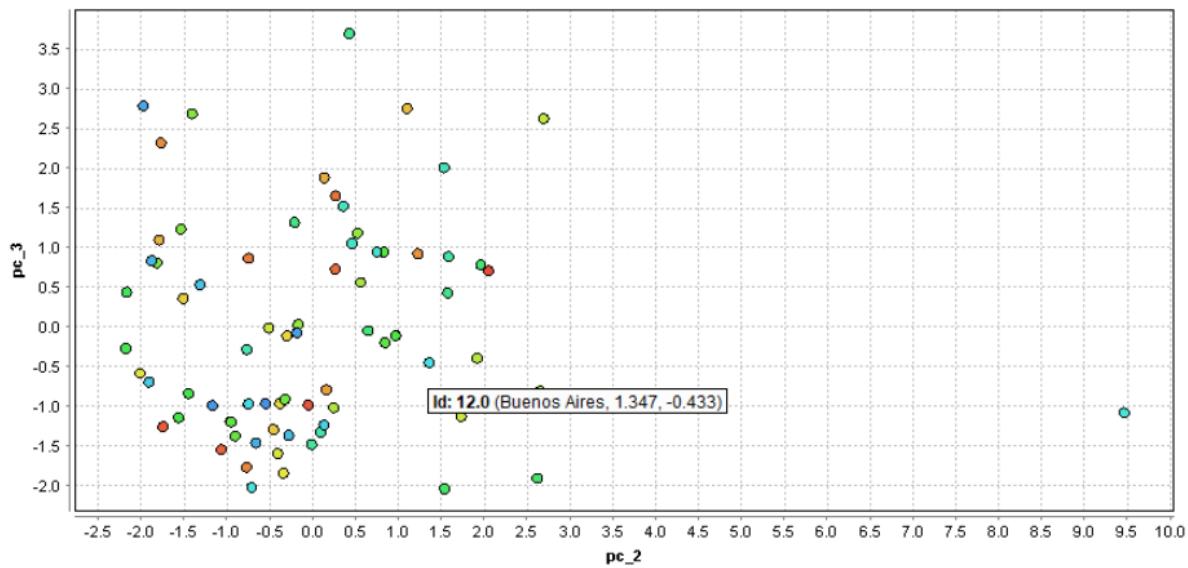
Ciudad	Distancia
Moscú	1,56
Estambul	6,42
Doha	6,594
Mumbai	8,526
Cairo	8,746

Componente 1 VS Componente 3



Ciudad	Distancia
Estambul	0,924
Moscú	0,934
Nairobi	1,49

Componente 2 VS Componente 3



Ciudad	Distancia
Londres	0,52
Nairobi	0,56
Luxemburgo	0,575
Johannesburgo	0,82
Moscú	1,343

Dado que la componente 1 y componente 2 son aquellas que captan más información, acumulando un 71,5% de la variabilidad explicada, utilizamos los resultados de la distancia euclídea de las observaciones de dichas componentes, lo cual concluimos que los mercados más similares a Buenos Aires corresponden a Moscú, Estambul y Doha.

Para verificar la conclusión anterior decidimos tomar los valores originales de algunas de las variables de las ciudades elegidas y compararlas con los valores de Buenos Aires. Como podemos observar, los valores son similares, lo cual respalda la decisión tomada anteriormente.

Ciudad	Buenos Aires	Doha	Estambul	Moscú
Horas de trabajo promedio [hs/año]	1830,000	2165,000	2139,000	1799,000
Inflación 2006	10,898	11,828	9,597	9,679
Inflación 2007	8,830	13,764	8,756	9,007
Sueldo promedio chofer colectivo [USD por año]	16300,000	10400,000	14600,000	18600,000
Sueldo promedio mecánico de automóviles [USD por año]	11900,000	9800,000	13500,000	15800,000
Sueldo promedio secretaria [USD por año]	15800,000	19800,000	13500,000	16800,000
Sueldo promedio vendedor [USD por año]	14600,000	10900,000	9500,000	12200,000