UNIVERSIDAD TECNOLÓGICA NACIONAL

FACULTAD REGIONAL ROSARIO

Cátedra: Minería de Datos

5º Año Ingeniería en Sistemas de Información

Trabajo Práctico Integrador

Comisión Nº: E504

Alumnos:

Bertone Andrés - Legajo N° 44989
Bianchi Ignacio - Legajo N° 44863
Labanca Francisco - Legajo N° 44843
Vives Camila - Legajo N° 44787

Profesores:

Ing. Cristian Bigatti
Mg. Juan Miguel Moine

Año 2022

Análisis del problema

Definición del problema

La empresa AllHome se dedica a la venta de una amplia gama de productos y busca impulsar la comercialización de bicicletas a través de un convenio con una reconocida marca nacional, la cual fabricará tres tipos de estas: para niños (Kinder), estándares (Basic) y deportivas (Sport).

Para esto, una de las estrategias propuestas se basa en una campaña de publicidad por correo electrónico. De esta forma, el problema a abordar consiste en caracterizar a los potenciales clientes de la empresa para establecer hacia cuáles de ellos focalizar la nueva estrategia de marketing. Para ello, contamos con un archivo con formato csv que contiene datos históricos de cada cliente y si alguna vez compró o no una bicicleta. Asimismo, tenemos a disposición otro archivo en formato txt con datos de 1.500 potenciales clientes para decidir si se le envía o no la publicidad. Sobre estos archivos debemos analizar y extraer información que nos resulte útil en el desarrollo del trabajo.

A su vez, se está evaluando la posibilidad de vender las bicicletas en mercados extranjeros de características sociales y económicas parecidas al nuestro. Se nos solicita que recomendemos cuáles serían los mercados candidatos para la comercialización de las bicicletas. Por lo tanto, haremos uso del archivo xlsx que se nos provee y que contiene información de diferentes mercados, analizando cada uno de ellos para determinar y seleccionar los tres principales en donde la campaña de comercialización será más exitosa.

Objetivos

El objetivo general que busca cumplir este trabajo consiste en aplicar los conocimientos adquiridos durante el transcurso de la materia, elaborando modelos y luego comparándolos entre sí para poder implementar el que, según nuestro criterio, resuelve de manera efectiva el caso en estudio.

Por otro lado, existen objetivos específicos que se desprenden del dominio del problema y que se enumeran a continuación:

- Analizar los datos registrados en los archivos disponibles para caracterizar y predecir los potenciales clientes de las nuevas líneas de bicicletas con el fin de direccionar correctamente la publicidad hacia esas personas.
- Determinar qué tipo de bicicleta (Kinder, Basic o Sport) le vamos a ofrecer a cada cliente.
- Recomendar estrategias de comercialización de las nuevas líneas de bicicletas al gerente de ventas a partir del estudio de los datos de los mercados extranjeros candidatos.

Técnicas a utilizar

Para el desarrollo del trabajo se utilizarán tres técnicas diferentes, las cuales se definen a continuación.

- **Árbol de decisión**

  Los árboles de decisión son una serie de condiciones organizadas en forma jerárquica, donde cada uno de los caminos lleva a una clasificación o decisión a tomar. Esta técnica permite determinar, a partir de la elección de un atributo como raíz, las variables más significativas para un nuevo elemento dado.

- **KNN**

  El método K-Nearest Neighbor se basa en clasificar un nuevo caso en función de la distancia con los casos vecinos, los cuales fueron etiquetados anteriormente. En este caso se utilizará la distancia euclídea para determinar la cercanía entre casos.

- **LDA (Análisis discriminante lineal)**

  El análisis discriminante es una técnica estadística que ayuda a identificar las características que diferencian a dos o más grupos. Esto se logra a través de la búsqueda de la combinación lineal de variables independientes que mejor permitan discriminar a dichos grupos (función discriminante). Una vez obtenida esta función discriminante, podrá ser utilizada para clasificar nuevos casos.

  Es importante aclarar que este método sólo permite utilizar variables independientes cuantitativas continuas, o que al menos admitan un tratamiento numérico significativo.

La elección de estas técnicas tiene que ver con que la tarea a realizar es predictiva, es decir, se requiere encontrar patrones en los datos de forma que se pueda clasificar una nueva instancia desconocida por el modelo. Todas las técnicas seleccionadas permiten realizar tareas de este tipo.

A su vez, existen diferencias entre las técnicas elegidas, las cuales permiten comparar los distintos modelos obtenidos, para luego poder seleccionar el que mejor se ajuste a los datos del problema.

En primer lugar, un árbol de decisión es anticipativo y KNN es perezoso o retardado. Que una técnica sea anticipativa significa que construye un modelo a partir de los datos provistos, por lo que para analizar un dato nuevo se consulta el modelo creado directamente. Por otro lado, que sea perezoso, implica que se va a buscar entre los datos existentes el más parecido a un nuevo caso y se va a actuar en función de ello.

A diferencia de los métodos de árboles y KNN, el análisis discriminante lineal (LDA) permite diferenciar grupos de observaciones que tienen características similares entre ellas y que a su vez son diferentes de las características de otros grupos. Los nuevos casos serán entonces clasificados en alguno de estos grupos diferenciados.

Pre-procesamiento de datos

Selección y transformación de datos

En esta fase del proceso de extracción de conocimiento nos encargamos de trabajar los datos que se nos proveen. Vale aclarar que la fase uno de “Integración y recopilación” fue omitida dado que ya contamos con los datos iniciales. Los datos provienen de diferentes fuentes, presentan gran volumen y hasta pueden contener errores o inconsistencias. Por eso, la etapa de selección y transformación adquiere una mayor relevancia.

A las variables **total de hijos, cantidad automóviles y propietario** decidimos cambiarle el tipo de dato de numérico a polinomial para realizar un análisis univariante más exacto. En el caso de la variable **compró bicicleta**, se hizo una transformación del tipo de dato numérico a binomial.

Análisis exploratorio

Para realizar un análisis óptimo y lograr el mejor entendimiento de los datos, decidimos seleccionar, desde el conjunto de datos del archivo **clientes.csv**, aquellas variables que creemos más importantes. A continuación, se enumeran estas variables:

- Estado civil
- Género
- Ingreso anual
- Total hijos
- Educación
- Ocupación
- Propietario
- Cantidad automóviles
- Distancia al trabajo
- Región
- Edad
- Compro bicicleta

La primera actividad que llevamos a cabo en el análisis exploratorio fue un vistazo general del archivo con el cual vamos a trabajar en este momento. Notamos que la cantidad de registros era de 6.500. Al abrir el archivo en RapidMiner como csv notamos que la cantidad de registros disminuye a 6.497. Esa particularidad nos llamó la atención, por lo que decidimos convertirlo a formato Excel (.xlsx). Después de esta conversión, abrimos el nuevo archivo en el software y logramos que se muestren los 6.500 registros originales, por lo cual seguiremos el desarrollo del trabajo práctico con este nuevo archivo.

El próximo paso a desarrollar es el análisis descriptivo, es decir univariante, de cada una de las variables nombradas con anterioridad y luego un análisis en donde relacionamos una o más variables, esto es el análisis multivariante. Para ambos, haremos uso de gráficos tales como histogramas, diagramas de caja y bigote (box plot) para las variables numéricas, y diagramas de barras y de torta para aquellas

que son cualitativas. Además, para el análisis multivariante, haremos uso de los gráficos de dispersión (scatter), de barras apiladas, histogramas con color para diferenciar cada variable, entre otros.

Un punto para destacar en esta etapa es que detectamos que existen 11 valores faltantes en la columna **ingreso anual**. Intuímos que una de las razones de la información faltante puede deberse a que el entrevistado no quiso brindar ese dato por cuestiones de privacidad.

A continuación, presentamos una tabla a modo de resumen de la cantidad de observaciones y las variables predictoras y predicha.

Observaciones	6.500
Variables predictoras	EstadoCivil, Genero, IngresoAnual, TotalHijos, Educación, Ocupación, Propietario, CantAutomoviles, Distancia, Region, Edad
Variable predicha	ComproBicicleta

Análisis univariante

Tabla para variables numéricas:

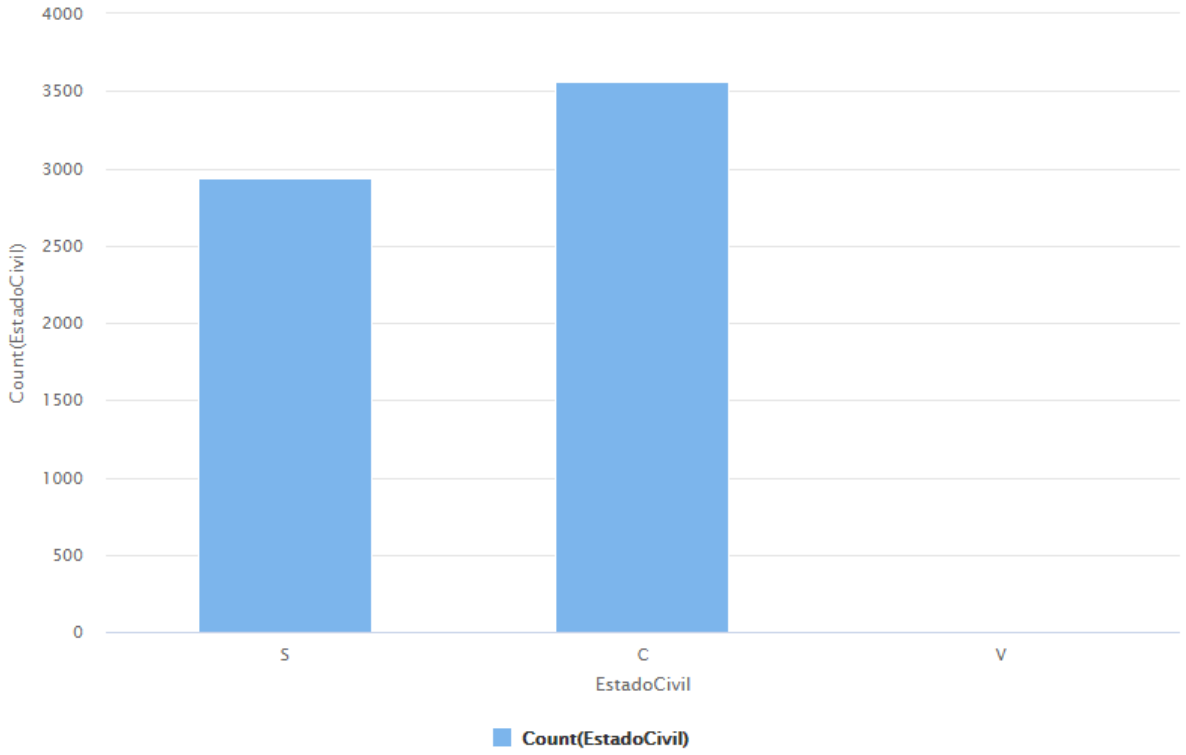
Variable	Unidad de medida	Media o promedio	Mediana	Desvío estándar	Asimetría
Ingreso Anual	Pesos	57.535,83	60.000	32.367,95	Hacia la derecha
Edad	Años	51.174	49	11.495	Hacia la derecha

Tabla para variables polinomiales:

Variable	Descripción	Valor con menos apariciones	Valor con más apariciones
Estado Civil	Variable que describe si el cliente que compró era Separado/a, Casado/a o Viudo/a	V (2)	C (3.562)
Género	Variable que describe el género del cliente	F (3.226)	M (3.274)
TotalHijos	Variable convertida a polinomial que indica 5 categorías, relativas a la cantidad de hijos que posee cada uno de los clientes registrados	5 (545)	0 (1.774)
Educación	Representa las diferentes categorías de estudio que posee cada cliente	'Educación secundaria (en curso)' (565)	Licenciatura (1.829)
Ocupación	Representa las diferentes categorías de ocupación que posee cada cliente	Obrero (841)	Profesional (1.975)
Propietario	Variable polinomial que indica con 1 si el cliente es propietario, de lo contrario es 0	0 (2.104)	1 (4.396)
CantAutomoviles	Variable para describir la cantidad de autos que posee cada cliente	4 (486)	2 (2.337)
Distancia	Representa la distancia al trabajo en km	‘10+ km’ (934)	‘0-1 km’ (2.202)
Región	Indica la región en la cual vive cada cliente	Noroeste (1)	Norte (3.359)

Variable **Estado Civil**

La variable Estado Civil tiene un tipo de datos polinomial, que acepta tres valores distintos: ‘S’ que indica soltero, ‘C’ que indica casado y ‘V’ que indica viudo. Su análisis se realizará con un diagrama de barras.

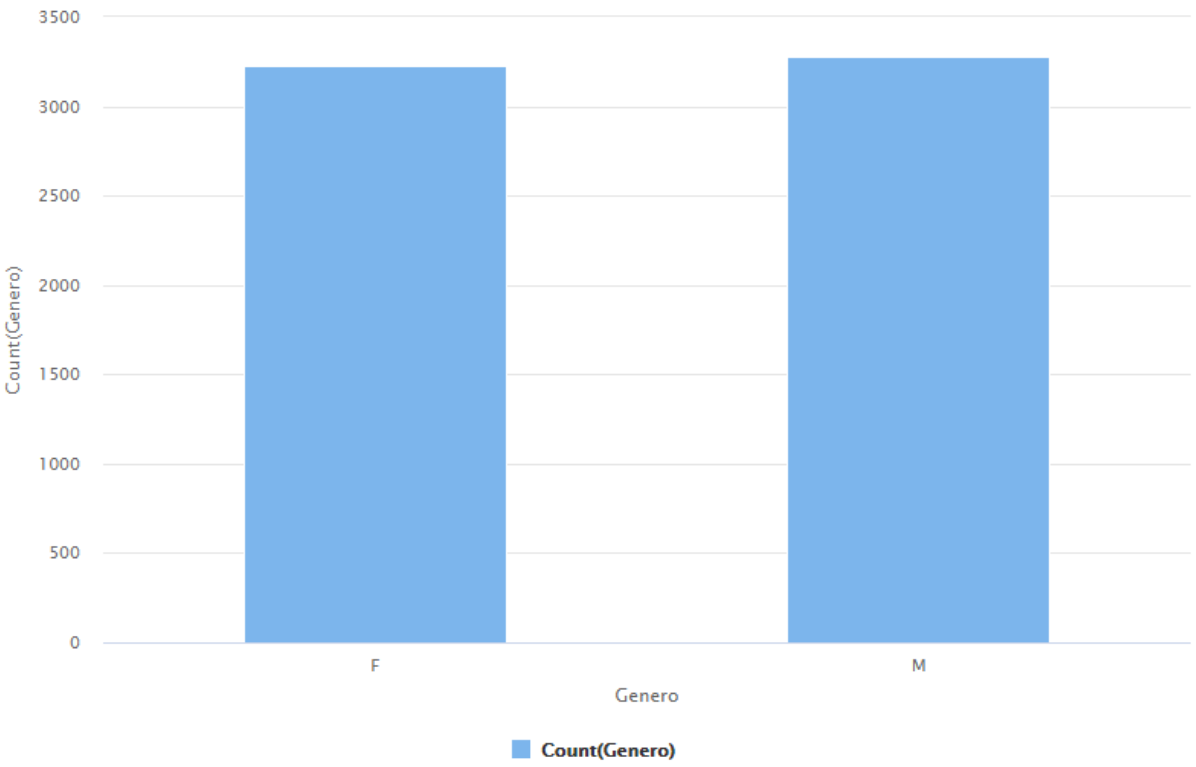


Del gráfico podemos concluir que, casi la totalidad de los clientes son **solteros** o **casados**. El porcentaje que representa a los clientes con estado civil **solteros** es de 45,1% (2.936 clientes) y el de **casados** es de 54,8% (3.562 clientes). Existe una cifra despreciable de clientes **viudos** con un valor de tan sólo 2 observaciones.

Dado que el conjunto de datos posee un total de 6.500 registros, y por lo tanto no causará un sesgo importante en los datos, hemos decidido suprimir los 2 registros que han sido clasificados en la categoría de viudos.

Variable **Género**

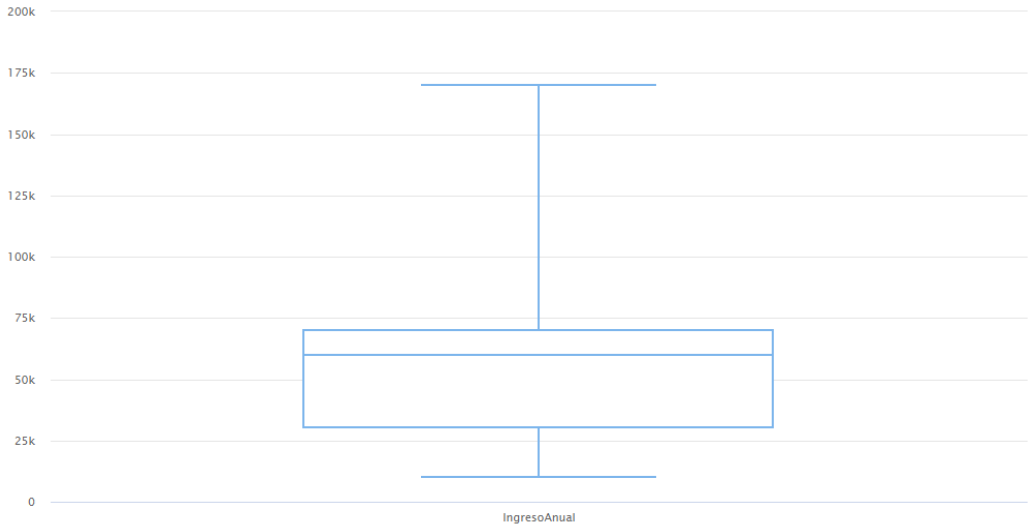
Al igual que la variable anterior, la variable Género es polinomial, indicando que las observaciones se clasifican con ‘M’ si el género del cliente es masculino y ‘F’ si es femenino.



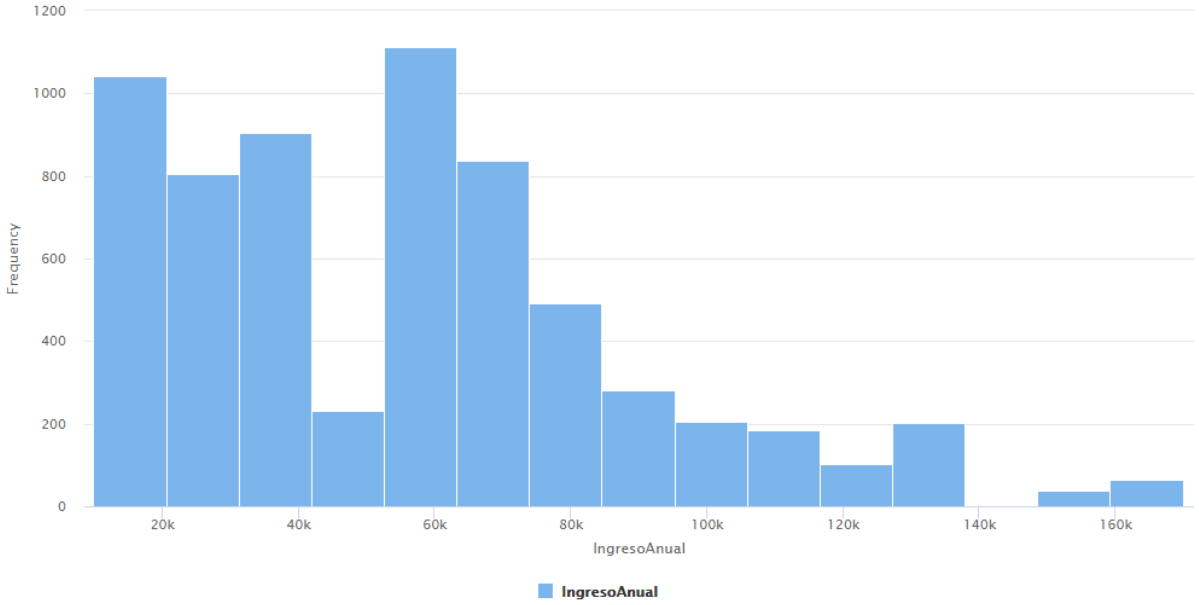
Viendo la proporción y cantidad de clientes **femeninos** (49,6%) y **masculinos** (50,3%) notamos que las cantidades son bastante similares, con una diferencia muy pequeña de únicamente 48 clientes a favor de aquellos clientes con género **masculino**.

Variable **Ingreso Anual**

Dado que la variable ingreso anual es un valor numérico continuo, el análisis se realizará a través de dos gráficos, que son el diagrama de caja y bigotes y el histograma. Es importante recordar que para esta variable se ha hecho un tratamiento de los datos faltantes reemplazando estos valores por la media del conjunto de datos.



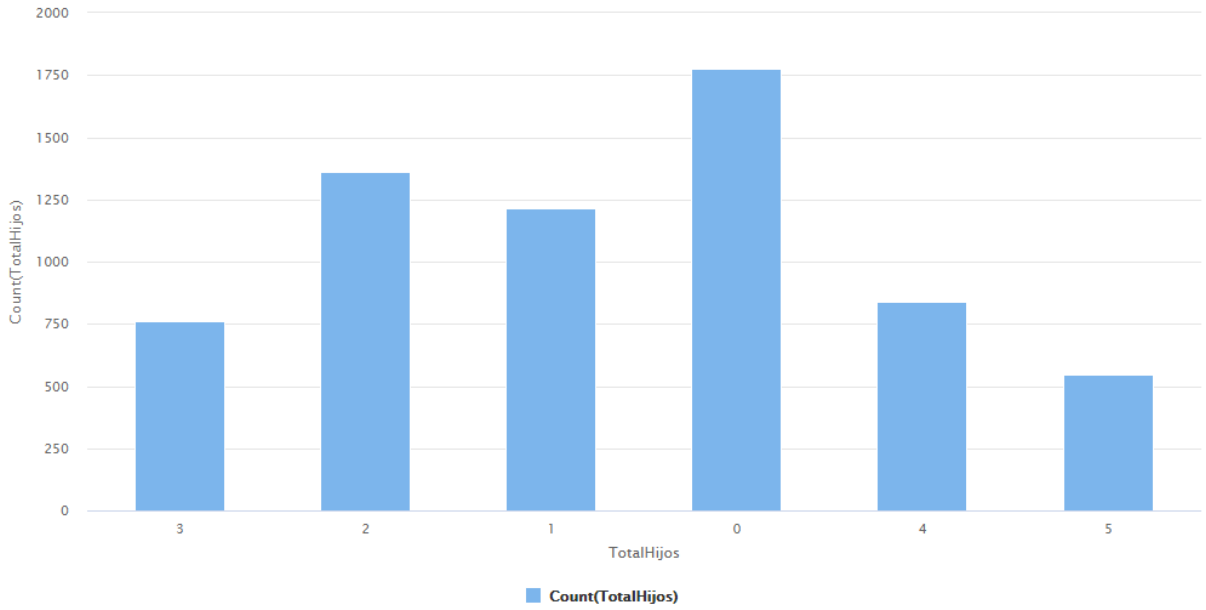
A partir del diagrama de la figura, se puede extraer información sobre los distintos valores estadísticos, que indican que el menor valor de ingreso anual es $10.000, el mayor de ellos es $170.000 y la mediana es igual a $60.000. Por otro lado se observa que la caja está dividida por la mediana de forma asimétrica, siendo la parte inferior mayor a la superior. Esto indica que los ingresos de los clientes registrados comprendidos entre el 25% y el 50% del valor tienen una mayor dispersión que los que se encuentran entre el 50% y el 75%. Además, dado que el bigote inferior es más corto que el superior, lo que indica que el 25% de los valores bajos de ingreso anual están más concentrados que el 25% de los valores altos.



Observando el histograma puede notarse que la distribución de la variable tiende a ser asimétrica hacia la derecha. Esto refuerza la hipótesis presentada junto al gráfico anterior, de que los valores de ingreso anual están más concentrados en valores bajos.

Variable **Total de Hijos**

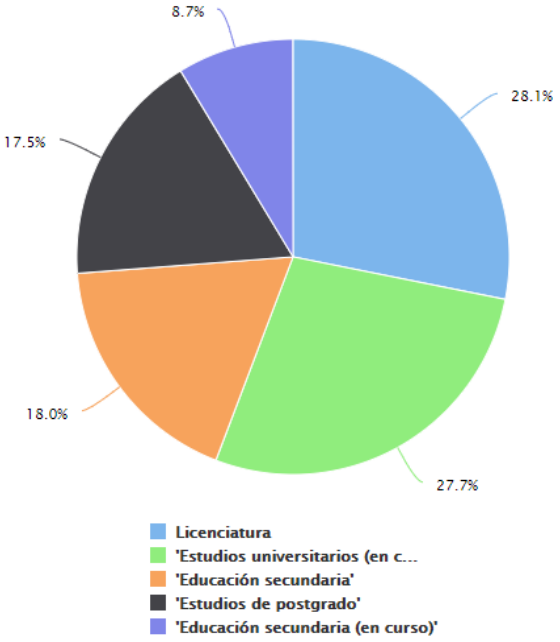
En este caso la variable convertida a polinomial indica 6 categorías, relativas a la cantidad de hijos que posea cada uno de los clientes registrados.



A partir del análisis de esta variable podemos notar que la mayor cantidad de clientes registrados no tiene hijos. Por otro lado, en el gráfico puede notarse que se da con mayor frecuencia que una observación se encuentre en las categorías 1 y 2 que en las categorías 3, 4 y 5.

Variable **Educación**

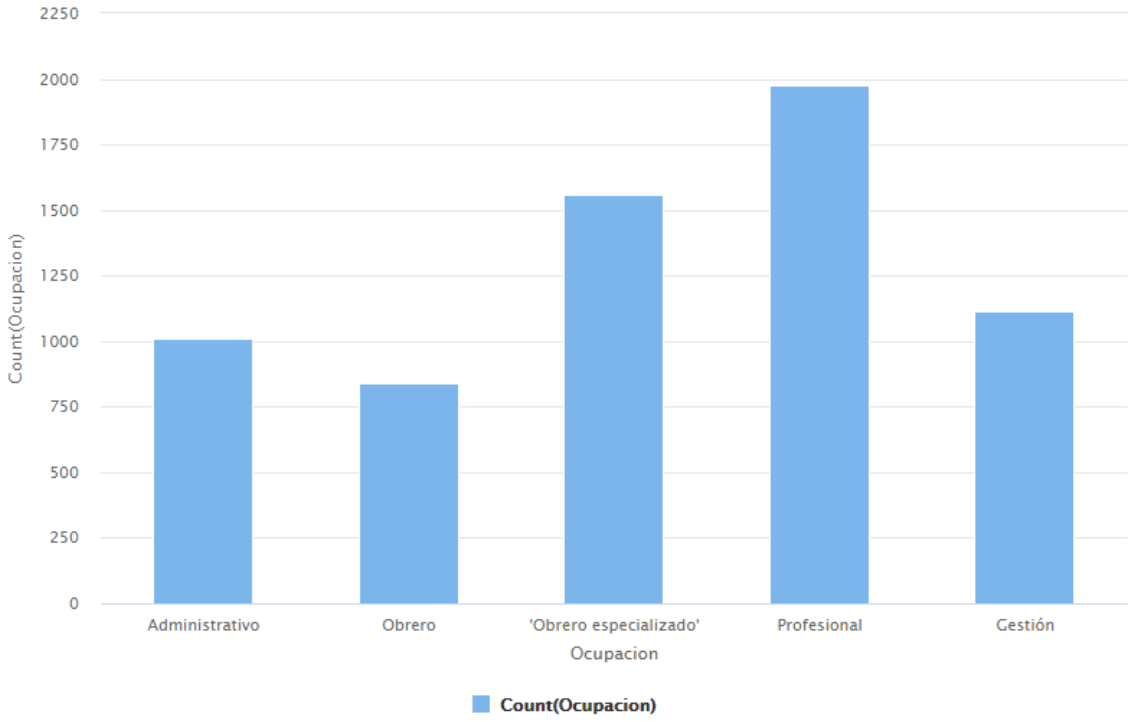
Para analizar la variable Educación se hará uso de un gráfico de torta, representando cada categoría.



Este gráfico nos muestra que casi el 75% de los clientes encuestados tiene formación universitaria o superior, en donde poco más de la mitad tiene un título vigente mientras que el resto está en proceso de obtenerlo.

Variable **Ocupación**

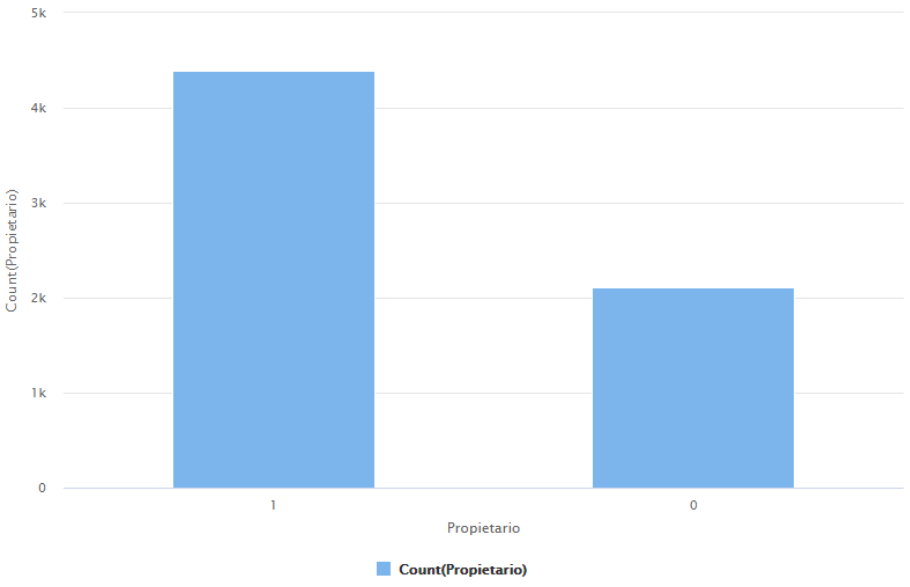
En este caso se utilizará un gráfico de barras para analizar la cantidad de observaciones que pertenecen a cada una de las categorías de ocupación.



El gráfico nos muestra la distribución de las profesiones de los clientes. Entre ellos la mayor parte son obreros (especializados o no) y existe una gran cantidad de profesionales. Completan la información ocupaciones de gestión y administrativos.

Variable **Propietario**

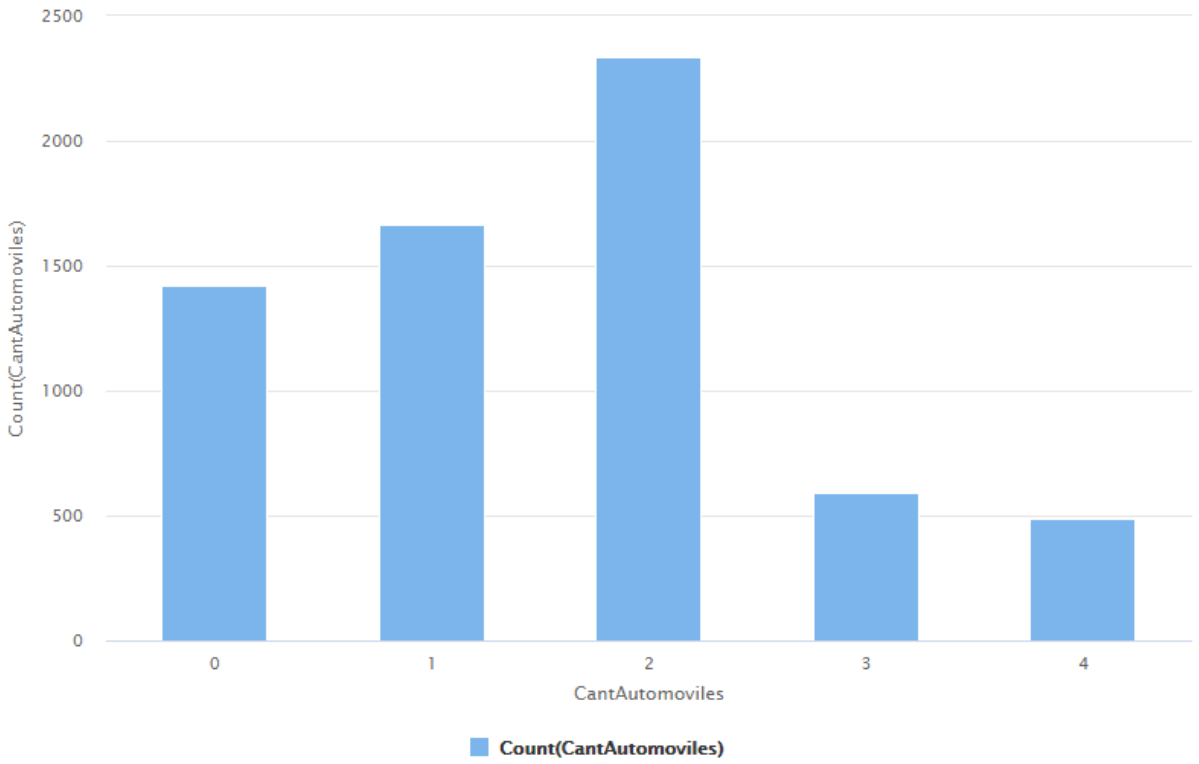
La variable propietario solo admite dos valores, siendo 0 para aquellos clientes que no poseen propiedades y 1 para los que sí lo hacen.



Interpretando el gráfico anterior, podemos ver, que de todos los clientes presentes en nuestro archivo “clientes.csv”, un total de 4.396 (67.63%) son propietarios, mientras que 2.104 (32,37%) no son propietarios y se podría considerar que son inquilinos, es decir, que alquilan.

Variable **Cantidad de automóviles**

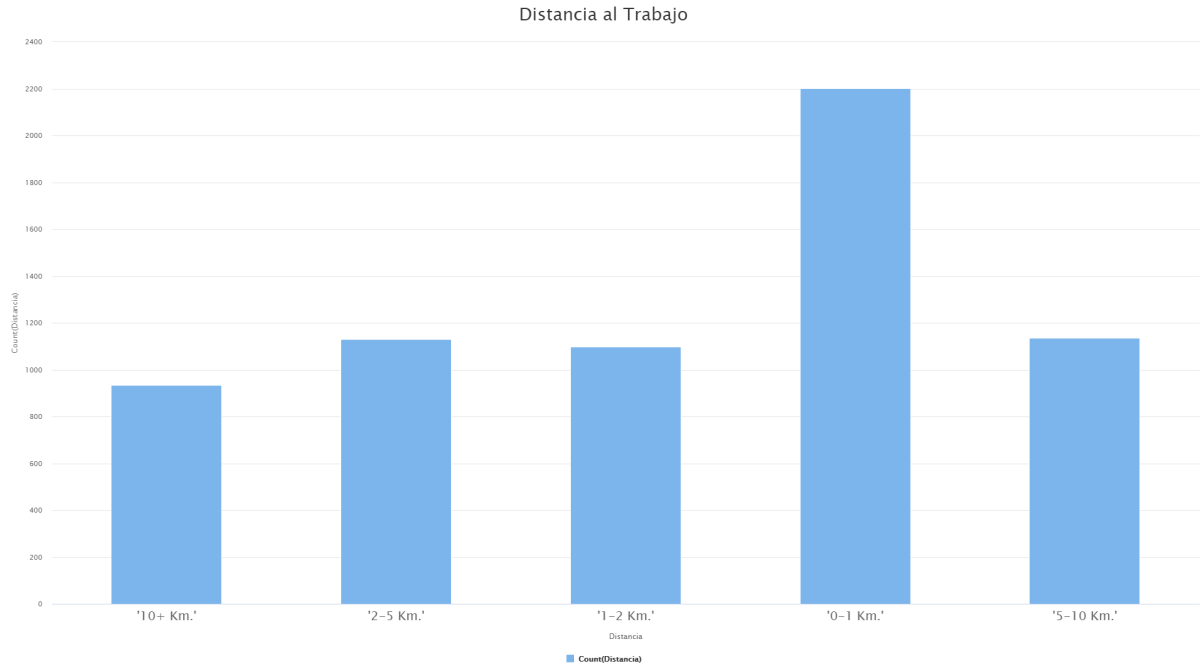
La variable cantidad de automóviles es del tipo polinomial y por lo tanto se representará a través de un gráfico de barras.



En este gráfico de barras, cada clase representa la cantidad total de automóviles que posee cada cliente, siendo estas: 0, 1, 2, 3 y 4 automóviles. Se aprecia a simple vista que los clientes que poseen 2 automóviles representan la mayoría de las observaciones, con un total de 2.337. Le siguen las clases que indican una cantidad de 1 automóvil con 1.665 observaciones y la clase 0 con 1.419. Las que menos datos presentan son las clases 3 y 4 que tienen 593 y 486 observaciones, respectivamente.

Variable **Distancia al trabajo**

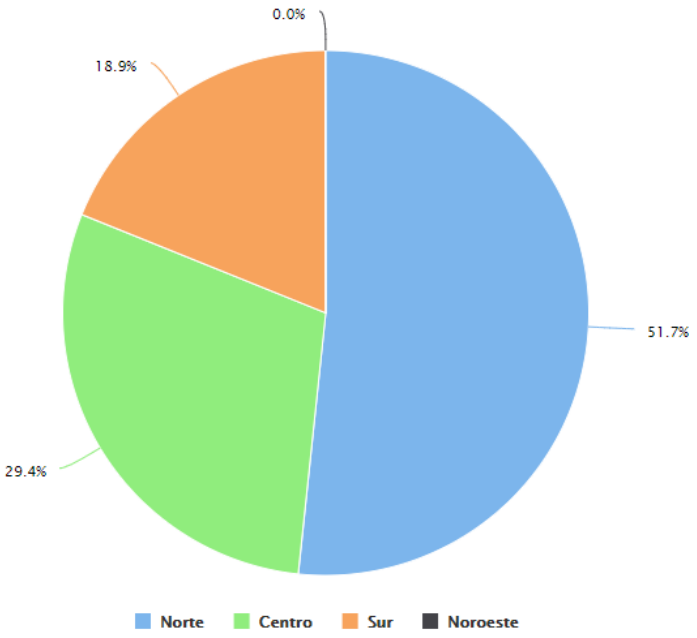
Dado que la variable distancia se encuentra registrada utilizando etiquetas que especifican intervalos de distancia desde su hogar hacia su lugar de trabajo, se ha decidido utilizar un gráfico de barras para analizar la frecuencia de cada categoría.



En este caso puede verse que la mayor cantidad de clientes deben recorrer una distancia menor a 1 km para llegar a sus trabajos. Por otro lado, las demás categorías se encuentran con valores similares entre ellas.

Variable **Región**

La variable región indica para cada instancia de cliente, si este vive en la región Norte, Sur, Centro o Noroeste. Para graficar la frecuencia de cada categoría en el conjunto de datos se utilizó un gráfico de torta.

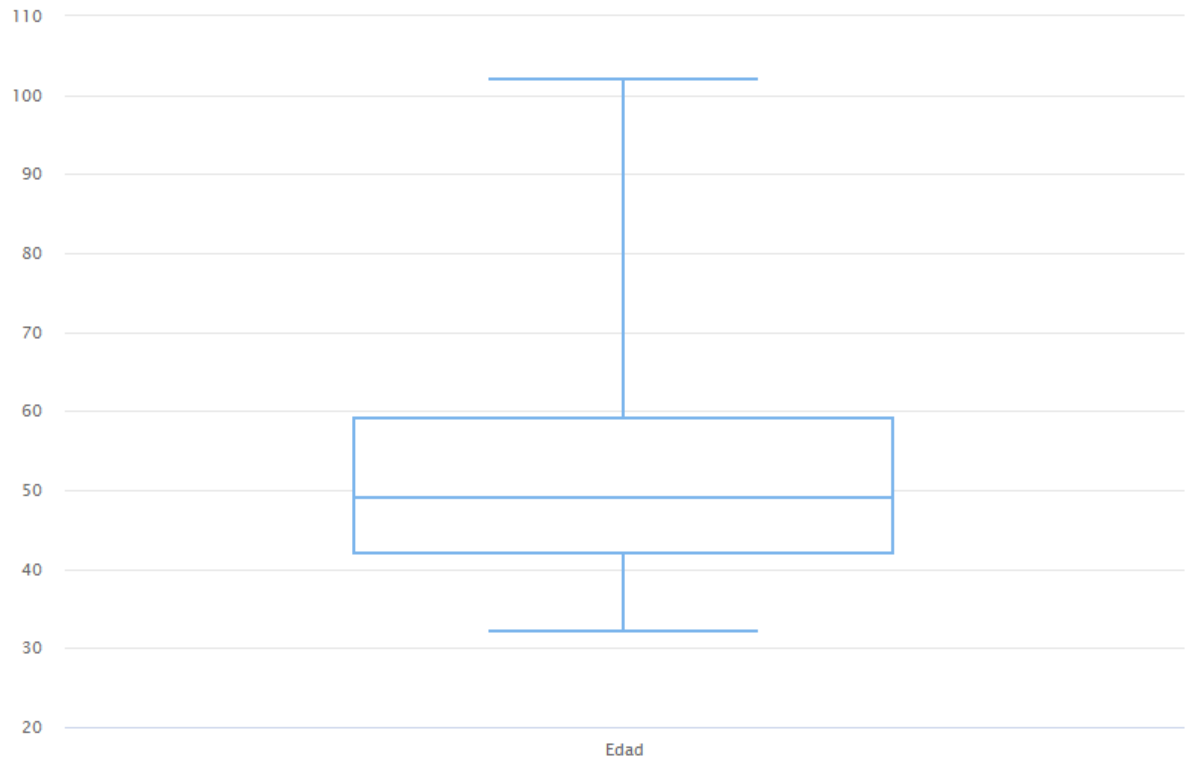


Analizando la gráfica puede notarse rápidamente que más del 50% de los clientes registrados viven en la región Norte. En el caso de la categoría Noroeste, se detecta que sólo existe 1 caso registrado, por lo tanto, se ha decidido imputar el dato por un

valor considerado cercano en cuanto al contexto del negocio. Es decir, se reemplazará el valor ‘Noroeste’ por ‘Norte’.

Variable **Edad**

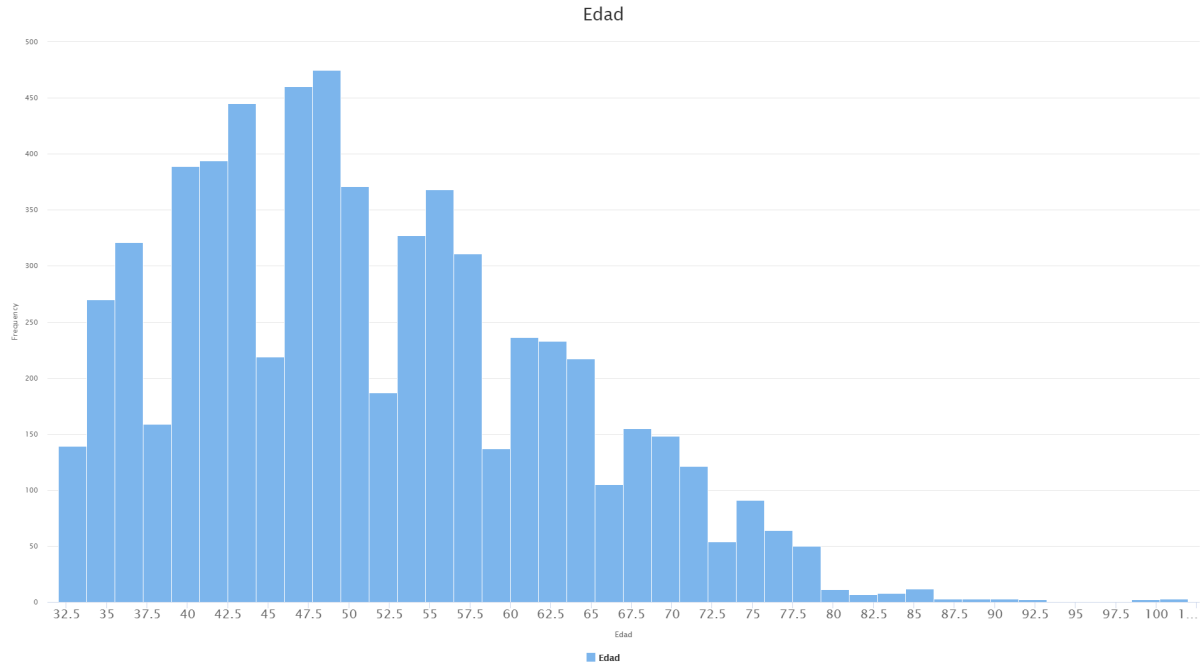
La variable Edad es una variable numérica y por lo tanto, es posible analizarla a través de un gráfico de caja y bigotes, así como también de un histograma.



A partir del diagrama de caja y bigotes se interpreta que los datos registrados no son simétricos dado que la mediana divide a la caja en dos partes desiguales. De esta forma, dado que la parte superior de la caja es mayor a la inferior, la asimetría será positiva y por lo tanto la media tendrá un valor superior a la mediana.

Los bigotes del diagrama indican que el rango de edades que se encuentran registradas es extenso, y que la variación entre la mediana, que es 49 años y el valor máximo que es 102 años es mayor a la variación entre el valor mínimo, es decir, 32 años y la mediana.

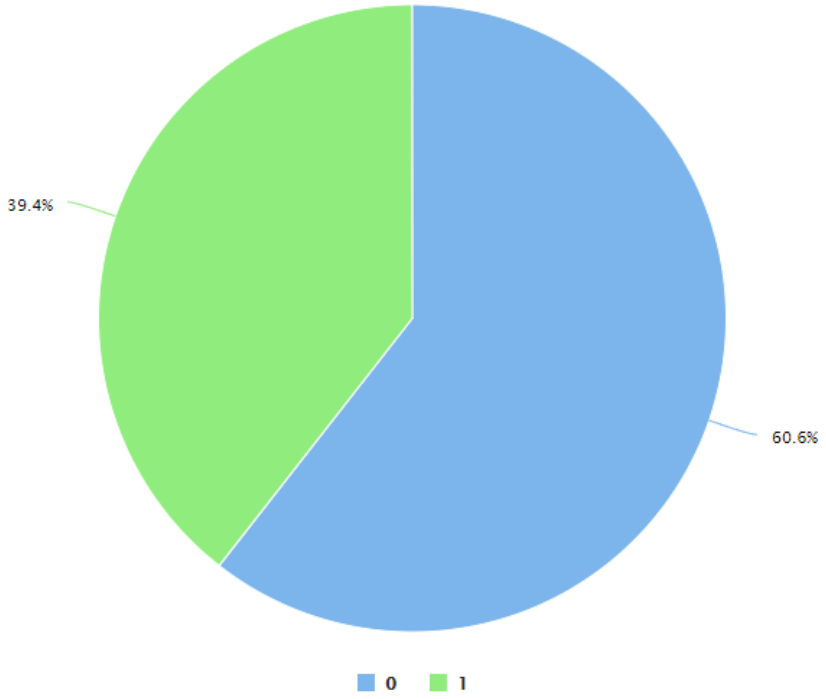
Consideramos que el valor mínimo de edad registrada es un dato a analizar con mayor profundidad en el contexto del negocio. Es decir, supongamos que sólo pueden comprar individuos mayores de edad, parecería inconsistente que no existan clientes registrados que se encuentren en el rango de edad de 18 a 32 años.



A partir del histograma, se puede ver que la distribución tiende a una asimetría a la derecha, confirmando que la mayor cantidad de clientes registrados se concentran en los valores de edad más bajos del conjunto de datos.

Variable **Compró bicicleta**

Esta variable indica si el cliente registrado realizó la compra de al menos una bicicleta, independientemente del tipo. Dado que es una variable binomial, realizamos un gráfico de torta que indique la proporción de cada una de las categorías, donde 0 indica que el cliente no compró una bicicleta y el 1 que sí lo hizo.



En el gráfico puede notarse rápidamente que la mayor cantidad de clientes no realizaron ninguna compra de una bicicleta al momento de registrar los datos. La cantidad de clientes que sí compraron bicicletas es de 2.564 sobre un total de 6.500 registros, es decir aproximadamente el 40% del conjunto de datos.

Análisis multivariante

Matriz S

La matriz S, también llamada matriz de varianza y covarianza, nos muestra las varianzas en la diagonal principal y las covarianzas en las demás celdas. Esta matriz la formamos únicamente con las variables numéricas.

Attributes	IngresoAnual	Edad
IngresoAnual	1047683938.999	57225.195
Edad	57225.195	132.140

Matriz R

La matriz R, conocida como matriz de correlación, mide el grado de relación lineal entre cada par de variables. Los valores de correlación se pueden ubicar entre -1 y +1.

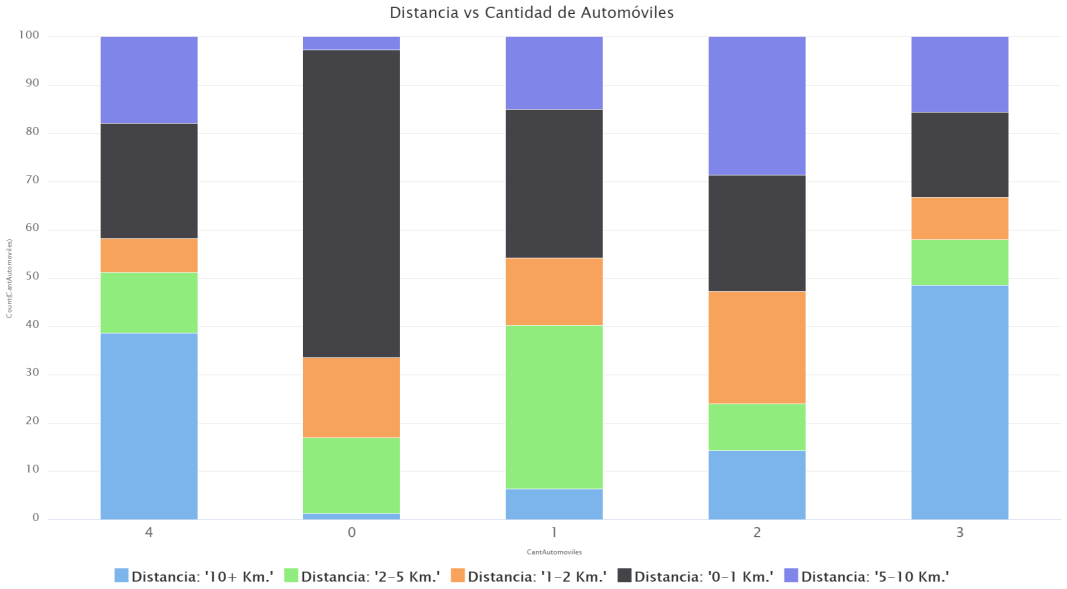
A continuación presentamos dicha matriz.

Attributes	IngresoAnual	Edad
IngresoAnual	1	0.154
Edad	0.154	1

A partir de la tabla podemos observar que existe un cierto grado de correlación positiva entre las variables Edad e Ingreso anual. Sin embargo, al acercarse su valor a 0, no es significativo dentro del contexto del problema.

Distancia al trabajo vs Cantidad de automóviles

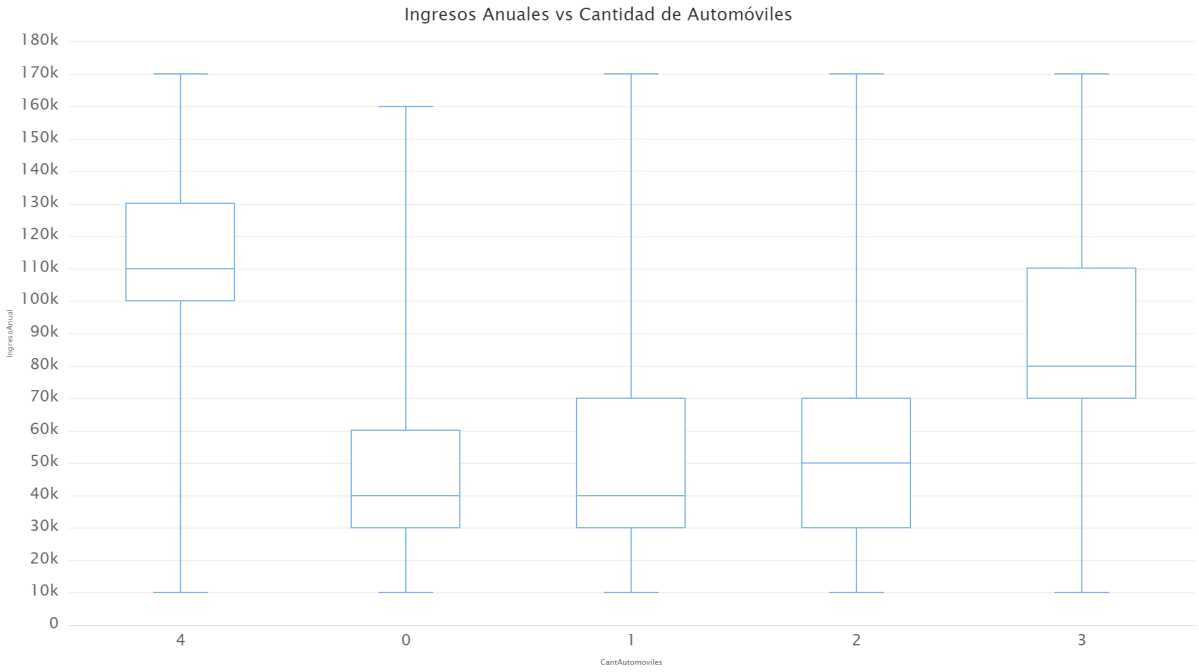
Al ser dos variables polinomiales, se realizará un gráfico de barras agrupado por color, con un stacking al 100%, de forma que puedan compararse las diferentes categorías.



Luego de observar el gráfico podemos notar que los clientes que viven a mayor distancia del trabajo tienden a tener una mayor cantidad de automóviles. Por otra parte, el 63,8% de los clientes registrados que no poseen ningún auto, viven a una distancia menor a 1 km, y sólo el 1,3% de ellos vive a más de 10 km de su lugar de trabajo.

Ingresos anuales vs Cantidad de automóviles

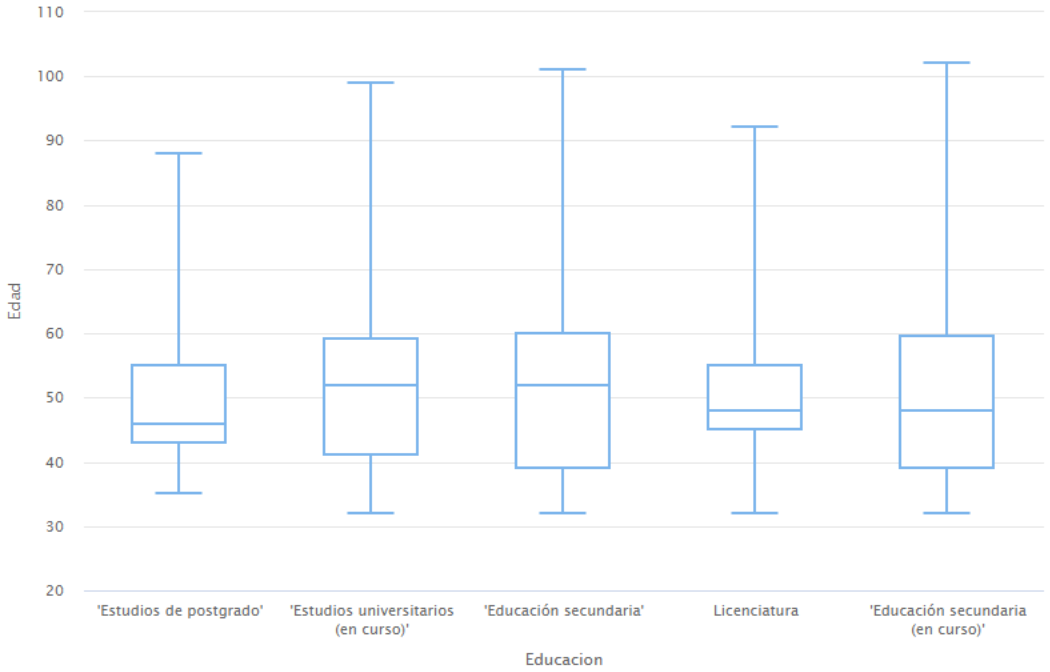
En este caso, la variable ingresos anuales tiene un tipo de dato numérico y la variable cantidad de automóviles tiene un tipo de datos nominal, por lo tanto para su análisis se realiza un diagrama de caja y bigotes.



En el diagrama puede notarse que en general, los clientes que no poseen autos o que tienen sólamente uno, reciben un menor ingreso anual que el resto de los clientes registrados. Por otro lado, las personas que disponen de 4 automóviles tienen un ingreso anual cuya mediana es $110.000, el cual resulta ser un valor muy superior a la mediana de todo el conjunto de datos.

Edad vs Educación

Al igual que el caso anterior, las variables edad y educación son del tipo numérico y polinomial respectivamente y por lo tanto se utilizará un diagrama de caja y bigotes para su análisis.



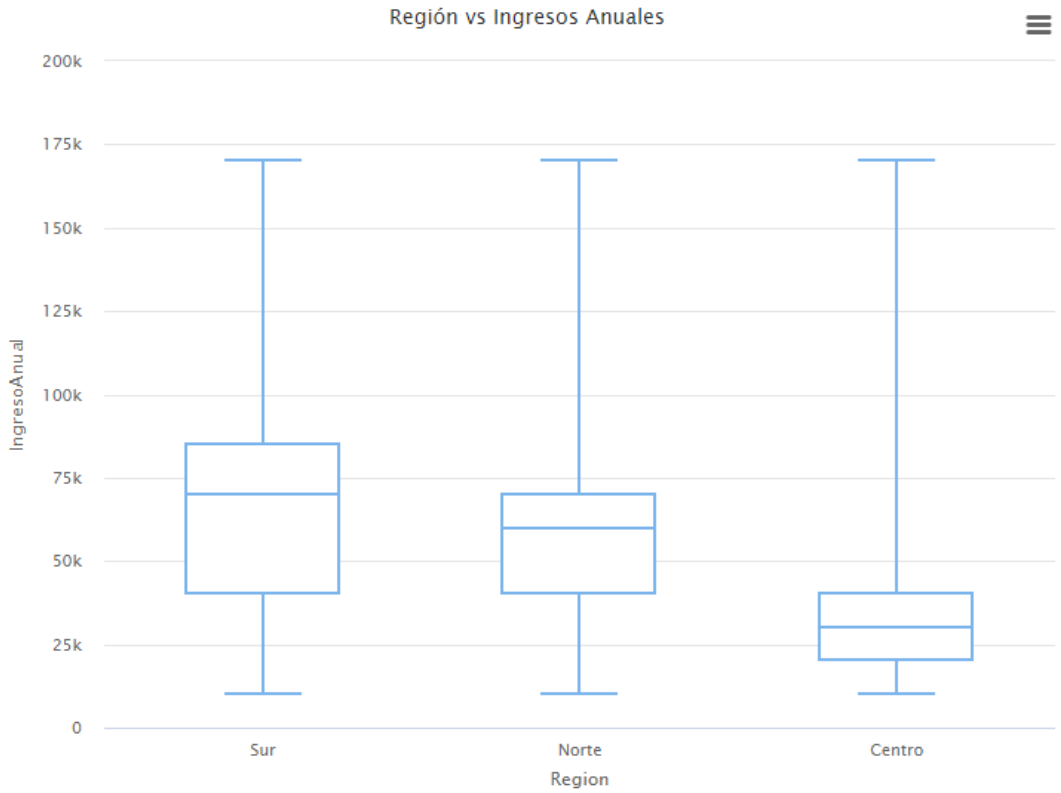
Interpretando este gráfico donde se ilustran diferentes diagramas de caja y bigotes que compara la variable “edad” contra cada categoría de “educación”, podemos analizar lo siguiente:

- Con respecto a “*Estudios de postgrado*”, vemos que el 50% de los datos están concentrados entre 35 y 46 años. También, que presenta una asimetría hacia la derecha ya que la mediana está situada más cerca del Q1. Por lo tanto, los datos están concentrados en edades bajas. Si miramos la clase “*Licenciatura*” notamos que presenta características similares con la categoría anterior, aunque el valor mínimo es de 32 años y los datos están más concentrados porque la caja es ligeramente más pequeña.
- Entre las clases “*Estudios universitarios (en curso)*”, “*Educación secundaria*” y “*Educación secundaria (en curso)*” interpretamos características muy similares, salvo que las 2 primeras tienen distribución asimétrica hacia la izquierda y la última una pequeña tendencia hacia una distribución asimétrica hacia la derecha.
- Por último, con respecto a las categorías de estudio que están “en curso” (“*Estudios universitarios*” y “*Educación secundaria*”), vemos que el valor

máximo es de 99 y 102 años, respectivamente. La conclusión a la que llegamos a partir de esto refiere a que esta categoría corresponde a las personas que no finalizaron sus estudios secundarios.

Región vs Ingresos Anuales

Dado que la variable región es polinomial y la variable ingresos anuales es numérica, se hará uso nuevamente de un diagrama de caja y bigotes.

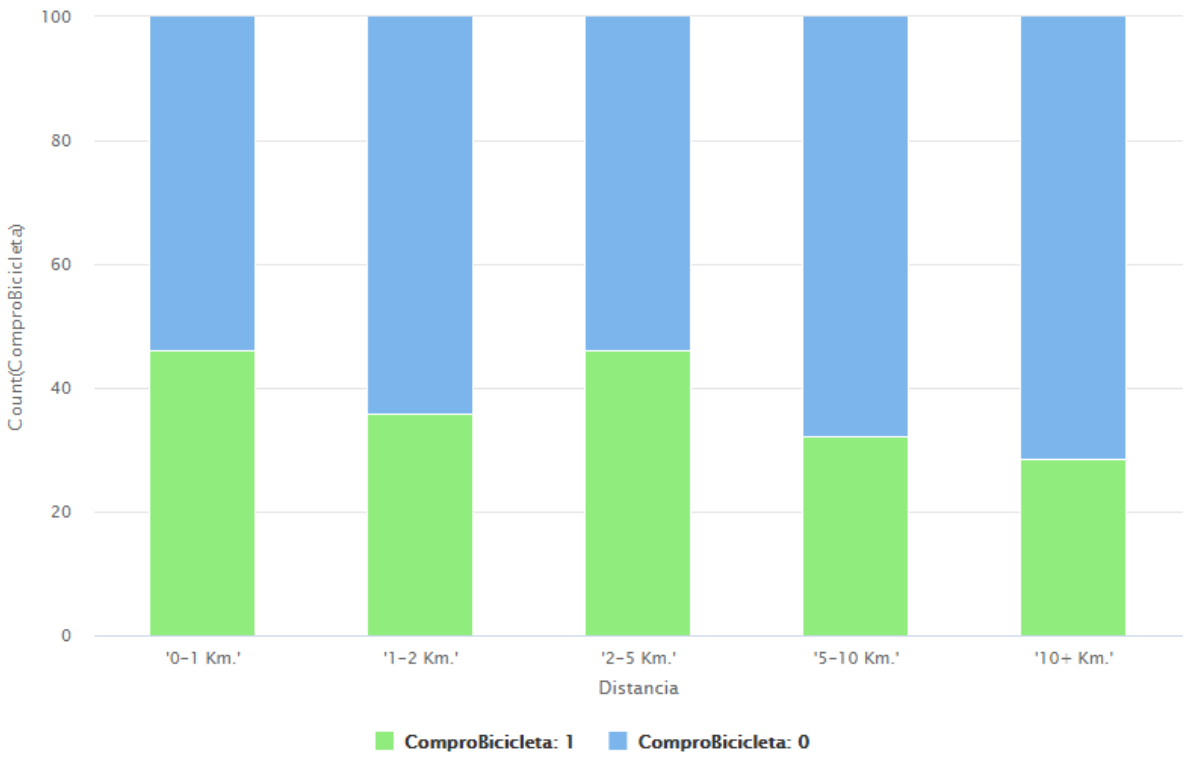


En este gráfico vemos la distribución del ingreso de los clientes con respecto a su región. Aquí observamos:

- Región Sur: la mayor parte de la gente que se encuentra en esta zona tiene ingresos medios con respecto al resto de los clientes.
- Región Norte: los clientes de esta zona se reparten entre ingresos medios y bajos.
- Región Sur: es una zona de mayoría de clientes de ingresos bajos con respecto a la muestra.

Distancia al trabajo vs Compró Bicicleta

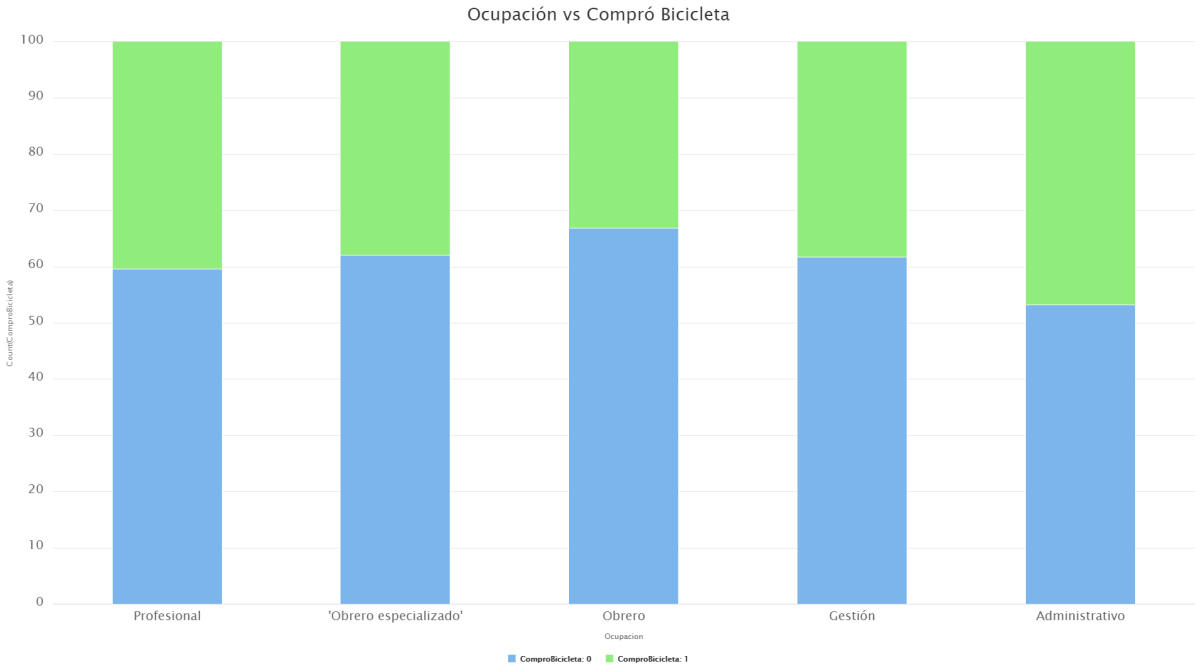
Al ser la variable distancia al trabajo polinomial y la variable compró bicicleta binomial, se analizarán ambas en conjunto utilizando un gráfico de barras con un stacking al 100% para poder comparar todas las categorías.



En el gráfico anterior notamos, por un lado, que aquellas personas que viven entre 0 y 1 km al trabajo, el 46,1% **si** compró bicicletas, mientras que el 53,9% restante **no** compró. Por otro lado, las personas que viven a una distancia mayor a 10km del trabajo, un porcentaje bastante bajo registran la compra de bicicletas (28,6%). En general, en cada categoría de distancia aquí mostradas, el porcentaje de compra de bicicleta es menor al porcentaje de no compra.

Ocupación vs Compró Bicicleta

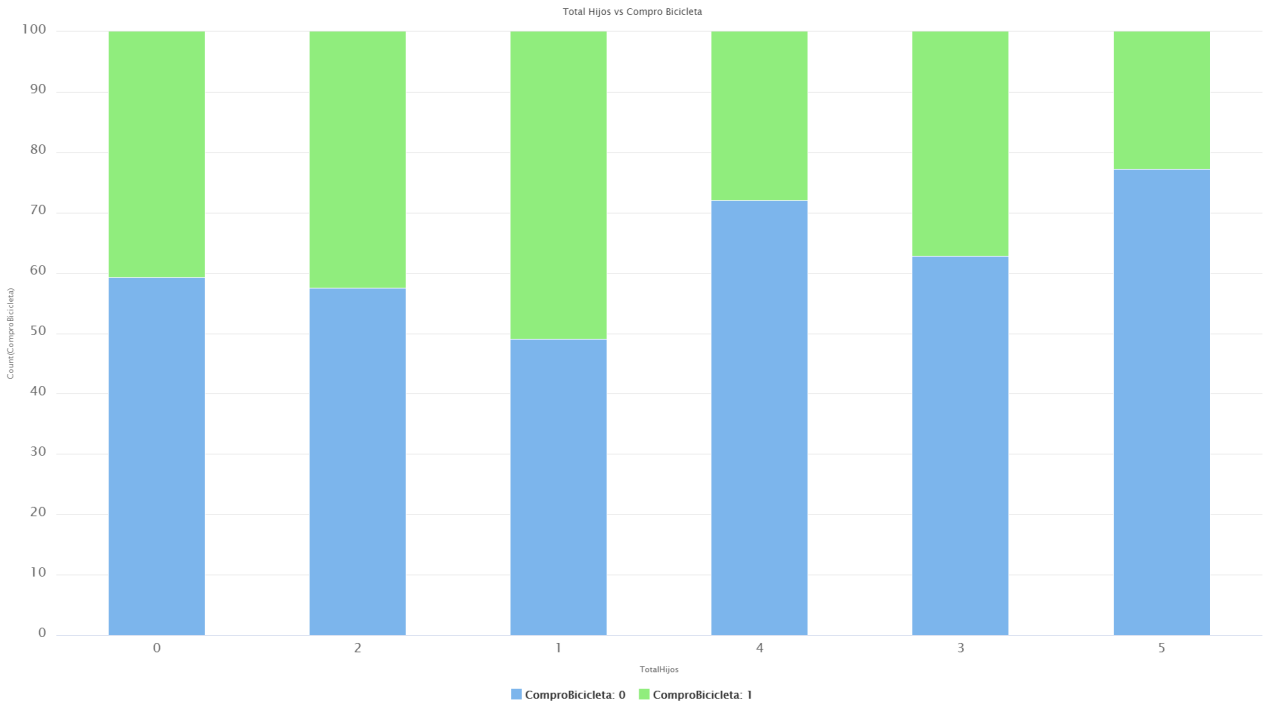
Tanto la variable ocupación como compró bicicleta son variables nominales, por lo tanto se realizó un gráfico de barras apilado para poder compararlas.



En este caso, puede verse que la ocupación que realizó la mayor compra de bicicletas, es administrativo. Sin embargo, a partir de la observación del gráfico puede decirse que no existen mayores diferencias entre las categorías de ocupación con respecto a si se compraron bicicletas o no. Esto quiere decir que ambas variables parecen ser independientes una de otra.

Total de hijos vs Compró Bicicleta

Para este análisis se realizó un gráfico de barras apiladas por colores, debido a que las variables analizadas son de tipo polinomiales.



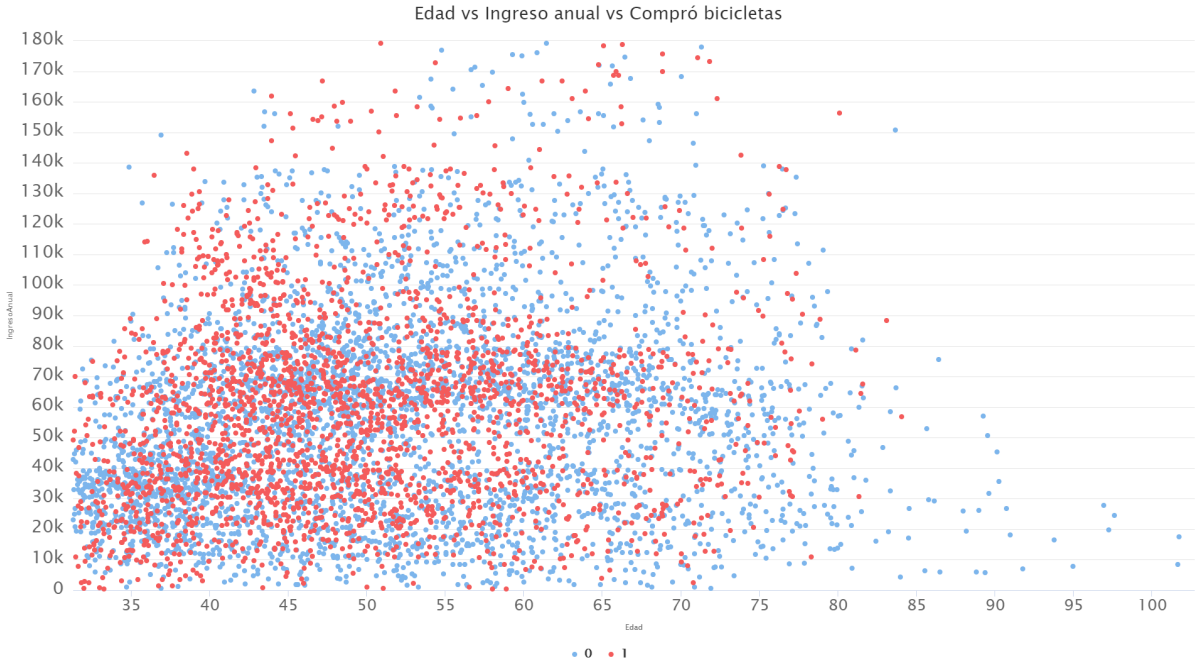
Dejando de lado las personas que no tienen hijos, donde la proporción es aproximadamente 60% no compran bicicletas; 40% si compran, podemos ver que existe una correlación entre la cantidad de hijos y la compra o no de bicicletas.

En ese sector del gráfico vemos cómo a medida que se incrementa la cantidad de hijos, disminuye la proporción de clientes que adquieren bicicletas.

Se aprecia que para clientes con un hijo, el porcentaje de los mismos que no compro bicicletas se acerca al 50%. Continuando con esto, para clientes con 2, 3, 4 y 5 hijos, la proporción de los mismos que no compraron bicicletas es de 57%, 63%, 72% y 77%, respectivamente.

Edad vs Ingresos Anuales vs Compró Bicicleta

En este caso se realizó un diagrama del tipo scatter, en el cual se representan las variables Edad en el eje X, Ingreso anual en el eje Y y Compró Bicicleta con colores rojo o azul dependiendo de si el cliente hizo la compra de una bicicleta o no, respectivamente.



A partir de la observación del gráfico puede notarse que todos los clientes que compraron bicicletas tienen una edad menor a 85 años. Por otra parte, respecto a los ingresos anuales, la mayor cantidad de clientes registrados que compraron bicicletas suelen tener ingresos medios o bajos.

Además puede notarse que las personas que tienen ingresos anuales altos se encuentran en el rango etario de entre 45 y 80 años, y que en general, al tener menor edad, también es menor su ingreso anual.

Limpieza de datos

En este apartado se resumen las modificaciones que se realizaron al conjunto de datos a partir del análisis exploratorio realizado en la etapa anterior. Esta limpieza y preparación nos ayudará a obtener datos de calidad relevantes para el negocio, de forma que el modelo a realizar utilice sólo información adecuada para el problema.

En primer lugar, se detectó que la variable **ingreso anual** registraba 11 datos faltantes. Para corregir el problema de los datos faltantes existen varias opciones:

- Ignorar, es decir, dejar pasar los datos sin valor. A pesar de que los modelos de árboles de decisión son robustos y permitirían en un principio ignorar los datos, consideramos que dado que queremos trabajar con el mismo dataset en todos los modelos, no se podría utilizar en KNN.

- Para este caso, eliminar toda la columna no sería una opción viable ya que resultaría una alternativa drástica y modificaría totalmente el conjunto de datos.

- Otra opción válida sería filtrar las filas que contienen los datos faltantes. De esta forma se eliminarían 11 registros entre un total de 6.500, es decir, un 0,16% de los datos. Si bien esta proporción es reducida, puede producir un sesgo en la información a analizar.

- Ya que el valor faltante siempre recae sobre la misma columna (Ingreso Anual), la cual es una variable numérica, reemplazar el valor faltante por otro que preserve la media o la varianza promete ser una muy buena opción para aplicar a este caso.

De esta forma, se ha decidido reemplazar los datos faltantes de la variable por el valor de la media, es decir $57.535,83.

Otra de las transformaciones realizadas tiene que ver con la variable **Estado Civil**, en la cual existía la categoría ‘Viudo’ (V) que sólo clasificaba a 2 observaciones. En este caso el tratamiento realizado fue descartar estos dos registros ya que representaban solo el 0,03% de los datos, y por lo tanto no se considera que puedan crear un sesgo en ellos.

Por último, hemos detectado que en la variable **Región** la categoría ‘Noroeste’ contiene una sola observación. Teniendo en cuenta que todos los datos del conjunto se clasifican entre las categorías ‘Norte’, ‘Centro’ y ‘Sur’, se decidió sustituir el dato por el valor ‘Norte’, que es el valor que consideramos más cercano teniendo en cuenta el contexto del problema.

Especificación de las vistas minables

Luego del proceso de selección, limpieza y transformación de las variables, se obtuvo una vista minable. En esta sección se describe cada una de estas variables, que nos serán útiles tanto para generar los diferentes modelos, como en las etapas posteriores del proceso.

- **Estado civil**: variable que describe el estado civil del cliente. Es de tipo polinomial y toma dos valores C (Casado) y S (Soltero).
- **Género**: variable de tipo polinomial que describe el género del cliente. Sus valores son F (Femenino) y M (Masculino).
- **Ingreso anual:** variable entera que representa el sueldo (en pesos) que obtiene cada cliente en el año. Los valores posibles se encuentran entre 10.000 y 170.000 pesos.
- **Total hijos:** es la cantidad de hijos que posee cada cliente. Se categorizó como polinomial para que el análisis gráfico de la variable sea más sencillo puesto que solo tiene valores que van del 0 al 5.
- **Educación**: variable que describe el nivel educativo máximo que alcanzó el cliente, es de tipo polinomial. Toma los valores: Licenciatura, Estudios de Posgrado, Educación Secundaria, Educación Secundaria (en curso) y Estudios Universitarios (en curso).
- **Ocupación:** se trata de una variable de tipo polinomial que describe el oficio del cliente. Sus valores pueden ser profesional, obrero especializado, obrero, gestión, administrativo.
- **Propietario:** variable binomial que indica si el cliente es propietario (valor 1) o no (valor 0) de una vivienda.
- **Cantidad automóviles:** variable entera que indica la cantidad de automóviles que posee el cliente en cuestión. Su valor mínimo es 0 y su máximo es 5.
- **Distancia al trabajo**: esta variable describe la distancia del cliente hacia su lugar de trabajo. Es de tipo polinomial y sus valores son: 10+ Km, 2-5 Km, 1-2 Km, 0-1 Km, 5-10 Km.
- **Región:** variable de tipo polinomial que indica en qué región vive el cliente. Las categorías en las cuales clasifica son ‘Norte’, ‘Centro’ y ‘Sur’.
- **Edad**: variable que indica la edad del cliente. Es de tipo entero y sus valores van desde 32 años hasta 102 años.
- **Compro bicicleta:** variable binomial, indicadora de si el cliente adquirió (valor 1) o no (valor 0) una bicicleta en el comercio. Se trata de la variable a predecir.

Modelado

A partir de esta sección comenzaremos con la fase de modelado, en donde describiremos la estrategia y especificaremos los parámetros utilizados para construir los diferentes modelos explicados en los apartados anteriores, para finalmente dejar a disposición los modelos obtenidos haciendo uso de diferentes softwares, como RapidMiner e IBM SPSS Modeler.

Los modelos que vamos a trabajar son los llamados predictivos y descriptivos. A continuación se realiza una breve descripción de ellos.

- Modelos predictivos: el objetivo de estos modelos es determinar o predecir resultados futuros. Para ello es necesario disponer de datos históricos. Si llevamos esto a nuestro trabajo, los datos históricos son aquellos que están

representados en el archivo “clientes.csv” y el resultado a predecir es la variable “ComproBicicleta”.

- Modelos descriptivos: también conocidos como modelos no supervisados, tienen como objetivo proporcionar información sobre las relaciones entre los datos y sus características de forma de poder clasificarlos en grupos según patrones de interés. La idea de aplicar este tipo de modelos en nuestro trabajo, es para resolver la siguiente pregunta: ¿qué tipo de bicicleta le vamos a ofrecer a cada cliente?.

Estrategia para construcción de modelos.

Para construir el modelo definitivo que luego será implementado, es necesario explorar modelos alternativos hasta encontrar el que resulte útil para resolver nuestro problema. Para esto puede que tengamos que retroceder a fases anteriores y hacer cambios en los datos que estamos usando o incluso modificar la definición del problema.

El proceso de construcción de modelos predictivos requiere tener bien definidas las etapas de entrenamiento y validación para asegurar que las predicciones serán robustas y precisas. Para esto, la idea es entrenar el modelo con una porción de los datos (training dataset) y luego validarlo con el resto de los datos (test dataset).

En este caso, se ha decidido dividir el conjunto de datos de forma aleatoria, tomando un 70% de los mismos como parte del training dataset y el restante 30% como el test dataset. Para esto, haremos uso del operador Split Data en el software Rapid Miner.

Especificación de parámetros utilizados.

Árboles de decisión

Para el árbol de decisión creado en RapidMiner, los parámetros utilizados son los siguientes:

- **Criterio:** Especifica el criterio con el cual los atributos van a ser seleccionados. Hay diferentes opciones y nosotros optamos por “gain_ratio”.

- **Máxima profundidad:** Este parámetro se utiliza para restringir la profundidad del árbol de decisión. Nosotros probaremos con diferentes valores y mantendremos aquel que nos proporcione un modelo aceptable.

- **Poda:** El modelo de árbol de decisión puede ser podado después de la generación, esto es conocido como post-poda. Aplicaremos este tipo de poda debido a que la pre-poda recorta el árbol a medida que este se construye y puede derivar en un modelo con mayor error. A la post-poda le asignaremos un nivel de confianza de 0,10.

Para el caso de los árboles generados con el software Modeler, se explicarán en sus respectivos apartados los parámetros utilizados.

KNN

- **k:** El primer paso del algoritmo KNN es encontrar los "**k**" casos de entrenamiento más cercanos al caso desconocido. Si k = 1, el caso se asigna simplemente a la clase de su vecino más cercano. "**k**" suele ser un número entero pequeño, positivo e impar. Para nuestro trabajo, aplicaremos valores de **k** que van desde 1 hasta 10.

- **Voto ponderado:** Si se establece este parámetro, los valores de distancia entre los casos también se tienen en cuenta para la predicción. Puede ser útil ponderar las contribuciones de los vecinos, de modo que los vecinos más cercanos contribuyan más que los más lejanos. En nuestro caso dejaremos seleccionado este parámetro.

- **Tipos de medida:** Este parámetro se usa para seleccionar el tipo de medida que se va a utilizar para encontrar los vecinos más cercanos. Los posibles valores a elegir son:
    - MixedMeasures: Es usada para calcular distancias en caso de tener tanto variables nominales como numéricas.
    - NominalMeasures: En el caso de que las variables sean sólo nominales, se pueden utilizar diferentes métricas de distancia para calcular las distancias en estos atributos nominales.
    - NumericalMeasures: En el caso de que las variables sean sólo numéricas, se pueden utilizar diferentes métricas de distancia para calcular las distancias en estos atributos numéricos.
    - BregmannDivergences: Las divergencias de Bregmann son tipos de medidas de "cercanía" más genéricas

  Como en nuestro trabajo contamos con variables nominales y numéricas, el tipo de medida a implementar será el MixedMeasures.

- **Medidas mezcladas:** La única opción disponible (y la cual utilizaremos) cuando el parámetro de tipo de medida se establece en "MixedMeasures" es la "Mixed Euclidean Distance" (Distancia euclídea mixta). Para los valores numéricos se calcula la distancia euclídea. Para los valores nominales, se toma una distancia de 0 si ambos valores son iguales y una distancia de 1 en caso contrario.

Modelos obtenidos

Árboles de decisión (RapidMiner)

Para este modelo decidimos tener como único parámetro la profundidad del árbol, de manera que sea más fácil la comparación entre los mismos. Por lo que la pre-poda y la post-poda no van a estar habilitadas en el programa construido.

Los resultados obtenidos fueron los siguientes:

	Class Recall		Accuracy
Depth	True 0	True 1	
1	100,00%	0,00%	60,75%
2	100,00%	0,00%	60,75%
3	100,00%	0,00%	60,75%
4	83,61%	26,80%	61,31%
5	82,26%	30,59%	61,98%
6	79,98%	35,95%	62,70%
7	79,90%	36,73%	62,96%
8	74,58%	56,08%	67,32%
9	78,29%	53,86%	68,70%
10	78,72%	55,16%	69,47%
11	78,63%	54,25%	69,06%
12	83,78%	49,02%	70,14%
13	82,43%	54,38%	71,42%
14	78,97%	64,97%	73,47%
15	79,05%	66,01%	73,94%
16	80,83%	65,62%	74,86%
17	80,24%	66,41%	74,81%
18	79,22%	66,80%	74,35%
19	79,90%	66,14%	74,50%
20	79,39%	66,41%	74,29%
21	80,66%	65,23%	74,60%
22	80,24%	66,27%	74,76%
23	80,74%	65,62%	74,81%
24	80,66%	64,58%	74,35%
25	80,32%	64,58%	74,14%
26	80,41%	64,97%	74,35%
27	80,83%	63,79%	74,14%
28	80,83%	63,79%	74,14%
29	80,83%	62,75%	73,73%
30	80,83%	62,75%	73,73%

Donde los colores rojos y amarillos son modelos que no cumplen con nuestros criterios de aceptación y los de color verde son los aceptados. La elección de uno de estos modelos se determinó por la diferencia entre los porcentajes de aciertos (Columnas True 0 y True 1) y la menor de ellas es el árbol con profundidad 18, cuya matriz se presenta en la siguiente imagen:

accuracy: 74.35%

	true 0	true 1	class precision
pred. 0	938	254	78.69%
pred. 1	246	511	67.50%
class recall	79.22%	66.80%	

Árboles de decisión (Modeler)

A partir de este momento presentaremos los 3 árboles de decisión generados con la herramienta Modeler. Ellos son el árbol Quest, CHAID y C5.0.

Quest

Los parámetros que configuramos para este árbol son los siguientes:

- Profundidad máxima = 5.
- Se selecciona la opción para podar el árbol.
- Para las "reglas de parada" se utiliza porcentaje, donde el mínimo de registros en la rama padre es de un 2% y en la rama hijo del 1%.

Las matrices de confusión que obtuvimos tanto para la partición de entrenamiento como para la de prueba se ilustran a continuación:

Entrenamiento

ComproBicicleta		0	1
0	Recuento	2210	532
	% de filas	80.598	19.402
1	Recuento	1044	750
	% de filas	58.194	41.806

Prueba

ComproBicicleta		0	1
0	Recuento	961	233
	% de filas	80.486	19.514
1	Recuento	448	320
	% de filas	58.333	41.667

La tasa general de acierto para la matriz de confusión en el conjunto de prueba es de 65,29%. Si miramos los porcentajes de las filas entre matrices, estos presentan una buena estabilidad. Sin embargo, cuando el valor real es que compró bicicleta, el porcentaje de acierto al predecir que compró bicicleta es menor con respecto al que no compró.

CHAID

Este árbol CHAID fue establecido con los parámetros siguientes:

- Profundidad máxima = 5.
- Se selecciona la opción para podar el árbol.
- Para las “reglas de parada” se utiliza porcentaje, donde el mínimo de registros en la rama padre es de un 2% y en la rama hijo del 1%.

Aquí se presentan las matrices de confusión para ambos conjunto de datos:

Entrenamiento

ComproBicicleta		0	1
0	Recuento	2238	504
	% de filas	81.619	18.381
1	Recuento	748	1046
	% de filas	41.695	58.305

Prueba

ComproBicicleta		0	1
0	Recuento	968	226
	% de filas	81.072	18.928
1	Recuento	368	400
	% de filas	47.917	52.083

La tasa general de acierto es igual a 69,72% para los datos de prueba. Si bien mejora este valor con respecto al anterior, mirando los porcentajes de las filas y comparándolo con el conjunto de datos de entrenamiento, vemos mayor desequilibrio cuando el valor real indica que compró bicicleta (tercera y cuarta fila). No obstante, mejora el porcentaje de acierto en la predicción que compró bicicleta cuando el valor real es que compró bicicleta.

C5.0

Este último árbol fue configurado con los siguientes parámetros:

- Modo = simple.
- Favorecer = Precisión.
- Ruido esperado = 0%.

A continuación ponemos a disposición las 2 matrices de confusión obtenidas para las diferentes particiones:

Entrenamiento

ComproBicicleta		0	1
0	Recuento	2427	315
	% de filas	88.512	11.488
1	Recuento	490	1304
	% de filas	27.313	72.687

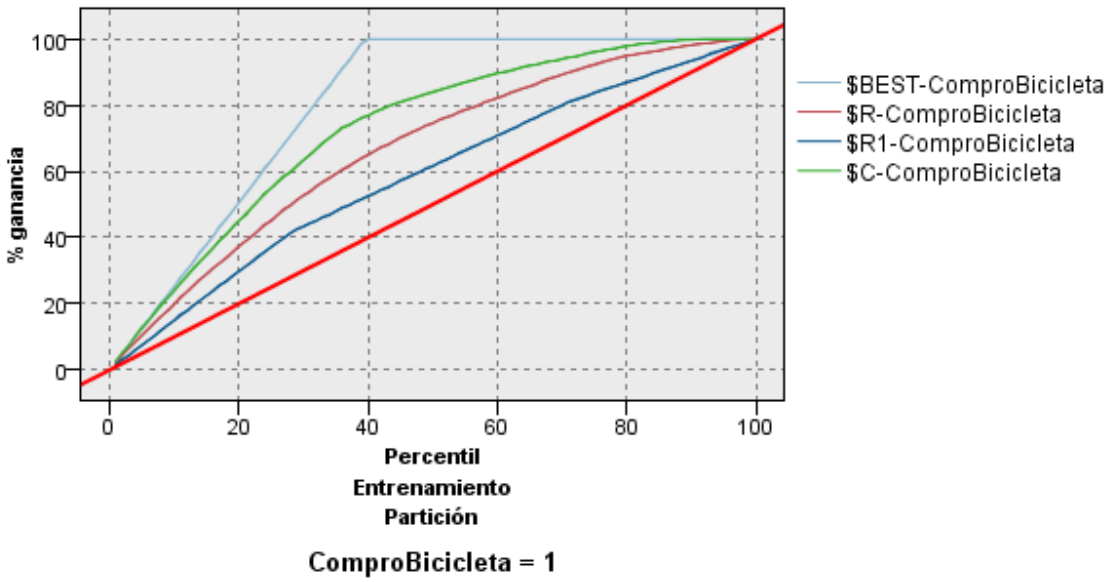
Prueba

ComproBicicleta		0	1
0	Recuento	1008	186
	% de filas	84.422	15.578
1	Recuento	286	482
	% de filas	37.240	62.760

Analizando los valores de “% de filas” en ambas matrices, se puede decir que estos están desbalanceados, pero con respecto a los anteriores, son mayores en el caso de la diagonal principal que son los valores que indican que el modelo está prediciendo bien. Calculando la tasa general de acierto del conjunto de prueba nos da que es igual a 75,94%.

Comparación entre modelos

A continuación, vamos a evaluar los modelos **Quest**, **CHAID** y **C5.0** para determinar cuál resulta mejor:



ComproBicicleta = 1

Aquí se pueden observar 5 líneas. La línea roja más gruesa representa un umbral que indica que aquellos modelos que estén cerca de esa línea no brindan buena información. Luego tenemos las líneas azul, roja y verde que pertenecen a los modelos Quest, CHAID y C5.0, respectivamente. La línea celeste indica el mejor modelo.

A partir de la interpretación de este gráfico podemos decir que el mejor modelo es el C5.0 ya que es el que más se acerca a la línea celeste del mejor modelo. A pesar

de ello, más adelante en la sección de evaluación de los modelos compararemos los 3 modelos para determinar, en base a la matriz de costos, cuál es el mejor.

KNN (RapidMiner)

Durante todo este apartado presentaremos los resultados obtenidos para el modelo de KNN realizado en el software RapidMiner. Los mismos fueron obtenidos para distintos valores de **k** el cual determina la cantidad de vecinos cercanos con los cuales se comparará la nueva observación. Además se hace uso de la estrategia de construcción del modelo expuesta con anterioridad.

Es importante destacar que, ya que contamos con variables numéricas y polinomiales, el tipo de medida aplicada es “Mixed Measures” y la distancia es “Mixed Euclidean Distance”. Dado que las variables numéricas (IngresoAnual y Edad) poseen diferentes unidades, decidimos normalizar los mismos para que puedan ser comparables.

A continuación analizaremos las diferentes matrices de confusión que corresponden a las salidas del modelo para cada valor de **k**. En cada matriz se muestra en las columnas los valores reales de la variable a predecir (ComproBicicleta), mientras que en las filas están los valores predichos. Adicionalmente, se encuentran una fila llamada “class recall” y una columna “class precision”, que identifican los porcentajes de aciertos para las distintas clases: 0 si no compró bicicleta y 1 si compró.

- **Con k=1:**

accuracy: 75.27%

	true 1	true 0	class precision
pred. 1	533	242	68.77%
pred. 0	240	934	79.56%
class recall	68.95%	79.42%	

- **Con k=2:**

accuracy: 73.63%

	true 1	true 0	class precision
pred. 1	555	296	65.22%
pred. 0	218	880	80.15%
class recall	71.80%	74.83%	

- **Con k=3:**

accuracy: 75.22%

	true 1	true 0	class precision
pred. 1	539	249	68.40%
pred. 0	234	927	79.84%
class recall	69.73%	78.83%	

- **Con k=4:**

accuracy: 75.63%

	true 1	true 0	class precision
pred. 1	524	226	69.87%
pred. 0	249	950	79.23%
class recall	67.79%	80.78%	

- **Con k=5:**

accuracy: 74.50%

	true 1	true 0	class precision
pred. 1	498	222	69.17%
pred. 0	275	954	77.62%
class recall	64.42%	81.12%	

- **Con k=6:**

accuracy: 75.17%

	true 1	true 0	class precision
pred. 1	498	209	70.44%
pred. 0	275	967	77.86%
class recall	64.42%	82.23%	

- **Con k=7:**

accuracy: 74.86%

	true 1	true 0	class precision
pred. 1	470	187	71.54%
pred. 0	303	989	76.55%
class recall	60.80%	84.10%	

- **Con k=8:**

accuracy: 75.42%

	true 1	true 0	class precision
pred. 1	477	183	72.27%
pred. 0	296	993	77.04%
class recall	61.71%	84.44%	

- **Con k=9:**

accuracy: 75.37%

	true 1	true 0	class precision
pred. 1	462	169	73.22%
pred. 0	311	1007	76.40%
class recall	59.77%	85.63%	

- **Con k=10:**

accuracy: 74.96%

	true 1	true 0	class precision
pred. 1	455	170	72.80%
pred. 0	318	1006	75.98%
class recall	58.86%	85.54%	

Para una mejor representación y poder comparar de manera prolija la tasa general de acierto (accuracy) de los diferentes resultados, se dispone de la siguiente tabla:

k	Tasa general de acierto (%)
1	75,27
2	73,63
3	75,22
4	75,63
5	74,50
6	75,17
7	74,86
8	75,42
9	75,37
10	74,96

Si miramos esta tabla y las diferentes matrices de confusión, podríamos decir que el mejor modelo es el que se obtiene con un **k=4** ya que es el que presenta una mayor tasa general de acierto frente a los demás. Sin embargo, este no es el único parámetro que debemos tener en cuenta, sino también comparar los porcentajes de acierto para cada clase y determinar cuál de todos los modelos posee mayor estabilidad.

Podemos comparar entre resultados y analizar qué sucede con el modelo obtenido con un **k=8** y el obtenido con un **k=9**. Estos son los modelos que le siguen al k=4 en cuanto a tasa general de acierto, con 75,42% y 75,37%, respectivamente. Ahora si observamos los porcentajes de acierto para cada clase nos encontramos con que son valores muy desparejos, lo cual no es bueno para un modelo productivo.

Tomemos ahora los modelos obtenidos con **k=1** y **k=3**. Comparando sus tasas generales de acierto vemos que son muy similares y además presentan un buen equilibrio en los porcentajes de acierto entre clases, siendo levemente mejor el modelo con k=1.

Por todo lo expuesto anteriormente, decidimos que el mejor modelo de KNN es el obtenido con un **k=1**. Sin embargo, para no descartar el modelo con **k=3**, que también presenta muy buenas características, vamos a considerar ambos para posteriormente ponerlos en perspectiva junto a los modelos obtenidos con las demás técnicas y analizar, en base a la matriz de costos, cuál modelo seleccionar para posteriormente implementarlo en el trabajo.

LDA (SPSS)

Para poder aplicar un modelo de análisis discriminante lineal, se debe tener en cuenta que sólo pueden utilizarse variables de tipo numérico. Dicho esto, a partir de nuestro conjunto de datos, se seleccionarán las variables Edad e Ingreso anual, y se modificará el tipo de las variables Total de Hijos y Cantidad de Automóviles de nominal a cuantitativa.

Una vez seleccionadas las variables, debemos proceder con la validación de los supuestos del modelo LDA, los cuales son:

1. La población debe tener una distribución normal multivariante.
2. Las variables del modelo deben ser independientes (no multicolinealidad).
3. Las matrices de varianzas-covarianzas intra-grupos deben ser iguales en todos los grupos.

Supuesto de normalidad

La verificación del primero de los supuestos se llevará adelante con el test de Kolmogorov-Smirnov, el cual propone las siguientes hipótesis:

- **Hipótesis nula (H_0):** los datos siguen una distribución normal.
- **Hipótesis alternativa (H_1):** al menos un valor no coincide con una distribución normal.

Al ser una prueba de bondad de ajuste de una distribución normal, lo que nos interesa es no rechazar la hipótesis nula. Para esto, lo que tiene que ocurrir es que, tomando un nivel de confianza del 95%, cada uno de los *p-values* debe ser mayor a 0,05.

A continuación se realiza el test con las variables indicadas anteriormente en el software SPSS, obteniendo la siguiente tabla:

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
TotalHijos	.169	6498	.000
CantAutomoviles	.181	6498	.000
Edad	.078	6498	.000
IngresoAnual_1	.129	6498	.000

a. Lilliefors Significance Correction

Observando los valores de la tabla se puede notar que ninguna variable tiene un *p-value* mayor a 0,05, por lo que se concluye que ninguna de las variables sigue una distribución normal.

Si bien los resultados obtenidos en el análisis del supuesto indican que el mismo no se cumple, se continuará con la construcción del modelo de todas formas, teniendo en cuenta que la distribución multivariada no es normal.

Supuesto de no multicolinealidad

La multicolinealidad es una condición que ocurre cuando algunas variables predictoras están correlacionadas con otras. Si se da esta condición en el modelo, puede incrementar la varianza de los coeficientes, haciéndolos inestables.

Al realizar el diagnóstico de colinealidad en el software, se obtiene la siguiente tabla:

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	TotalHijos	CantAutomoviles	Edad	IngresoAnual_1
1	1	4.277	1.000	.00	.01	.01	.00	.01
	2	.318	3.666	.00	.70	.12	.00	.06
	3	.247	4.163	.04	.08	.53	.03	.00
	4	.138	5.571	.02	.03	.34	.02	.92
	5	.020	14.562	.93	.18	.00	.95	.01

a. Dependent Variable: IdCliente

Para que se cumpla el supuesto de no multicolinealidad, el índice de condición de la última fila de la tabla debe ser menor a 30. En este caso, ese valor es 14,562, por lo que el supuesto se verifica. La conclusión es, entonces, que las variables seleccionadas no están correlacionadas y por lo tanto pasan la prueba de no multicolinealidad.

Supuesto de matrices de varianza-covarianza iguales

Por último, se intentará verificar que las matrices de varianza-covarianza entre los grupos son iguales. Para esto, se utilizará el estadístico M de Box, el cual contrasta la igualdad de dichas matrices. Si el estadístico indica un *p-value* no significativo, entonces no habrá suficiente evidencia de que las varianzas sean iguales.

Al realizar el test en el software se obtiene la siguiente tabla, en donde el *p-value* se indica en la última fila.

Test Results

Box's M		344.287
F	Approx.	34.404
	df1	10
	df2	140129725.7
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Dado que el *p-value* no es significativo, entonces no se cumple la hipótesis nula y no puede asegurarse que las matrices de varianza-covarianza sean iguales. Es importante recordar que este test es sensible a las desviaciones de la normalidad multivariada, y que nuestro modelo no superó ese supuesto.

En la siguiente tabla se detallan las matrices de covarianza de cada uno de los grupos y puede verse que el supuesto no se cumple, dado que existen diferencias entre ellas.

Covariance Matrices[a]

ComproBicicleta		IngresoAnual	TotalHijos	CantAutomoviles	Edad
0	IngresoAnual	1036720551	15459.408	17488.746	63439.604
	TotalHijos	15459.408	2.916	.574	9.928
	CantAutomoviles	17488.746	.574	1.272	1.551
	Edad	63439.604	9.928	1.551	151.121
1	IngresoAnual	1061983812	6714.406	18401.084	53234.505
	TotalHijos	6714.406	2.149	.304	7.652
	CantAutomoviles	18401.084	.304	1.275	2.707
	Edad	53234.505	7.652	2.707	99.849
Total	IngresoAnual	1049758936	11627.434	17478.144	57327.224
	TotalHijos	11627.434	2.658	.511	9.273
	CantAutomoviles	17478.144	.511	1.314	2.243
	Edad	57327.224	9.273	2.243	132.211

a. The total covariance matrix has 6486 degrees of freedom.

Test de medias (Lambda de Wilks)

Antes de continuar con el análisis, se debe verificar que los grupos puedan diferenciarse entre sí, de forma que no existan solapamientos en la clasificación. Para esto, utilizaremos el estadístico Lambda de Wilks, el cual es una medida de qué tan bien cada función separa los casos entre grupos. En este caso, la hipótesis nula es que las medias multivariantes de los grupos son iguales.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.935	307.446	4	.000

Al realizar el test con el software, se obtiene que el p-value no es significativo, y por lo tanto la hipótesis nula se rechaza. A partir de esto se puede concluir que la función tiene un gran poder discriminatorio de los grupos.

Función discriminante

Finalmente, el objetivo del análisis discriminante es calcular una función discriminante de forma que se pueda determinar la probabilidad de que un nuevo caso pertenezca a un grupo. Para nuestro problema, los coeficientes de la función se presentan en la siguiente tabla.

Canonical Discriminant Function Coefficients

	Function 1
IngresoAnual	.000
TotalHijos	.204
CantAutomoviles	.787
Edad	.021
(Constant)	-1.334

Unstandardized coefficients

De esta manera, la función discriminante se construye como:

$$D = 0 * IngresoAnual + 0,204 * TotalHijos + 0,787 * CantAutomoviles + 0,021 * Edad - 1,334$$

Los valores de los coeficientes de la función discriminante se pueden interpretar como la importancia que se le da a cada variable asociada para diferenciar los distintos grupos.

Matriz de confusión

A continuación se presenta la matriz de confusión del modelo desarrollado, en donde se puede comparar el porcentaje de acierto de la predicción tanto de los datos de entrenamiento como de los datos de prueba.

Classification Results[a,b]

				Predicted Group Membership		
			ComproBicicleta	0	1	Total
Cases Selected	Original	Count	0	1694	1066	2760
			1	662	1122	1784
		%	0	61.4	38.6	100.0
			1	37.1	62.9	100.0
Cases Not Selected	Original	Count	0	713	455	1168
			1	281	494	775
		%	0	61.0	39.0	100.0
			1	36.3	63.7	100.0

a. 62.0% of selected original grouped cases correctly classified.

b. 62.1% of unselected original grouped cases correctly classified.

De la tabla se puede extraer que los casos en los que el cliente no compró bicicletas son correctamente clasificados el 61,4% de las veces y los casos en que sí compraron bicicletas son correctamente clasificados el 62,9%. En total, la función puede clasificar correctamente el 62% de los casos. Dado que sólo se tienen dos grupos, la elección aleatoria de un grupo para un nuevo caso corresponde con el

50% de probabilidad, por lo cual el modelo, al superar este porcentaje, puede considerarse aceptable.

Por otro lado, si se comparan las matrices de confusión de los casos seleccionados y no seleccionados, puede verse que ambas arrojan porcentajes similares de acierto. Esto quiere decir que el modelo generado es estable.

Evaluación

Criterio de evaluación

Antes de proceder a la comparación de los distintos modelos, se debe recordar que en este caso existe una condición que pondera las predicciones. Por lo tanto, debe considerarse el hecho de que se prefiere enviar un mail innecesario a una persona que luego no realice la compra, a que la empresa pierda una venta por no haber enviado el correo a un potencial cliente.

A partir de esto, decidimos crear la siguiente matriz de costos, la cual penaliza los casos en los que se predice que el cliente no compra bicicletas (*ComproBicicleta = 0*), cuando el dato real es que sí lo hace (*ComproBicicleta = 1*).

		Datos Predichos	
		0	1
Datos Reales	0	0	1
	1	2	0

Evaluación y selección de los modelos construidos.

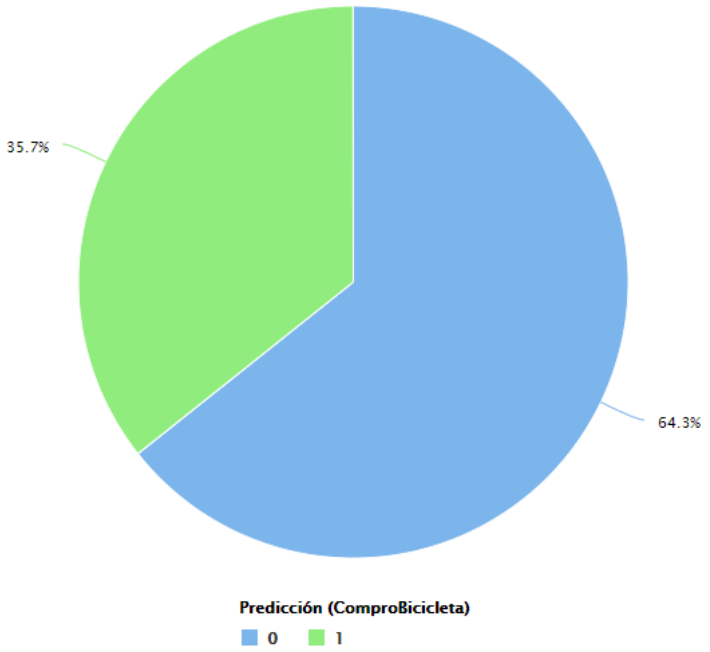
A continuación se presentará una tabla en la que se analizan y comparan los distintos modelos presentados en la sección anterior teniendo en cuenta los costos, con el fin de poder elegir el que mejor se adapta al problema.

Modelo	Real 0 - Predice 1	Real 1 - Predice 0	Costo total
Árbol de Decisión (Rapid Miner)	246	508	754
Árbol Quest	233	896	1129
Árbol CHAID	226	736	962
Árbol C5.0	186	572	758
KNN (K=1)	242	436	678
KNN (K=3)	249	468	717
LDA	455	562	1017

Observando la columna de costo total en la tabla, podemos observar que el modelo que representa el menor costo en el contexto del problema es el modelo KNN con un K = 1. Por lo tanto, en el apartado siguiente de implementación, haremos uso de este modelo para poder cumplir con el objetivo propuesto, determinando a qué cliente se enviará la publicidad.

Implementación

Una vez implementado el modelo KNN, con K=1, utilizando los datos del archivo destinatarios.txt para poder predecir la variable *ComproBicicleta* de cada uno de los nuevos casos, se obtienen los siguientes resultados.



En el gráfico podemos observar que los potenciales clientes a los cuales se les enviará el email con la publicidad propuesta por el departamento de marketing, representan un 35,7%, es decir, 535 clientes del total contenido en el nuevo dataset.

Caracterización del tipo de bicicleta a promocionar

Para la construcción de los distintos clusters se ha decidido trabajar directamente sobre el archivo 'destinatarios.csv' en el cual se filtraron aquellos individuos que no eran potenciales clientes. Es decir, sólo se analizarán las características de los nuevos potenciales clientes.

Clustering Jerárquico

El algoritmo de clustering que desarrollaremos en este apartado tiene la particularidad de que trabaja con una matriz de distancias y no es necesario que le indiquemos al inicio la cantidad de grupos o clusters que queremos formar. Por tal motivo, vamos a generar resultados para 2, 3 y 4 clusters y luego los compararemos a través de una tabla para decidir con cuál valor obtenemos el mejor modelo. Queremos hallar grupos de observaciones tales que las observaciones en un grupo sean similares a otros y, a su vez, diferentes a las observaciones de otros grupos. Esto lo podemos corroborar gráficamente viendo que los intervalos de confianza entre grupos no se solapen. Además, otra característica importante es que el promedio de los datos del cluster sea diferente a la media poblacional, ya que de otro modo no podríamos afirmar con certeza de que los datos no son una muestra de la población.

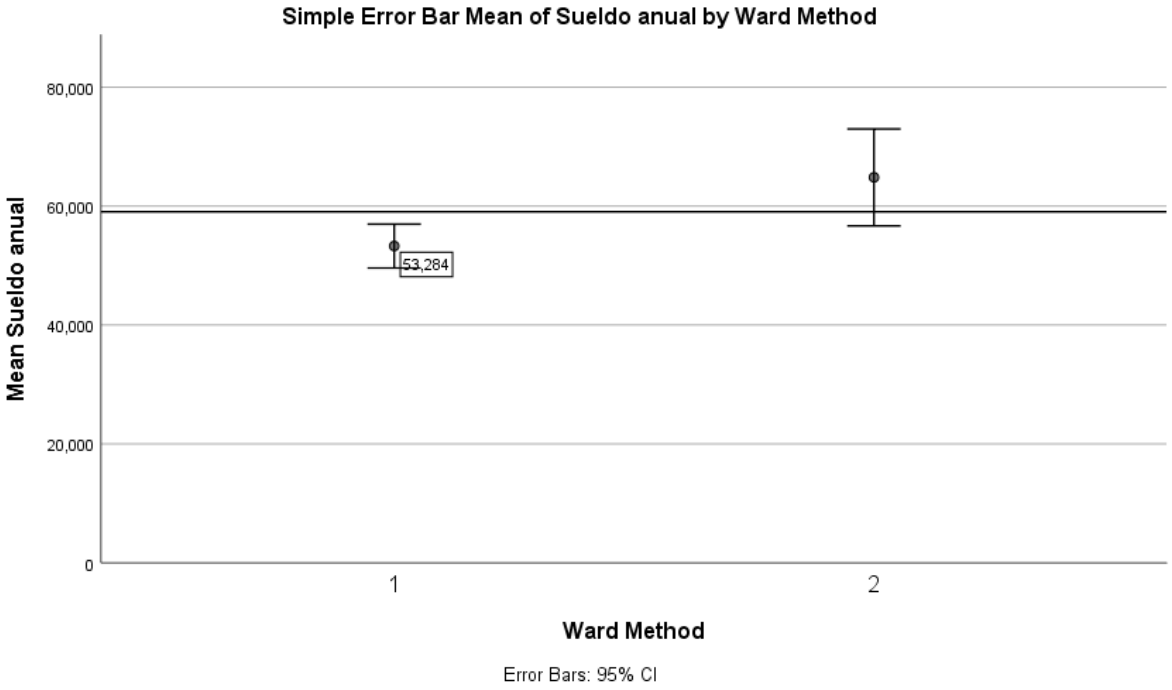
Para comenzar, primero haremos los diferentes clusters para las variables numéricas y luego continuaremos con las variables nominales o categóricas.

Variables numéricas

Para realizar el clustering jerárquico de las variables numéricas se ha decidido estandarizar las variables en un rango de 0 a 1, de forma que los ingresos, los cuales tienen un valor muy superior, se equiparen con los valores de las edades. De esta forma se evita que se de mayor importancia a sólo una de las variables. Como método de clasificación se utilizará el *Ward Method.*

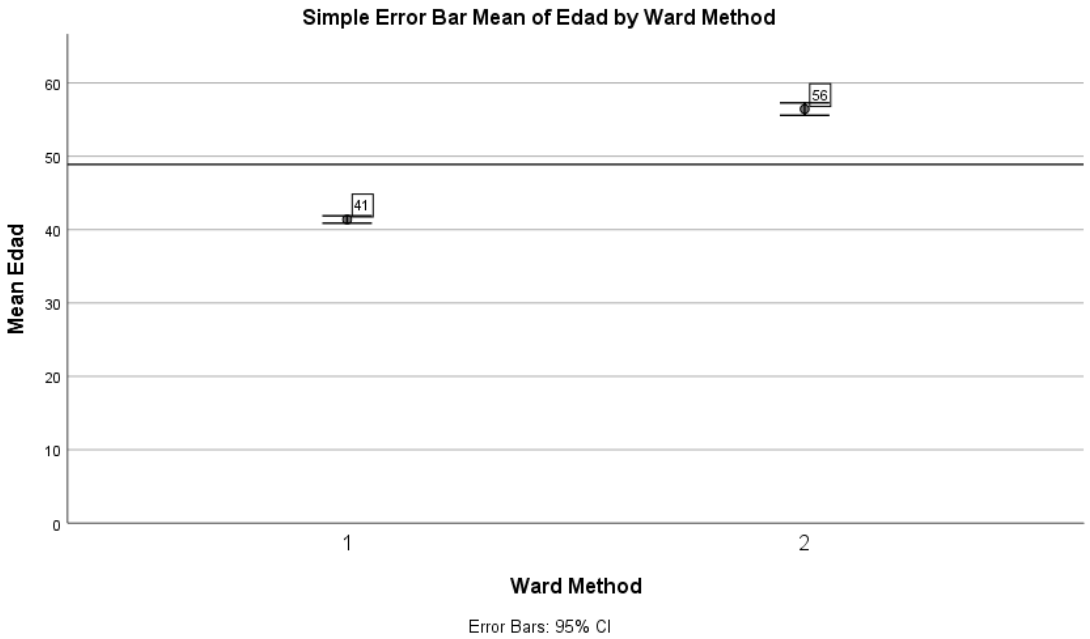
Resultados con 2 clusters

Ingreso Anual



En el gráfico superior se puede observar que el grupo 2 se encuentra atravesando la línea horizontal que indica la media de la población. Por lo tanto, puede afirmarse que el cluster 2 posee un comportamiento similar al de la población, lo cual indica que no se diferencian las características del grupo con el resto de los datos, y por lo tanto no es un buen modelo de clasificación.

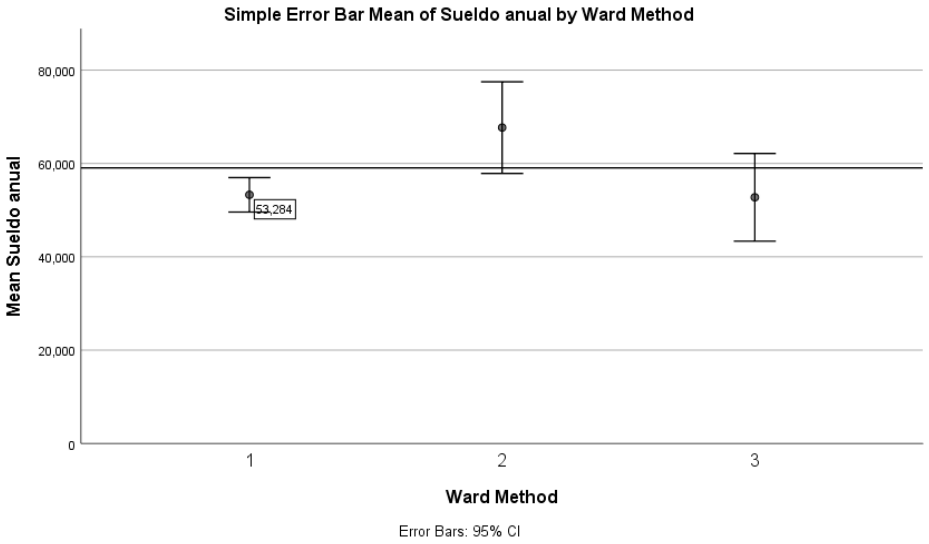
Edad



En el caso de la edad con dos clusters, a través del gráfico se puede ver que los intervalos de confianza no se solapan, y que además ningún grupo atraviesa la línea de la media poblacional. De esta forma, se puede concluir que ambos grupos son diferentes uno de otro, y además son distintos a la población en general, por lo cual aportan información que nos permite clasificar los datos.

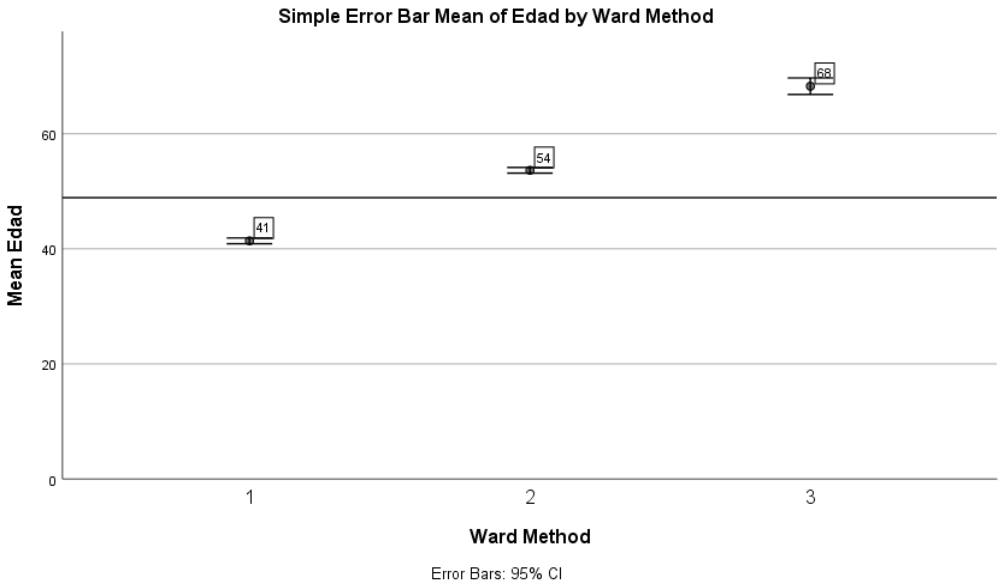
Resultados con 3 clusters

Ingreso Anual



A partir del gráfico, podemos observar que tanto el cluster 2, como el cluster 3 intersecan la línea de la media poblacional, y por lo tanto no son útiles para clasificar. De esta manera, para el análisis se tendrá en cuenta sólo el valor de la media del cluster 1.

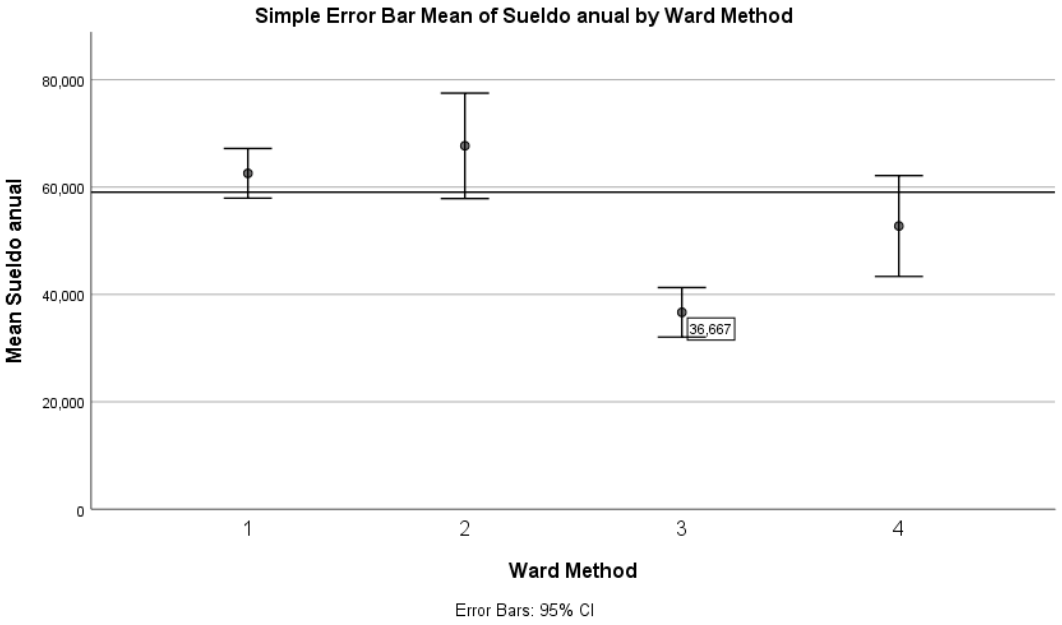
Edad



Analizando el gráfico se puede concluir que ninguno de los clusters se solapa con otro, y que además ninguno atraviesa la línea media poblacional. Por lo tanto, se tienen tres grupos de edad bien diferenciados que serán de utilidad para clasificar los datos de la población.

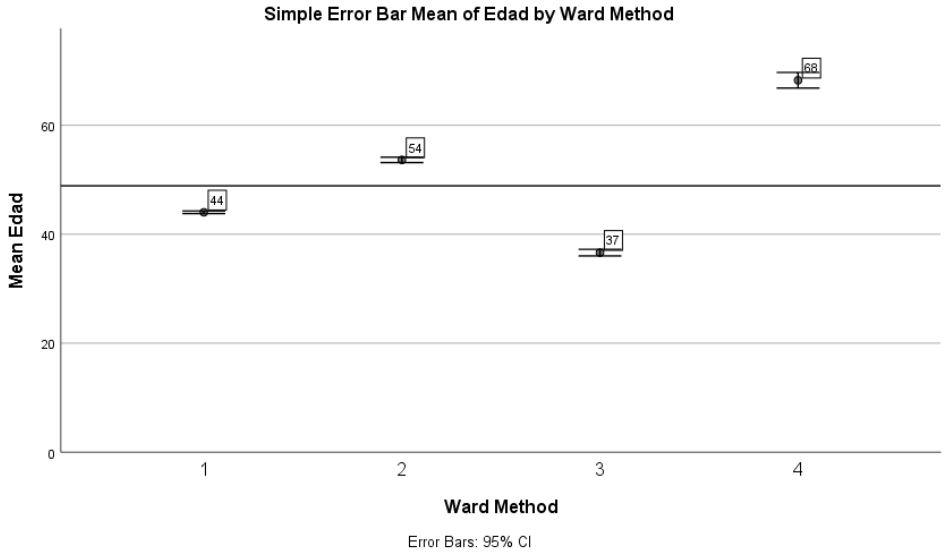
Resultados con 4 clusters

Ingreso Anual



Nuevamente, existen clusters en los que la variable sueldo anual tiene un comportamiento similar al de la población. Esto quiere decir que los grupos 1, 2 y 4 no aportan información de utilidad para clasificar, por lo cual se descartarán en el posterior análisis.

Edad



Para el caso de la edad con cuatro grupos, se tiene que ninguno de ellos cruza la línea poblacional y tampoco existe solapamiento entre ellos. Esto quiere decir que los cuatro clusters se diferencian entre sí y por lo tanto pueden utilizarse para clasificar a la población.

Tabla resumen

En el siguiente cuadro se resumen los resultados de cada uno de los casos para cada variable.

		Grupos								
		2		3			4			
Media poblacional	Variables	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 3	Grupo 1	Grupo 2	Grupo 3	Grupo 4
59046,73	IngresoAnual	53284	--	53284	--	--	--	--	36667	--
48,89	Edad	41	56	41	54	68	44	54	37	68

En ella se puede notar que la variable ingreso anual no es un buen parámetro para clasificar dado que no divide adecuadamente a la población en distintos grupos. Por otra parte, si se descarta dicha variable y se utiliza sólo la variable edad para clasificar, consideramos que el mejor modelo corresponde al de tres grupos.

Esto se debe a que con dos grupos, la población se divide en clusters muy grandes que pueden no presentar suficientes diferencias dentro de los mismos. Por otro lado, si se eligiera trabajar con cuatro grupos, la diferencia entre las medias de 37 (cluster 3) y 44 (cluster 4) no parece ser significativa.

Variables categóricas

En este apartado realizaremos el algoritmo de clustering jerárquico para las variables categóricas o nominales. Presentaremos los resultados obtenidos para 2, 3 y 4 clusters.

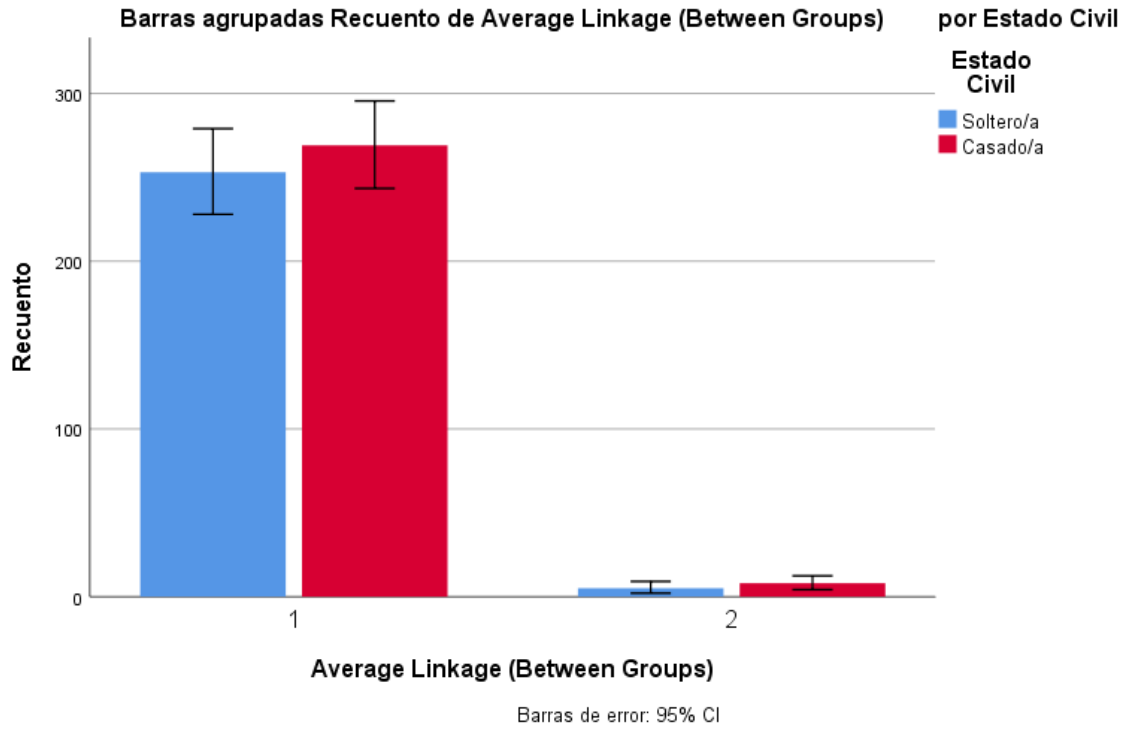
Las variables como “Estado Civil”, “Género”, “Educación”, “Ocupación”, “Distancia” y “Región” se recodificaron como numéricas para poder trabajarlas. Por ejemplo, la variable “Estado Civil” se recodificó con un valor de 1 para “S” (Soltero/a) y el valor de 2 para “C” (Casado/a). El mismo criterio se siguió con las demás variables nombradas.

Los parámetros usados para generar los diferentes clusters fueron los siguientes:

- Método de agrupación en clústeres: enlace entre grupos.
- Medida: recuento a través de la medida de chi cuadrado.

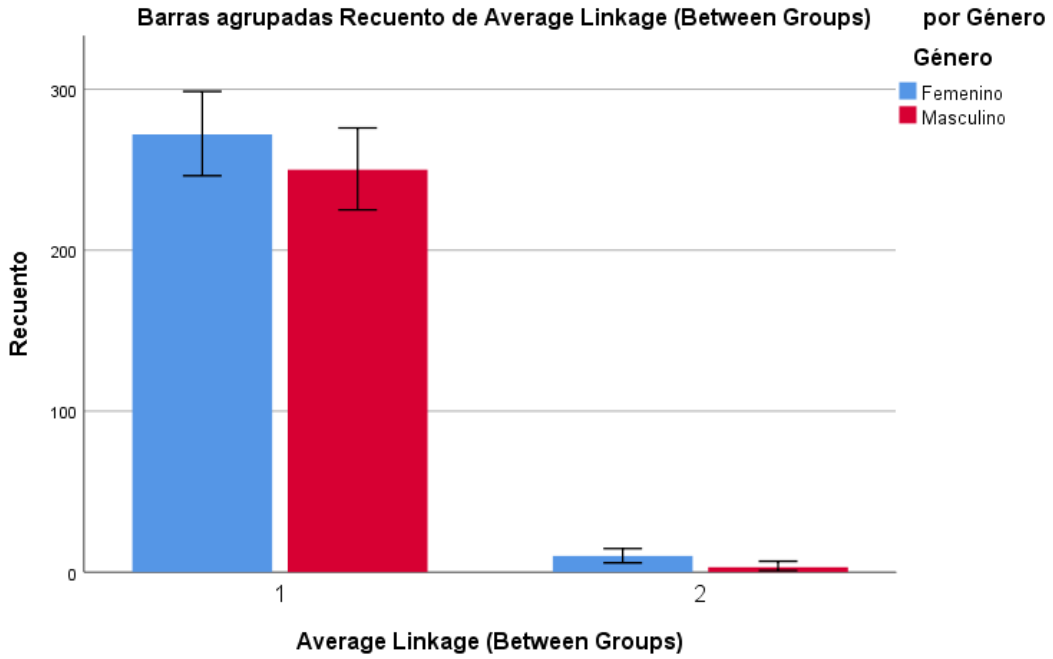
Resultados con 2 clusters

Estado Civil



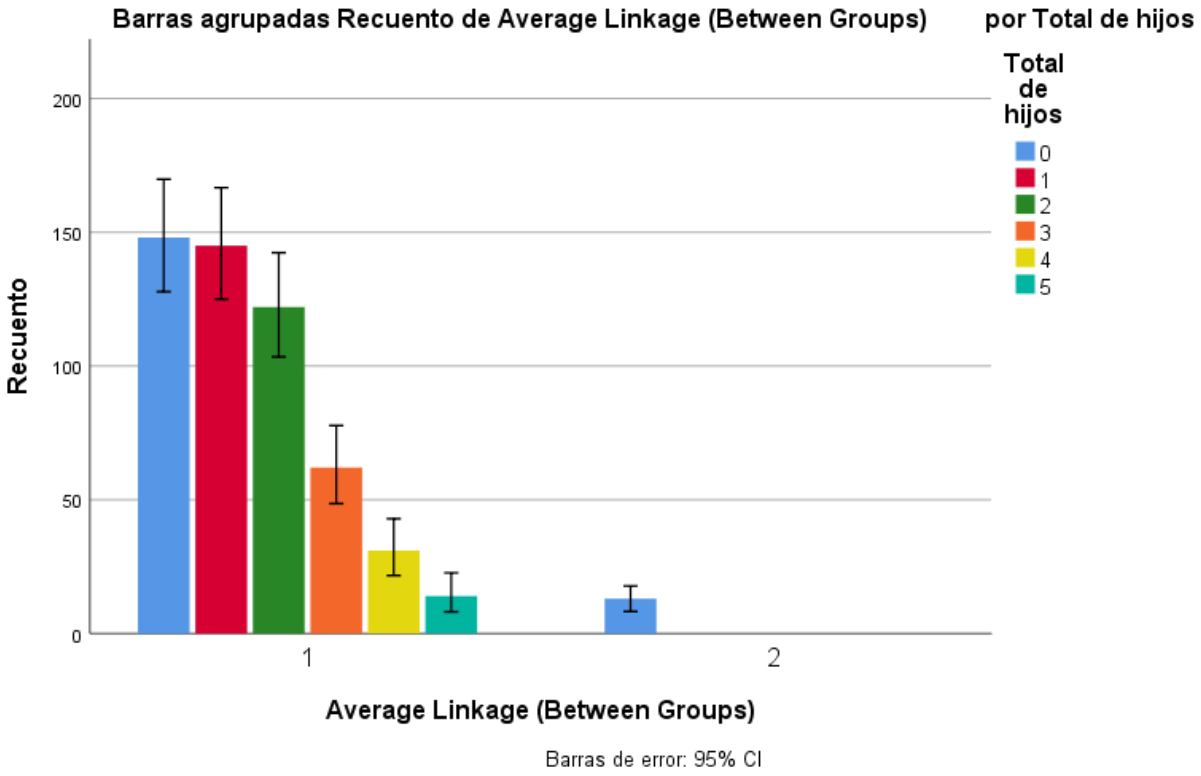
En el gráfico podemos observar que para ambos clusters, los intervalos de confianza se superponen. Esto quiere decir que los clusters no caracterizan correctamente a la variable y por lo tanto no podemos concluir.

Género



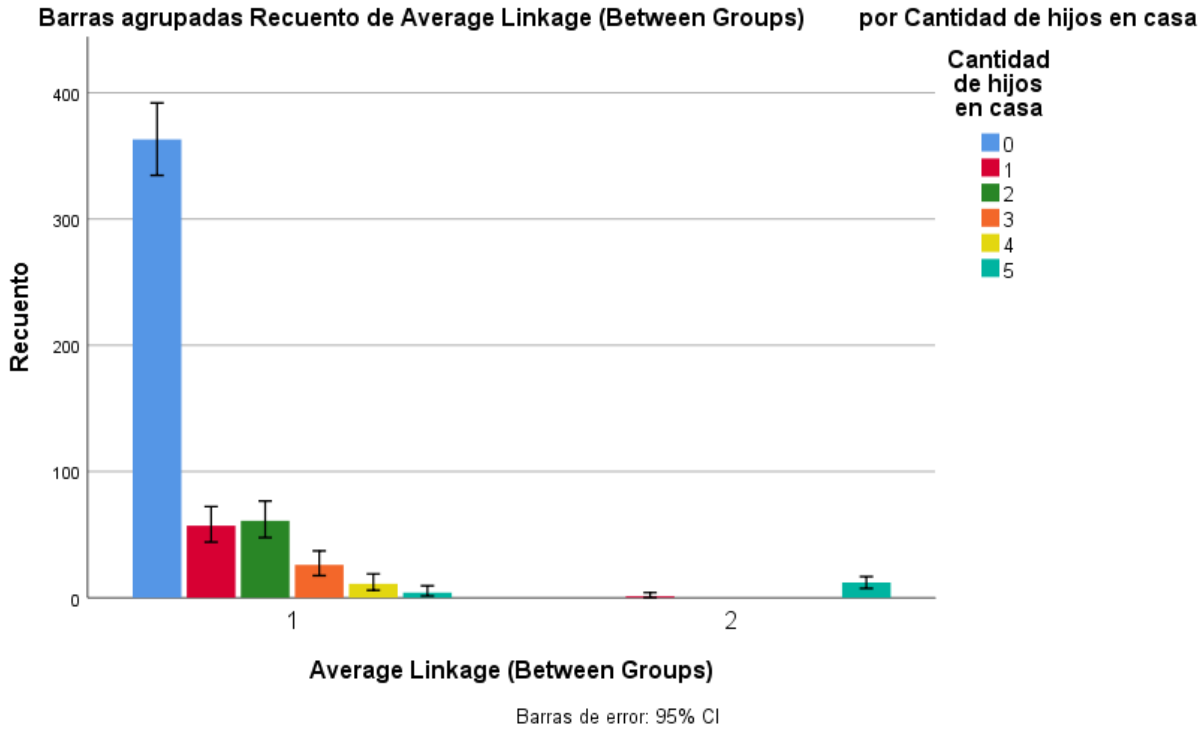
De forma similar a lo ocurrido con el caso anterior, podemos ver que los clusters no caracterizan a la variable de alguna forma específica, por lo tanto no aporta información que ayude a encontrar patrones dentro de las observaciones.

Total Hijos



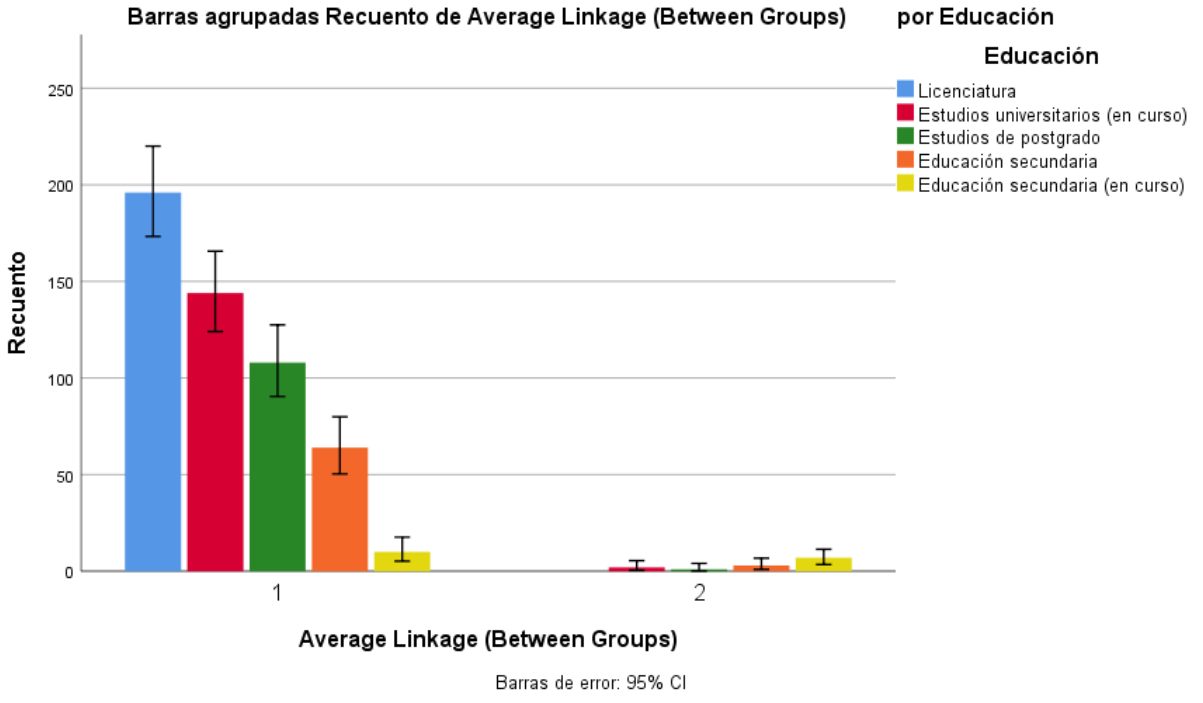
Para esta variable, viendo el primer cluster notamos que los intervalos se superponen. Sin embargo, al contar con muchas categorías y las observaciones de las 3 primeras son las que destacan en el cluster, decidimos que este mismo destaca aquellas observaciones con baja cantidad total de hijos (0, 1 y 2). En el cluster 2 se computan las observaciones con total de hijos igual a 0.

Cantidad Hijos en Casa



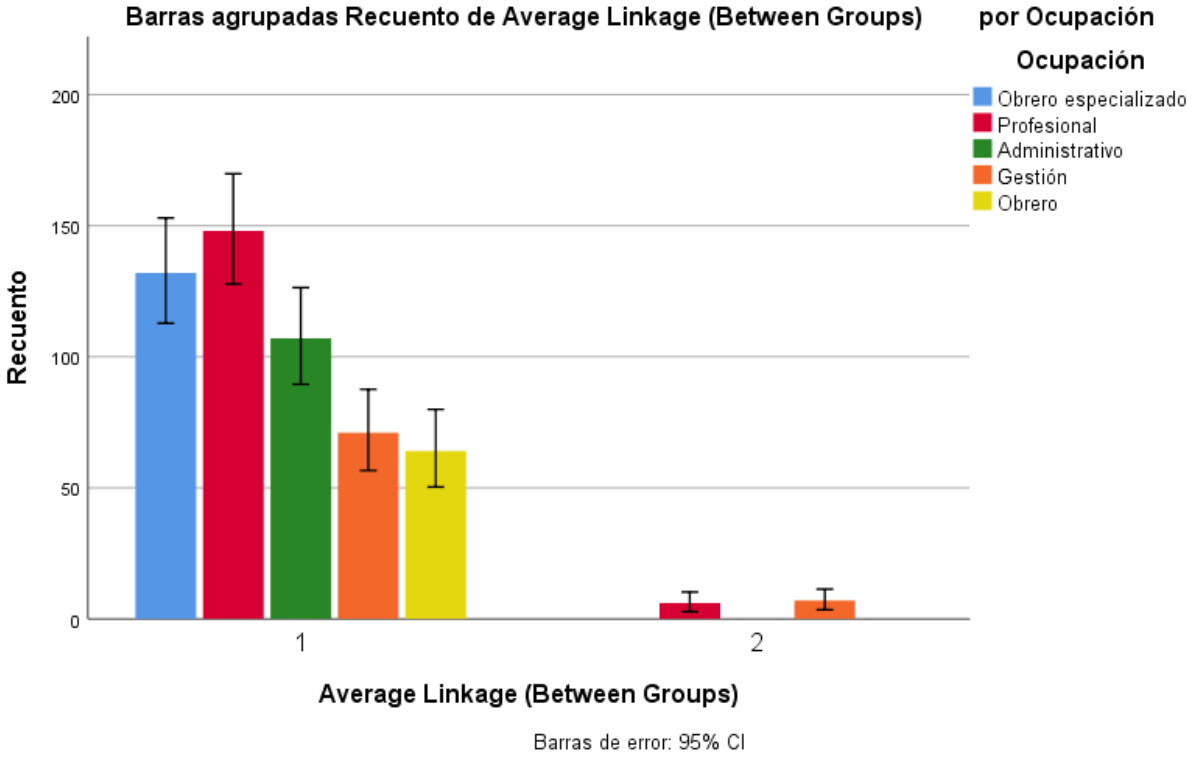
En el gráfico anterior se observa que en el cluster 1 la mayoría de los casos corresponden a 0 hijos en casa. Por otro lado, para el cluster 2, si bien puede haber un solapamiento entre los intervalos de confianza, predominan los individuos con una cantidad alta de hijos en casa la cual es igual a 5.

Educación



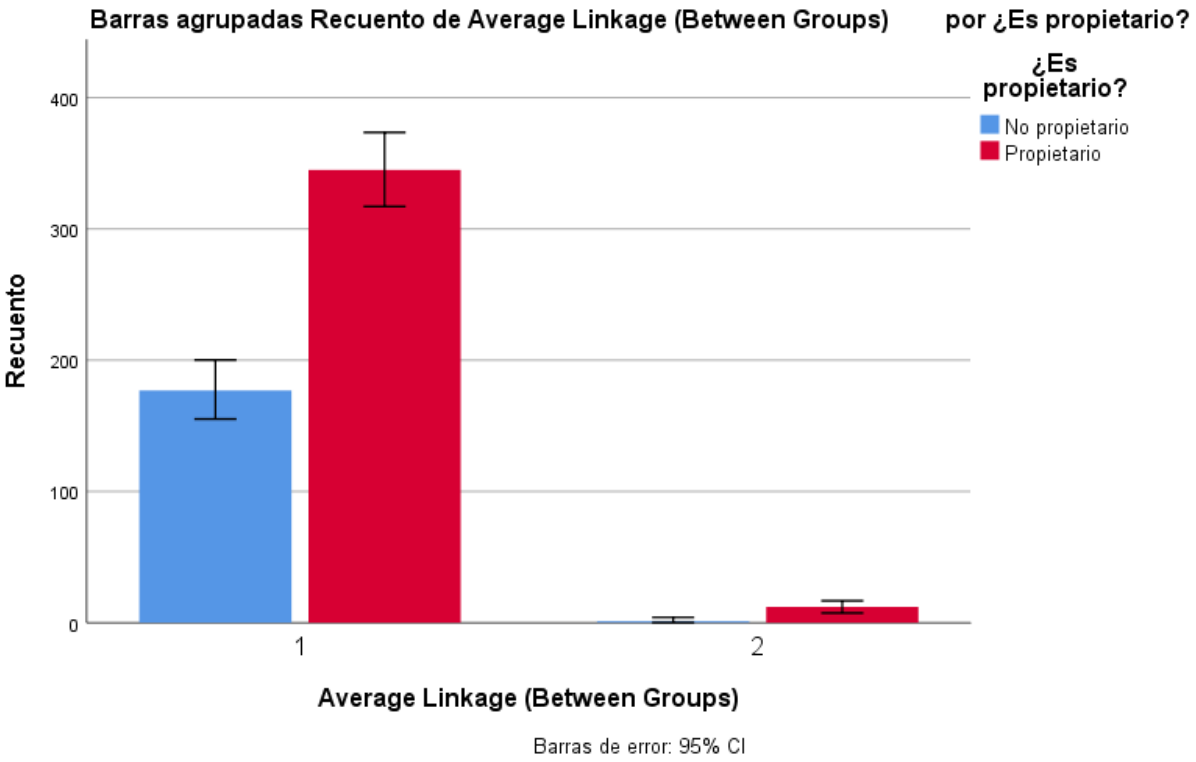
Podemos observar como la totalidad de las observaciones correspondientes al nivel de educación licenciatura son categorizadas únicamente en el cluster 1, dejando de lado esta situación, la categorización de los demás niveles educativos no aportan información, debido al solapamiento entre sus intervalos de confianza.

Ocupación



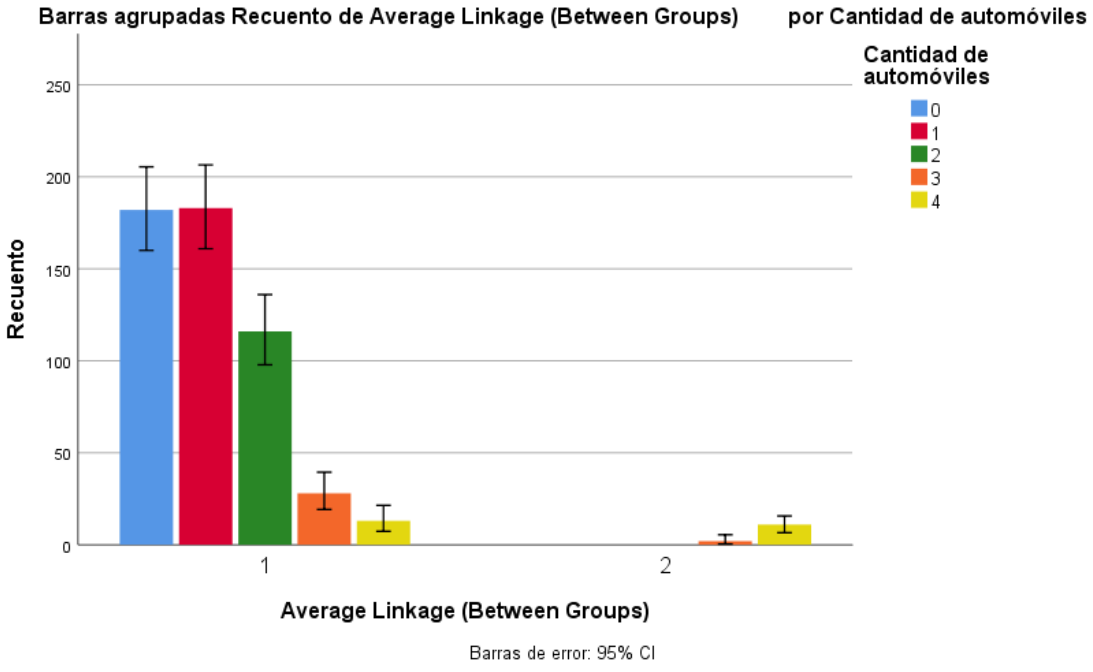
Interpretando este gráfico, podemos notar que en el cluster 1 hay solapamiento en los intervalos de confianza, es decir que este cluster no caracteriza bien a la variable. En el caso del segundo cluster, notamos un solapamiento, pero al haber dos clases únicamente, podemos decir que este cluster caracteriza a las observaciones con ocupación Profesional y Gestión.

Propietario



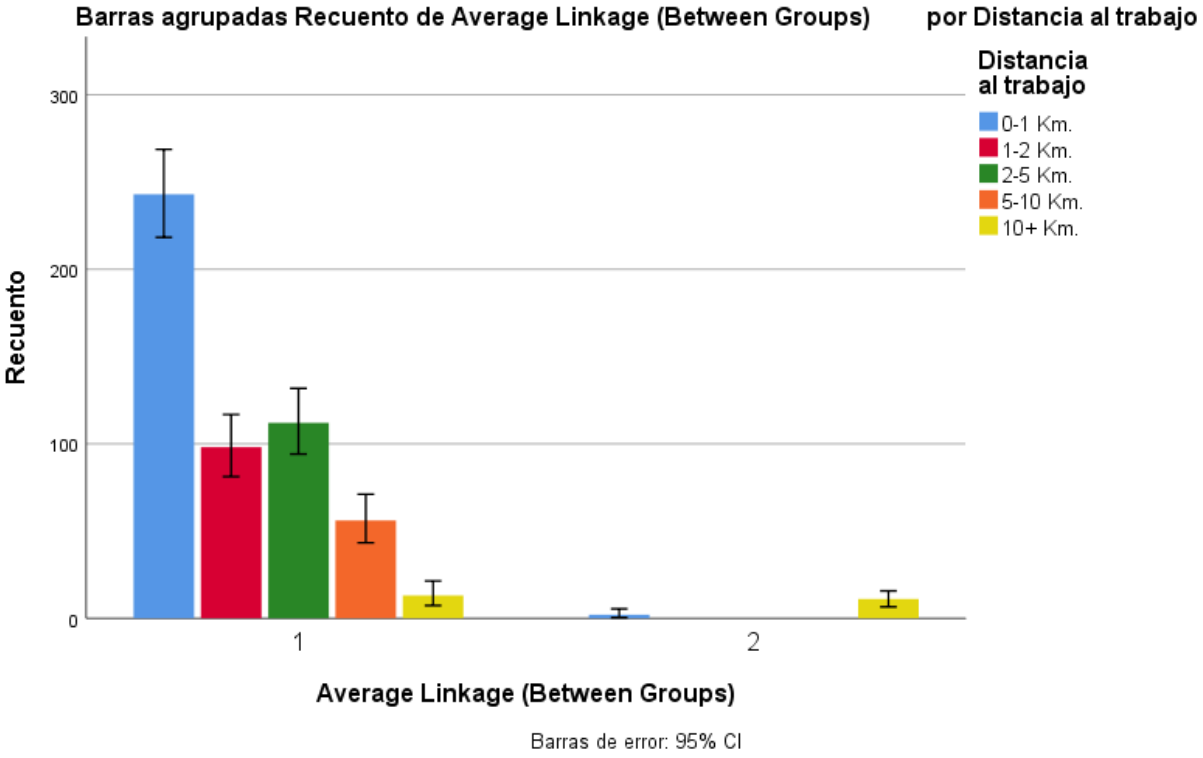
En este caso, por un lado, podemos caracterizar al cluster 1 como aquel en el que predominan los individuos que son propietarios. Por otro lado, el solapamiento existente en el cluster 2 no permite que podamos caracterizar con certeza a las observaciones.

Cantidad Automóviles



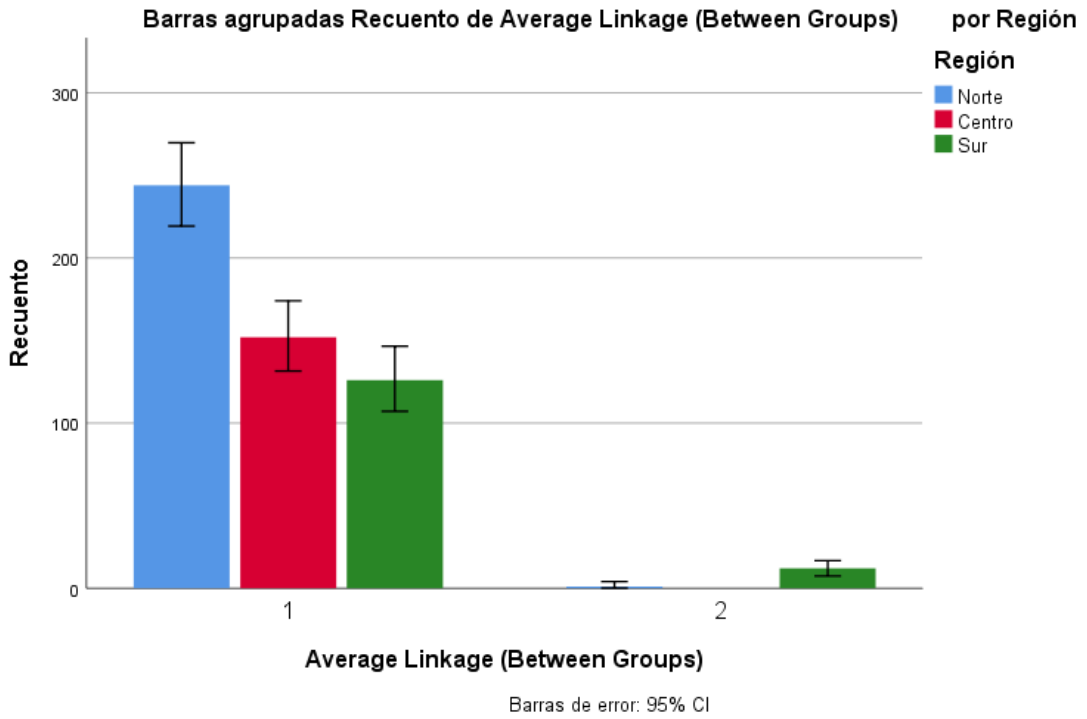
Mirando el gráfico anterior, podemos ver una superposición entre los intervalos de las 2 primeras clases del primer cluster. De todos modos, como la mayoría de las observaciones están presentes en esas clases, podemos decir que este cluster incluye a las observaciones que poseen baja cantidad de automóviles (0 y 1). Para el caso del segundo cluster, predominan los individuos con una cantidad alta de automóviles (3 y 4).

Distancia



A partir del gráfico, podemos observar como gran cantidad de los individuos, cuya distancia al trabajo está entre 0 y 10 km, son caracterizados bajo el primer cluster; con una distinción amplia para la categoría que representa a una distancia entre 0 y 1 km, que representa la mayoría de las observaciones. Para la categoría de 10 km o más, vemos como está representada en ambos clusters con similar recuento, lo que indica que no existe una distinción entre ambos clusters con respecto a dicha categoría.

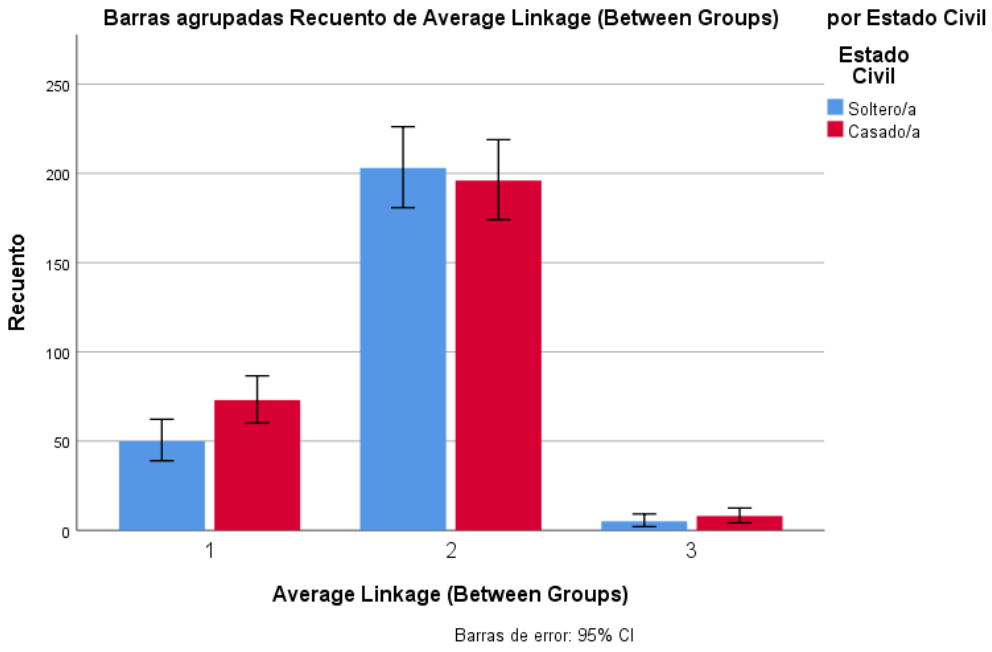
Región



Para esta variable región, en el primer cluster notamos que aunque hay superposición entre los intervalos de Centro y Sur, el valor que más predomina es Norte, entonces podemos afirmar que el cluster 1 caracteriza a los clientes que residen en la región Norte. En el segundo cluster prevalecen los clientes que viven en la región Sur.

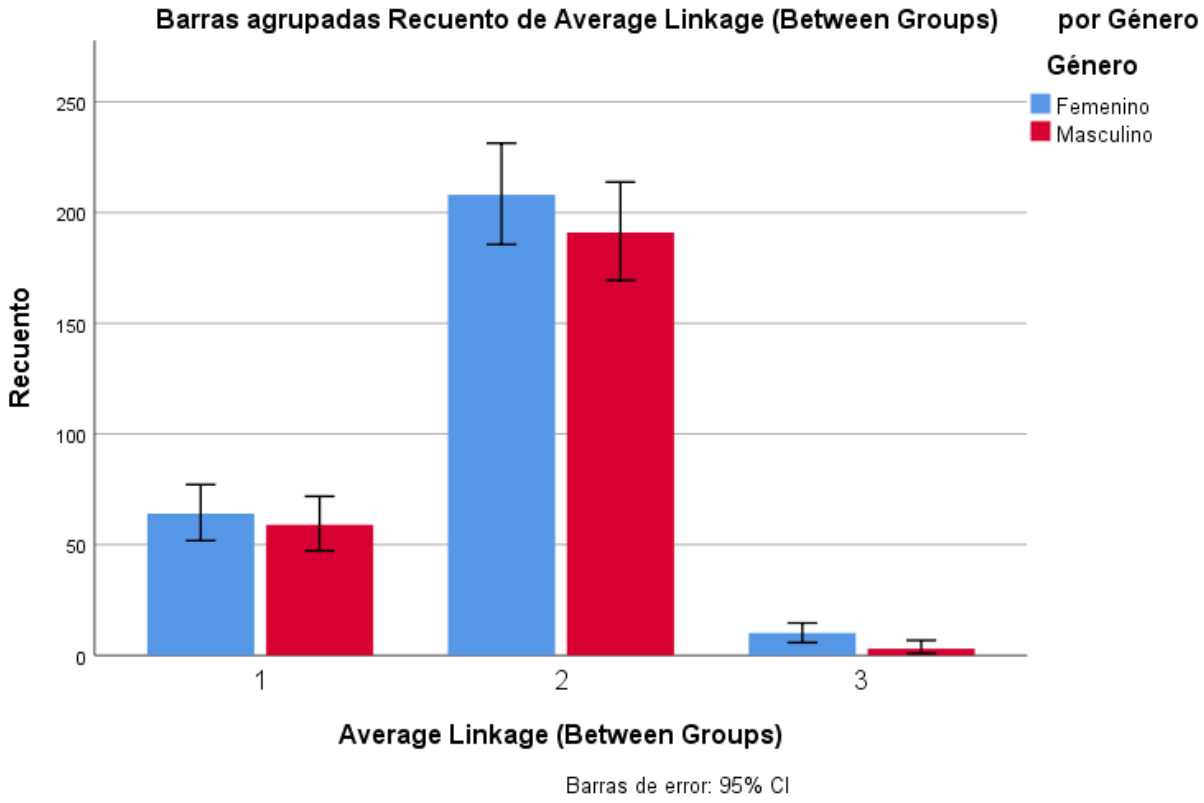
Resultados con 3 clusters

Estado Civil



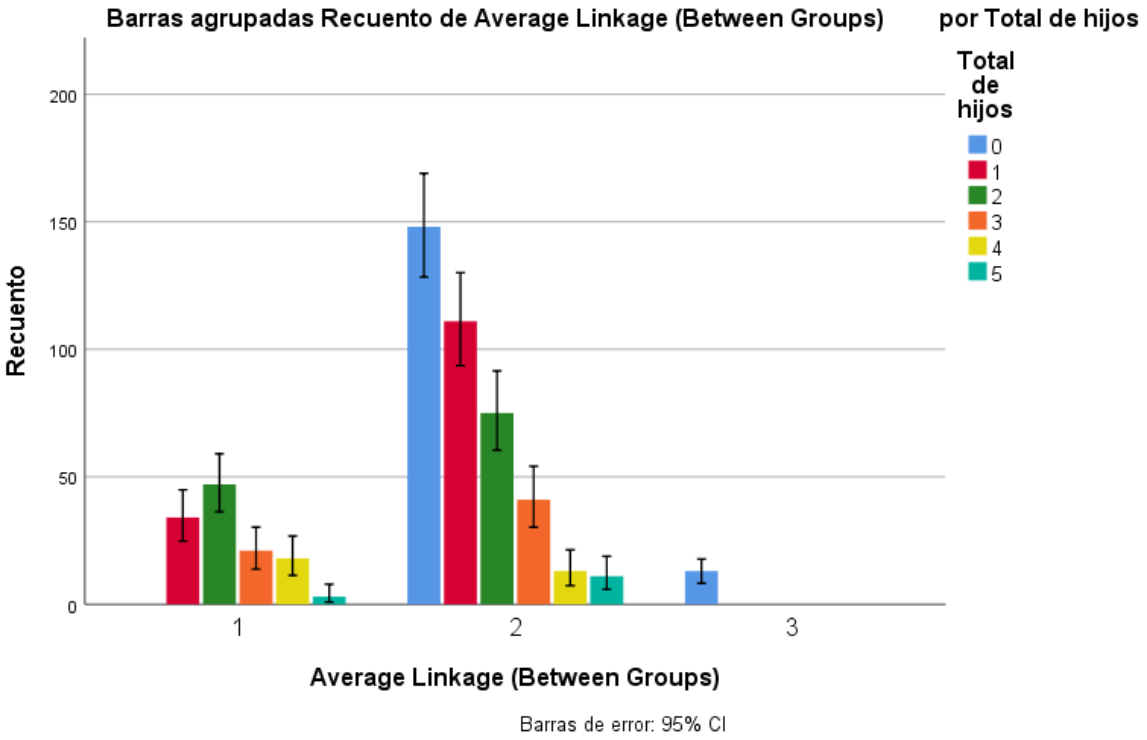
De la misma manera que en el modelo con 2 clusters, dentro de cada cluster hay solapamiento de intervalos de confianza, por tanto, no es posible caracterizar con alguno de estos valores a cada uno de los clusters.

Género



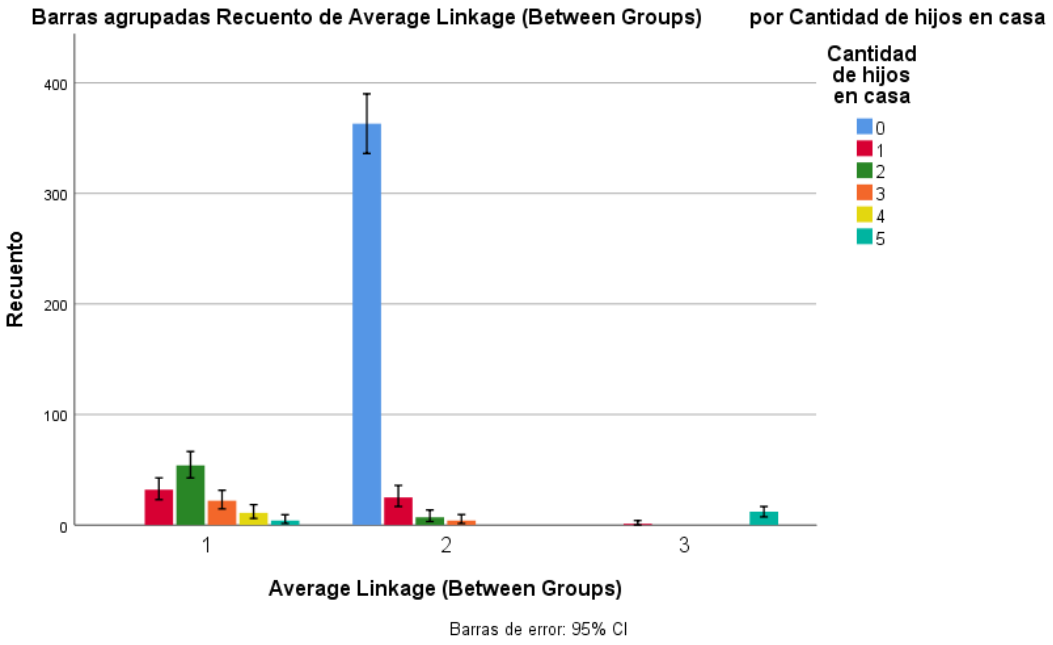
Para la variable Género tenemos una situación similar a la anterior, dentro de cada cluster vemos una superposición en sus respectivos intervalos de confianza. De este modo, no es posible caracterizar a cada cluster.

Total Hijos



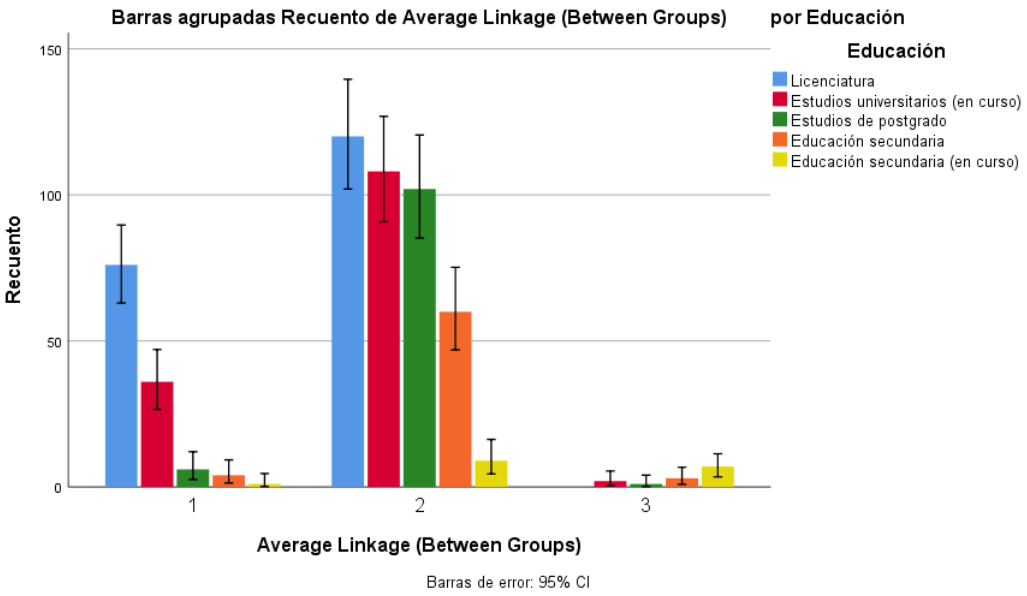
Si miramos el primer cluster, notamos que hay nuevamente superposición de intervalos de confianza y no podemos caracterizar dicho cluster. El tercer cluster está formado por aquellos clientes que no poseen hijos. Por último, el segundo cluster, si bien hay superposición de intervalos, hay una buena cantidad de clientes que no poseen hijos, por ende decidimos tomar que en este cluster predominan los que no poseen hijos.

Cantidad Hijos en Casa



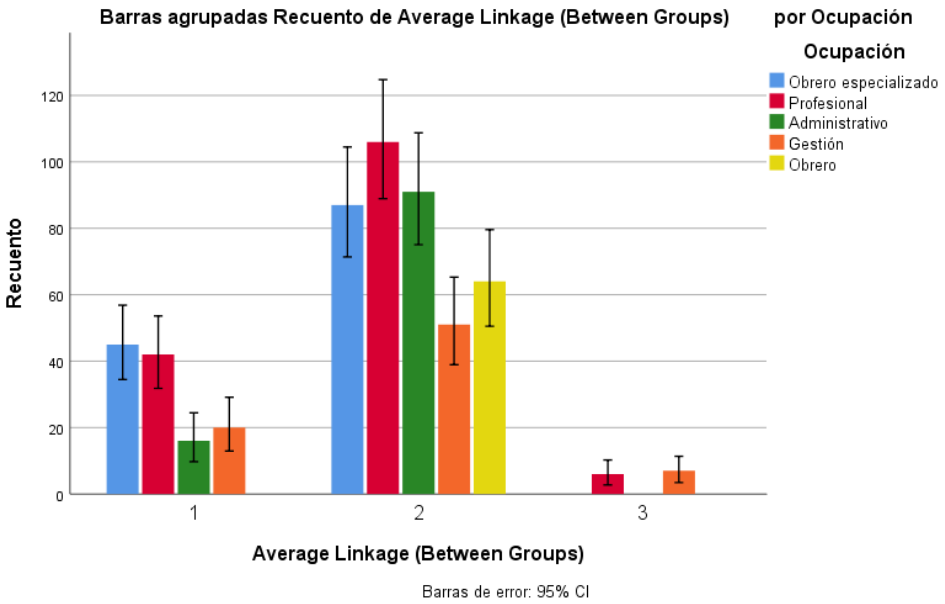
Para el primer cluster no podemos determinar con certeza debido a que hay superposición de intervalos. Para el cluster 2 y 3 sí podemos caracterizar y en el cluster 2 claramente predominan aquellos clientes que no poseen hijos en casa y en el tercer cluster aquellos que poseen 5 hijos en casa.

Educación



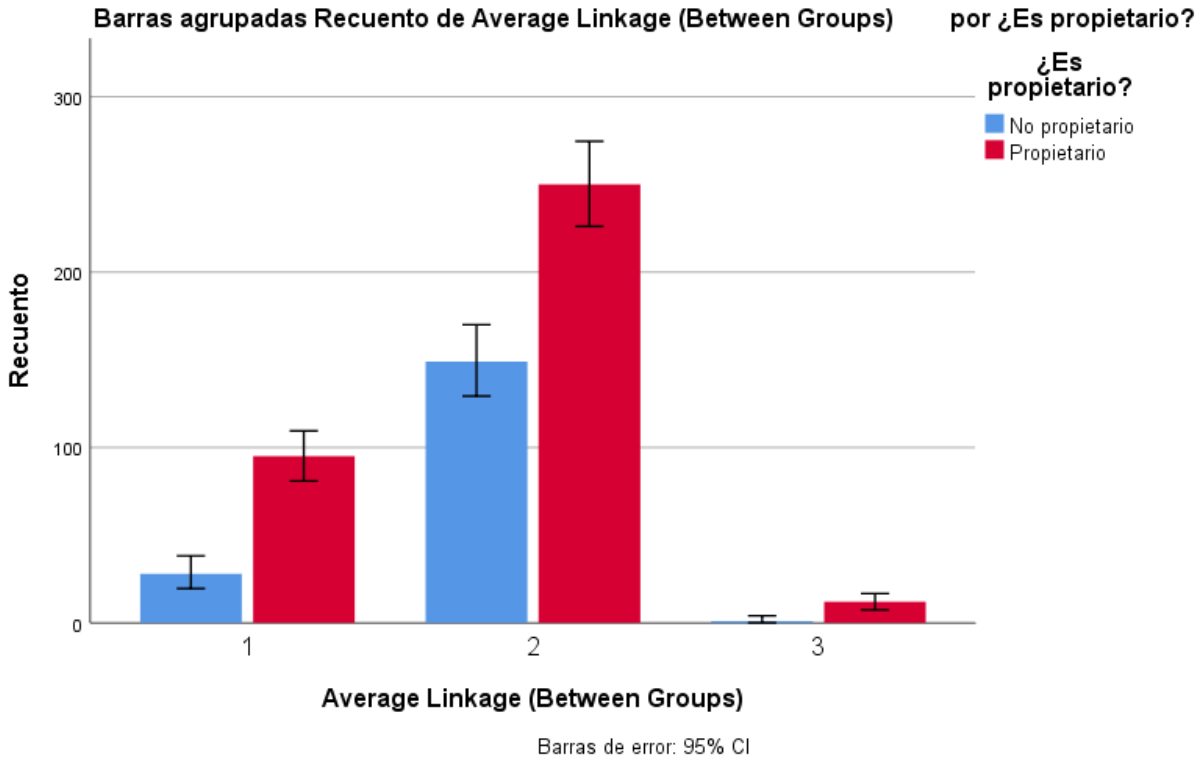
En el gráfico anterior vemos que el primer cluster caracteriza a aquellos individuos con el tipo de educación “Licenciatura”. En el cluster 2, si bien hay superposición, podemos igualmente decir que este cluster caracteriza a los clientes con tipo de educación “Licenciatura”, “Estudios universitarios (en curso)” y “Estudios de postgrado”. En el tercer cluster vemos mucha superposición, así que no es posible concluir acerca de la caracterización.

Ocupación



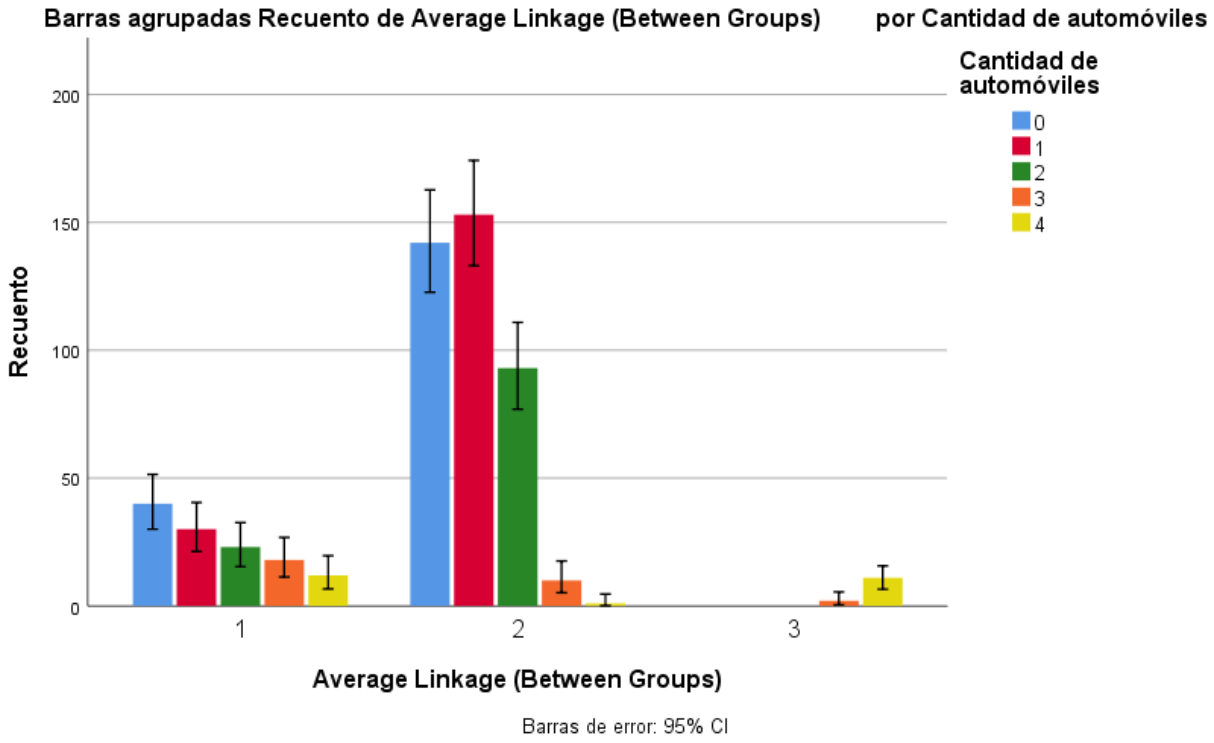
En este gráfico vemos que en todos los clusters tenemos solapamiento de los intervalos de confianza. Sin embargo, para el primer grupo podemos decir que predominan los clientes con ocupación de “Obrero especializado” y “Profesional”. En el segundo grupo, no podemos decidir con certeza ya que hay un mayor solapamiento y el recuento es bastante cercano entre las diferentes ocupaciones. En el grupo número 3, también hay superposición, pero al haber recuento de sólo dos tipos de ocupaciones, podemos afirmar que en este cluster se encuentra los clientes con ocupación “Profesional” y “Gestión”.

Propietario



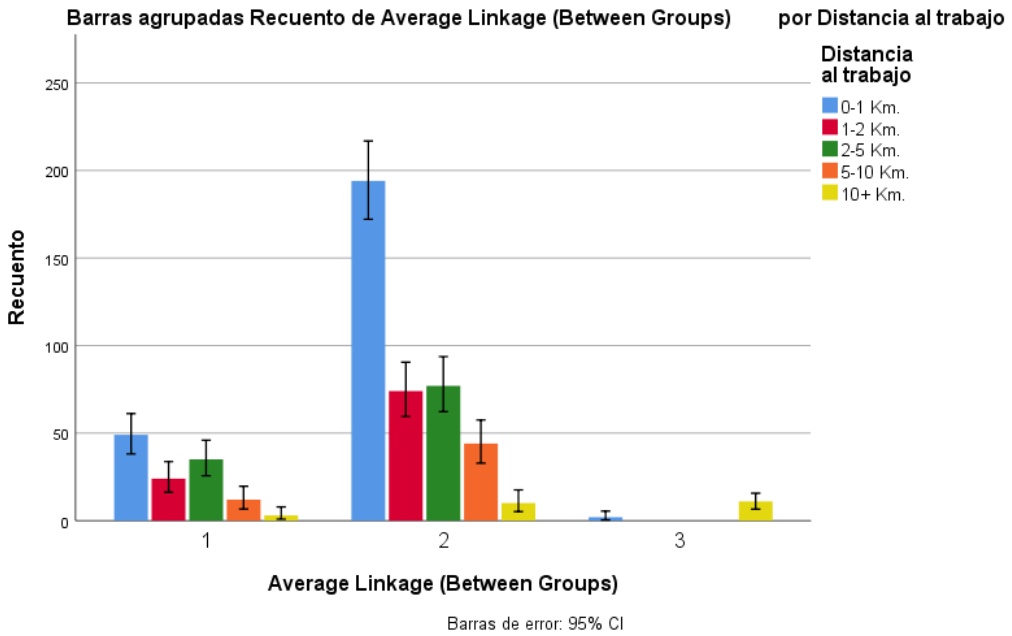
Tanto para el cluster 1 como para el 2, se caracterizan los clientes que son propietarios. Para el cluster no podemos afirmar ya que se observa un solapamiento entre los intervalos de confianza.

Cantidad Automóviles



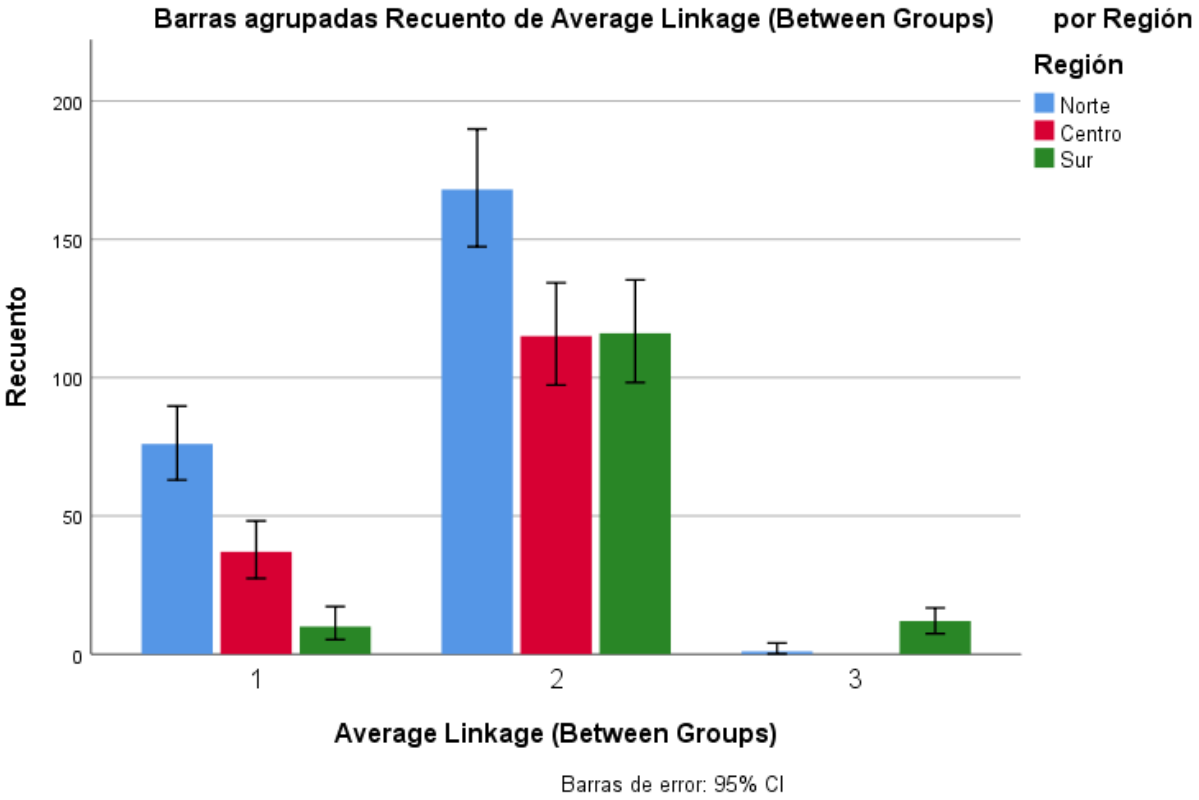
En el cluster número 2 vemos que predominan los clientes que poseen una cantidad baja de automóviles (0 y 1). En el cluster número 3, en cambio, predominan aquellos con una alta cantidad de automóviles (3 y 4). Para el caso del cluster número 1, no podemos caracterizar correctamente debido a la superposición de intervalos de confianza.

Distancia



En este gráfico, por un lado, para los clusters 1 y 3, la superposición de los intervalos no nos permite concluir con certeza respecto a la caracterización. Por otro lado, el segundo cluster caracteriza a los clientes que recorren entre “0-1 Km” de distancia al trabajo.

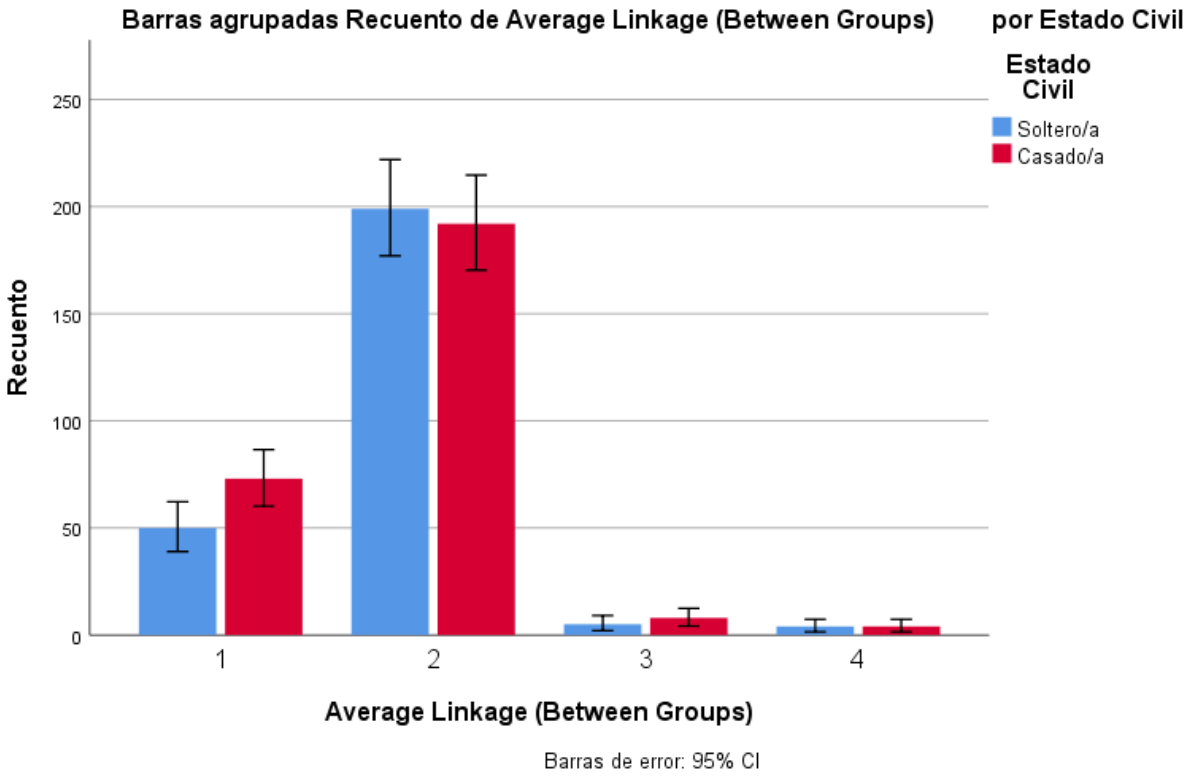
Región



En los clusters formados para esta variable, vemos que en los 2 primeros se encuentran los clientes que viven en la región etiquetada como “Norte”, mientras que para el tercer cluster encontramos predominantemente los que residen en la región “Sur”.

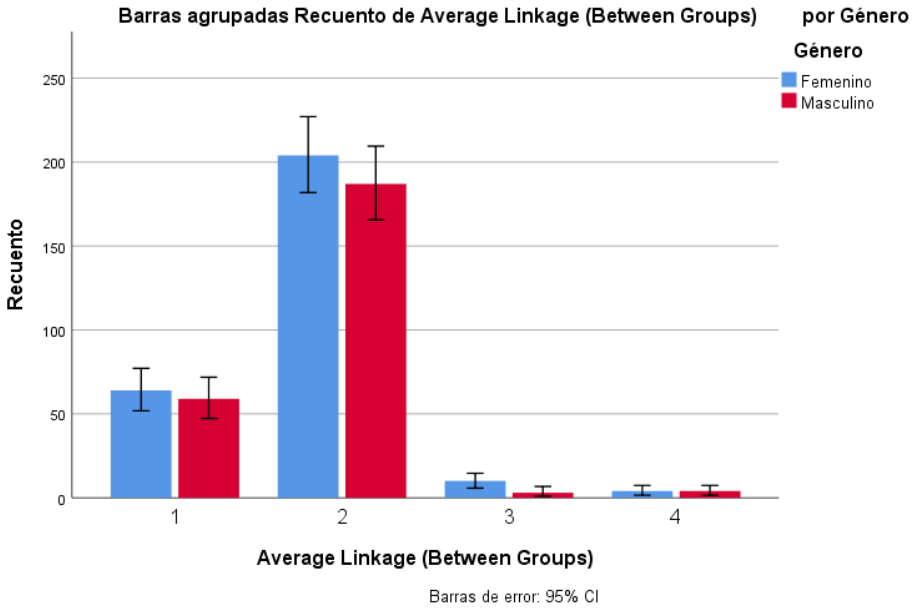
Resultados con 4 clusters

Estado Civil



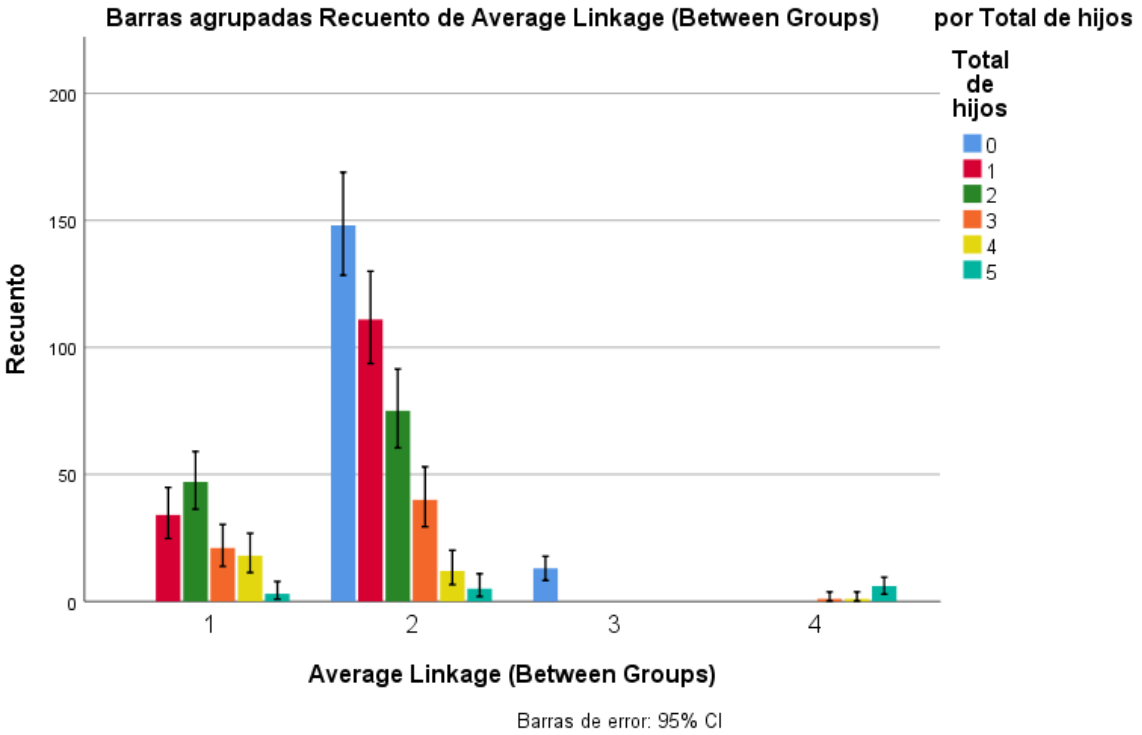
Para esta variable y para 4 clusters, sucede la misma situación que analizamos anteriormente para la misma variable; los intervalos de confianza se superponen para todos los clusters, por lo tanto no podemos caracterizar a cada uno de ellos.

Género



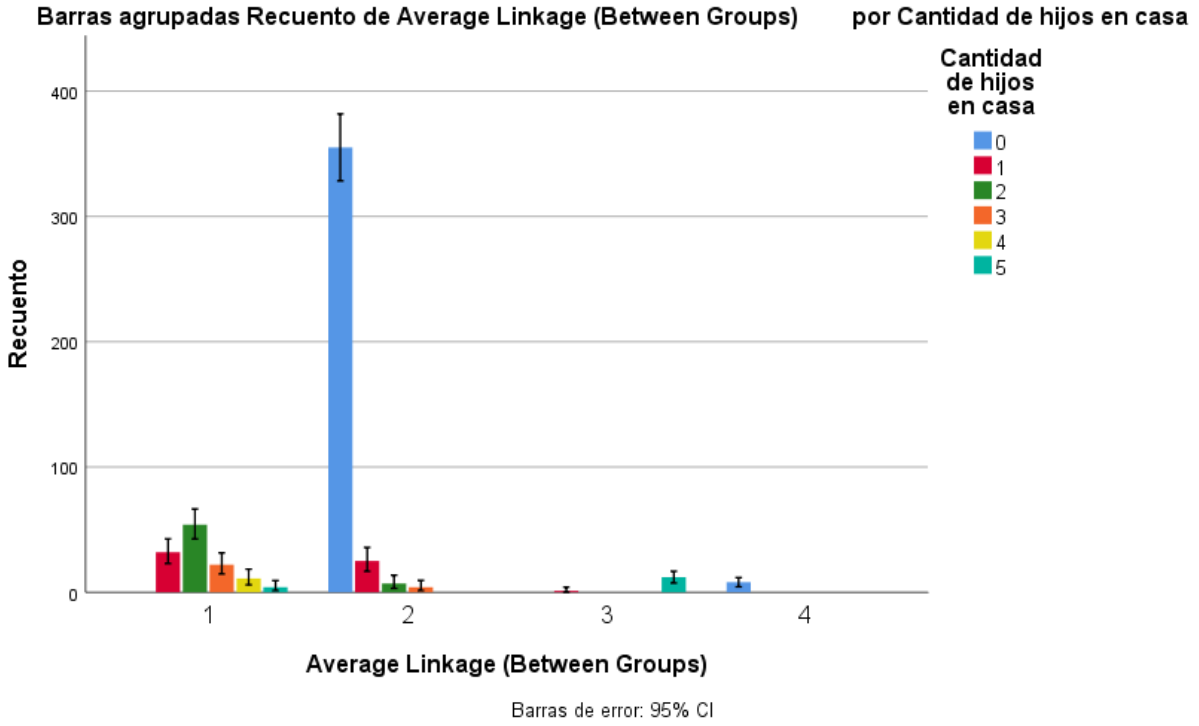
Al igual que las conclusiones anteriores, en el gráfico observamos el solapamiento de los intervalos para todos los clusters. De esta manera, resulta imposible caracterizar a cada cluster.

Total Hijos



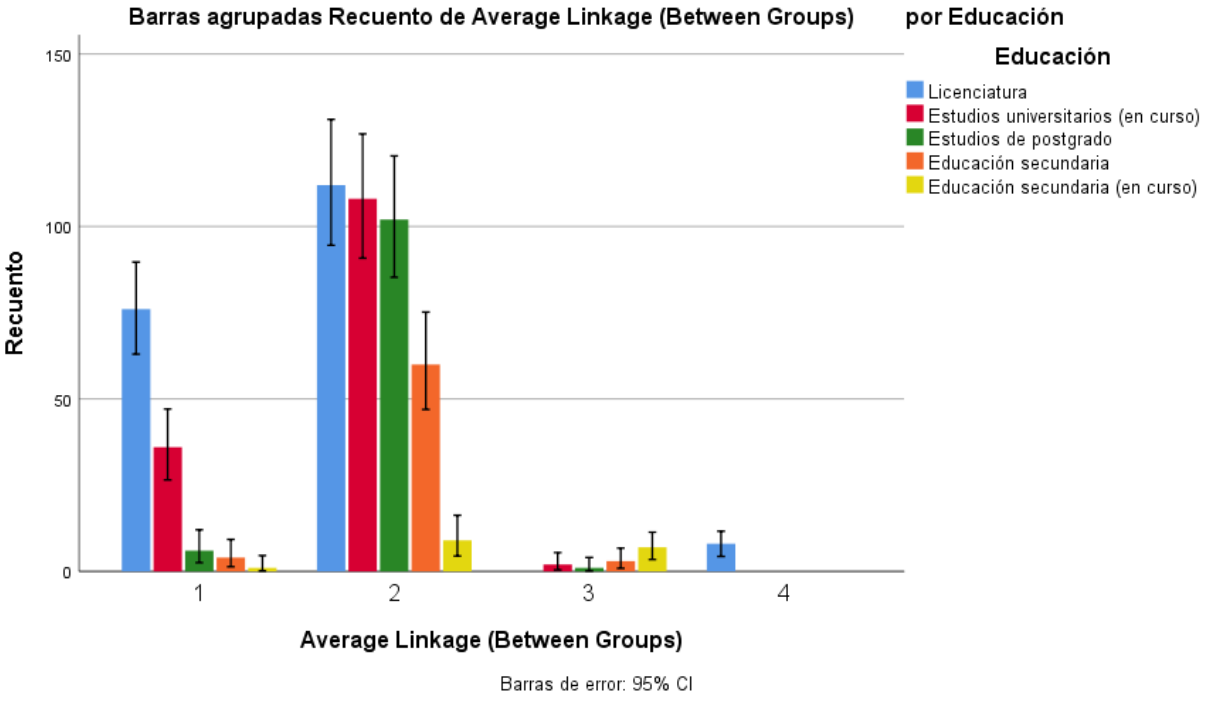
Para el primer cluster, no podemos caracterizarlo de manera correcta por el solapamiento. En el segundo cluster, hay un poco de superposición pero se observa que el recuento del total de hijos igual a 0 es el que más destaca con respecto al resto. Así que podemos decir que en el cluster 2 predominan los clientes con un total de hijos igual a 0. En el cluster 3 se observa la misma caracterización (total de hijos igual a 0). Por último, el cuarto cluster al tener solapamiento y bajo número de recuento, no nos permite concluir correctamente respecto de la caracterización.

Cantidad Hijos en Casa



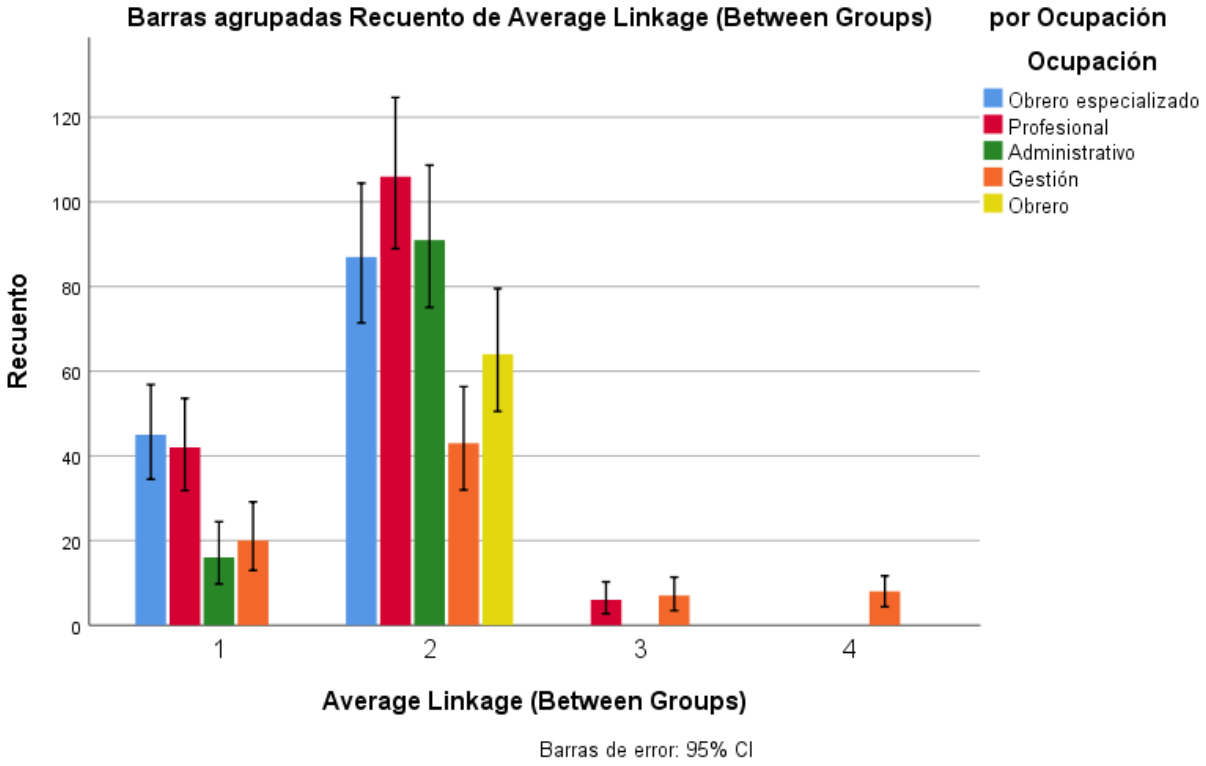
En este gráfico vemos claramente que en el cluster 2 y 4 predominan los clientes que poseen baja cantidad de hijos en casa. Los clusters 1 y 3, no caracterizan bien a las observaciones porque hay un solapamiento de intervalos y un bajo recuento, respectivamente.

Educación



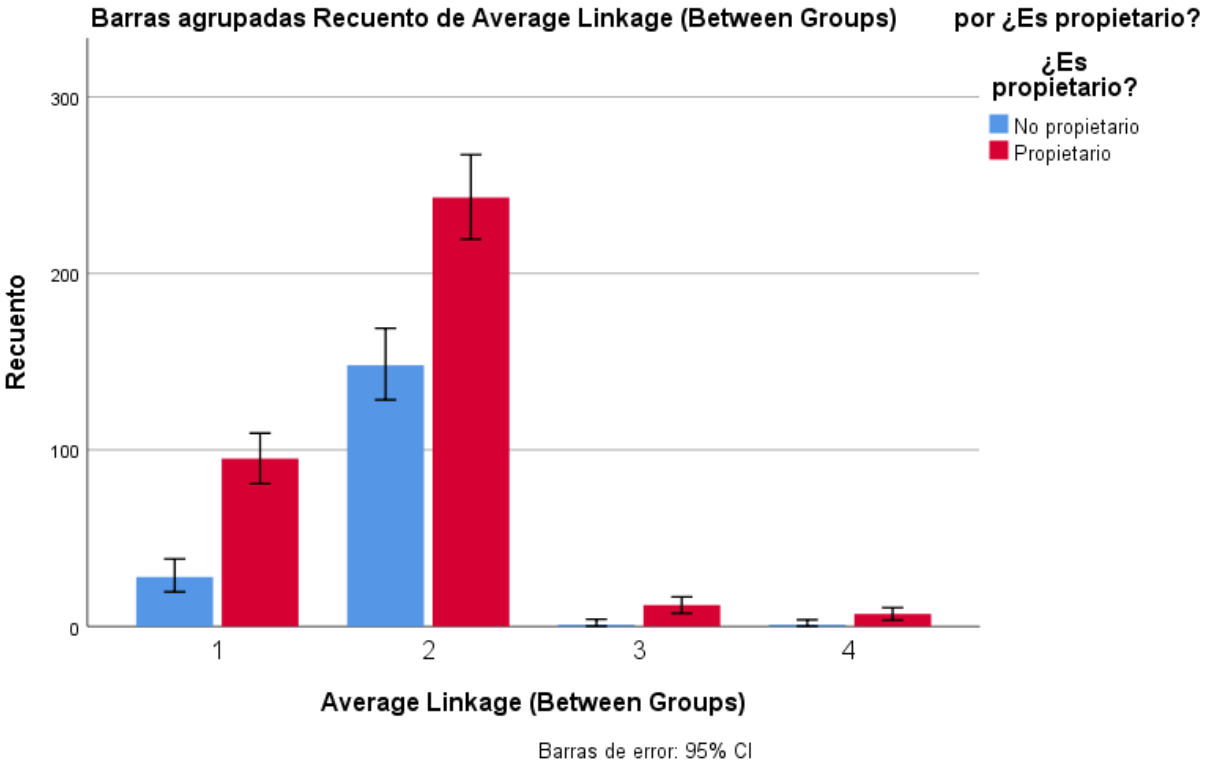
Este gráfico nos muestra que en el cluster 1 y 4 se encuentran aquellos clientes con educación “Licenciatura”. En el cluster número 2 predominan los clientes con educación “Licenciatura”, “Estudios universitarios (en curso)” y “Estudios de postgrado”. El cluster 3 no caracteriza correctamente a la variable porque presenta solapamiento y un bajo recuento de observaciones.

Ocupación



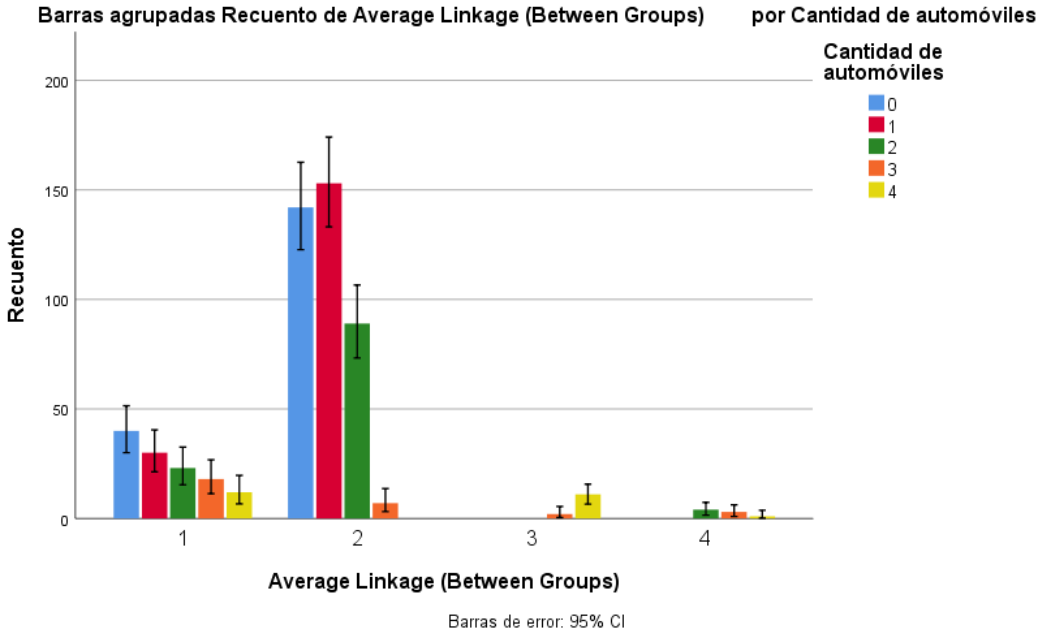
Viendo el gráfico podemos decir que en el cluster 1 encontramos a los clientes que poseen una ocupación de “Obrero especializado” y “Profesional”. Para el cluster 2, si bien se destacan las ocupaciones de “Obrero especializado”, “Profesional” y “Administrativo”, no podemos concluir con certeza porque hay demasiado solapamiento en los intervalos de confianza. Con respecto al cluster 3, encontramos los clientes con ocupación “Profesional” y los que se dedican a la “Gestión”. En el cluster 4, están presentes sólo los que se dedican a la “Gestión”.

Propietario



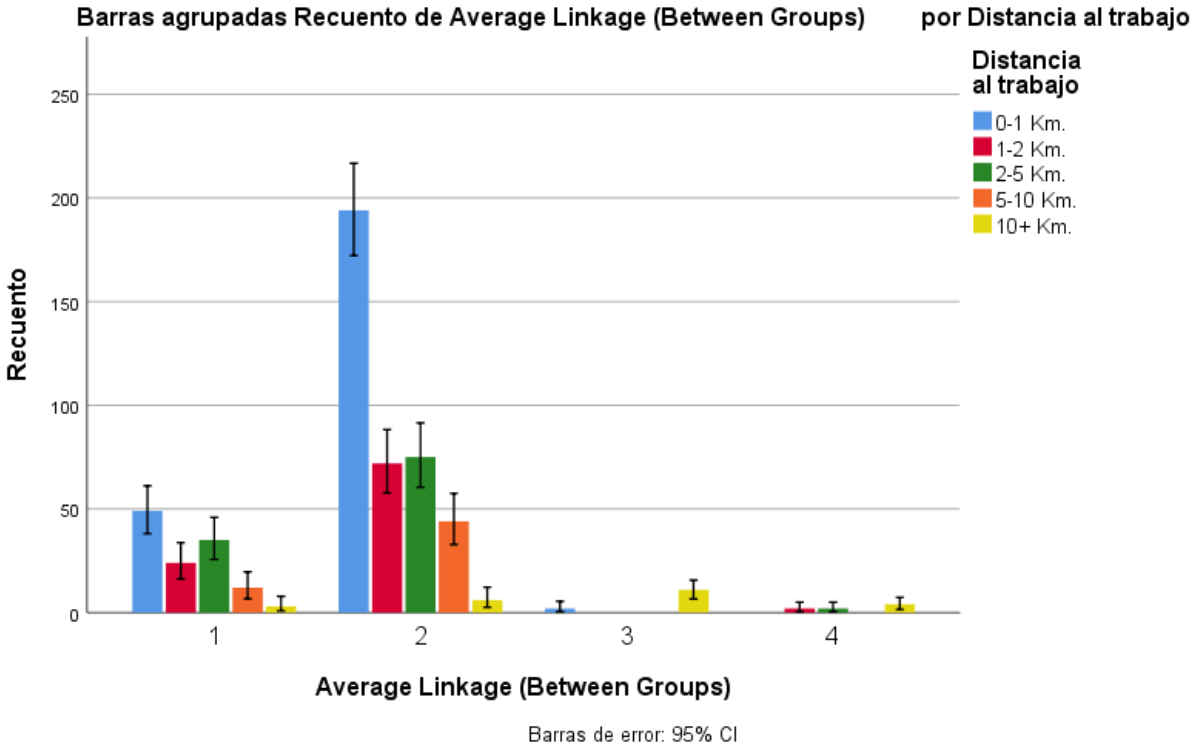
En el gráfico notamos, por un lado, que en los clusters 1 y 2 se encuentran los clientes que son “Propietarios”. Por otro lado, en los clusters 3 y 4 hay solapamiento en sus respectivos intervalos de confianza, por lo que no es posible caracterizar con un valor a cada cluster.

Cantidad Automóviles



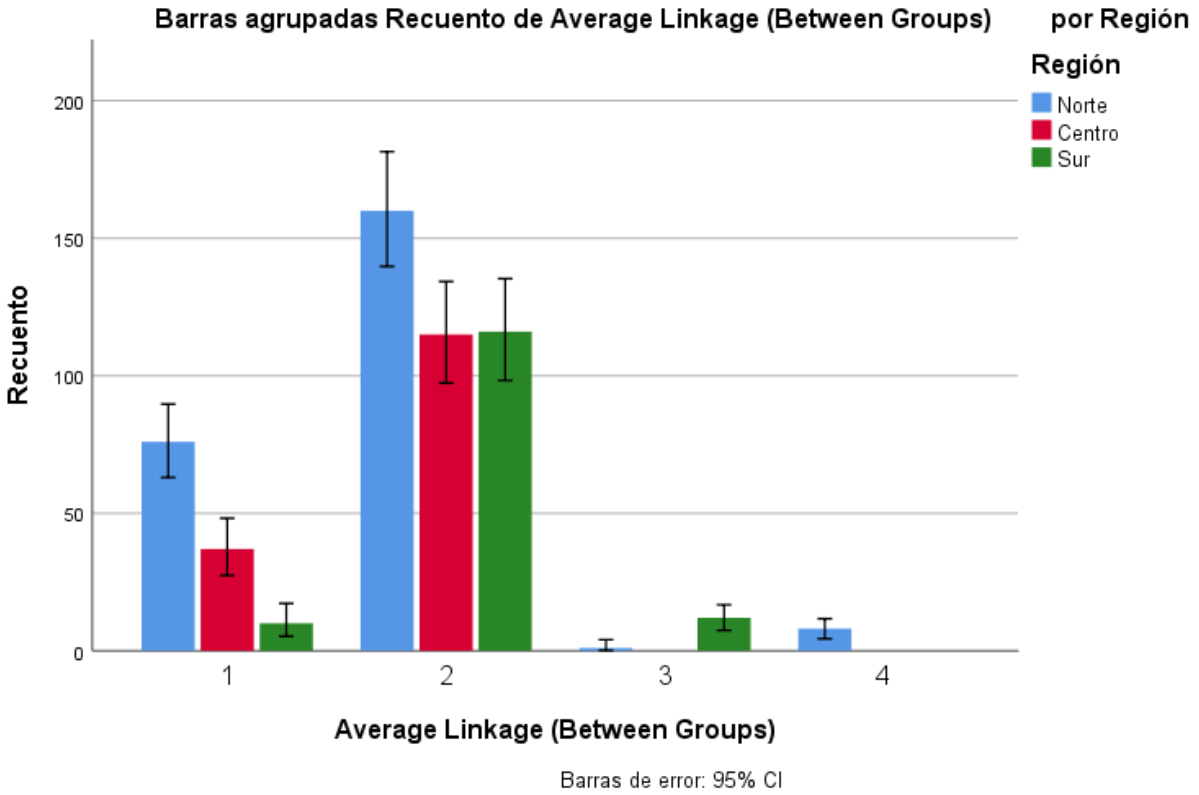
Viendo el gráfico podemos afirmar que para los clusters 1 y 4 no es posible caracterizarlos con un valor concreto debido a la superposición de los intervalos y al bajo recuento registrado. En el segundo cluster predominan los clientes con baja cantidad de automóviles (0 y 1) y en el tercer cluster aquellos con una alta cantidad de automóviles (3 y 4).

Distancia



Para este caso, el único cluster que caracteriza correctamente a las observaciones es el segundo en donde predominan los clientes que viven a una distancia de “0-1 Km” al trabajo. El resto de los clusters poseen superposición de intervalos y en otros se observa un bajo recuento que no nos permite caracterizarlos.

Región



En este gráfico observamos que en los clusters 1, 2 y 4 predominan aquellos clientes que viven en la región “Norte”, mientras que en el cluster 3 se encuentran los que residen en la región “Sur”.

Tabla resumen

A continuación, dejamos a disposición una tabla a modo de resumen y comparación entre los diferentes clusters para las variables nominales o categóricas.

	Grupos								
	2		3			4			
Variables	**Grupo 1**	**Grupo 2**	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**
EstadoCivil	-	-	-	-	-	-	-	-	-
Género	-	-	-	-	-	-	-	-	-
TotalHijos	Bajo (0, 1, 2)	Bajo (0)	-	Bajo (0)	Bajo (0)	-	Bajo (0)	Bajo (0)	-
CantHijosEnCasa	Bajo (0)	Alto (5)	-	Bajo (0)	Alto (5)	-	Bajo (0)	-	Bajo (0)
Educación	Licenciatura	-	Licenciatura	Licenciatura, Estudios universitarios (en curso) y Estudios de postgrado	-	Licenciatura	Licenciatura, Estudios universitarios (en curso) y Estudios de postgrado	-	Licenciatura
Ocupación	-	Profesional y Gestión	Obrero especializado y Profesional	-	Profesional y Gestión	Obrero especializado y Profesional	-	Profesional y Gestión	Gestión
Propietario	Propietario	-	Propietario	Propietario	-	Propietario	Propietario	-	-
CantAutomóviles	Bajo (0 y 1)	Alto (3 y 4)	-	Bajo (0 y 1)	Alto (3 y 4)	-	Bajo (0 y 1)	Alto (3 y 4)	-
Distancia	0-1 Km	-	-	0-1 Km	-	-	0-1 Km	-	-
Región	Norte	Sur	Norte	Norte	Sur	Norte	Norte	Sur	Norte

Como vemos en esta tabla resumen, las variables “Estado Civil” y “Género” no proporcionan ninguna información relevante para ningún cluster. Un trabajo adicional (el cual no llevaremos a cabo) que se puede hacer es eliminar estas variables en cuestión, rehacer los clusters y volver a realizar la interpretación.

Si bien las diferentes cantidades de clusters formados no son muy destacables, el modelo con 2 grupos parece ser el más aceptable entre los tres. Comparando el modelo generado para 3 grupos con el obtenido para 4 grupos, no se aprecian mayores diferencias, salvo en algunos grupos. Por esta razón es que elegimos el modelo obtenido con 2 grupos.

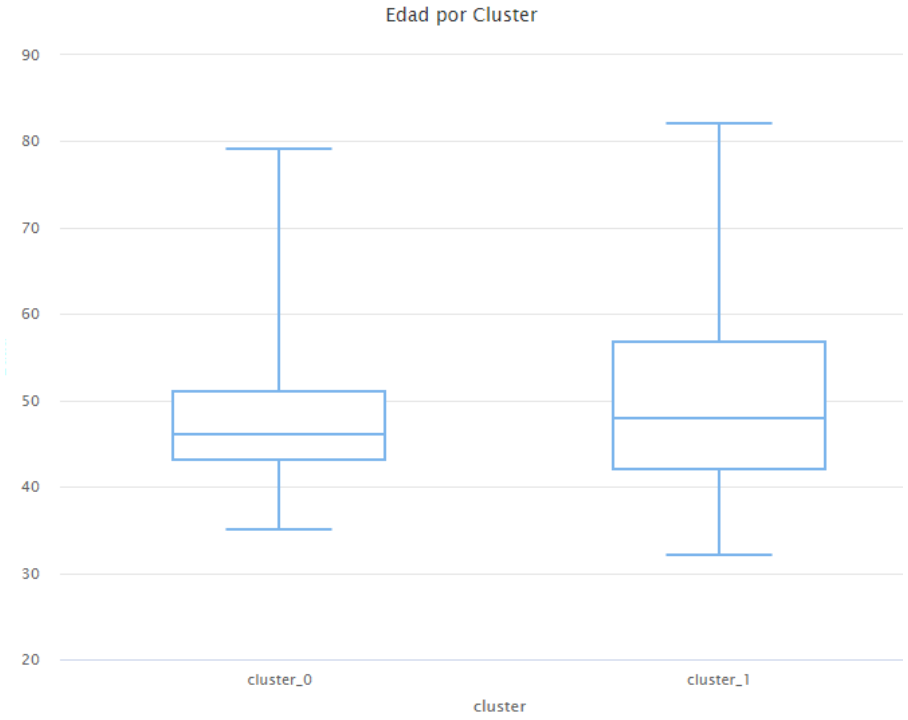
Algoritmo K-medias

El algoritmo k medias se trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar. Este método trabaja solamente con atributos numéricos por lo que en nuestro caso utilizaremos las variables **Cantidad de Hijos en Casa**, **Edad** e **Ingreso Anual**. Además, dado que dichas variables tienen distintas unidades de medición y valores muy distintos, para emparejar los pesos de las mismas normalizaremos el conjunto de datos. Por último, y al igual que el método anterior, formaremos divisiones de 2, 3 y 4 grupos para la muestra.

Resultados con 2 Clusters

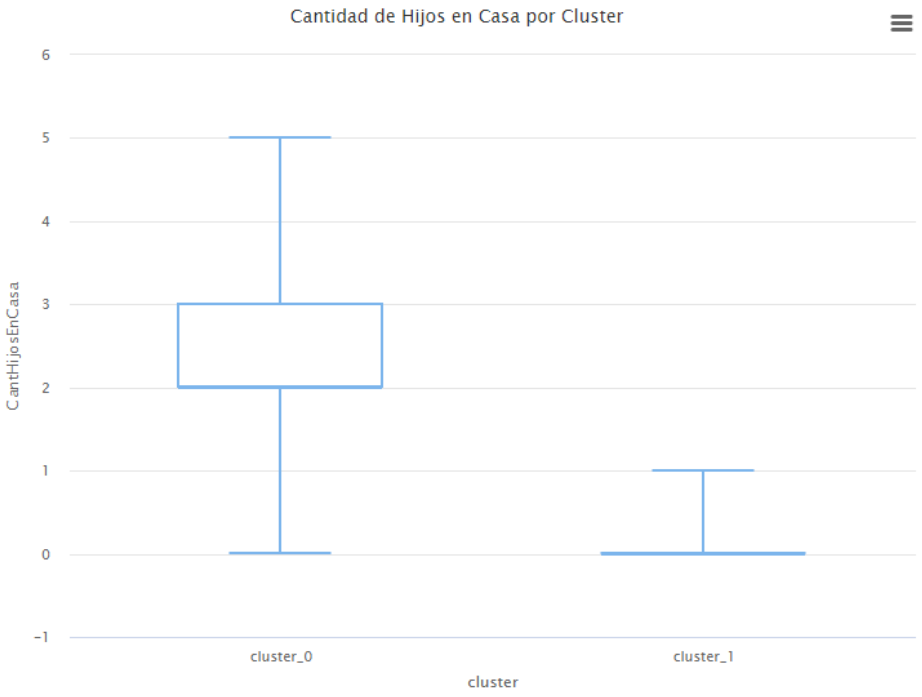
En esta primera instancia se divide la muestra en dos grupos, los cuales rapidminer denomina como cluster_0 y cluster_1 que quedan con 119 y 416 personas respectivamente. En las siguientes subsecciones describiremos los atributos que toman valores particulares en cada cluster.

Edad



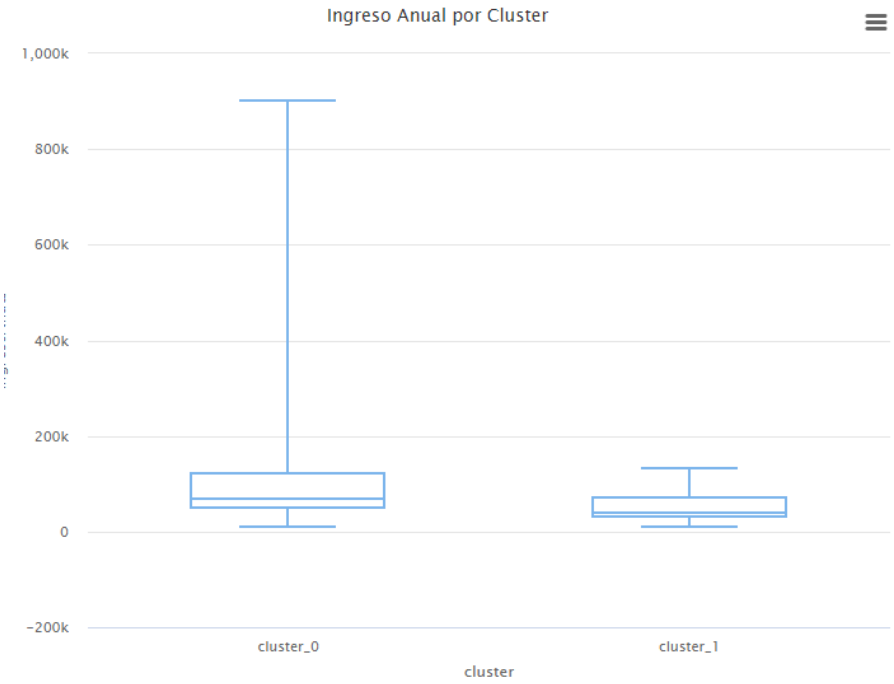
Con respecto a la edad se observa que en el cluster_0 predominan aquellas personas que tienen entre 43 y 51 años estableciendo una de las características principales de este grupo. Por otro lado, el cluster_1 tiene un rango más amplio que a su vez incluye al del otro grupo por lo que no podemos establecer una propiedad que caracterice a este grupo.

Cantidad de Hijos en Casa



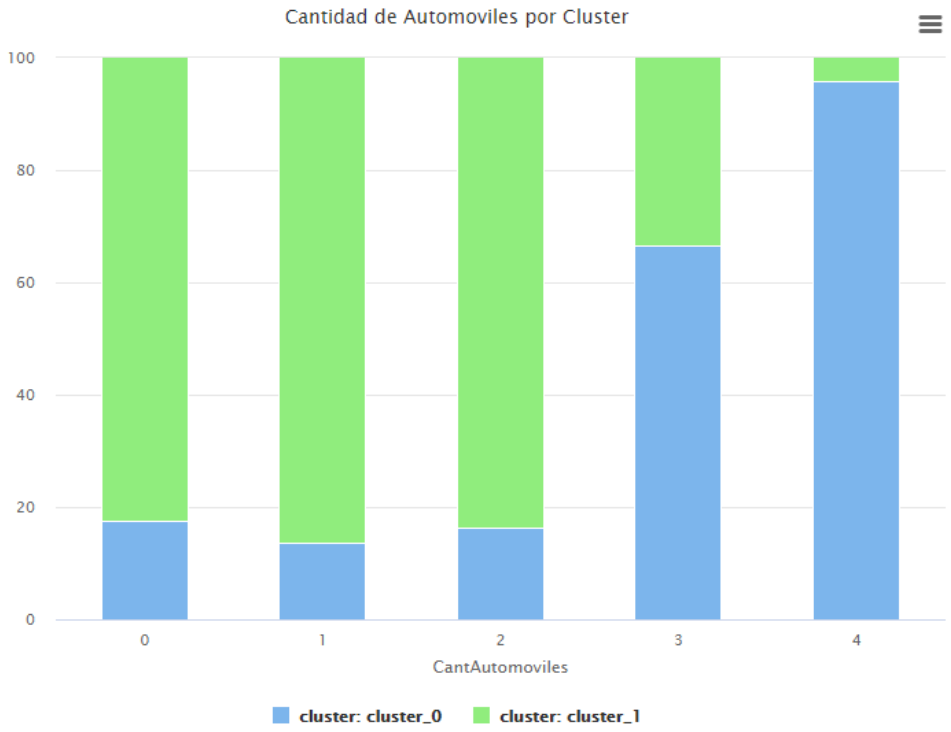
En esta variable si notamos tendencias claras y diferentes en ambos grupos. Por el lado del cluster_0 vemos que la mayoría de los casos se concentran entre 2 y 3 hijos en casa. En cambio, en el cluster_1, prácticamente todos los casos no presentan hijos en casa, salvo algunas excepciones con uno solo.

Ingreso Anual



Por el lado del ingreso anual se nos presenta que el cluster_0 ganan más en promedio con respecto al otro cluster. Sin embargo, en el gráfico se muestra claramente un caso que contiene un ingreso anual muy superior a la media de la muestra, lo que puede estar empujando hacia arriba las mediciones de dicho cluster.

Cantidad de Automóviles



La variable cantidad de automóviles presenta tendencias muy claras con respecto a los grupos. Vemos que las personas que pertenecen al cluster_0 tienen mayor cantidad de automóviles a su disposición, confirmando quizás lo descrito en la variable anterior (que el cluster_0 tiene un mayor ingreso que el cluster_1).

Conclusiones

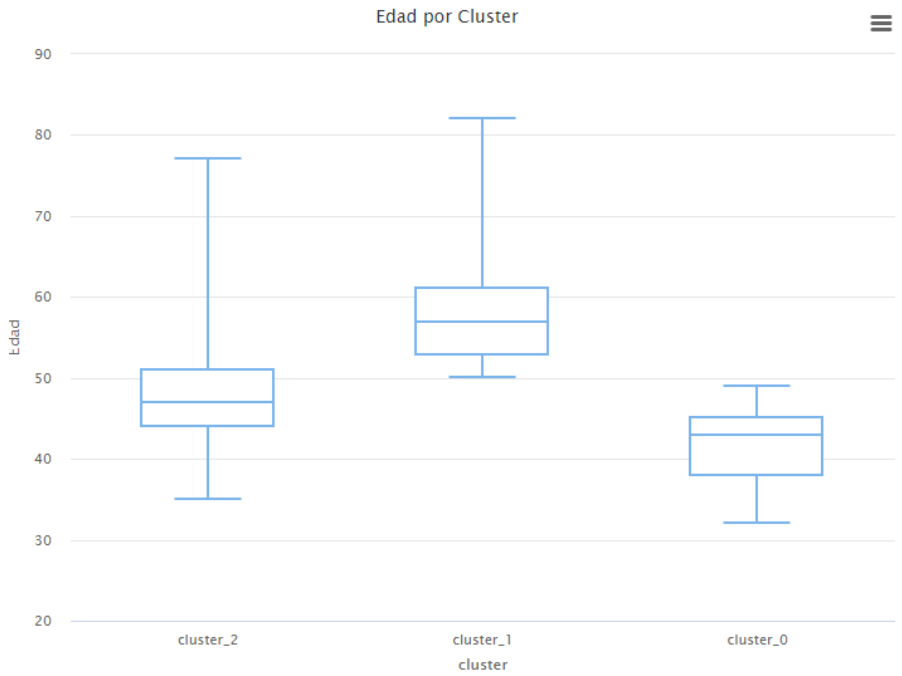
El resto de los atributos no presentados en las subsecciones anteriores no se les encontraron tendencias apreciables, en la mayoría de los casos el cluster_1 ocupaba la mayoría de las variables pero esto se debe a que dicho cluster contiene casi 3,5 veces más de personas. Dicho esto, y sumándole lo expuesto anteriormente, presentamos un cuadro con las características distintivas de cada grupo.

	Cluster_0	**Cluster_1**
Edad	Entre 43 y 51 años	Sin tendencia apreciable
Cantidad de Hijos en Casa	2 y 3	0 y 1
Ingreso Anual	50.000 - 120.000	30.000 - 70.000
Cantidad de Automóviles	3 y 4	0, 1 y 2

Resultados con 3 Clusters

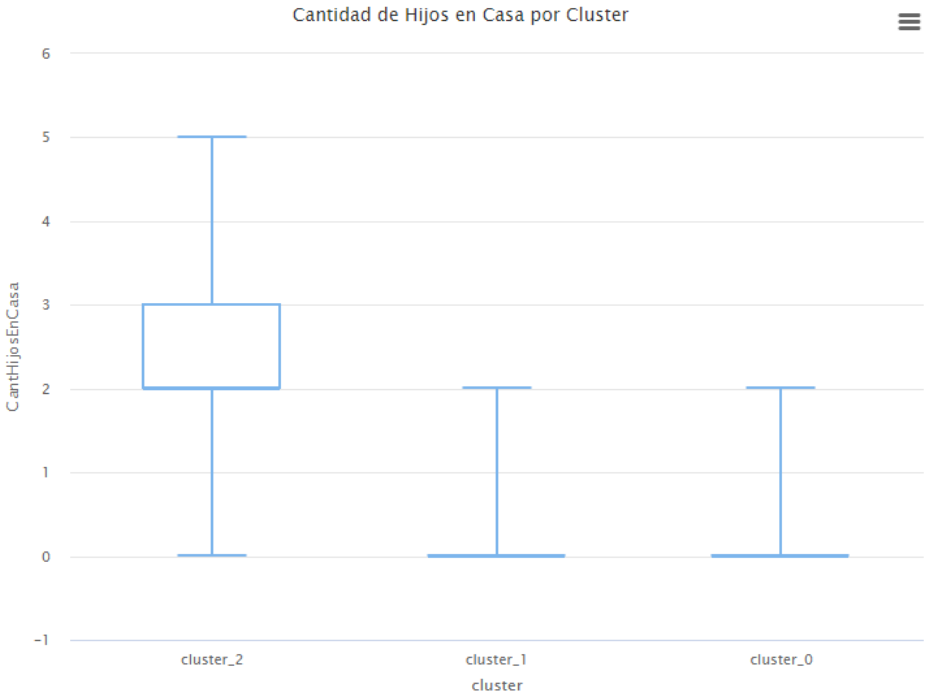
Con k igual a 3, el algoritmo nos entrega 3 grupos: cluster_0 con 224 personas, cluster_1 con 198 personas y cluster_2 con 113 personas. A continuación pasaremos a describir las características distintivas de cada grupo.

Edad



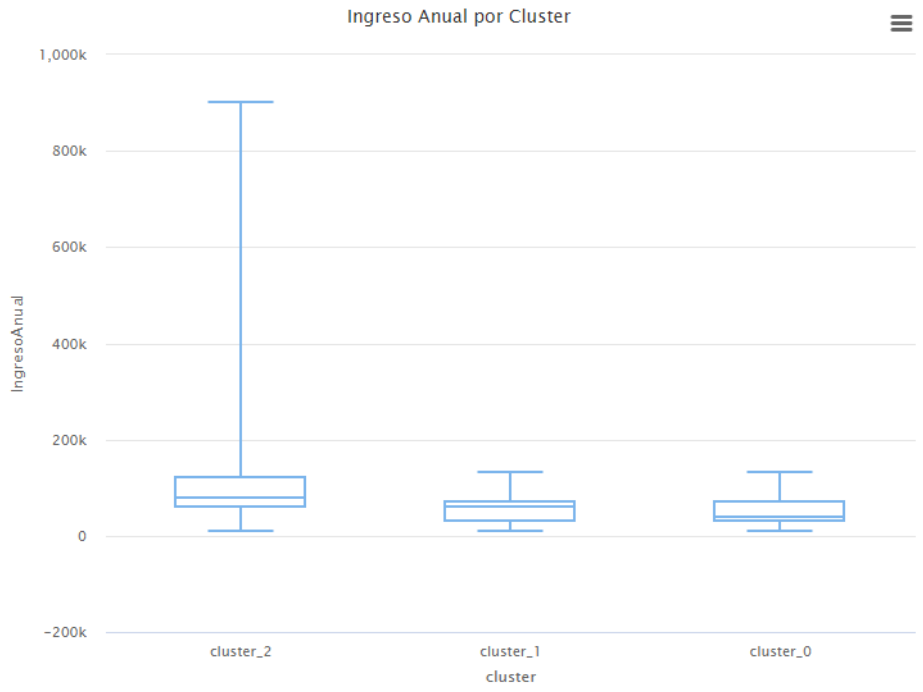
En esta variables vemos tendencias claras con respecto a las edades de cada grupo, siendo el cluster_1 el de mayor edad (entre 53 y 61) , seguido del cluster_2 (entre 44 y 51) y, por último, el cluster_0 (entre 38 y 45).

Cantidad de Hijos en Casa



La variable Cantidad de Hijos en casa presenta la particularidad de que el cluster_1 y el cluster_0 presentan exactamente la misma distribución, ambos con 0 hijos en casa. Por el lado del cluster restante, la tendencia parece ser entre 2 y 3 hijos en casa.

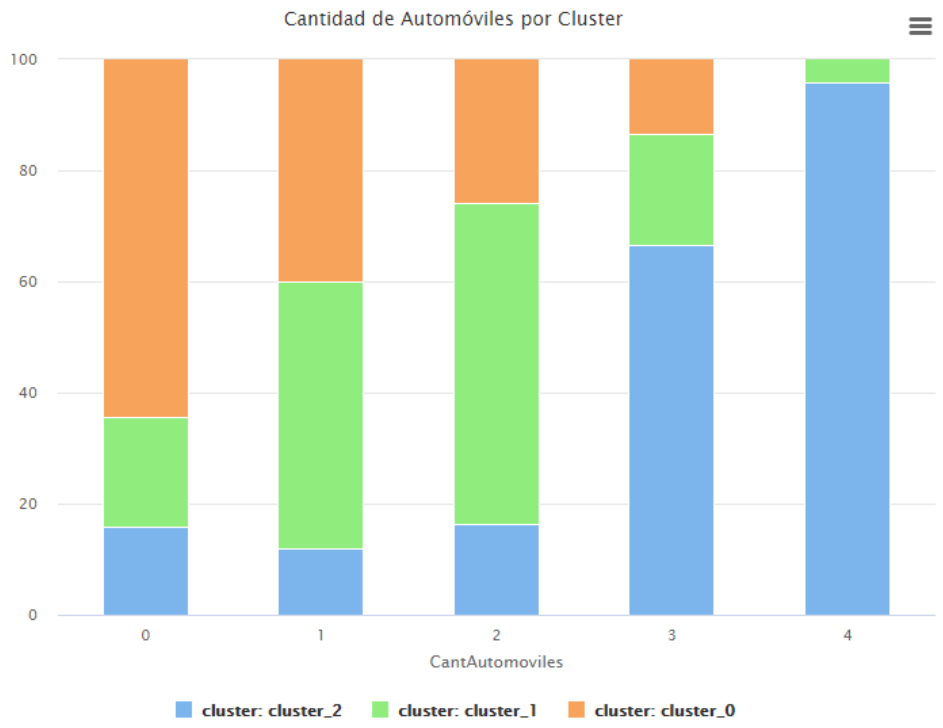
Ingreso Anual



Al igual que el caso anterior el cluster_1 y el cluster_0 presentan similaridades, el grueso de sus datos se corresponden con el mismo intervalo de ingresos anuales de

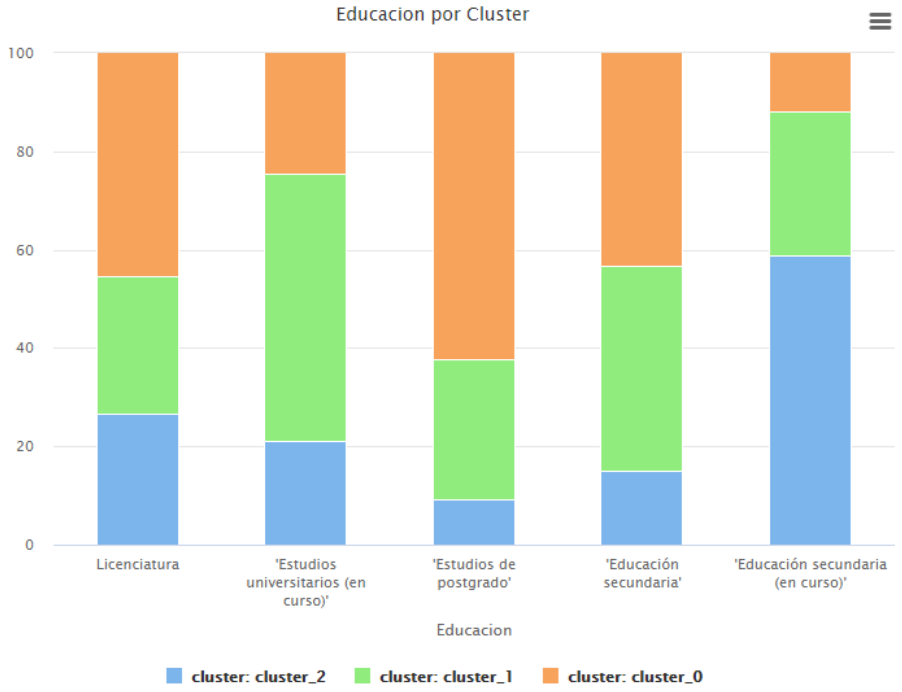
la muestra (30.000 y 70.000) pero el cluster_1 presenta una mediana más alta por lo tanto este grupo tiene un ingreso anual ligeramente mayor. Por otro lado, las personas pertenecientes al cluster_2 tienen un ingreso mayor que el resto de la muestra.

Cantidad de Automóviles



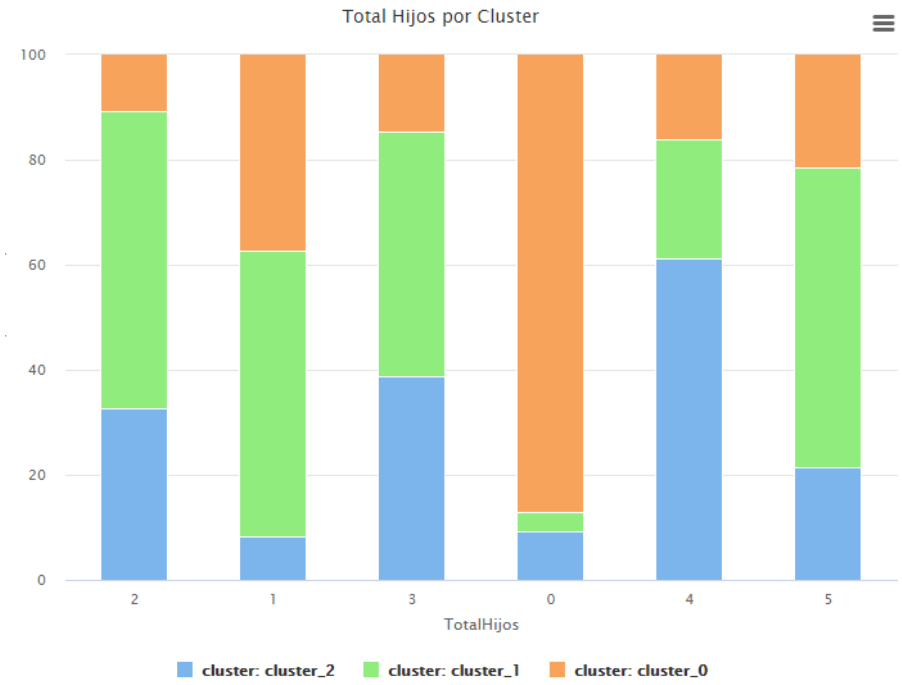
A diferencia de las variables anteriores, acá se aprecian tendencias más claras. El cluster_2 definitivamente está compuesto por personas con mayor cantidad de automóviles (3 y 4), el cluster_1 son personas con uno o dos automóviles mientras las personas sin automóviles pertenecen al cluster_0.

Educación



Por el lado de la educación las tendencias se marcan en las personas con estudios de posgrado donde el cluster_0 es mayoría, con la secundaria en curso siendo el cluster_2 el dominante y con estudios universitarios en curso ocupado principalmente por el cluster_1. En los otros casos ningún cluster llega a obtener la mayoría por sí solo.

Total Hijos



El total de hijos por persona también presenta ciertas características que distinguen a cada grupo. Por el lado del cluster_0 claramente los que no tienen hijos tienen grandes chances de pertenecer a este grupo. El cluster_1 presenta más variabilidad teniendo la mayoría en personas con 1, 2 y 5 hijos. Por último, el cluster_2 contiene a las personas con 4 hijos.

Conclusiones

A continuación presentamos una tabla resumiendo lo descrito anteriormente.

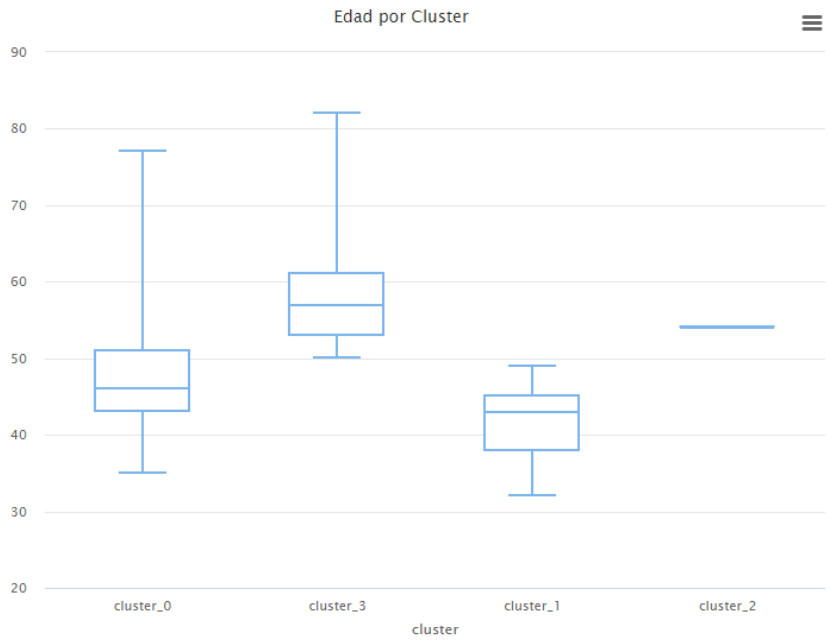
	Cluster_0	**Cluster_1**	**Cluster_2**
Edad	38 - 45	53 - 61	44 - 51
Cantidad de Hijos en Casa	0 o 2	0 o 2	2 - 3
Ingreso Anual	30.000 - 70.000	30.000 - 70.000	60.000 - 120.000
Cantidad de Automóviles	0	1 y 2	4
Educación	Posgrado	Universitario en curso	Secundario en curso
Total Hijos	0	1, 2 y 5	4

Resultados con 4 Clusters

La distribución de los grupos obtenidos es la siguiente:

- Cluster_0: 115 personas.
- Cluster_1: 219 personas.
- Cluster_2: 2 personas.
- Cluster_3: 199 personas.

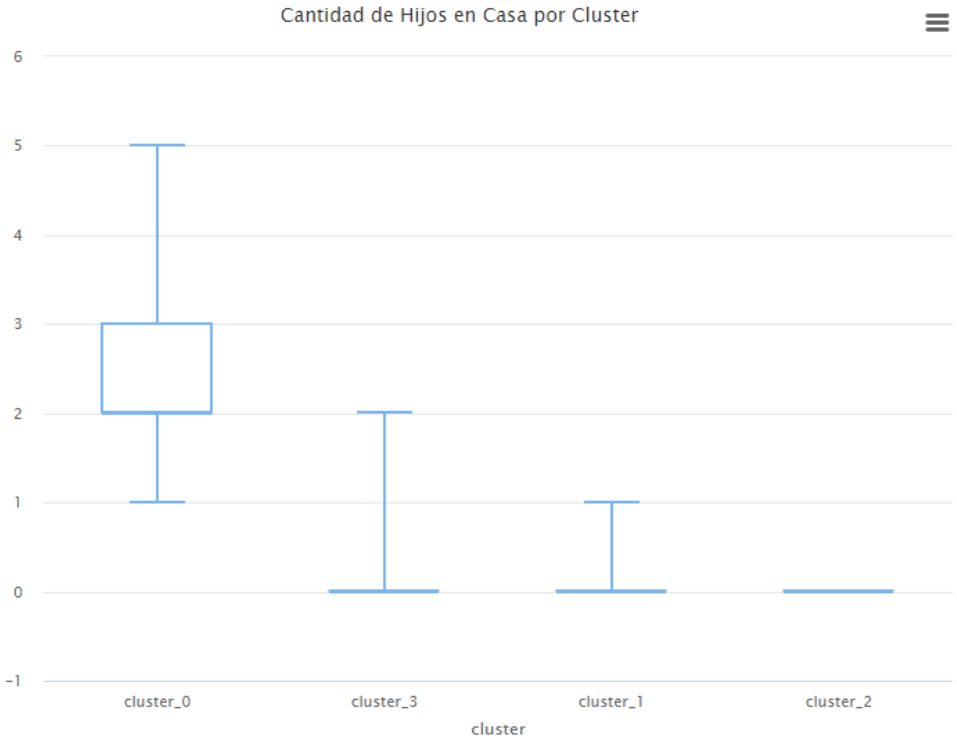
Edad



En la variable edad encontramos las siguientes características en cada grupo:

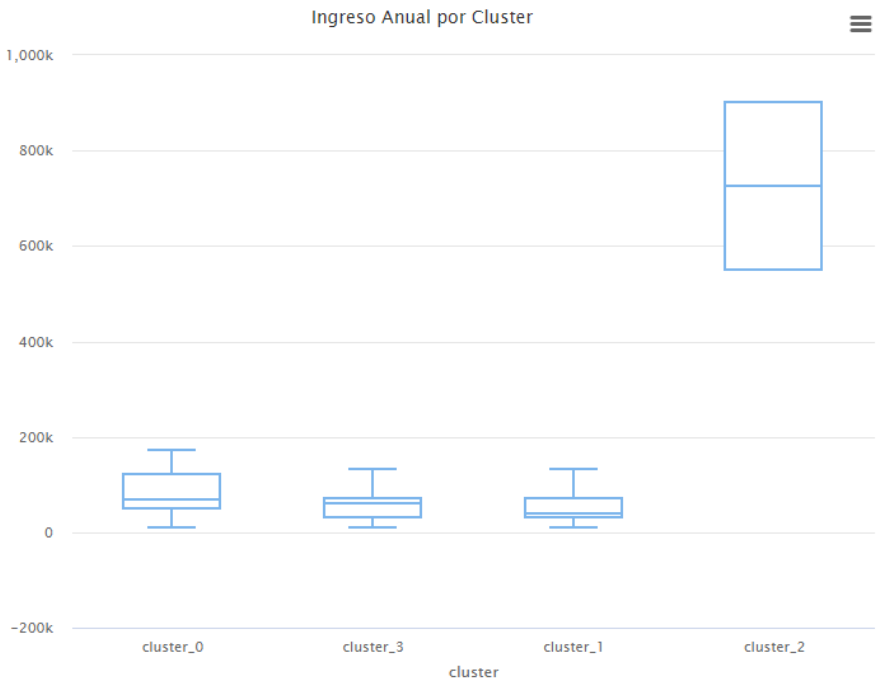
- Cluster_0: su edad mínima es de 35 y su máxima de 77 años pero un gran porcentaje de los datos se concentra entre los 43 y 51 años.
- Cluster_1: ocupa un rango más chico que el cluster anterior, la edad mínima es de 32, la máxima es 49 y la mayoría de sus datos se encuentran en el rango de los 38 a los 45 años.
- Cluster_2: este grupo tiene la particularidad de tener solamente 2 integrantes y ambos tienen la edad de 54 años.
- Cluster_3: este cluster es el que presenta las personas con mayores edades. Con un mínimo de 50, un máximo de 82 y el grueso de los datos concentrados entre 53 y 61 años.

Cantidad de Hijos en Casa



Con respecto a la cantidad de hijos en casa los grupos cluster_1, cluster_2 y cluster_3 tienen como tendencia el valor 0 en dicha variable, con variabilidad hasta 1 en el primero y hasta dos en el último. Por otro lado, en el cluster_0, presentan de 1 a 5 hijos siendo los valores 2 y 3 como la tendencia del grupo.

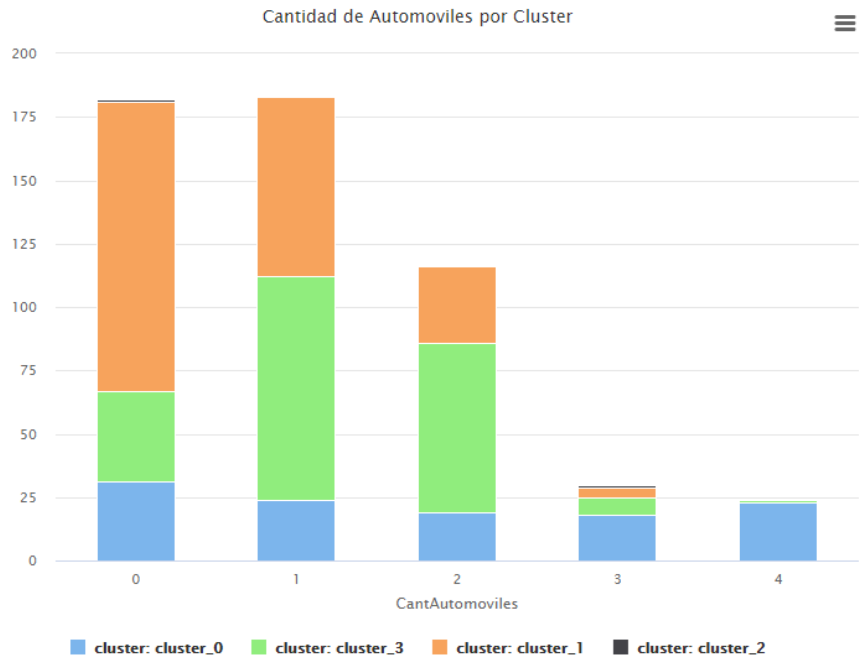
Ingreso Anual



Aquí observamos una clara diferenciación del cluster_2 con los demás, sus dos integrantes tienen un ingreso anual muy superior al del resto de la muestra por lo que se concluye que esta es la principal característica de una persona perteneciente a este grupo.

Con respecto a los tres grupos restantes siguen una tendencia muy parecida al análisis de esta variable con k igual a 3. Tenemos un cluster (cluster_0) con un ingreso anual promedio ligeramente superior a los otros dos, mientras que el cluster_3 y el cluster_1 son prácticamente iguales.

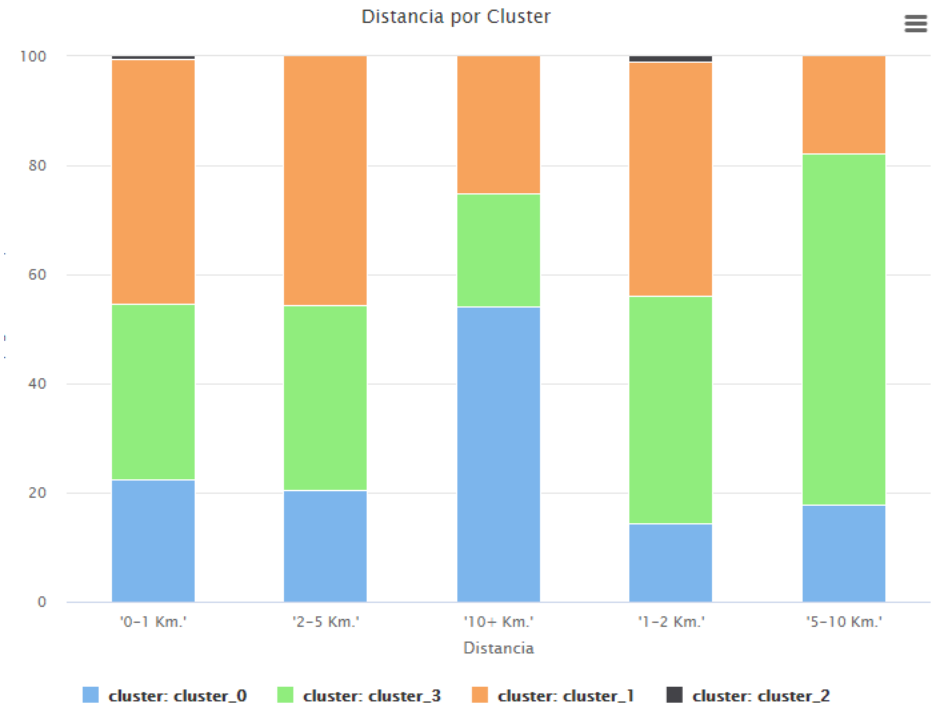
Cantidad de Automóviles



En esta variable observamos las siguientes tendencias:

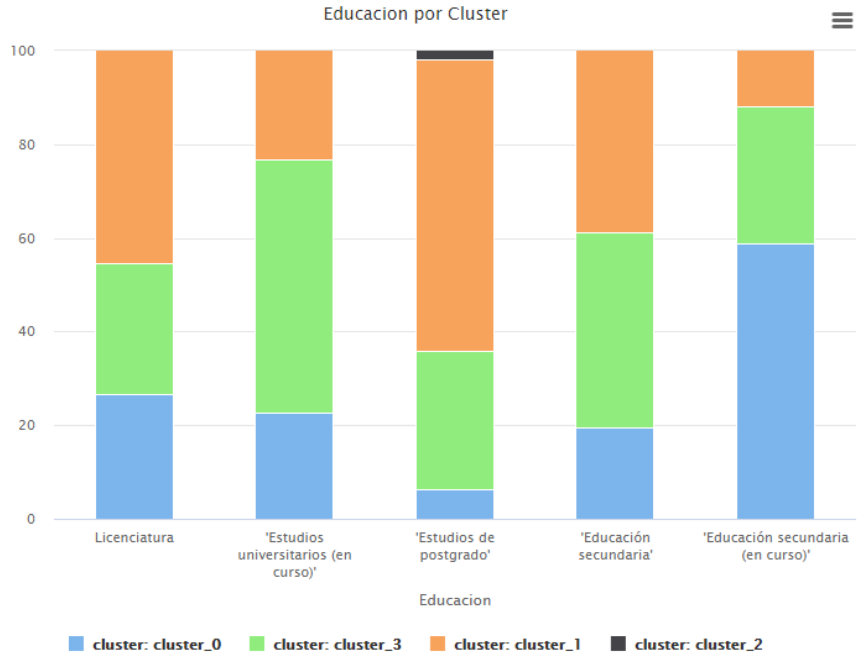
- El cluster_0 ocupa casi la totalidad de las personas con 4 automóviles.
- El cluster_1 es mayoría cuando la variable toma el valor 0.
- Y las personas en el cluster_3 suelen tener entre 1 y 2 autos.

Distancia



En la variable distancia observamos al cluster_0 siendo mayoritario en las personas que viven más lejos de su trabajo. Por el lado del cluster_1 es el representante mayor en las distancias 0-1 km y 2-5 km sin llegar a ocupar el 50% de los datos. En cambio, en las distancias de entre 5 y 10 km, el cluster_3 es el más presente. Por último, el cluster_2, al tener 2 casos y que estén en valores distintos no marca ninguna característica en especial.

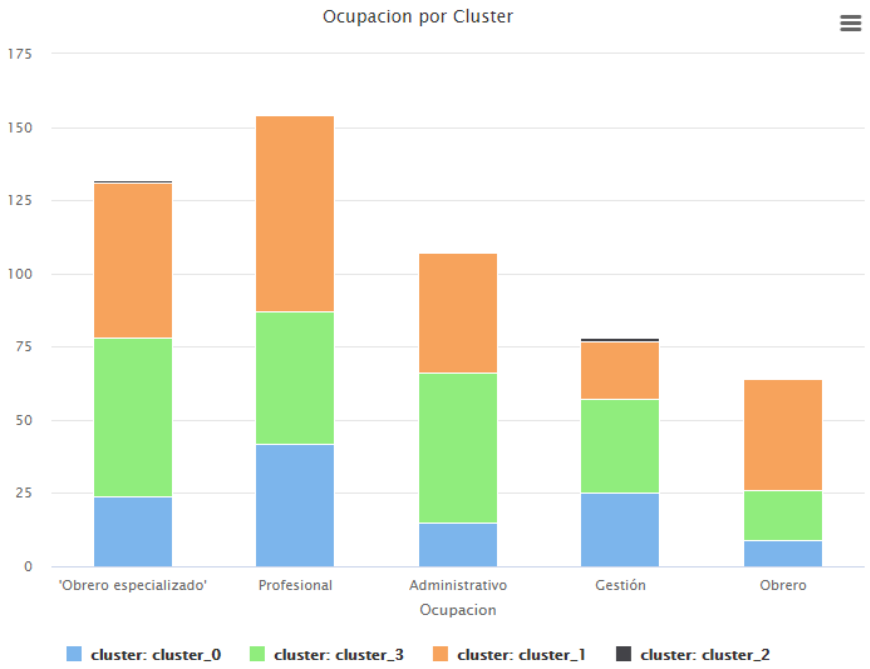
Educación



En la variable Educación realizamos las siguientes observaciones:

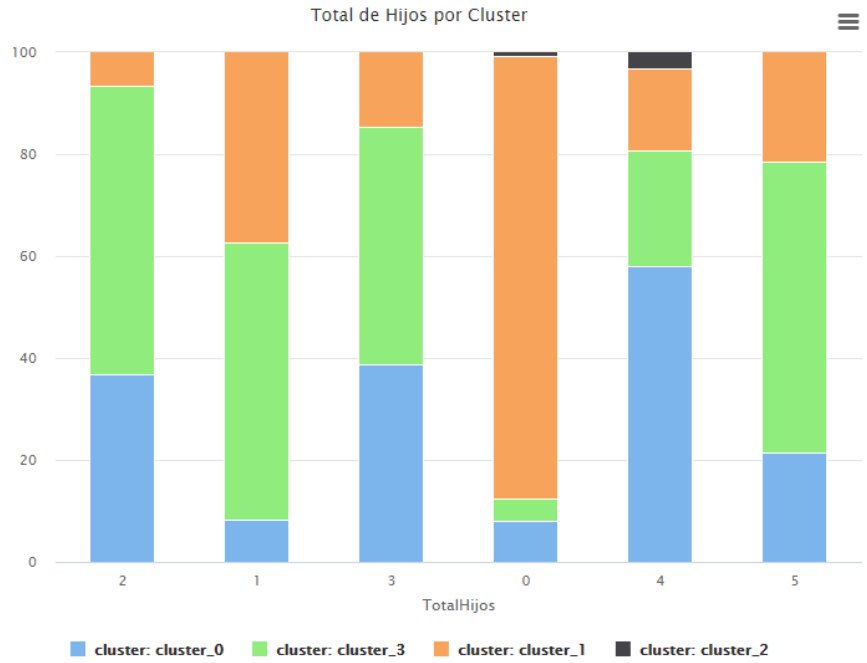
- Aquellos que tienen la educación secundaria en curso, es muy probable que pertenezcan al cluster_0.
- En el cluster_1 se encuentran las personas con estudios de posgrado y licenciaturas.
- Los dos casos del cluster_2 tienen estudios de posgrado.
- Si están realizando sus estudios universitarios tienen grandes chances de pertenecer al cluster_3.

Ocupación



La variable ocupación no entrega mucha información, todos los datos están bastante repartidos por todos los valores posibles con la excepción de los obreros donde las personas pertenecientes al cluster_1 ocupan un gran porcentaje.

Total de Hijos por Cluster



El total de hijos por personas presenta las siguientes características:

- El cluster_0 sólo es mayoritario en las personas con 4 hijos.
- El cluster_1 ocupa casi la totalidad de las personas con 0 hijos
- Cuando las personas tienen 1, 2 y 5 hijos probablemente pertenezcan al cluster_3

Conclusiones

A continuación presentamos la tabla resumen de lo descrito en cada variable.

	Cluster_0	**Cluster_1**	**Cluster_2**	**Cluster_3**
Edad	43 - 51	38 - 45	54	53 - 61
Cantidad de Hijos en Casa	1 - 5, sobre todo 2 y 3	0 - 1	0	0 - 2
Ingreso Anual	50.000 - 120.000	30.000 - 70.000	550.000 y 900.000	30.000 - 70.000
Cantidad de Automóviles	4	0	Sin tendencia apreciable	1 - 2
Distancia	10+ km	0 - 1 km y 2 - 5 km	Sin tendencia apreciable	5 - 10 km
Educación	Educación Secundaria en curso	Estudios de Posgrado y Licenciatura	Estudios de Posgrado	Estudios Universitarios en curso
Ocupación	Sin tendencia apreciable	Obreros	Sin tendencia apreciable	Sin tendencia apreciable
Total de Hijos	4 Hijos	0 Hijos	Sin tendencia apreciable	1, 2 y 5 Hijos

Conclusiones Finales

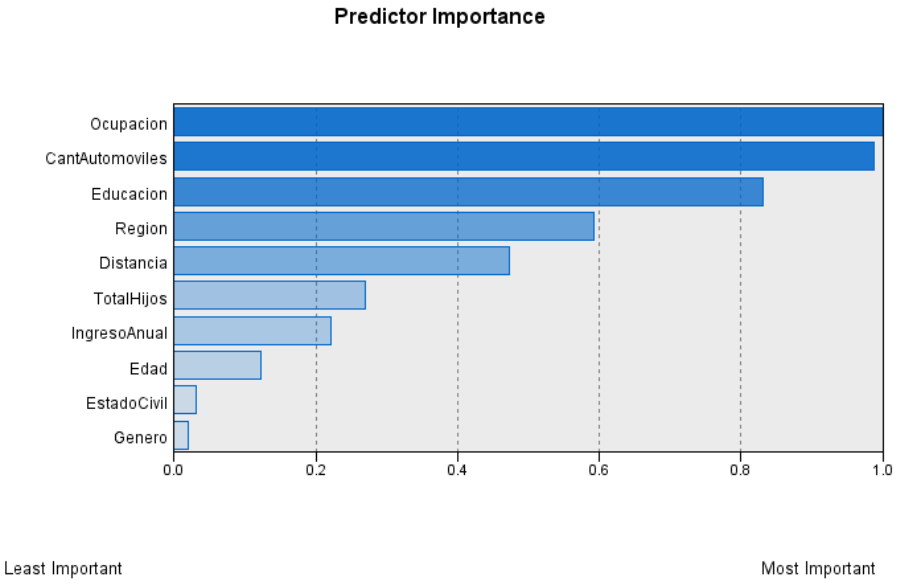
Tras analizar cada caso, decidimos que la mejor separación de la muestra es con 2 grupos ya que cada uno tiene características bien definidas. En el caso con 3 clústers, el primer y segundo grupo tienen algunas características similares y con 4 clústers se nos forma un grupo de solamente 2 personas, lo que nos resulta inconveniente para el análisis.

Clustering Bietápico

El clustering Bietápico es un método de agrupación en clusters que se realiza en dos pasos. En el primero de ellos, se ejecuta un algoritmo similar al K-medias, donde se comprimen los datos en subclusters, y en el segundo paso se utiliza un procedimiento jerárquico para combinar observaciones de forma que todos los clusters formados sean homogéneos.

A diferencia de los métodos anteriores, el clustering bietápico permite trabajar con variables mixtas, es decir, tanto variables nominales como numéricas. Además, en este caso, el algoritmo determina de forma automática la cantidad óptima de grupos, por lo que no es necesario hacer comparaciones con diferentes salidas.

Para poder obtener un modelo en el cual cada clase contenga elementos similares entre sus individuos y a su vez, sea distinta a las demás clases, se procede a analizar cada uno de los clusters y eliminar las variables que no definen ninguna diferencia entre grupos, comenzando con las de menor importancia. De esta forma, para la construcción del modelo final, se eliminaron las variables *propietario* y *cantidad de hijos en casa*.



Finalmente, el modelo óptimo obtenido define cuatro clusters, en los cuales sus variables predictoras de mayor importancia son *ocupación*, *cantidad de automóviles,* y *educación*. A continuación se realizará un análisis de cada uno de los clusters para poder identificar las características en común de cada una de las observaciones.

Análisis de clusters

Cluster 1

Los clientes que se encuentran en el primer clúster en general trabajan como administrativos, no tienen automóviles, son licenciados, viven en la región centro, su distancia al trabajo es menor a 1 km, no tienen hijos, su ingreso anual ($39.000 aproximadamente) es bajo a comparación del de la población total, la mediana de su edad es de 45 años, se encuentran casados, y son mujeres.

Cluster 2

En el cluster 2 están contenidos los clientes que en general tienen las siguientes características: su ocupación es una profesión, tienen 1 solo automóvil, su educación es una licenciatura, viven en la región Norte, la distancia desde su hogar al trabajo es menor a 1 km, no tienen hijos, su ingreso anual ($69.000 aproximadamente) es superior al de la mayor parte de la población estudiada, su edad se encuentra alrededor de los 45 años, su estado civil es soltero y son mujeres.

Cluster 3

En este cluster los individuos son, en general, obreros, tienen 1 solo automóvil, se encuentran cursando sus estudios universitarios, viven en la región centro, la distancia a su trabajo es menor a 1 km, tienen 1 hijo, la mediana de su ingreso anual es la menor de todos los grupos (aproximadamente $28.000), su edad ronda los 47 años, su estado civil es soltero, y son hombres.

Cluster 4

Finalmente, en el cluster número 4, las observaciones presentan las siguientes características de los clientes: son en su mayoría profesionales, tienen 2 automóviles, se encuentran cursando sus estudios universitarios, viven en la región norte, la distancia a su trabajo está entre 5 y 10 km, tienen dos hijos, la mediana de su ingreso anual ($68.000 aproximadamente) es superior a la mediana de la población y similar a la del cluster 2, su edad ronda los 57 años, siendo mucho mayor que la de los demás grupos, su estado civil es soltero, y son mujeres.

Conclusiones e implementación

El problema planteado tiene como objetivo indicar cuáles de los grupos de potenciales clientes con características similares es más probable que consideren comprar tres tipos distintos de bicicletas: para niños (Kinder), estándares (Basic) y deportivas (Sport).

Para esto, y a partir de los análisis anteriores, se ha decidido implementar el modelo obtenido con el clustering bietápico el cual define una cantidad de clústeres igual a 4. Por lo tanto se concluye que los grupos y los tipos de bicicletas estarán asociados de la siguiente manera:

- Para el primer clúster, en el cual se encuentran clientes que no tienen hijos y su ingreso es en general, bajo, no tienen autos y la distancia a su trabajo es menor a 1 km, se recomienda ofrecer el tipo de bicicleta **Basic**.

- Los clientes del segundo clúster, caracterizados por no tener hijos, que su ingreso sea el más alto de la población en general, y que su edad ronde los 45 años, podrían tener un mayor interés en comprar bicicletas del tipo **Sport.**

- Al igual que los clientes del grupo 1, los individuos del cluster 3 presentan ingresos anuales bajos, y además tienen en general 1 hijo. Por lo tanto, consideramos que podrían ofrecerse tanto las bicicletas del tipo **Basic** como las bicicletas del tipo **Kinder**.

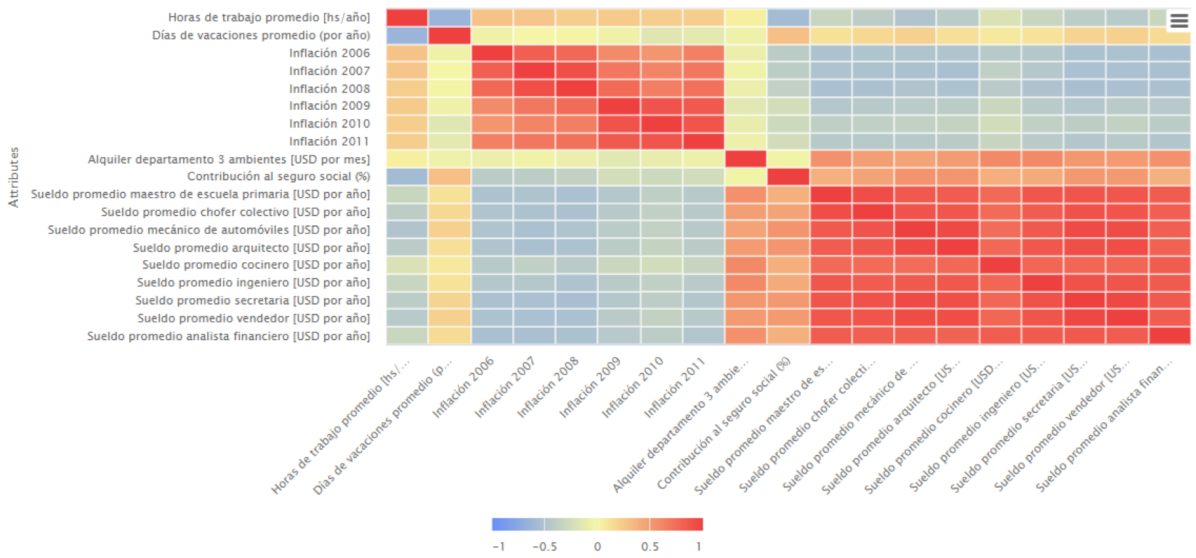
- Por último, los potenciales clientes del grupo 4 presentan un ingreso anual mayor a la media de la población, la distancia a su trabajo se encuentra entre 5 y 10 km, poseen 2 autos y su edad ronda los 57 años. Por tanto, recomendamos ofrecer un tipo de bicicleta **Sport**.

Análisis de componentes principales

Como último objetivo de este trabajo, se nos pide que recomendemos tres países con características sociales y económicas similares a Argentina para comercializar la nueva línea de bicicletas. Para cumplir con lo propuesto, haremos uso de un archivo Excel llamado “datosMercados.xlsx” sobre el cual aplicaremos la técnica de análisis de componentes principales.

Para comenzar, presentaremos un mapa de calor que muestra la correlación entre las variables. Es equivalente a la matriz de correlación, sin embargo, para mayor legibilidad preferimos mostrar este tipo de gráfico.

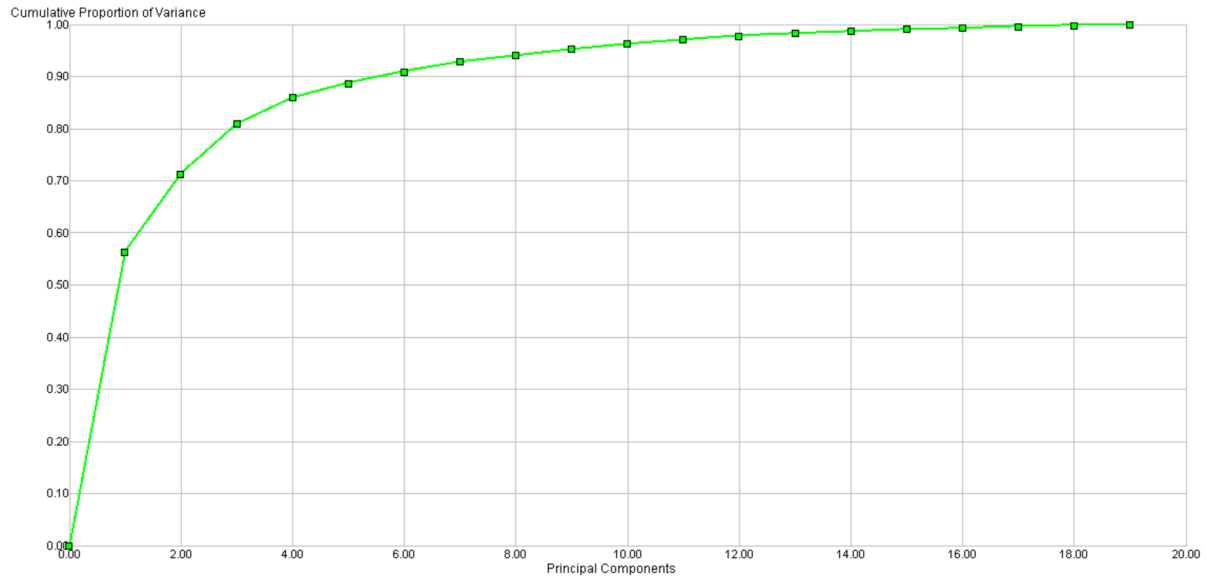
Como observamos en la escala, aquellas variables que están correladas positivamente se muestran en color rojo y las que se correlacionan negativamente en azul. Lo que buscamos para poder aplicar esta técnica es que las variables estén correlacionadas. Si bien vemos que hay variables que no presentan mucha correlación, la mayoría si lo hace. Dicho esto, podemos avanzar en la aplicación del método.



Para continuar con el análisis, es necesario aclarar que se han normalizado los valores de las variables, dado que ellas contienen distintas unidades de medida y por lo tanto se les daba distintos niveles de importancia dentro del análisis. De esta forma se logra una mayor coherencia para la comparación de las ciudades.

A partir del porcentaje de variabilidad de las componentes podemos notar que las primeras 9 explican aproximadamente el 95% de variabilidad en los datos. Si bien podríamos decidir trabajar con 3 variables para poder realizar un scatter en tres dimensiones, esto no sería recomendable ya que, si observamos el gráfico de abajo, vemos que estas sólo representan poco más del 80% de la información de las variables originales y consideramos que no es representativo para este caso.

En el siguiente gráfico puede verse la variabilidad acumulada mencionada.



Como mencionamos, vamos a trabajar con 9 componentes principales y por tanto no podremos graficarlas. Sin embargo, a continuación presentaremos cómo quedaron formadas las componentes:

$$CP_1 = -0.139 * X_1 + 0.062 * X_2 - 0.210 * X_3 - 0.217 * X_4 - 0.220 * X_5 -$$
$$- 0.199 * X_6 - 0.181 * X_7 - 0.200 * X_8 + 0.143 * X_9 + 0.168 * X_{10} +$$
$$+ 0.274 * X_{11} + 0.276 * X_{12} + 0.280 * X_{13} + 0.278 * X_{14} + 0.244 * X_{15} +$$
$$+ 0.272 * X_{16} + 0.286 * X_{17} + 0.283 * X_{18} + 0.271 * X_{19}$$

$$CP_2 = 0.010 * X_1 + 0.006 * X_2 + 0.267 * X_3 + 0.320 * X_4 + 0.324 * X_5 +$$
$$0.367 * X_6 + 0.378 * X_7 + 0.372 * X_8 + 0.301 * X_9 - 0.026 * X_{10} +$$

$$+ 0.141 * X_{11} + 0.141 * X_{12} + 0.151 * X_{13} + 0.153 * X_{14} + 0.203 * X_{15} +$$
$$+ 0.157 * X_{16} + 0.141 * X_{17} + 0.153 * X_{18} + 0.123X_{19}$$

$$CP_3 = 0.582 * X_1 - 0.576 * X_2 - 0.024 * X_3 - 0.053 * X_4 - 0.095 * X_5 -$$
$$- 0.111 * X_6 - 0.052 * X_7 - 0.076 * X_8 + 0.310 * X_9 - 0.400 * X_{10} +$$
$$0.081 * X_{11} + 0.002 * X_{12} - 0.076 * X_{13} - 0.017 * X_{14} + 0.117 * X_{15} +$$
$$+ 0.076 * X_{16} - 0.009 * X_{17} - 0.034 * X_{18} + 0.078 * X_{19}$$

$$CP_4 = - 0.077 * X_1 + 0.433 * X_2 + 0.325 * X_3 + 0.298 * X_4 + 0.204 * X_5 -$$
$$- 0.279 * X_6 - 0.387 * X_7 - 0.268 * X_8 + 0.353 * X_9 - 0.357 * X_{10} +$$
$$+ 0.028 * X_{11} - 0.012 * X_{12} - 0.050 * X_{13} - 0.098 * X_{14} - 0.024 * X_{15} +$$
$$+ 0.045 * X_{16} - 0.008 * X_{17} - 0.033 * X_{18} + 0.049 * X_{19}$$

$$CP_5 = - 0.051 * X_1 + 0.447 * X_2 - 0.420 * X_3 - 0.258 * X_4 - 0.227 * X_5 +$$
$$+ 0.197 * X_6 + 0.295 * X_7 + 0.120 * X_8 + 0.154 * X_9 - 0.500 * X_{10} +$$
$$+ 0.039 * X_{11} - 0.033 * X_{12} - 0.129 * X_{13} - 0.148 * X_{14} + 0.104 * X_{15} -$$
$$- 0.050 * X_{16} - 0.078 * X_{17} - 0.027 * X_{18} + 0.153 * X_{19}$$

$$CP_6 = - 0.064 * X_1 - 0.119 * X_2 + 0.190 * X_3 - 0.058 * X_4 - 0.018 * X_5 -$$
$$- 0.047 * X_6 + 0.138 * X_7 - 0.027 * X_8 - 0.402 * X_9 - 0.549 * X_{10} +$$
$$+ 0.210 * X_{11} + 0.271 * X_{12} + 0.218 * X_{13} + 0.144 * X_{14} - 0.444 * X_{15} +$$
$$+ 0.025 * X_{16} + 0.113 * X_{17} + 0.142 * X_{18} - 0.133 X_{19}$$

$$CP_7 = 0.501 * X_1 + 0.282 * X_2 + 0.083 * X_3 + 0.156 * X_4 + 0.047 * X_5 +$$
$$+ 0.195 * X_6 - 0.049 * X_7 - 0.233 * X_8 - 0.598 * X_9 - 0.032 * X_{10} +$$
$$+ 0.050 * X_{11} + 0.109 * X_{12} - 0.003 * X_{13} - 0.052 * X_{14} + 0.342 * X_{15} -$$

$$- 0.011 * X_{16} + 0.040 * X_{17} + 0.036 * X_{18} + 0.213 * X_{19}$$

$$CP_8 = 0.396 * X_1 + 0.338 * X_2 + 0.466 * X_3 - 0.278 * X_4 - 0.448 * X_5 -$$
$$- 0.036 * X_6 + 0.046 * X_7 + 0.217 * X_8 + 0.152 * X_9 + 0.162 * X_{10} -$$
$$- 0.041 * X_{11} - 0.010 * X_{12} + 0.011 * X_{13} + 0.076 * X_{14} - 0.164 * X_{15} +$$
$$+ 0.020 * X_{16} + 0.105 * X_{17} + 0.125 * X_{18} - 0.269 * X_{19}$$

$$CP_9 = - 0.360 * X_1 - 0.175 * X_2 + 0.475 * X_3 - 0.103 * X_4 - 0.298 * X_5 -$$
$$- 0.213 * X_6 + 0.100 * X_7 + 0.137 * X_8 - 0.169 * X_9 - 0.150 * X_{10} +$$
$$+ 0.202 * X_{11} - 0.110 * X_{12} - 0.056 * X_{13} + 0.008 * X_{14} + 0.463 * X_{15} -$$
$$- 0.028 * X_{16} - 0.126 * X_{17} - 0.089 * X_{18} + 0.302 * X_{19}$$

Siendo:

- X_1: horas de trabajo promedio[hs/año]
- X_2: Días de vacaciones promedio (por año)
- X_3: Inflación 2006
- X_4: Inflación 2007
- X_5: Inflación 2008
- X_6: Inflación 2009
- X_7: Inflación 2010
- X_8: Inflación 2011
- X_9: Alquiler departamento 3 ambientes [USD por mes]
- X_{10}: Contribución del seguro social (%)
- X_{11}: Sueldo promedio maestro de escuela primaria [USD por año]
- X_{12}: Sueldo promedio chofer colectivo [USD por año]
- X_{13}: Sueldo promedio mecánico de automóviles [USD por año]
- X_{14}: Sueldo promedio arquitecto [USD por año]
- X_{15}: Sueldo promedio cocinero [USD por año]
- X_{16}: Sueldo promedio ingeniero [USD por año]
- X_{17}: Sueldo promedio secretaria [USD por año]
- X_{18}: Sueldo promedio vendedor [USD por año]
- X_{19}: Sueldo promedio analista financiero [USD por año]

Además, ya que debemos comparar países con características similares a Argentina, hicimos un cálculo de la distancia euclídea entre la ciudad de Buenos Aires y las demás ciudades. La tabla que mostramos a continuación es una porción de la misma, ya que nos interesa mostrar principalmente aquellos 3 países más parecidos a Argentina.

Row No.	FIRST_ID	SECOND_ID	DISTANCE ↑
12	Buenos Aires	Buenos Aires	0
25	Buenos Aires	Jakarta	2.964
24	Buenos Aires	Istanbul	2.967
26	Buenos Aires	Johannesburg	3.353
48	Buenos Aires	Nairobi	3.355
16	Buenos Aires	Delhi	3.358
46	Buenos Aires	Mumbai	3.389
45	Buenos Aires	Moscow	3.441
13	Buenos Aires	Bucharest	3.445
62	Buenos Aires	Sofia	3.460
8	Buenos Aires	Bogotá	3.666
11	Buenos Aires	Budapest	3.765
41	Buenos Aires	Manila	3.930
66	Buenos Aires	Tallinn	4.040
70	Buenos Aires	Vilnius	4.104

ExampleSet (73 examples, 0 special attributes, 3 regular attributes)

Como conclusión a partir del análisis de la tabla de distancias, se obtiene que los tres países que presentan más similitudes con Argentina son **Indonesia** (Jakarta), **Turquía** (Istanbul) y **Sudáfrica** (Johannesburg). Por lo tanto, nuestra recomendación para el gerente de ventas de la empresa AllHome es desplegar el mercado internacional hacia estos tres países.