Sección de Matemáticas
Universidad de La Laguna

Helí Alonso Afonso

Modelos de agrupamiento aplicados al conocimiento de los usuarios del transporte público de pasajeros

Clustering models applied to the knowledge of the users of public passenger transportation

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Marzo de 2023

Dirigido por
Julio Antonio Brito Santana

Julio Antonio Brito Santana
Ingeniería Informática y de Sistemas
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

En primer lugar, me gustaría agradecer a mi familia por apoyarme durante todo mi camino. También me gustaría agradecer a mis amigos, incluidas a todas las personas que he conocido durante la carrera y me han apoyo incluso en los momentos más complicados. Y, finalmente, me gustaría agradecer a mi tutor Julio por ayudarme, apoyarme y guiarme para poder llevar a cabo este proyecto.

Helí Alonso Afonso
La Laguna, 10 de marzo de 2023

Resumen · Abstract

Resumen

Las técnicas de agrupamiento son procedimiento usados para agrupar datos en función de su similitud o proximidad, e identificar patrones o estructuras subyacentes en los datos. El agrupamiento se utiliza habitualmente en marketing para segmentar a los clientes en grupos en función de su comportamiento o preferencias. Esto ayuda a las empresas a adaptar sus estrategias de marketing y mejorar la captación de clientes. El objetivo de este proyecto consiste en realizar una segmentación de clientes con datos de la empresa de transportes público de pasajeros de Tenerife TITSA, utilizando para ello técnicas de agrupamiento. Se estudian los distintos modelos de segmentación más utilizados por las empresas con el fin de determinar el modelo que resulte de interés para resolver el problema de segmentación en TITSA. Se investigan las diferentes técnicas de agrupamiento que existen para identificar el algoritmo de agrupamiento que resulte más útil en la obtención de la segmentación. Se analizan los resultados del agrupamiento para la obtención de conocimientos de los clientes y sus comportamientos, los cuales son de utilidad para la toma de decisiones en esta entidad

Palabras clave: *Técnicas de Agrupamiento – Segmentación – K-medias –Transporte Público de Pasajeros.*

Abstract

Clustering techniques are procedures to group data based on similarity or proximity and identify underlying patterns or structures. Clustering is commonly used in marketing to segment customers into groups based on their behavior or preferences. This helps companies to adapt their marketing strategies and improve customer acquisition. This project aims to carry out a customer segmentation with data from the Tenerife public passenger transport company TITSA. Clustering techniques are used to perform segmentation. Companies' different segmentation models are studied to determine the interesting model to solve the segmentation problem in TITSA. The different clustering techniques are investigated to identify the clustering algorithm that is most useful in obtaining segmentation. The clustering results are analyzed to acquire knowledge of the clients and their behaviors, which are useful for decision-making in this entity.

Keywords: *Clustering – Segmentation – K-means – Public Passenger Transportation*

Contenido

Agradecimientos III

Resumen/Abstract V

Introducción IX

1. Segmentación de clientes y mercados 1
1.1. Segmentación y beneficios 1
1.2. Tipos de segmentación 2
1.3. Modelos de segmentación mas utilizados 3
1.4. Valoración para su aplicación 4

2. Técnicas de Agrupamiento 7
2.1. Técnicas de Agrupamiento 7
2.2. Medidas de Similitud 8
2.2.1. Distancia de Minkowski 8
2.2.2. Distancia de Mahalanobis 9
2.2.3. Distancia de Levenshtein 9
2.2.4. La función de similitud de Tversky 9
2.2.5. Similitud coseno 10
2.3. Tipos de Agrupamiento 10
2.3.1. Agrupamiento por particiones 10
2.3.2. Agrupamientos basados en densidad 12
2.3.3. Agrupamiento jerárquico 13
2.3.4. Otras técnicas de Agrupamiento 14
2.4. Algoritmo de agrupamiento utilizado 15

3. Aplicación y resultados del Agrupamiento 17
3.1. Proceso de creación del dataset 17
3.1.1. Depuración del conjunto de datos 19
3.2. Análisis descriptivo del conjunto de datos 21

3.3. Obtención del número de agrupamientos 25
3.4. Aplicación del algoritmo de K-medias 26
3.5. Análisis de los resultados 29

Bibliografía 37

Poster 39

Introducción

La transformación digital es un proceso que consiste en reorientar una compañía hacía la aplicación y el uso de las tecnologías emergentes. Este proceso no supone simplemente aplicar tecnología a los departamentos, se trata de darle un sentido a la transformación digital en beneficio de la organización. Un proceso de cambio cultural, cambio organizacional y finalmente de aplicación de las nuevas tecnologías en toda la organización. Así esta transformación implica repensar completamente una organización, para adaptarla integralmente a las demandas del mundo actual [1].

La digitalización ha cambiado nuestra forma de trabajar y de conectarnos. Este cambio de cultura lleva a las personas y organizaciones a evolucionar y a adaptarse, implementando nuevos cambios para avanzar hacía la sociedad del conocimiento. La sociedad del conocimiento y la sociedad de la información son dos términos para denominar los cambios que tienen matices diferentes. Por un lado la información es un instrumento del conocimiento, que se compone de hechos y sucesos, y que a su vez son aquellos elementos asociados con los datos generados en las organizaciones y el entorno. El conocimiento va un poco más allá, y es aquel que puede ser comprendido por cualquier mente humana razonable y tiene que ver con la interpretación de hechos dentro de un contexto, encaminada a alguna finalidad.

Siempre se ha dicho que la información es poder, y quien la posee, se encuentra en una ventaja competitiva respecto al resto. Hoy en día una de las fuentes más valiosas de información se encuentra en los datos generados y recopilados. Debido a la digitalización, la cantidad de datos disponibles para las organizaciones ha aumentado exponencialmente. En este nuevo paradigma de gestión de los datos, las organizaciones han tenido que repasar y renovar los procesos, tecnologías y flujos de trabajo para adaptarse a estas grandes cantidades de información y poder transformarlas en conocimiento de forma efectiva.

Dentro de la denominada Inteligencia Empresarial o Business Intelligence (BI) se incluyen varios procesos y métodos para recopilar, almacenar, transfor-

mar y analizar datos que utilizan técnicas de minería de datos, análisis estadísticos y visualización de datos. En este contexto se enmarca este proyecto, resolver un problema de segmentación. La segmentación consiste en dividir un grupo en subgrupos más pequeños o segmentos que comparten alguna característica, facilitando la obtención de conocimiento. En concreto aplicamos la segmentación de clientes aplicado a los usuarios de una compañía de transporte.

El objetivo de este proyecto consiste en realizar una segmentación de clientes con datos de la empresa de transportes público de pasajeros TITSA. Previamente se estudian los distintos modelos de segmentación más utilizados por las empresas, para analizar y seleccionar aquellos que puedan ser factibles con los intereses de TITSA y los datos disponibles. Posteriormente se indagará sobre diferentes técnicas de Agrupamiento, que puedan ser implementadas, en nuestro caso en código python, las cuales pueden ser de utilidad para la obtención de dichas segmentaciones. Finalmente se analizarán los resultados de aplicar dichas técnicas de agrupamiento y se expondrán las conclusiones.

Cabe destacar que la empresa no dispone de trabajos previos que pudieran servir como guía o referencia , y por tanto no hay una idea a priori de los resultados que se puedan obtener de este proceso.

El contenido de esta memoria, además de esta introducción, las conclusiones finales, apéndices y bibliografía, está estructurado en 3 capítulos principales:

- **Capítulo 1: Segmentación** En este capítulo se hace una introducción a la segmentación, los tipos que existen y se explican brevemente algunos de los modelos más utilizados.
- **Capítulo 2: Técnicas de Agrupamiento** Aquí se describen diferentes técnicas de agrupamiento junto con algunas de sus características y ejemplos de algunas de ellas.
- **Capítulo 3: Aplicando agrupamiento a los datos** En este capítulo se detalla la creación del dataset así como se describen los datos que contiene, se aplican técnicas de agrupamiento sobre estos datos y se analizan los resultados obtenidos.

1

Segmentación de clientes y mercados

Kotler y otros autores definen el término marketing como «un proceso social y de gestión, a través del cual individuos y grupos obtienen lo que necesitan y desean, creando, ofreciendo e intercambiando productos u otras entidades con valor para los otros». Las necesidades humanas son estados de carencia percibida. Incluyen las necesidades físicas de comida, vestido, calor y seguridad; las necesidades sociales de pertenencia y afecto; y las necesidades individuales de conocimiento y autoexpresión. Los deseos son la forma que toman las necesidades humanas a medida que son procesadas por la cultura y la personalidad individual.Los deseos son moldeados por la sociedad y se describen en términos de los objetos que satisfarán esas necesidades. Cuando están respaldados por el poder de compra, los deseos se convierten en demandas [2].

Desde las áreas de marketing, las empresas dedican importantes esfuerzos por entender las necesidades, deseos y demandas de los consumidores. En la actualidad, gracias al desarrollo de las tecnologías de la información y comunicaciones, el uso masivo de las redes sociales y la tecnología disponible de analítica de datos e IA es posible almacenar y analizar una gran cantidad de datos, o que facilita los estudios y análisis de consumidores.

Las empresas que operan en mercados amplios, normalmente no pueden atender a las necesidades de todos sus clientes, ya que son demasiado numerosos y dispersos. Por ello, en lugar de competir en todos los campos y a ciegas, donde normalmente se enfrentan a competidores, necesitan identificar los segmentos de mercado más atractivos donde concentrarse y servir de forma eficaz.

1.1. Segmentación y beneficios

El término “segmentación” fue acuñado en Marketing por primera vez por Smith [3]. Segmentar es dividir el mercado total de un producto o servicio en diferentes grupos de consumidores, homogéneos entre sí y diferentes a los demás,

en cuanto a hábitos, necesidades y gustos, que podrían requerir servicios o productos diferentes. Estos grupos se denominan segmentos y se obtienen mediante diferentes procedimientos estadísticos, a fin de poder aplicar a cada segmento las estrategias de marketing más adecuadas para cumplir los objetivos de la empresa [4].

Algunos de los beneficios que aporta tener a los usuarios clasificados en segmentos se detallan a continuación:

- Facilita la identificación de necesidades específicas. Para los usuarios de diferentes segmentos permite definir canales de comunicación específicos y efectivos para cada grupo, haciendo más sencillo mantener una interacción constante.

- Realiza la asignación de recursos de marketing con mayor nivel de eficacia, ya que se adaptan las estrategias y las acciones emprendidas a las características de cada segmento. De esta manera, se ajustan los procesos comerciales a cada uno de estos grupos.

- Simplifica encontrar un nicho propio donde no se tenga competencia directa. Es decir, aumenta las posibilidades de crecer rápidamente en segmentos del mercado donde no haya competidores.

1.2. Tipos de segmentación

Hay diferentes enfoques para realizar la segmentación de clientes. En concreto se definen cuatro tipos de segmentación, estos son:

- **Segmentación demográfica**
  La segmentación demográfica supone que la gente con características en común, tendrán unos intereses, comportamientos y gustos similares, los cuales pueden influir en sus hábitos de compra. Este tipo de segmentación se suele complementar con otros tipos de segmentación para obtener mayor certeza a la hora de identificar objetivos de mercado. Este tipo de segmentación tiene en cuenta aspectos tales como la edad, el género, la profesión, la educación y el sueldo, entre otros.

- **Segmentación psicográfica**
  La segmentación psicográfica divide a los usuarios en grupos basándose en: sus personalidades, estatus social, estilo de vida, intereses, actividades, opiniones, intereses y actitudes. Este tipo de segmentación funciona muy bien junto con

las segmentaciones demográficas ya que sirven para identificar qué es lo que motiva a los clientes a tomar ciertas decisiones

- **Segmentación geográfica**
  La segmentación geográfica permite agrupar a los usuarios basándose en el lugar donde viven, trabajan, o viajan, ya que estos aspectos tienen mucha influencia en los hábitos de compra de los clientes. También se tienen en cuenta variables tales como el país de residencia, la región, la provincia, ciudad, el clima, la cultura y la densidad de población, todos ellos aspectos que dan información de problemas y necesidades que afectan al espacio físico que circunscriben que productos o servicios necesitan.
- **Segmentación por conducta**
  El objetivo de este tipo de segmentación es encontrar el por qué detrás de las compras de los clientes. Esta segmentación considera cuestiones como el motivo de la compra, si es por valoraciones positivas del producto, por encontrar un buen precio o por otro motivo. El impacto positivo que tiene el producto en el cliente, si es un producto seguro, de confianza o si es un producto de última tecnología. Si el cliente es usuario habitual de la marca o quiere probarla por primera vez, así como la sensibilidad al precio que puedan tener los clientes.

1.3. Modelos de segmentación mas utilizados

Los siguientes modelos de segmentación son los que más han sido adoptados por los profesionales de marketing:

- **Segmentación por valor del cliente (CLV)**
  CLV son las siglas de "client lifetime-value". Esta segmentación se basa en el concepto del valor monetario que representa el cliente durante todo el ciclo de vida del cliente con la marca. Se considera el CLV como el valor actual del futuro flujo monetario que representa un cliente, es decir, es la suma de los ingresos que supone el cliente durante toda su relación con la empresa tras descontarle los gastos que suponen la atracción, venta y servicios del cliente [5].Esta segmentación tiene como finalidad agrupar los clientes en alto, medio o bajo valor desde un punto de vista puramente monetario de rentabilidad e inversión por cada cliente a largo plazo.
- **Segmentación por satisfacción (NPS)**
  NPS (Net Promoter Score) agrupa a los clientes según el grado de satisfacción que tienen con la empresa o con los servicios que ésta ofrece, basándose en la respuesta de estos a una simple pregunta: ¿Qué tan probable es que recomiendes los servicios de la empresa a un amigo o conocido?. Las respuestas a esta pregunta vienen dadas en una escala del 0 al 10 y son agrupadas en tres categorías: promotor si la respuesta es 9 o 10, neutral si la respuesta es 7 o 8,

y detractor si la respuesta está por debajo de 7. Esta calcular el NPS de una empresa simplemente tomando la diferencia entre promotores y detractores y dividiendo esta diferencia entre el total de clientes. Este modelo ha sido mayoritariamente adoptado por su sencillez, además de permitir comparar diferentes compañías, sirve para hacerse una idea de la forma de pensar de los clientes respecto a la marca y para predecir el posible crecimiento de las ventas [6].

- **Segmentación Transaccional (RFM)**
  El modelo RFM diferencia la importancia de los clientes a partir de tres variables, proximidad de las últimas compras, frecuencia de compra y valor monetario de las compras. Las siglas de RFM provienen de "recency", "frequency" y "monetary". La primera hace referencia a que tan reciente es la última compra realizada, "frequency" es la frecuencia de consumo en un determinado periodo de tiempo y "monetary" se refiere a la cantidad monetaria gastada en las compras en un determinado periodo de tiempo [7]. Esta mecánica es simple de implementar y permite decidir el nivel de segmentación que se desea conseguir. Es un modelo muy intuitivo porque tipifica a los clientes exclusivamente en base a sus hábitos de compra.
- **Segmentación por Valor/Risego**
  Esta segmentación tiene la finalidad de dividir los clientes en 4 cuadrantes considerando dos de las variables más importantes para el tratamiento de clientes, el valor que supone el cliente a largo plazo y el riesgo de abandono. Esto nos deja una segmentación bastante sencilla pero también efectiva que separa a los clientes en: (valor alto y riesgo alto), (valor alto y riesgo bajo), (valor bajo y riesgo alto) y (valor bajo y riesgo bajo).

1.4. Valoración para su aplicación

Una vez planteados los diferentes modelos de segmentación más comúnmente adoptados por las empresas, en coordinación con los miembros encargados del análisis de datos de la empresa TITSA, se realiza una valoración de los modelos. Teniendo en cuenta el interés y las necesidades de la empresa, así como la viabilidad del estudio a partidr de los datos disponible, se determinar qué aspectos que resultan de mayor interés para la empresa.

Com conclusión de la valoaración se determina agrupar a los clientes principalmente tomando en cuenta los datos que representan el valor monetario que estos aportan a la empresa, así como por los datos que puedan representar el esfuerzo que cuesta a los clientes utilizar los servicios de la empresa. La combinación de estos datos daría información bastante útil para tomar decisiones operativas y de marketing. Este es un planteamiento inicial teniendo en cuenta los datos disponibles. Posteriormente una vez analizados en profundidad los

datos y obtenidos algunos resultados preliminares del uso de las técnicas utilizadas, se pueden reducir o aumentar las variables utilizadas, y los objetivos y resultados esperados.

Una vez confirmados los datos que vamos a utilizar para la segmentación y el enfoque de análisis con los mismos, es en le tercer capítulo donde se describe el conjunto de datos utilizados en detalle, incluyendo carácterística que son importantes para nuestro estudio. A continuación en el siguiente capítulo se introduce a las técnicas de agrupamiento en general y específicamente la que utilizaremos para encontrar los segmentos de clientes propuestos.

2

Técnicas de Agrupamiento

En este capítulo se definirá las técnicas de agrupamiento y algunas de las utilidades que tienen. El capítulo consta de una introducción a estas técnicas y a las distintas medidas de similitud que se pueden utilizar según el tipo de datos que manejemos. Así como las características que queramos considerar a la hora de determinar si dos objetos son similares o diferentes. También se presenta una clasificación de los diferentes tipos de técnicas de agrupamiento con algunos ejemplos representativos de cada uno de ellas. Para concluir el capítulo se aporta una valoración sobre cuáles de estas técnicas pueden resultar de utilidad para realizar la segmentación con los datos de TITSA y cuáles se van a implementar.

2.1. Técnicas de Agrupamiento

El agrupamiento o clustering es un proceso de aprendizaje automático, que consiste en realizar una clasificación de observaciones o datos en distintos grupos (cluster). Es un tipo de aprendizaje no supervisado, lo que quiere decir que los datos a los que se le aplica no están etiquetados, es decir, no se dispone de ningún tipo de información sobre los valores de salida de los datos.

Generalmente se utiliza agrupamiento para encontrar algún tipo de estructura en un conjunto de datos sin etiquetar. Los algoritmos de agrupamiento buscan organizar objetos en distintos grupos, de forma que los miembros de un mismo grupo tienen características similares entre sí y diferentes respecto a los miembros de otros grupos.

El agrupamiento es una técnica popularmente utilizada en el análisis de datos, se emplea comúnmente en tareas como la minería de datos, el reconocimiento de patrones o para encontrar grupos significativos dentro del conjunto de datos. Encontramos también algunas de sus utilidades a la hora de formar una idea de la estructura que tienen los datos, para mejorar tareas de predicción [8] o en la toma de decisiones. En particular en este trabajo el agrupamiento será

la herramienta principal para obtener una segmentación de los clientes a partir de los datos que se van a utilizar.

2.2. Medidas de Similitud

Como se mencionó anteriormente, el agrupamiento es el proceso por el que se clasifican los distintos datos de un conjunto en grupos, de forma que los elementos de un mismo grupo sean similares entre sí y diferentes de los elementos de otros grupos. La manera de determinar si dos objetos son similares entre sí, son las medidas de similitud. Son muchas las medidas que podemos considerar según el tipo de datos que tengamos, la distribución de los mismos o las características que se quieran considerar. La manera más popular de evaluar la similitud es utilizando distancias, de forma que cuanto menor es la distancia entre dos objetos, mayor su similitud, entonces a la hora de agrupar el objetivo es minimizar la distancia dentro del grupo y maximizar la distancia entre grupos. [9]

Para que una medida de similitud sea considerada una distancia debe verificar las siguientes propiedades:

Reflexiva $d(i,j) = 0 \Leftrightarrow i = j$
Simétrica $d(i,j) = d(j,i)$
Desigualdad triangular $d(i,j) \leq d(i,k) + d(k,j)$

En el proceso de agrupamiento es importante utilizar medidas de similitud adecuadas a los datos para obtener buenos resultados, pero también hay otro factores que influyen en la efectividad del agrupamiento, como normalizar las variables para evitar que unas dominen sobre otras o utilizar solo los elementos más representativos de nuestro conjunto de datos son normalmente minimizará posibles errores y disminuye el gasto computacional.

A continuación se describirán y se explicarán brevemente algunas de las medidas de similitud más comúnmente utilizadas.

2.2.1. Distancia de Minkowski

La distancia Minkowski es una métrica que mide la distancia entre dos puntos de un espacio vectorial N-dimensional. Básicamente es una generalización tanto de la distancia euclídea como de la distancia Manhattan.

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Los tres casos particulares más conocidos de esta distancia son:

Distancia de Manhattan/taxi ($p = 1$)

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|)$$

Distancia euclídea ($p = 2$)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Distancia Chebyshev ($p \to \infty$)

$$d(x, y) = \max_{i=1...n} |x_i - y_i|$$

2.2.2. Distancia de Mahalanobis

La distancia de mahalanobis es una métrica que se utiliza para determinar la similitud entre dos variables aleatorias multidimensionales, esta tiene en cuenta la correlación entre variables aleatorias. La distancia mahalanobis entre dos variables aleatorias que tienen la misma distribución de probabilidad y con matriz de covarianza M se define de la siguiente manera:

$$d(x, y) = \sqrt{(x - y)^T M^{-1} (x - y)}$$

2.2.3. Distancia de Levenshtein

La distancia de Levenshtein, también llamada distancia de edición o distancia entre palabras, es el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra, donde una operación puede ser una inserción, sustitución o eliminación de un carácter. Por ejemplo la distancia entre 'nuevo' y 'nieve' es de 2, ya que se necesitan como mínimo dos operaciones para transformar uno en el otro

nuev**o** → nuev**e** (sustitución de 'o' por 'e')

n**u**eve → n**i**eve (sustitución de 'u' por 'i')

2.2.4. La función de similitud de Tversky

La función de similitud de Tversky se utiliza para medir la similitud entre conjuntos, y se define como:

$$\sigma_{\beta,\gamma}(A, B) = \frac{|A \cap B|}{|A \cap B| + \beta|A - B| + \gamma|B - A|}$$

Donde $\beta, \gamma \geq 0$ son pesos que le dan mayor o menor importancia a los elementos no comunes de ambos conjuntos.

2.2.5. Similitud coseno

La similitud coseno mide la similitud que hay entre dos vectores de un espacio vectorial con producto interior a partir del valor del coseno del ángulo que forman ambos vectores. Esta medida devuelve valores en el intervalo [-1,1]. La similitud coseno no es una métrica ya que no cumple la desigualdad triangular. Es una de las medidas más populares que se utilizan para medir la similitud entre documentos de texto, cuando estos se representan como vectores, se utiliza esta medida para medir su similitud.

$$cos(x, y) = \sum_i \frac{x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

Las medidas detalladas anteriormente son solo algunas de las más popularmente utilizadas, hay muchas otras medidas de similitud que se pueden encontrar en la literatura. Se puede encontrar más información en las lecturas siguientes [10] [11] [12]

2.3. Tipos de Agrupamiento

Son muchas las técnicas de agrupamiento que podemos encontrar en la literatura, cada una con sus puntos fuertes y desventajas. Muchas comparten características comunes, por lo que resulta difícil realizar una categorización completa de los distintos tipos de agrupamiento. A continuación se presentan algunos de los tipos de agrupamiento que existen según [13] de los cuales veremos algunos ejemplos.

2.3.1. Agrupamiento por particiones

La idea principal del agrupamiento por particiones es la siguiente: dado un conjunto de datos construir un número dado de particiones “k” , empezando por una partición inicial donde tendremos k grupos, cada uno con un centro o centroide. Estas técnicas se centran en estos centros de los grupos, a los que se les aplican técnicas de relocalización iterativa de forma que se obtendrán nuevos grupos. El proceso se realiza para intentar mejorar la partición, y se repite hasta que se cumpla cierto criterio. Generalmente el criterio para considerar una partición como buena consiste en que elementos del mismo grupo estén cerca o estén relacionados, mientras que los elementos de grupos diferentes estén lejos o sean diferentes entre sí [14]

Algunos algoritmos relevantes de agrupamiento por particiones son:

- **K-Medias**

  Este algoritmo tiene como objetivo particionar un conjunto de datos en 'k' grupos. El proceso comienza con 'k' grupos iniciales, compuestos únicamente por un punto aleatorio cada uno. A partir de ahí, se añade cada nuevo punto al grupo cuya media sea más cercana al punto. Cada vez que se añade un nuevo punto a un grupo se reajusta la media del grupo teniendo en cuenta ese nuevo punto. Por tanto en cada etapa las k-Medias son de hecho las medias de los grupos que representan, de ahí proviene su nombre [15].

  En primera instancia no se conoce el valor de 'k' para que el agrupamiento separe los elementos lo mejor posible. Este valor dependerá de los datos con los que se trabajen. Existen varios métodos para determinar buenos valores para la K, por ejemplo el método del codo. Este método consiste en realizar K-medias para distintos valores de K y asignarles una puntuación, normalmente esto se hace utilizando la distancia de los puntos de un grupo a su centroide. Posteriormente, se representa gráficamente la puntuación respecto al número de grupos (K). Esta representación tendrá una forma similar a la de un brazo, y será el cambio brusco en la pendiente (el codo) lo que nos indicará un número apropiado para K.

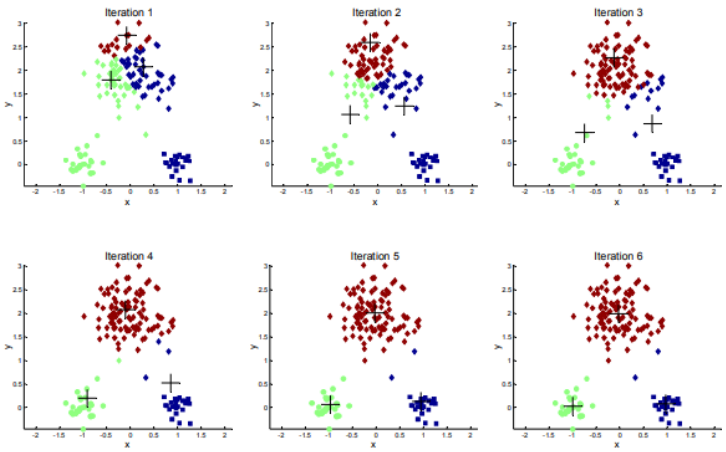


Figura 2.1. Resultado del algoritmo de k-medias

- **CLARANS**

  CLARANS (Clustering Large Applications based on RANdomized Search) es un algoritmo de agrupamiento por particiones. Este método es particularmente útil en la minería de datos espaciales, es decir, para descubrir relaciones y características que podrían existir de forma implícita en las bases de datos espaciales. El principal objetivo de este algoritmo es identificar estructuras espaciales que podrían encontrarse en los datos [16].

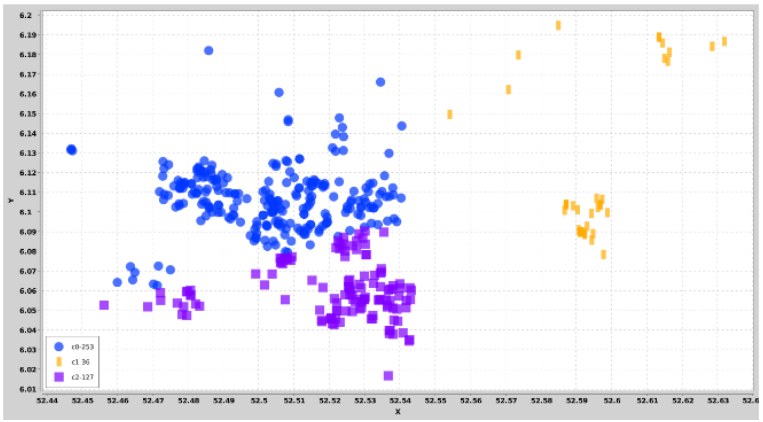


Figura 2.2. Resultado del algoritmo CLARANS

2.3.2. Agrupamientos basados en densidad

En los agrupamientos basados en densidad, se toman como un grupo, los elementos que se reparten en el espacio dentro de una misma región de puntos con alta densidad de datos, de forma que los elementos de un mismo grupo tienen puntos del mismo grupo colindantes o contiguos a él. Estos grupos se separan entre sí por regiones de baja densidad de datos, los puntos que se encuentran en estas regiones de baja densidad normalmente se consideran ruido o outliers [17].

Estos métodos se pueden utilizar para filtrar ruidos o descubrir grupos de formas arbitrarias. Tienen como ventaja respecto a otros que no requieren como valor inicial el número de grupos y no hacen suposiciones respecto a la densidad de los datos o la varianza entre los grupos. Por estas razones el resultado de realizar agrupación por densidad no son necesariamente grupos de puntos similares, por esto mismo los grupos no tienen necesariamente forma convexa, sino que pueden tener formas arbitrarias.

El algoritmo más característico cuando se habla de agrupamiento por densidad es:

- **DBSCAN**

DBSCAN (Density Based Spatial Clustering of Applications with Noise) es un algoritmo de agrupamiento por densidad, es uno de los más utilizados y citados, está diseñado para encontrar tanto los posibles grupos como el ruido en bases de datos espaciales.

La idea principal del algoritmo es que para cada punto de un grupo, el entorno del punto de un radio dado, deben contener al menos un número mínimo de puntos. Es decir, la densidad de puntos en su entorno tiene que superar cierto umbral. El proceso comienza tomando un punto arbitrario y comprobando si

cumple este criterio de densidad. Si lo cumple se construye un grupo añadiendo puntos cercanos y cuando acaba de construir el grupo, salta a otro punto. Si el punto inicial no supera el umbral de densidad se clasificará como ruido. Se debe tener en cuenta que la forma de los entornos de un punto viene determinada por la distancia que se utilice [18].

A continuación se muestra el resultado de aplicar DBSCAN sobre puntos con una distribución espacial sobre la que el algoritmo funciona particularmente bien.

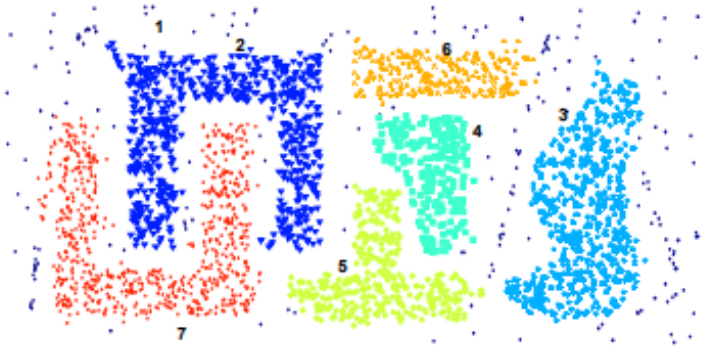


Figura 2.3. Resultado del algoritmo DBSCAN

2.3.3. Agrupamiento jerárquico

El agrupamiento jerárquico, al contrario que otros tipos de agrupamiento, no clasifica los elementos en grupos directamente, sino que crea una estructura jerárquica como un árbol de categorías llamado dendrograma. El dendrograma muestra posibles agrupaciones de los elementos según el nivel del mismo donde nos encontremos. El nivel más alto es un grupo con todos los elementos y el nivel más bajo, donde cada elemento conforma su propio grupo [19].

Los procesos de agrupamiento jerárquico se pueden dividir en dos categorías, aglomerativos y divisivos. Los procesos aglomerativos o de abajo hacia arriba comienzan con grupos unitarios conformados por cada elemento, que suponen el nivel más bajo de la jerarquía. Sucesivamente agrupan los elementos hasta que todos los elementos están fusionados en un solo grupo, que supone el nivel más alto de la jerarquía. Inversamente los procesos divisivos o de arriba hacia abajo comienzan con un grupo que contiene todos los elementos y divide en grupos más pequeños sucesivamente hasta que cada elemento es un grupo.

Algunos ejemplos de algoritmos de agrupamiento jerárquico son:

- **Chameleon**

  Chameleon es un algoritmo de agrupamiento jerárquico aglomerativo, este consta de dos fases para encontrar los grupos. En la primera fase se agrupan

los datos en subgrupos relativamente pequeños usando algoritmos de particionamiento de grafos. Durante la segunda fase, se realizan combinaciones de estos subgrupos hasta encontrar los grupos definitivos. Una de las claves del algoritmo es que tiene en cuenta la interconectividad y cercanía a la hora de identificar los pares de subgrupos más similares [20].

- **BIRCH**

  BIRCH (Balanced Iterative Reducing and Agrupamiento using Hierarchies) es un algoritmo de agrupamiento jerárquico que funciona especialmente bien con conjuntos de datos muy grandes, BIRCH normalmente puede encontrar buenas agrupaciones con recorrer los datos una sola vez, y puede mejorar la calidad recorriendo los datos algunas veces más.

  Este algoritmo se divide en cuatro fases que se detallan a continuación. En la primera fase, se cargan los datos y se construye un árbol de Características de Agrupamiento (Clustering Feature Tree) inicial. La segunda fase es opcional, y consiste en reconstruir el árbol inicial en uno más pequeño que se adapte mejor al rango deseado. La tercera fase ejecuta algoritmos de agrupamiento global o semi-global para agrupar las entradas de todas las hojas del árbol. La cuarta y última fase es opcional y consiste en refinar la agrupación para obtener mejores resultados[21].

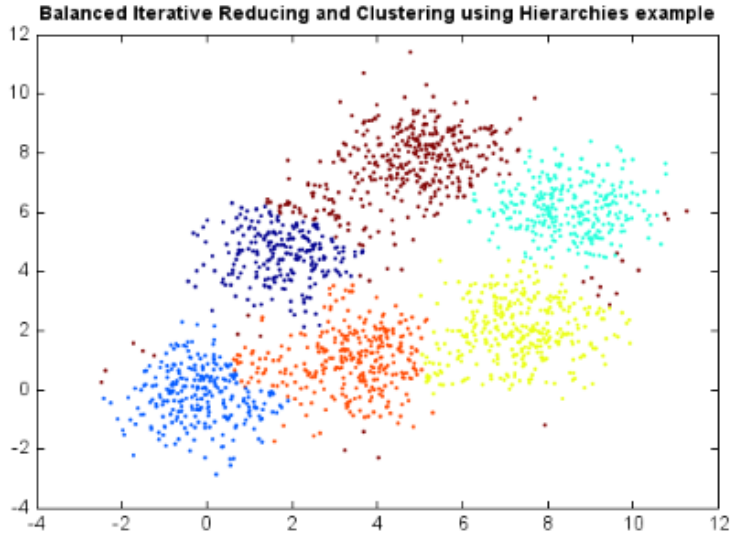


Figura 2.4. Resultado del algoritmo BIRCH

2.3.4. Otras técnicas de Agrupamiento

- **Métodos basados en cuadrícula**

La idea de este tipo de algoritmos de agrupamiento consiste en cambiar el espacio original de los datos a una estructura de rejilla o cuadrícula y trabajar con este espacio de rejilla en lugar del original. El proceso consiste en que una vez se divide el espacio en celdas se buscan las regiones densas. Para ello, se identifican las celdas que contienen más de cierto número de datos como densas y a partir de aquí se forman los grupos conectando estas celdas densas[22].

La principal ventaja de este tipo de agrupaciones es su velocidad de procesamiento, ya que normalmente no depende del número de datos, sino del número de celdas en cada dimensión del espacio. Por otro lado, pueden perder efectividad a medida que aumenta la dimensión de los datos.

Algunos ejemplos de algoritmos basados en cuadrícula son STING y CLIQUE.

- **Métodos basados en modelos**

  Aunque la mayoría de algoritmos encuentran los grupos optimizando ciertos criterios basados en la distancia de los datos, los métodos basados en modelos seleccionan un modelo generador para los datos y a partir de esto utilizan similitud o posteriormente probabilidad derivada de este modelo como criterio a optimizar para encontrar los grupos. Este tipo de agrupamiento permite identificar los grupos basándose en su forma o estructura en lugar de la proximidad de los datos [23].

2.4. Algoritmo de agrupamiento utilizado

Los datos sobre los que aplicamos el agrupamiento son son todo variables numéricas, esto hace que el problema sea bastante sencillo. Es por ello, que vamos a utilizar el agrupamiento de K-medias. Este algoritmo es de los más simples y rápidos, trabaja bien con conjuntos de datos grandes y es fácil de interpretar los resultados.

En un principio se planteó aplicar distintos tipos de agrupamiento, ya que al combinar varios tipos de agrupamiento normalmente se obtiene una información más completa, pero dadas las limitaciones, tanto de tiempo, como límite en la extensión del trabajo, se ha optado por realizar únicamente K-medias para poder profundizar en su desarrollo y en el análisis de los resultados.

Se descartaron otros tipos de algoritmos como los basados en densidad ya que observando la distribución de los datos se concluye que los resultados obtenidos por este tipo de algoritmos serían probablemente grupos muy evidentes o de poco interés. También los agrupamientos jerárquicos se descartaron ya que algunos tienen un coste computacional muy alto con conjuntos grandes de datos. Estos son más sensibles a valores extremos, los cuales hay en el conjunto de datos disponibles, y además suelen ser más complejos de interpretar.

3

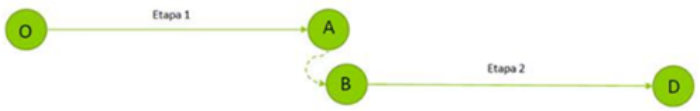
Aplicación y resultados del Agrupamiento

El contenido de este capítulo de experimentación aplicada en este trabajo se estructura en cuatro partes. La primera parte explica la creación del dataset sobre el que se ha trabajado el agrupamiento, los orígenes de datos y los datos que se utilizan. La seguda parte, describe el proceso aplicado a los datos previamente a la aplicación de las técnicas de agrupamiento. A continuación en un tercer apartado se presenta el proceso realizado de aplicación de la técnica de agrupamiento. Para finalizar, describimos los resultados de la experimentación realizada y el análisis dichos resultados.

3.1. Proceso de creación del dataset

Para obtener el conjunto de datos para la experimentación, se han tenido acceso a las bases de datos de TITSA mediante consultas SQL. Esto ha permitido identificar datos de interés y obtener diferentes tablas para realizar el agrupamiento. También se realizan uniones de varias tablas para obtener los datos requeridos. A continuación se describen cada una de las tablas generadas y algunos aspectos y características a tener en cuenta sobre ellas.

La tabla principal de la que se obtienen los datos es la matriz origen-destino (OD), La matriz OD permite conocer para cada usuario el conjunto de etapas, es decir, de líneas de transporte que han sido utilizados por este para llegar desde su origen 'O' al destino 'D'. Este conjunto de etapas, que denominamos viaje, puede comprender desde 1 a 4 etapas con las que el usuario puede alcanzar su destino final.



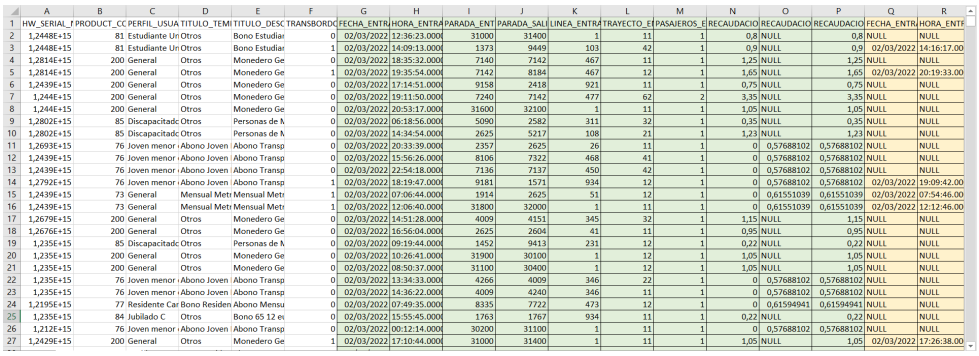
Parada subida	Parada bajada	Paradas subida	Paradas bajada	Número transbordos
O	D	O, B	A, D	1

Figura 3.1. Representación de un viaje en la matriz Origen-Destino

Tambien se dispones de datos en otras tablas que se unen a la anterior, con datos procedentes de la empresa que contienen detalles sobre las líneas, las zonas financieras, los trayectos, las fechas, los horarios, las recaudaciones de tarifas temporales y políticas de compensación. Además hemos necesitado una tabla específica que contiene datos referentes a los títulos de transporte de los usuarios, necesaria para el proceso de enlazar las distintas tablas.

Cabe destacar que una de las uniones a la tabla principal se realizan utilizando la fecha de operación en lugar de la fecha de inserción, lo cual genera un ligero error respecto al valor real en la recaudación para un pequeño número de viajes realizados con títulos de tarifas temporales, se ha comprobado que no produce cambios significativos, pero es un punto a mejorar. En otro orden de cosas, se planteó añadir datos referentes a la latitud y longitud de las paradas, lo que nos permitiría calcular las distancias entre paradas. Información que sirve para medir el esfuerzo de los clientes (como tiempo invertido). Aunque finalmente se ha descartado esta opción en el análisis, queda como posible mejora en el futuro para su implementación.

Finalizado este proceso, obtenemos el conjunto de datos de la experimentación en una tabla con los datos requeridos para aplicar las técnicas de agrupamiento. Esta tabla se guarda en un formato '.csv' y contiene los siguientes datos estructurados: identificación de los clientes (número de tarjeta, el título y perfil de usuario), etapas del viaje que realiza, cada viaje puede estar compuesto hasta por cuatro etapas (fecha, hora, parada de entrada, parada de salida, línea utilizada, número de pasajeros y recaudación de la etapa), políticas de compensación de la empresa (ingreso monetario indirecto del viaje que cumplen determinadas condiciones). A continuación se muestra la tabla de datos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HW_SERIAL_I	PRODUCT_CC	PERFIL_USUA	TITULO_TEMI	TITULO_DESC	TRANSBORDC	FECHA_ENTR	HORA_ENTRA	PARADA_ENT	PARADA_SALI	LINEA_ENTRA	TRAYECTO_EI	PASAJEROS_E	RECAUDACIO	RECAUDACIO	RECAUDACIO	FECHA_ENTR	HORA_ENTF
2	1,2448E+15	81	Estudiante Un	Otros	Bono Estudiar	0	02/03/2022	12:36:23.000	31000	31400	1	11	1	0,8	NULL	0,8	NULL	NULL
3	1,2448E+15	81	Estudiante Un	Otros	Bono Estudiar	1	02/03/2022	14:09:13.000	1373	9449	103	42	1	0,9	NULL	0,9	02/03/2022	14:16:17.00
4	1,2814E+15	200	General	Otros	Monedero Ge	0	02/03/2022	18:35:32.000	7140	7142	467	11	1	1,25	NULL	1,25	NULL	NULL
5	1,2814E+15	200	General	Otros	Monedero Ge	1	02/03/2022	19:35:54.000	7142	8184	467	12	1	1,65	NULL	1,65	02/03/2022	20:19:33.00
6	1,2439E+15	200	General	Otros	Monedero Ge	0	02/03/2022	17:14:51.000	9158	2418	921	11	1	0,75	NULL	0,75	NULL	NULL
7	1,244E+15	200	General	Otros	Monedero Ge	0	02/03/2022	19:11:50.000	7240	7142	477	62	2	3,35	NULL	3,35	NULL	NULL
8	1,244E+15	200	General	Otros	Monedero Ge	0	02/03/2022	20:53:17.000	31600	32100	1	11	1	1,05	NULL	1,05	NULL	NULL
9	1,2802E+15	85	Discapacitado	Otros	Personas de M	0	02/03/2022	06:18:56.000	5090	2582	311	32	1	0,35	NULL	0,35	NULL	NULL
10	1,2802E+15	85	Discapacitado	Otros	Personas de M	0	02/03/2022	14:34:54.000	2625	5217	108	21	1	1,23	NULL	1,23	NULL	NULL
11	1,2693E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	20:33:39.000	2357	2625	26	11	1	0	0,57688102	0,57688102	NULL	NULL
12	1,2439E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	15:56:26.000	8106	7322	468	41	1	0	0,57688102	0,57688102	NULL	NULL
13	1,2439E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	22:54:18.000	7136	7137	450	42	1	0	0,57688102	0,57688102	NULL	NULL
14	1,2792E+15	76	Joven menor	Abono Joven	Abono Transp	1	02/03/2022	18:19:47.000	9181	1571	934	12	1	0	0,57688102	0,57688102	02/03/2022	19:09:42.00
15	1,2439E+15	73	General	Mensual Metr	Mensual Metr	1	02/03/2022	07:06:44.000	1914	2625	51	12	1	0	0,61551039	0,61551039	02/03/2022	07:54:46.00
16	1,2439E+15	73	General	Mensual Metr	Mensual Metr	1	02/03/2022	12:06:40.000	31800	32000	1	11	1	0	0,61551039	0,61551039	02/03/2022	12:12:46.00
17	1,2679E+15	200	General	Otros	Monedero Ge	0	02/03/2022	14:51:28.000	4009	4151	345	32	1	1,15	NULL	1,15	NULL	NULL
18	1,2676E+15	200	General	Otros	Monedero Ge	0	02/03/2022	16:56:04.000	2625	2604	41	11	1	0,95	NULL	0,95	NULL	NULL
19	1,235E+15	85	Discapacitado	Otros	Personas de M	0	02/03/2022	09:19:44.000	1452	9413	231	12	1	0,22	NULL	0,22	NULL	NULL
20	1,235E+15	200	General	Otros	Monedero Ge	0	02/03/2022	10:26:41.000	31900	30100	1	12	1	1,05	NULL	1,05	NULL	NULL
21	1,235E+15	200	General	Otros	Monedero Ge	0	02/03/2022	08:50:37.000	31100	30400	1	12	1	1,05	NULL	1,05	NULL	NULL
22	1,235E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	13:34:33.000	4266	4009	346	22	1	0	0,57688102	0,57688102	NULL	NULL
23	1,235E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	14:36:22.000	4009	4240	346	11	1	0	0,57688102	0,57688102	NULL	NULL
24	1,2195E+15	77	Residente Car	Bono Residen	Abono Mensu	0	02/03/2022	07:49:35.000	8335	7722	473	12	1	0	0,61594941	0,61594941	NULL	NULL
25	1,235E+15	84	Jubilado C	Otros	Bono 65 12 e	0	02/03/2022	15:55:45.000	1763	1767	934	11	1	0,22	NULL	0,22	NULL	NULL
26	1,212E+15	76	Joven menor	Abono Joven	Abono Transp	0	02/03/2022	00:12:14.000	30200	31100	1	11	1	0	0,57688102	0,57688102	NULL	NULL
27	1,2429E+15	200	General	Otros	Monedero Ge	1	02/03/2022	17:10:44.000	31000	31400	1	11	1	1,05	NULL	1,05	02/03/2022	17:26:38.00

Figura 3.2. Tabla de datos

En una revisión inicial de la tabla de datos no se han encontrado un gran número de errores o valores vacíos, aun así el siguiente paso consiste en refinar el conjunto de datos de la experimentación para prepararlo para el proceso de agrupamiento. Se trata de minimizar el número de posibles errores para mejorar los resultados. Para esto se eliminan filas que puedan inducir errores o no aporten información, se calcularán nuevos valores o se harán los cambios necesarios.

Antes de explicar el proceso de depuración señalar que aunque las tablas contienen variables nominales esto no influye, ya que solo se trabajará con los datos numéricos al aplicar el algoritmo. Por otro lado, inicialmente se habia decidido utilizar los datos de un mes entero, lo que suponía tratar aproxximadamente 4.000.000 de filas. Esto se descartó ya que el volumen de datos hace que los tiempos de ejecución del algoritmo sean excesivamente largo. Finalmente se opta por trabajar con los datos de un día, de un tamaño aproximado de 141.000 filas. En caso de se querer aplicar los procedimientos a una mayor cantidad de datos el procedimiento es similar y se puede aplicar sin cambios facilmente.

3.1.1. Depuración del conjunto de datos

A continuación se explican algunos de los cambios que se realizan en el proceso de depuración sobre el conjunto de datos de la experimentación.

En primer lugar se eliminan las filas compuestas únicamente por líneas que no corresponden a TITSA. Estos datos no interesan y sólo aportan ruido. Lo mismo ocurre con las filas cuya recaudación es negativa ya que están asociadas a viajes cancelados, por tanto también se eliminan estas filas. De igual modo se quitan aquellas filas donde un viaje lo realice más de un pasajero, nos interesa tratar individuos. Estas filas compuestas por varios usuarios pueden alterar los grupos que se obtiene al aplicar las técnicas de agrupamiento, ya que contienen valores mucho mayores en algunos casos que distorsionan los resultados.

```
Dimensiones antes de realizar ningun cambio:
(141010, 53)
Dimensiones despues de haber eliminado las filas mencionadas:
(106712, 53)
```

Figura 3.3. Se eliminan las filas que no se van a usar del dataset

Como se puede ver, con este primer paso de la depuración, pasamos de tener 141.000 filas de datos en la tabla a tener algo menos de 107.000 filas.

Como segundo paso se reajustan los índices de cada fila después de haber borrado algunas para evitar posibles errores. También se calcula una nueva variable de transbordos, que nos indica si el viaje tiene etapas extra, el cual puede tomar valores desde 0 si el viaje se completa en una etapa hasta 3 si el viaje se compone de 4 etapas. Este dato es de gran de utilidad para medir el esfuerzo de los usuarios el realizar un viaje. Cuanto mayor sea el valor, mayor el esfuerzo. Como se mencionó anteriormente hubiese sido importante para medir el esfuerzo diponer de información sobre el tiempo de duración del viaje, pero no lo tenemos entre el conjunto de datos disponible.

El siguiente paso es eliminar las recaudaciones asociadas a líneas que no son de TITSA y que por tanto no aportan información útil para el análisis. Al cambiar los datos de recaudaciones, es necesario recalcular las recaudaciones totales de los viajes. También se crean, a partir de las variables categóricas de título de usuario y líneas utilizadas, nuevas columnas de variables indicadoras ("dummy") que serán útiles más adelante para la visualización de los datos.

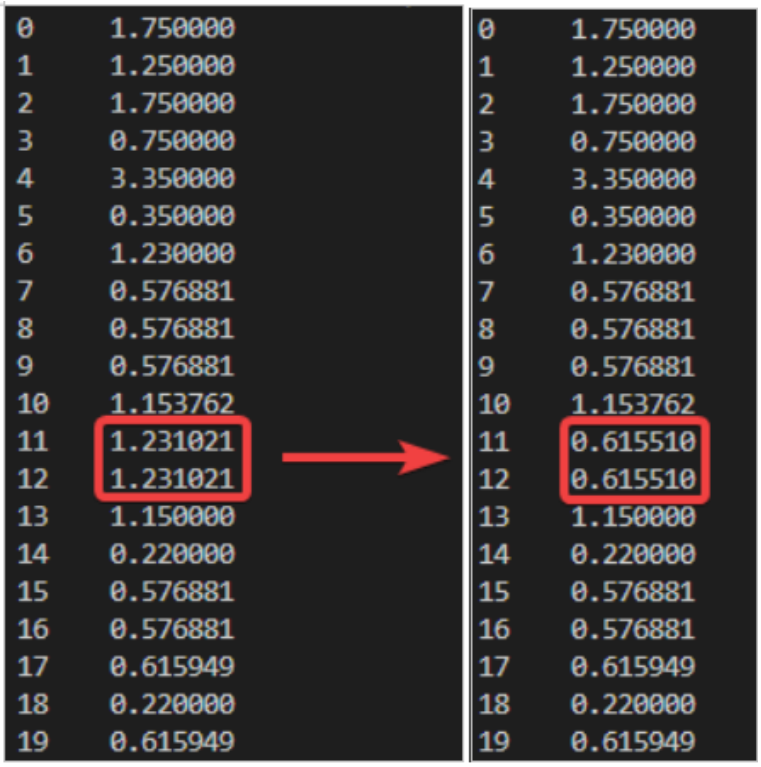


Figura 3.4. Resultado de eliminar las recaudaciones no asociadas a Titsa

3.2. Análisis descriptivo del conjunto de datos

Una vez depurados el conjunto de datos de la experimentación, se realiza un análisis descrptivo previo sobre estos. Para ello, se utilizan algunos procedimientos y algunas medidas estadísticas y algunas representaciones gráficas para tener un mayor conocimiento a priori de los datos, variables y situación de partida, antes de aplicar específicamente las técnicas de agrupamiento.

Lo primero que se hace es agrupar los datos según el título de usuario para ver la cantidad de viajes que realiza cada tipo de usuario, tal y como se muestra a continuación.

TITULO_DESCRIPCION	
5 viajes líneas Metropolitanas	506
Abono Mensual Residente Canario	19193
Abono Niño menor de 10 años	1155
Abono Transporte Joven (Mensual - Insular)	25523
Abono Turístico 1 Día	911
Abono Turístico 1 Semana	344
Bono 65 12 euros.	2172
Bono Estudiante 15 euros	461
Bono de 15 euros. Se sustituye por Monedero	2330
Bono de 25 euros. Se sustituye por el Monedero	493
EMV billete sencillo. SaldoMaximo=1 => titulo EMV por defecto (maxPurseBalance)	149
Empleado TITSA Urbano	31
Empleados/Familiar TITSA Interurbano	143
Familia Numerosa Cabildo	531
Familiar Titsa Interurbano.	896
Familiar Titsa Urbano	7
IASS Baja Renta	6
Ida/Vuelta TITSA (válido hoy)	17
Jubilado de Titsa	65
Mensual Discapacitado General (Discapacidad >= 50%)	783
Mensual Mayor de 65 (SENIOR)	1508
Mensual Metropolitanas	758
Monedero General	35103
Personas de Movilidad Reducida (PMR) Cabildo	3210
Semanal Metropolitanas Joven	121
Sencillo	9791
Social Ayuntamiento Santa Cruz	487
Vuelta TITSA	18

Figura 3.5. Cantidad de viajes realizados por titulo de usuario

A continuación se muestran dos tablas con algunas medidas estadísticas que nos permiten hacer un análisis y estudio de la situación de partida. La primera tabla contiene información de las variables de recaudación, las cuales hacen referencia al aporte monetario proveniente directamente del usuario. La segunda tiene información de las variables de compensación, que suponen un ingreso indirecto cuando los usuario cumplen ciertos criterios.

Tabla de recaudación

	RECAUDACION_TOTAL_1	RECAUDACION_TOTAL_2	RECAUDACION_TOTAL_3	RECAUDACION_TOTAL_4	RECAUDACION_TOTAL
count	101943.000000	102262.000000	106172.000000	106686.000000	106712.000000
mean	1.008866	0.145104	0.014295	0.001029	1.118084
std	0.945635	0.444731	0.136134	0.039602	1.060047
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.576881	0.000000	0.000000	0.000000	0.576881
50%	0.615949	0.000000	0.000000	0.000000	0.750000
75%	1.150000	0.000000	0.000000	0.000000	1.250000
max	18.700000	8.550000	6.200000	5.200000	18.700000

Figura 3.6. estadísticas de las variables de recaudación

Podemos observar información referentes a la recaudación de cada etapa y la total, el número de estas que se corresponden con viajes de TITSA, el valor medio, mínimo y máximo que toman.

Tabla de compensación

	POLITICAS_CPI	POLITICAS_RESIDENTE	POLITICAS_OTRAS	COMPENSACION_TOTAL
count	106712.000000	106712.000000	106712.000000	106712.000000
mean	0.235415	0.439894	0.312410	0.987718
std	0.321326	1.073001	0.274587	0.972406
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.426651
50%	0.000000	0.000000	0.426651	0.980049
75%	0.553398	0.000000	0.426651	0.980049
max	2.213592	7.560000	1.706606	7.560000

Figura 3.7. estadísticas de las variables de compensación

Se muestra información idéntica a la anterior sobre los valores de compensación de los viajes, detallando los tres tipos de compensación que existen y la compensación total.

Una representación gráfica de algunas variables resultan de interés para conocer su distribución y sacar algunas conclusiones previas al agrupamiento.

En la primera gráfica se representan los datos por valor monetario. En el eje X representamos los valores de recaudación (valor directo) mientras que en el eje Y tenemos los valores de compensación (valor indirecto). Conocer la distribución de estos datos resulta útil ya que son algunos de los que utilizará el algoritmo para agrupar. Por otro lado, la segunda gráfica tridimensional representa la recaudación y compensación en los ejes X e Y respectivamente, añadiendo el número de transbordos en el eje Z, otra variable de interés para el agrupamiento. Esta gráfica muestra información interesante de cómo se comportan estos datos en conjunto.

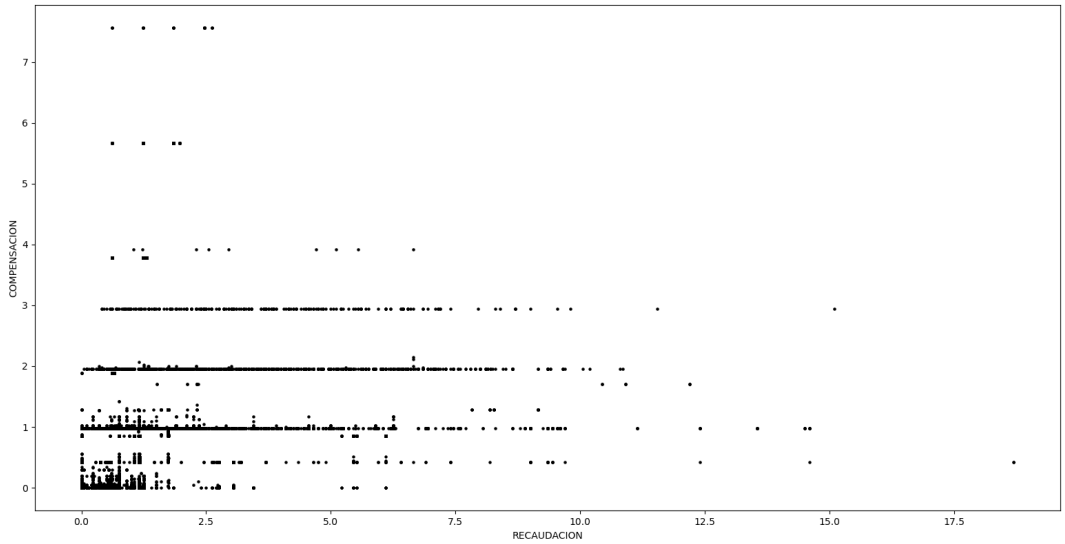


Figura 3.8. Gráfica de recaudación y compensación

Se puede observar que la mayoría de valores se concentran en las cercanías del (0,0) y en tres rectas de puntos que dejan fijo el valor de compensación en los valores 1, 2 y 3 mientras varían los valores de recaudación. Por otro lado se observan algunos puntos alejados que representan valores altos de recaudación o compensación.

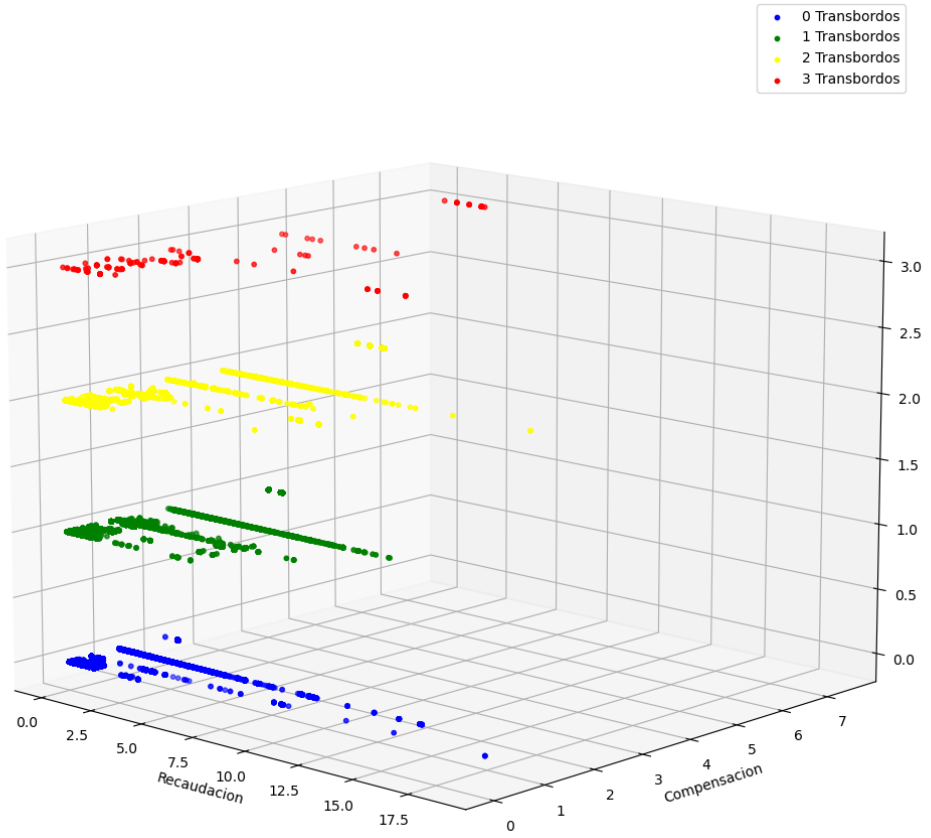


Figura 3.9. Gráfica de recaudación, compensación y transbordo

En esta gráfica podemos ver que contrariamente a lo que se podría esperar, un mayor número de transbordos y por tanto un mayor número de etapas o líneas utilizadas por viaje no implica un mayor valor de recaudación. Sin embargo sí que se puede observar que a medida que aumenta el número de transbordos también aumentan los valores que toma la compensación.

3.3. Obtención del número de agrupamientos

A continuación se proceda a encontrar el número óptimo de agrupamiento, el valor de K. Esto es necesario previamente para que los resultado al aplicar la técnica K-medias nos dé buenos resultados. Para encontrar dicho valor se utiliza la curva de codo que se mencionó anterior capítulo. Se trata de encontrar mediante la gráfica el punto de codo en el que disminuye la pendiente y que indicará un buen valor de k.

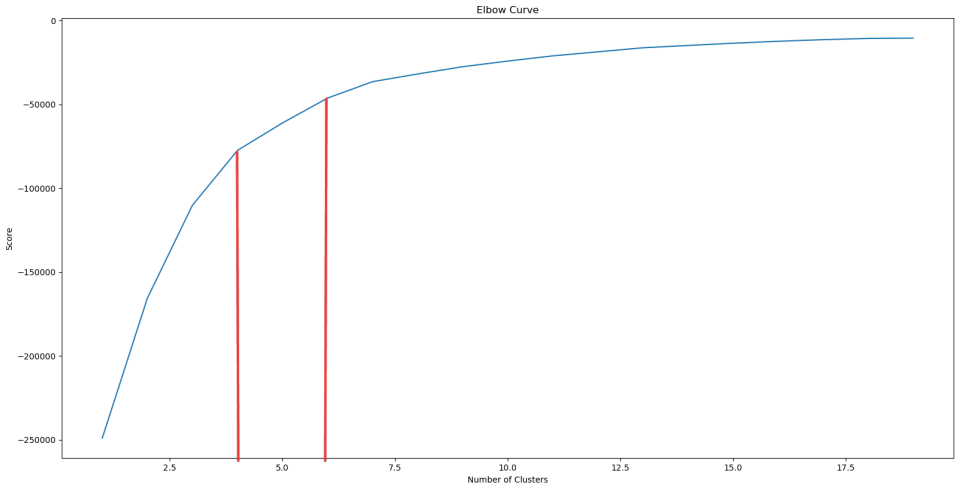


Figura 3.10. Gráfica de la curva de codo

La curva tiene un crecimiento bastante suave. Fijándonos en los puntos de codo de la curva el k=6 puede ser un buen valor, al igual que k=4. Con el valor k=4, al tener un menor número de clusters es más fácil su interpretación. Se ejecutará el algoritmo con ambos valores para comparar resultados.

3.4. Aplicación del algoritmo de K-medias

Una vez los datos preparados y encontrados dos buenos valores para k, se aplica el algoritmo de k-medias sobre el conjunto de datos de la experimentación.

A continuación se muestran las representaciones gráficas de los datos ya segmentados. Se preseentan las mismas gráficas que anteriormente pero diferenciando los datos de cada grupo. En primer lugar se muestran las imágenes de los datos agrupados en 6 grupos y en segundo lugar se muestran los resultados del agrupamiento en 4 grupos.

Representaciones con K=6

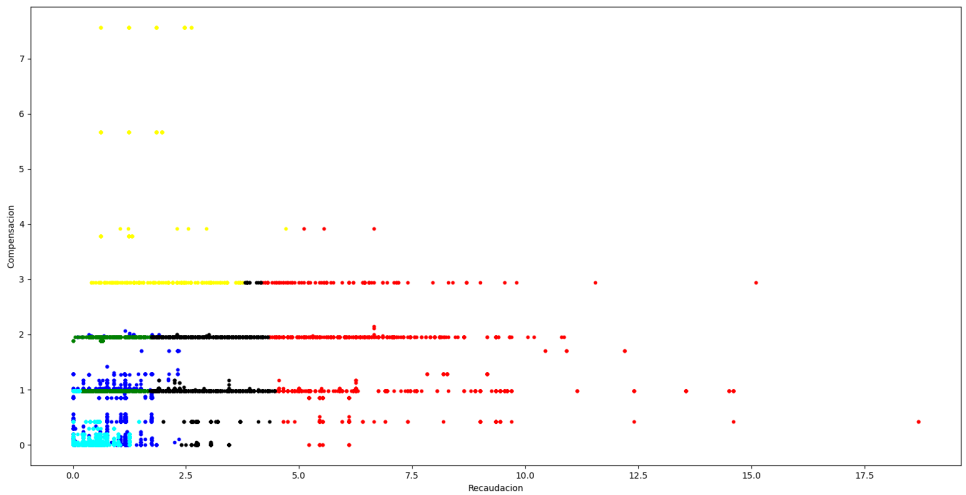


Figura 3.11. Datos de recaudación y compensación divididos en 6 grupos

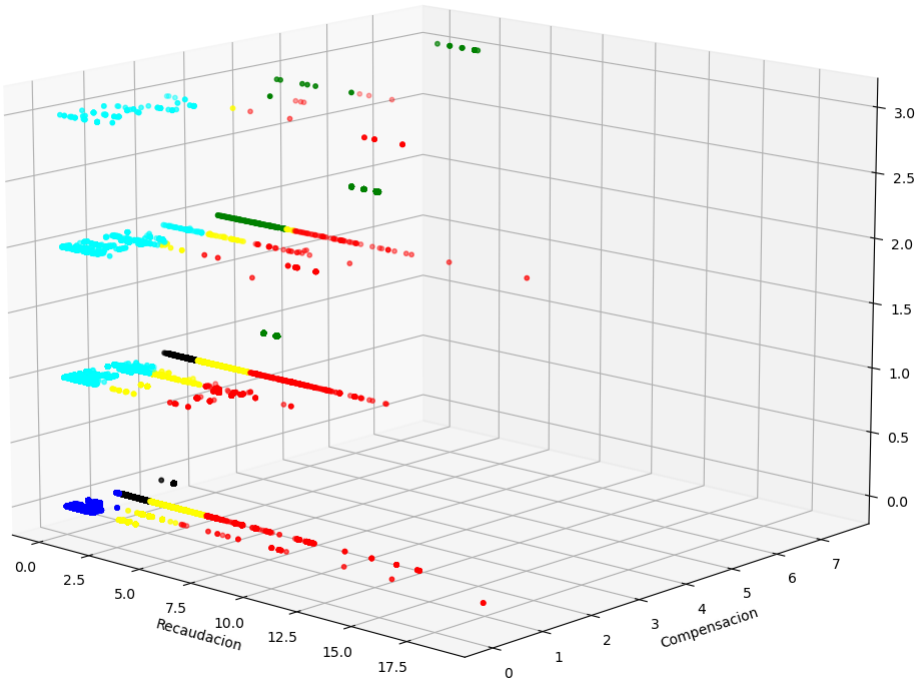


Figura 3.12. Representación tridimensional diferenciando en los 6 grupos

Representaciones con K=4

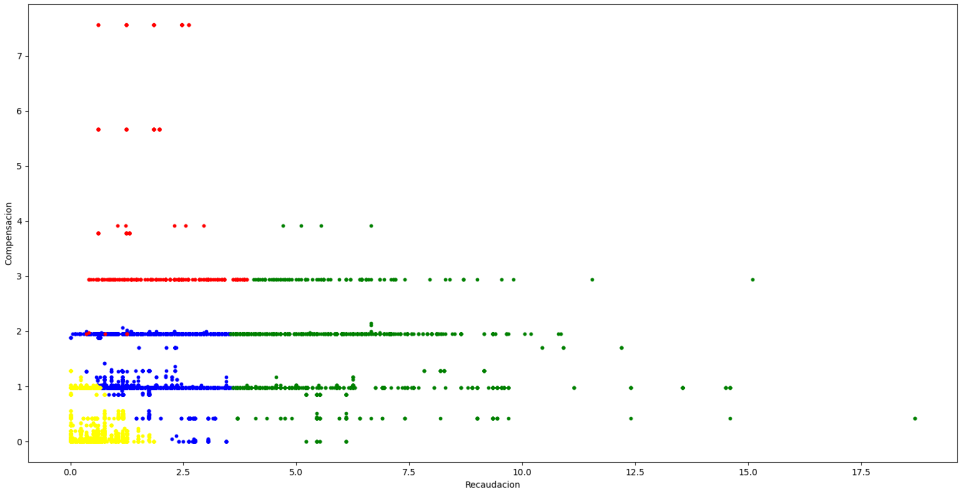


Figura 3.13. Datos de recaudación y compensación divididos en 4 grupos

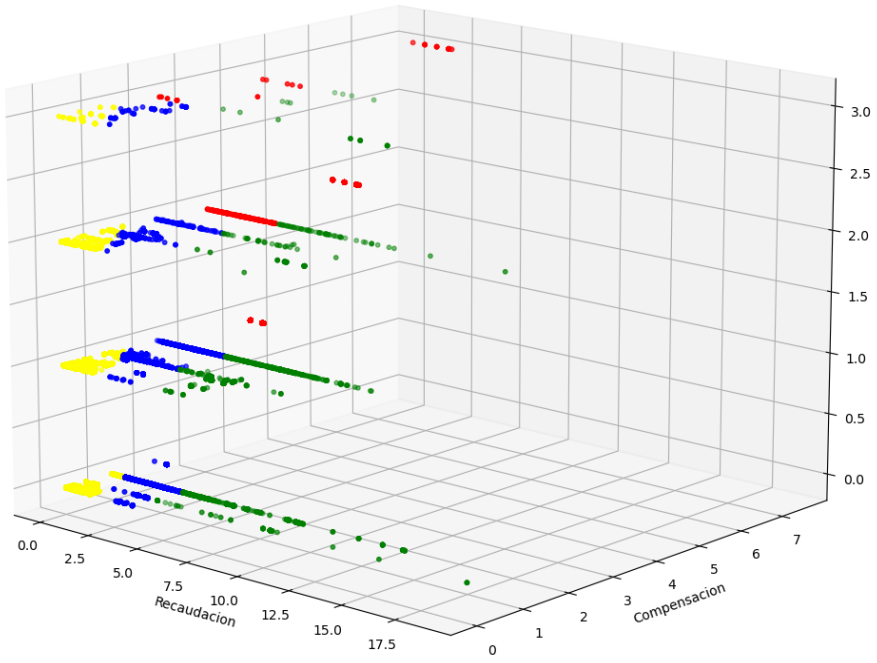


Figura 3.14. Representación tridimensional diferenciando en los 4 grupos

Fijándose en las representaciones anteriores se puede observar que en el caso en el que se dividen los datos en seis grupos (K=6), los grupos se entremezclan

más, y parecen más complejos de interpretar y analizar los distintos grupos. En el caso de 4 grupos, parece que hay una división más clara e interpretable, es por esto que se va a estudiar esta agrupación, los diferentes agrupamientos y encontrar si existen variables dominantes en alguno o patrones que puedan ser de interés.

En una primera observación de los datos ya agrupados, una posible lectura de los distintos agrupamiento sería:

- **Grupo 1** (amarillo)

  En este grupo se concentran los datos con valores de recaudación y compensación más cercanos al '0'. En este segmento se concentran los usuarios con un bajo valor desde un punto de vista monetario. Podría ser útil para encontrar los títulos de usuario o las líneas con menor rentabilidad.

- **Grupo 2** (azul)

  Esta región parece tener datos que aportan mayor valor monetarios que la anterior. Estos datos parecen mantener cierto equilibrio entre el valor de recaudación y compensación. Parece que se distribuyen de forma independiente al número de transbordos al igual que la región anterior.

- **Grupo 3** (verde)

  Los datos de este grupo parecen ser aquellos con mayores valores de recaudación y se extienden en la dirección en la que está aumenta. Aunque se encuentran algunos puntos en las posiciones que indican mayor número de transbordos, los que toman mayores valores de recaudación parecen estar en zonas que indican un o ningún transbordo.

- **Grupo 4** (rojo)

  Este grupo parece concentrar los datos que tienen mayores valores de compensación. Estos se concentran a medida que el número de transbordos aumenta, de hecho no hay ningún dato en este grupo con 0 transbordos y hay muy pocos puntos con valor de un transbordo.

3.5. Análisis de los resultados

A continuación vamos a realizar un análisis de los distintos segmentos para ver qué información de interés podemos obtener. Para esto se van a mostrar algunas visualizaciones y algunas medidas estadísticas de los datos de cada grupo y posteriormente profundizaremos en los detalles de las agrupaciones.

Se muestran dos visualizaciones realizadas con Power BI. En la primera podemos ver los títulos de usuario que componen cada grupo en forma de porcentajes. En la segunda se muestran las líneas más utilizadas por los usuarios de cada grupo.

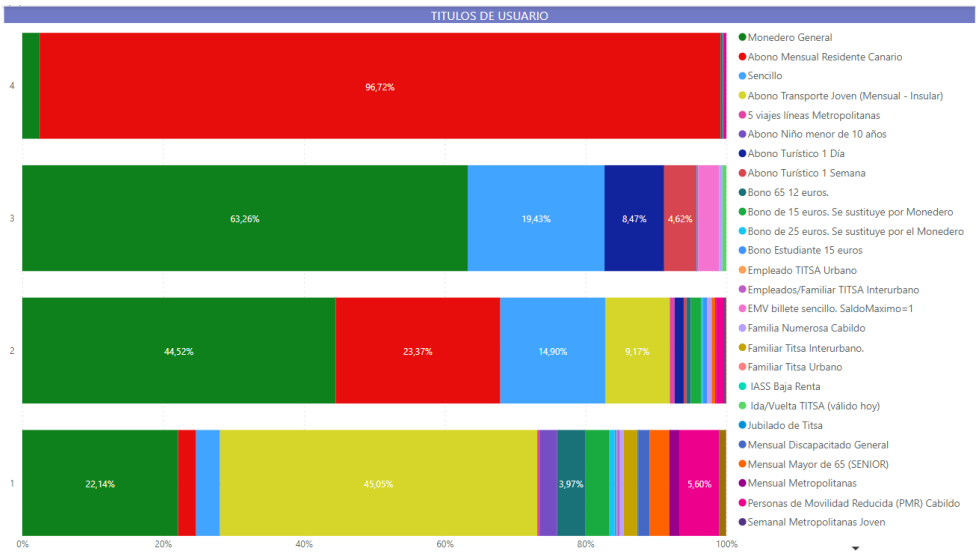


Figura 3.15. Títulos de usuario por grupo

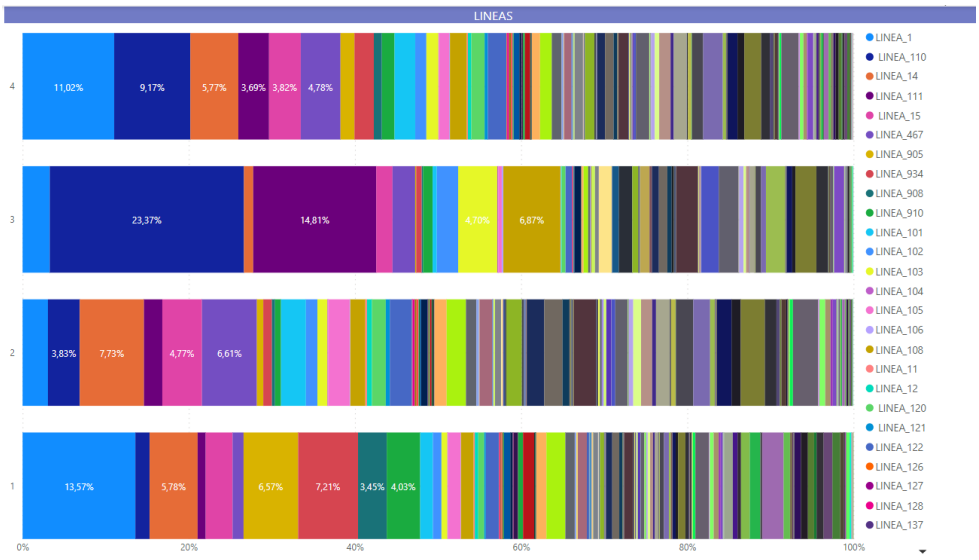


Figura 3.16. Lineas utilizadas por grupo

- **Grupo 1**

```
TITULO_DESCRIPCION
5 viajes líneas Metropolitanas                          152
Abono Mensual Residente Canario                        1178
Abono Niño menor de 10 años                            1155
Abono Transporte Joven (Mensual - Insular)            20819
Bono 65 12 euros.                                      1834
Bono Estudiante 15 euros                                136
Bono de 15 euros. Se sustituye por Monedero            1602
Bono de 25 euros. Se sustituye por el Monedero          343
Empleado TITSA Urbano                                    31
Empleados/Familiar TITSA Interurbano                    143
Familia Numerosa Cabildo                                265
Familiar Titsa Interurbano.                             896
Familiar Titsa Urbano                                     7
Jubilado de Titsa                                        65
Mensual Discapacitado General (Discapacidad >= 50%)     756
Mensual Mayor de 65 (SENIOR)                           1274
Mensual Metropolitanas                                  699
Monedero General                                      10230
Personas de Movilidad Reducida (PMR) Cabildo           2590
Semanal Metropolitanas Joven                             22
Sencillo                                               1570
Social Ayuntamiento Santa Cruz                          424
Vuelta TITSA                                             18
dtype: int64
       RECAUDACION_TOTAL  COMPENSACION_TOTAL   TRANSBORDOS
count       46209.000000       46209.000000  46209.000000
mean            0.579926           0.286769      0.221342
std             0.257232           0.300616      0.451081
min             0.000000           0.000000      0.000000
25%             0.576881           0.000000      0.000000
50%             0.576881           0.426651      0.000000
75%             0.750000           0.426651      0.000000
max             1.847848           1.279954      3.000000
```

Figura 3.17. Datos del grupo 1

En el grupo 1 tenemos 46.210 datos, es el segundo grupo más grande. Recordamos que tal y como se vio anteriormente este es el grupo que tiene los viajes con menor rentabilidad monetaria. En las estadísticas se observa que más del 75 % de los viajes de este grupo tienen 0 transbordos, es decir que en su mayoría son usuarios que realizan viajes de una etapa. Las líneas más utilizadas por los usuarios de este grupo son las líneas 14, 15, 905, 908, 910, 934 y la línea 1 que es externa a TITSA.

Este agrupamiento contiene la totalidad de los usuarios de determinados títulos, como los empleados, familiares y jubilados de TITSA, vuelta TITSA y niños menores de 10 años, estos títulos tienen en común que su valor de recaudación es 0 y tienen bajo valor de compensación. Varios de los títulos con características similares a los anteriores tienen muchos de sus usuarios en este grupo, como son:los abonos mensuales de discapacitados, mayores de 65, joven y metropolitanas, los abonos para PMR o los abonos del ayuntamiento. Tambien estan la mayoria de usuarios que utilizan bonos de 65, 25 y 15 euros y bastantes de los usuarios que pagan en efectivo, estos principalmente realizan viajes de una etapa que aportan bajo valor monetario, aunque algunos

realizan viajes de varias etapas que tienen un valor ligeramente mayor, los cuales suponen los valores más altos de este grupo.

- **Grupo 2**

```
TITULO_DESCRIPCION
5 viajes líneas Metropolitanas                                                    354
Abono Mensual Residente Canario                                                 11994
Abono Transporte Joven (Mensual - Insular)                                       4704
Abono Turístico 1 Día                                                             660
Abono Turístico 1 Semana                                                          207
Bono 65 12 euros.                                                                 318
Bono Estudiante 15 euros                                                          318
Bono de 15 euros. Se sustituye por Monedero                                       728
Bono de 25 euros. Se sustituye por el Monedero                                    150
EMV billete sencillo. SaldoMaximo=1 => titulo EMV por defecto (maxPurseBalance)    61
Familia Numerosa Cabildo                                                          250
IASS Baja Renta                                                                     6
Mensual Discapacitado General (Discapacidad >= 50%)                                27
Mensual Mayor de 65 (SENIOR)                                                      234
Mensual Metropolitanas                                                             59
Monedero General                                                                22845
Personas de Movilidad Reducida (PMR) Cabildo                                      592
Semanal Metropolitanas Joven                                                       99
Sencillo                                                                         7645
Social Ayuntamiento Santa Cruz                                                     63
dtype: int64
       RECAUDACION_TOTAL  COMPENSACION_TOTAL   TRANSBORDOS
count       51314.000000        51314.000000  51314.000000
mean            1.311437            1.238722      0.228534
std             0.663395            0.447552      0.471902
min             0.000000            0.000000      0.000000
25%             0.750000            0.980049      0.000000
50%             1.153762            0.980049      0.000000
75%             1.450000            1.890000      0.000000
max             3.600000            2.069865      3.000000
```

Figura 3.18. Datos del grupo 2

El segundo es el mayor de los grupos, contiene 51315 elementos de los 106713 totales. Encontramos que los viajes en este segmento tienen un valor monetario mayor que el del grupo anterior. Tal y como y se reflejaba en las medias de las variables recaudación y compensación. Al igual que el anterior este también tiene algo más del 75 % de viajes de una sola etapa y las líneas más utilizadas en este caso son las líneas 14, 15, 110, 467.

Está compuesto principalmente por usuarios que pagan en efectivo o con billetes sencillos, los cuales se reúnen mayormente en este grupo, usuarios con abono mensual de residente canario y con abono joven que también conforman un porcentaje notable del grupo. Además también contiene la mayoría de usuarios con abonos turísticos, bonos de estudiante y bonos para líneas metropolitanas.

- **Grupo 3**

```
TITULO_DESCRIPCION
Abono Turístico 1 Día                                                    251
Abono Turístico 1 Semana                                                 137
Bono Estudiante 15 euros                                                   6
EMV billete sencillo. SaldoMaximo=1 => titulo EMV por defecto (maxPurseBalance)    88
Familia Numerosa Cabildo                                                  14
Ida/Vuelta TITSA (válido hoy)                                             17
Monedero General                                                        1875
Sencillo                                                                 576
dtype: int64
       RECAUDACION_TOTAL  COMPENSACION_TOTAL  TRANSBORDOS
count        2964.000000         2964.000000  2964.000000
mean            5.922306            1.253108     0.542848
std             1.789349            0.560244     0.628268
min             3.550000            0.000000     0.000000
25%             4.650000            0.980049     0.000000
50%             5.900000            0.980049     0.000000
75%             6.350000            1.960099     1.000000
max            18.700000            3.920198     3.000000
```

Figura 3.19. Datos del grupo 3

Este tercer grupo es el más pequeño de los agrupamientos obtenidos, cuenta con 2965 elementos. Tiene valores de compensación relativamente bajos, similares a los del grupo anterior. Por otro lado, encontramos que los valores de recaudación aumentan notablemente, de hecho los viajes con mayores valores de recaudación se encuentran en este grupo.

Encontramos un mayor número de transbordos, los viajes con varias etapas suponen casi la mitad. Predominan las líneas 110, 111, 103 y 108. Los títulos de los usuarios más presentes son los que utilizan billetes sencillos o pagan en efectivo, seguidos de usuarios con abonos turísticos y usuarios que realizan el pago con tarjeta (EMV).

- **Grupo 4**

```
TITULO_DESCRIPCION
Abono Mensual Residente Canario                    6021
Bono 65 12 euros.                                    20
Bono Estudiante 15 euros                              1
Familia Numerosa Cabildo                              2
Monedero General                                    153
Personas de Movilidad Reducida (PMR) Cabildo         28
dtype: int64
       RECAUDACION_TOTAL  COMPENSACION_TOTAL  TRANSBORDOS
count        6225.000000         6225.000000  6225.000000
mean            1.231552            3.995518     1.163373
std             0.399288            0.700805     0.394135
min             0.350000            1.960099     1.000000
25%             1.231899            3.780000     1.000000
50%             1.231899            3.780000     1.000000
75%             1.231899            3.780000     1.000000
max             3.900000            7.560000     3.000000
```

Figura 3.20. Datos del grupo 4

Encontramos 6226 elementos en este último grupo. Los valores de recaudación son relativamente bajos mientras que los valores de compensación son los mayores de entre todos los datos. En este caso todos los viajes tienen al menos un transbordo, en particular los viajes de dos etapas son los más abundantes.

Las más utilizadas por estos usuarios son las líneas 14, 15, 110, 111, 467 y 1, siendo esta última externa a TITSA. Los usuarios con abono mensual de residente canario conforman prácticamente en su totalidad este grupo.

Conclusiones

El proyecto tiene como propósito la aplicación de técnicas de agrupamiento para realizar la segmentación de clientes de la empresa de transporte público de pasajeros de Tenerife, TITSA. El objetivo ha sido identificar grupos de clientes en función de su comportamiento y su rentabilidad. De este modo, TITSA puede adaptar sus estrategias de marketing y mejorar la captación de clientes.

La obtención de segmentos para una empresas de transporte utilizando técnicas de agrupamiento no ha sido un proceso sencillo, depende de la disponibilidad de datos, de ahí la importancia de la cultura del dato en la organización. Además los datos disponibles deben estar bien estructurados. Por otro lado, la preparación y depuración de los datos son dos tareas que ocupan una gran cantidad de tiempo. La obtención de resultados de calidad que aporten valor para la segmentación de clientes de la compañía es totalmente dependiente de estas dos tareas.

El trabajo ha estado centrado en la aplicación de la técnica de agrupamiento K-medias. Para su aplicación ha sido necesario el uso del método del codo, para la obtención del número óptimo de agrupaciones para el estudio. La experimentación se ha desarrollado concretamente con cuatro agrupaciones, así como el análisis de los resultados. Sobre estos grupos se ha realizado un análisis haciendo uso de algunos estadísticos, así como la obtención de gráficas para facilitar la visualización de información y la extración de conocimientos sobre los grupos, identificando las características y el tipo de usuario encontrado en cada segmento.

El sector del transporte tiene un gran interés en la aplicación de técnicas de analítica de datos. Es un campo reconocido para el desarrollo profesional de los matemáticos. En concreto en la empresa TITSA, tienen grandes perspectivas en el departamento de nueva creación de ciencia de datos dentro de la empresas, con su capacidad para abordar y resolver problemas y aportar conocimiento para mejorar la empresa.

Como trabajos futuros se proponen los siguientes:

- la ampliación del análisis utilizando un conjunto de datos para la experimentación más amplio, tal y como se comentó. Un número mayor de datos disponibles y un mayor número de variables que recogan aspectos de interés complementario como los aspectos de duración de los viajes.
- la implementación de otros agrupamientos auxiliares, como podría ser un agrupamiento jerárquico, que en combinación con lo realizado en este proyecto podría dar una visión más global de los comportamientos de los usuarios.
- la aplicación de otras técnicas de inteligencia artificial asociadas con la analítica de datos para determinar patrones de comportamiento y hacer algunas previsión futuras asociadas con comportamiento de clientes, previsión de ingresos y gastos, rentabilidad, necesidades de equipamientos y nuevas líneas. Más aún con los cambios que se están dando, con la reducción de tarifas y el crecimiento de viajeros.

Para finalizar comentar algunos aspectos relacionados con los resultados del aprendizaje, una vez desarrollado el trabajo. A lo largo de este trabajo se han aplicado conocimientos adquiridos durante el grado, principalmente los relacionados con la informática, estadística y modelización. El trabajo ha permitido adquirir nuevos conocimientos sobre los modelos de segmentación, las técnicas de agrupamientos y el tratamiento de los datos. Además, de estos conocimientos, durante la realización del trabajo se han adquirido habilidades en el uso de herramientas de visualización como Power BI y ampliado las de programación en Python, con el uso de diferentes librerías y funciones avanzadas.

Bibliografía

[1] Slotnisky, D. (2016). Transformación digital: cómo las empresas y los profesionales deben adaptarse a esta revolución. Digital House. Coding School.

[2] Kotler, Philip.; ARMSTRONG, Gary; ANG, Swee Hoon; LEONG, Siew Meng; TAN, Chin Tiong; and YAU, Oliver. Principles of marketing: An global perspective. (2008). Research Collection Lee Kong Chian School Of Business.

[3] Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of marketing, 21(1), 3-8.

[4] Monferrer Tirado, D. (2013). Fundamentos de marketing.

[5] Kahreh, M. S., Tive, M., Babania, A., and Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia-Social and Behavioral Sciences, 109, 590-594.

[6] Baehre, S., O'Dwyer, M., O'Malley, L., and Lee, N. (2022). The use of Net Promoter Score (NPS) to predict sales growth: insights from an empirical investigation. Journal of the Academy of Marketing Science, 50(1), 67-84.

[7] Wu, J., and Lin, Z. (2005, August). Research on customer segmentation model by clustering. In Proceedings of the 7th international conference on Electronic commerce (pp. 316-318).

[8] Trivedi, S., Pardos, Z. A., and Heffernan, N. T. (2015). The utility of clustering in prediction tasks. arXiv preprint arXiv:1509.06163.

[9] Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

[10] Yih, W. T., and Meek, C. (2007, July). Improving similarity measures for short segments of text. In AAAI (Vol. 7, No. 7, pp. 1489-1494).

[11] Kaufman, L., and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley and Sons.

[12] Alberca, A. S. (2018). Una nueva taxonomía de colecciones y de funciones de similitud para su comparación. Pensamiento Matemático, 8(2), 1.

[13] Xu, D., and Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.

[14] Boley, D., Gini, M., Gross, R., Han, E. H. S., Hastings, K., Karypis, G., ... and Moore, J. (1999). Partitioning-based clustering for web document categorization. Decision Support Systems, 27(3), 329-341.

[15] MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297).

[16] Ng, R. T., and Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016.

[17] Kriegel, H. P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(3), 231-240.

[18] Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[19] Duda, R. O., and Hart, P. E. (1973). Pattern classification and scene analysis (Vol. 3, pp. 731-739). New York: Wiley.

[20] Karypis, G., Han, E. H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75.

[21] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. ACM sigmod record, 25(2), 103-114.

[22] Pilevar, A. H., and Sukumar, M. (2005). GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern recognition letters, 26(7), 999-1010.

[23] Meila, M., and Heckerman, D. (2001). An experimental comparison of model-based clustering methods. Machine learning, 42(1), 9-29.

Clustering models applied to the knowledge of the users of public passenger transportation

Helí Alonso Afonso
Facultad de Ciencias • Sección de Matemáticas
Universidad de La Laguna
alu0100985245@ull.edu.es

Abstract

THE project involves using clustering techniques to perform customer segmentation for TITSA, a public passenger transport company in Tenerife. The objective is to identify groups of customers based on their behavior or preferences. So, TITSA can tailor its marketing strategies and improve customer acquisition. The project studies different segmentation models commonly used by companies and investigates various clustering techniques to identify the most suitable algorithm for our purpose. The clustering results will be analyzed to gain insights into customer behaviors, which can be used for decision-making within TITSA.

1. segmentation and clustering

WE want to obtain a customer segmentation for the public passenger transport company TITSA. For this purpose, we first study segmentation models most commonly used by companies to determine which model interests the company.

In order to determine such segmentation, we will use clustering, which is used in machine learning to group similar data points together.

Between the different clustering algorithms, we decided to use K-means clustering, which is one of the most commonly used and simple algorithms; it partitions the data into k clusters based on the mean distance between the data points.

The experimentation dataset contains the following structured data: customer identification (card number, title and user profile) and trip stages performed. Each trip comprises four stages (date, time, entry stop, exit stop, line used, number of passengers and stage revenue) and company compensation policies (indirect monetary revenue from the trip meeting certain conditions).

We preprocess the data, which involves cleaning, transforming and preparing the data for clustering. K-means is required to specify the number of clusters (k) beforehand. To find the optimal number of clusters the elbow method is used.

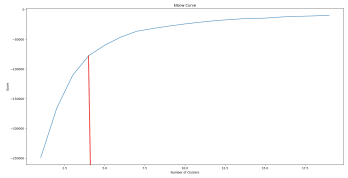


Figure 1: *Elbow curve*

As shown above, the optimal number of clusters for the algorithm is four (k=4).

2. Experimentation results

THE k-means algorithm has been applied to the data, and four groups have been obtained, as shown in the following figures

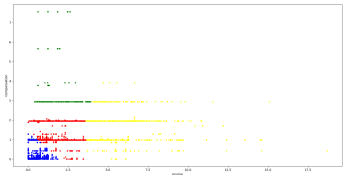


Figure 2: *income and compensation graph*

Here is a representation of monetary value data (Income and compensation) separated by clusters.

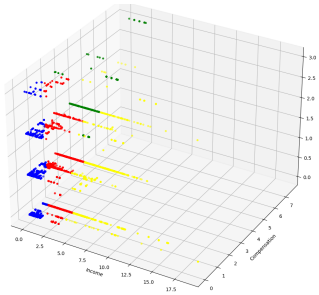


Figure 3: *three-dimensional graph of clusters*

And a tridimensional representation of clustered data adding transshipments.

The four groups obtained are studied using visualizations and Statistics to explore each in-depth and extract information.

References

[1] Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing, 21(1), 3-8.

[2] MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297).