# Contexto del Proyecto

Hace casi cuatro años, easyMoney fue fundada por Carol Denver, una profesional consolidada en el sector de la banca de inversión. Tras más de una década trabajando para grandes firmas, Carol decidió emprender su propio proyecto: una plataforma multicanal enfocada en la comercialización de productos financieros (ahorro, inversión y financiación). El objetivo de easyMoney era ofrecer una interfaz amigable y accesible, donde los clientes pudieran encontrar soluciones financieras personalizadas y contratarlas de forma sencilla.

El primer producto lanzado fue la cuenta easyMoney, un servicio innovador que permite a los usuarios acumular dinero automáticamente mediante el redondeo de sus compras. Este producto fue un éxito rotundo, lo que impulsó la ampliación de su cartera de soluciones, incluyendo productos de inversión y tarjetas, entre otros. Para facilitar su operación, y debido a la falta de licencia bancaria propia, easyMoney estableció una alianza con easyBanking S.A., una entidad regulada por el Banco de España que actúa como proveedor de productos financieros. Sin embargo, la fuerte influencia de easyBanking en la estrategia comercial ha generado tensiones en la relación entre ambas empresas.

Gracias a la perseverancia y red de contactos de Carol, easyMoney logró superar sus objetivos iniciales tras dos exitosas rondas de financiación. Esto permitió a la empresa captar un volumen significativo de clientes durante su primer año de operación, alcanzando una facturación millonaria y una plantilla cercana a las 100 personas. No obstante, tras cuatro años de actividad, la empresa enfrenta varios desafíos que amenazan su sostenibilidad.

Entre los principales retos se encuentran la creciente presión de easyBanking para incorporar productos complejos, lo que ha desviado a la empresa de su visión original de ofrecer soluciones sencillas y enfocadas en las necesidades del cliente. Además, los fondos obtenidos en las rondas de financiación están cerca de agotarse y, a pesar de su crecimiento, easyMoney no ha alcanzado aún un EBITDA positivo que le permita operar de manera independiente. Su principal inversor, Lion Global Management, ha demandado una mayor rentabilidad de la base de clientes actual antes de comprometer más capital.

Asimismo, la alta rotación en el equipo de tecnología y la falta de inversión en infraestructura tecnológica están generando ineficiencias que afectan a diversas áreas de la empresa. Las tensiones internas han comenzado a erosionar el espíritu ágil que caracterizó los primeros años de desarrollo. Ante este panorama, y tras la salida de un miembro clave del equipo de marketing, la dirección de easyMoney ha decidido contratar a un Data Scientist. Este profesional tendrá la misión de ayudar a la empresa a mejorar la rentabilidad de su cartera de clientes, abordando los retos actuales y acompañando a easyMoney en su nueva etapa de crecimiento.

# Preprocesamiento y Análisis

El preprocesamiento de datos es un paso fundamental en cualquier proyecto de **Data Science**. Consiste en preparar y limpiar los datos antes de su análisis, asegurando que la información con la que se trabajará sea precisa, completa y consistente. En el caso de *easyMoney*, el dataset proporcionado está dividido en tres secciones principales que agrupan distintos tipos de información:

1. **Características sociodemográficas:** Incluye variables relacionadas con los clientes, como su edad, país de residencia, género, provincia, y salario.
2. **Productos contratados:** Recoge información sobre los distintos productos financieros que cada cliente ha adquirido, como depósitos, tarjetas, y préstamos.
3. **Actividad comercial:** Proporciona datos tales como el segmento comercial de los clientes, canal de entrada, estado en la aplicación de EasyMoney y primera fecha de contratación.

Cada uno de estos conjuntos de datos contiene información valiosa y complementaria que, en conjunto, permitirá a *easyMoney* mejorar la segmentación de clientes y definir estrategias personalizadas. Sin embargo, antes de realizar cualquier tipo de análisis avanzado, es crucial llevar a cabo un preprocesamiento que permita:

* **Identificar y corregir errores:** Como datos faltantes, duplicados o inconsistentes.
* **Establecer la coherencia temporal:** Alineando las fechas y la evolución de las variables a lo largo del tiempo.
* **Transformar y estandarizar:** Adaptar las variables categóricas y numéricas a un formato adecuado para los modelos de machine learning.

El análisis de los datos comienza con una limpieza exhaustiva y una exploración preliminar. Este paso inicial no solo facilita la comprensión del estado actual de la base de datos, sino que también define el camino hacia la implementación de los modelos predictivos y el descubrimiento de patrones clave.

## Sociodemographic DataBase

En esta primera parte del análisis, nos enfocamos en la sección de **características sociodemográficas** del dataset de *easyMoney*, que incluye información clave sobre los clientes, como su edad, país de residencia, género, región y salario. El objetivo de este apartado es realizar un preprocesamiento adecuado de estos datos, asegurando la calidad de la información antes de llevar a cabo análisis más detallados.

### 1.1.1. Carga y Ajuste del Dataset

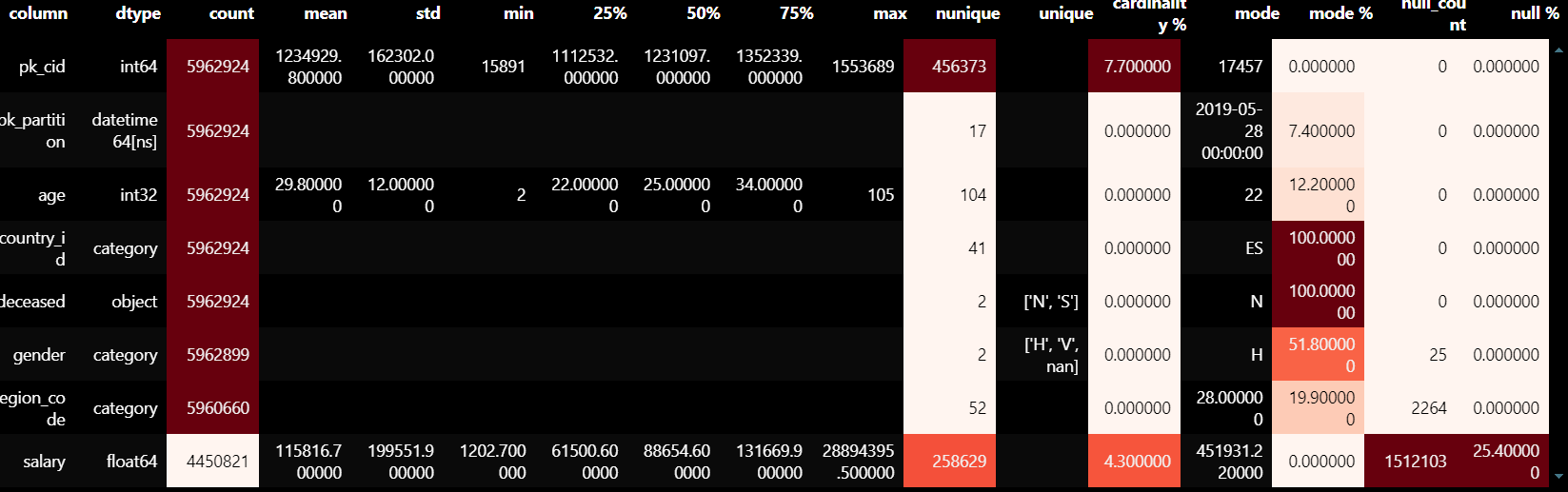
El archivo fue cargado y posteriormente se realizaron algunas limpiezas y ajustes iniciales:

* **Eliminación de columnas innecesarias:** Se eliminaron columnas que no aportaban información relevante, como 'Unnamed: 0\_x', 'Unnamed: 0\_y', y otras que eran resultado de procesos previos o simplemente no contribuían al análisis.
* **Cambio de tipo de datos:** Se ajustaron los tipos de datos de varias columnas para asegurar una correcta manipulación posterior. Por ejemplo, se convirtieron fechas a datetime, algunas columnas a categoría, y variables numéricas a int32.

### 1.1.2. Exploración Preliminar

Se realizó un análisis preliminar que proporciona una visión detallada de las principales características de cada columna del dataset, incluyendo estadísticas descriptivas, número de valores únicos, porcentaje de valores nulos, entre otros.

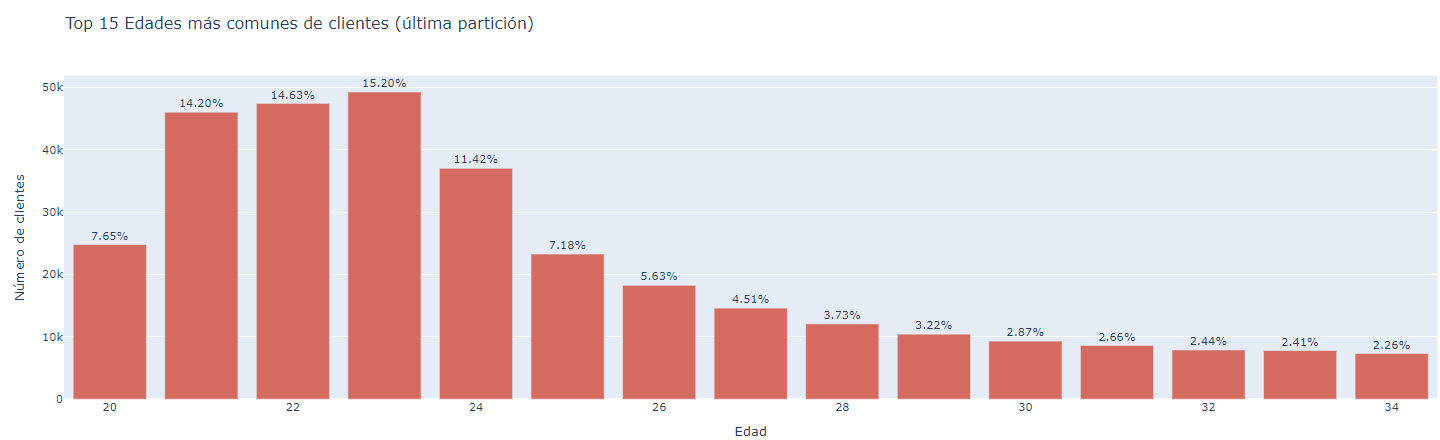
* **Nulos:** Se identificaron valores nulos en las variables **gender** (25 casos), **region\_code** (2,264 casos) y **salary** (más de 1.5 millones de casos), lo cual requerirá un tratamiento posterior.
* **Country y Deceased:** Se observó que la mayoría de los clientes residen en España y casi ningún cliente está marcado como fallecido, por lo que estas variables presentan poca variabilidad.



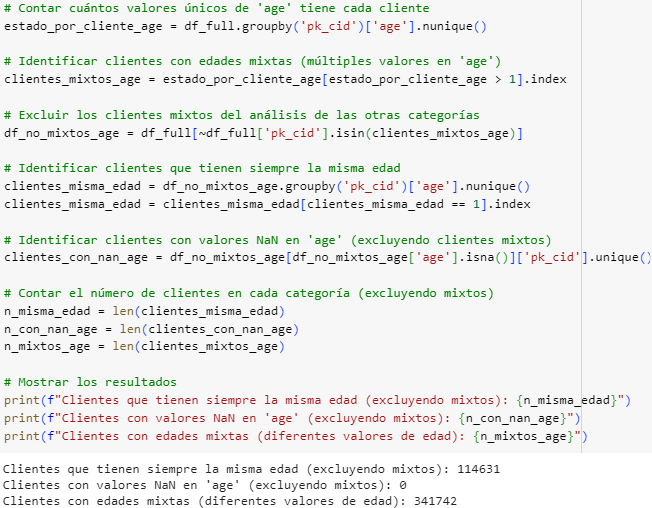
### 1.1.3. Análisis de la Edad

La variable **age** es una de las más importantes en el análisis de clientes. Se realizó un análisis detallado de su distribución:

* **Distribución de edades:** Se analizaron las edades más comunes de los clientes en la partición más reciente (mayo de 2019). Un gráfico interactivo mostró las 15 edades más frecuentes, con una **concentración significativa en torno a los 30 años**.



* **Clientes con edades mixtas:** Se detectaron **341,742 clientes** con múltiples valores de edad, lo que indica inconsistencias en la entrada de datos. Estos casos serán analizados y corregidos en las siguientes etapas.



Al analizar la variable **edad** en el dataset, se detectaron dos tipos de anomalías: clientes que se mantienen con una **edad constante**, siendo menores de edad, y aquellos con **edades mixtas**, donde se observan cambios incoherentes en la edad reportada a lo largo del tiempo. A continuación, se describe cómo se abordaron ambas situaciones.

**Clientes Menores**

En este caso, se identificaron clientes cuya edad es **menor a 18 años** y que mantienen la misma edad a lo largo del tiempo. Esto plantea un posible problema, ya que estos clientes, por su edad, no deberían tener acceso a ciertos productos financieros, lo cual podría ser indicativo de errores en los registros. Se procedió con el siguiente análisis:

* **Filtro de clientes menores de 18 años:** Se seleccionaron aquellos clientes cuya edad es inferior a los 18 años, obteniendo un resultado de 7368 clientes.
* **Análisis de productos activos:** Se evaluó cuántos de estos clientes tenían productos financieros contratados. De los productos disponibles, se detectaron muy pocos casos, 2 clientes con **nómina**, 9 con **planes de pensiones** y 3 con **cuenta EasyMoney.**
* **Clientes activos:** Se evaluó cuántos de estos menores eran marcados como **active\_customer**, es decir, que habían realizado alguna transacción reciente. El análisis reveló un número pequeño de casos activos.

Dado que el volumen de usuarios menores de edad no representa una cantidad relevante en comparación al volumen del dataset completo, y considerando que es posible que estas cuentas estén siendo gestionadas por los padres como cuentas de ahorro para el futuro de los menores, se ha decidido mantener estos registros en el análisis.

**Edad Mixta**

La segunda anomalía encontrada se refiere a los clientes con **edades mixtas**, es decir, aquellos cuyos registros muestran varias edades distintas a lo largo del tiempo, lo cual podría deberse a errores de tipificación o de captura en el sistema. Esta situación se aborda desde dos perspectivas:

**a) Variaciones Mayores a 3 Años**

Para detectar errores en la tipificación de la edad, se analizó la diferencia entre la edad más reciente y la más antigua registrada para cada cliente. Dado que los datos cubren un periodo de 3 años, se estableció un **umbral de 3 años** como la variación máxima aceptable. Los resultados mostraron que **1,093 clientes** tenían una diferencia mayor a 3 años, con una variación promedio de **25 años**, lo cual es claramente incoherente.

Se decidió **corregir estas inconsistencias** utilizando la última edad registrada para cada cliente, bajo la premisa de que las entradas más recientes son más fiables.

**b) Variaciones Negativas**

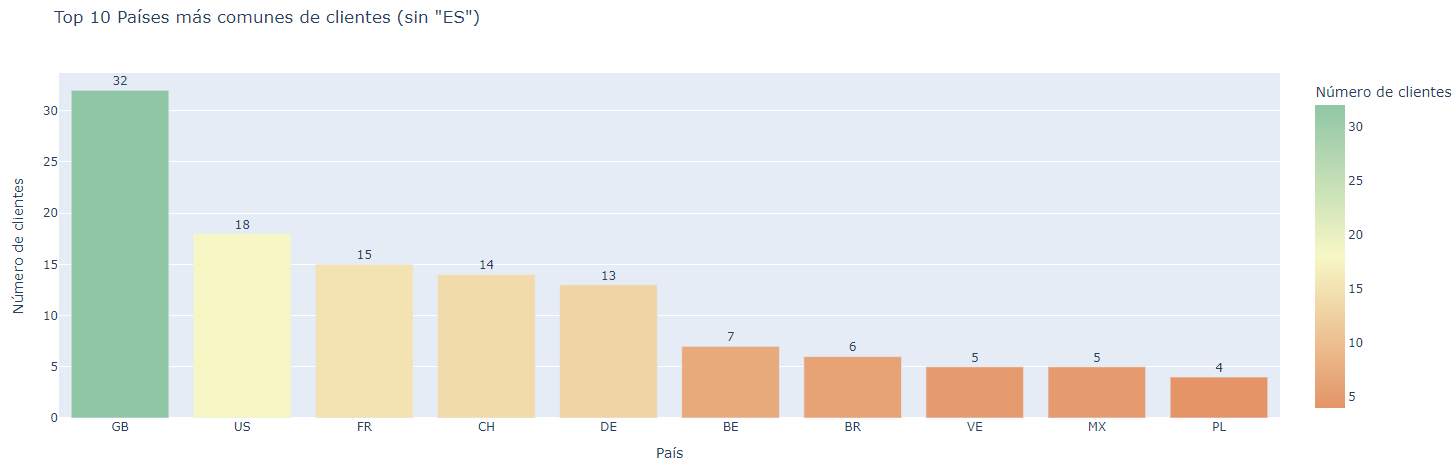
También se detectaron casos en los que la edad de los clientes **disminuía** en registros sucesivos, algo que es evidentemente incorrecto. En estos casos, se procedió de manera similar: se sustituyó la edad de estos clientes por la última edad conocida, ya que esta es probablemente la más correcta.

### 1.1.4 Country\_ID

La variable country\_id representa el país de residencia de los clientes y, en este dataset, muestra una predominancia clara de España ('ES'), con más de 5.9 millones de clientes registrados en este país. Esta variable es crucial para entender la distribución geográfica de los clientes, aunque con un claro sesgo hacia España. A continuación, se detallan los pasos realizados para la limpieza y preprocesamiento de esta variable.

**Distribución General de country\_id**

Inicialmente, se analizaron los **10 países más frecuentes** excluyendo España, con el objetivo de identificar si había alguna tendencia geográfica notable fuera de este país. En este caso, el Reino Unido ('GB'), Francia ('FR'), y Alemania ('DE') figuran entre los países más comunes después de España, pero el volumen de clientes de estos países es significativamente menor (con menos de 500 clientes en cada uno).



**Análisis de Clientes con country\_id Mixto**

Durante el preprocesamiento, se encontraron **25 clientes con valores mixtos** en la columna country\_id, lo que significa que en diferentes registros un mismo cliente aparecía con distintos países de residencia. Estos casos pueden tener diversas explicaciones:

* **Errores en la captura de datos**, donde el país de residencia podría haber sido registrado incorrectamente.
* **Cambios reales de residencia**, aunque este tipo de movimientos suelen implicar cambios en otras variables, como el salario, lo cual no se observó en estos casos.

Para resolver esta inconsistencia, se tomó la decisión de **mantener el último país registrado** para cada cliente mixto. Esta decisión se basa en la lógica de que el último valor es probablemente el más actualizado y preciso. Además, dado que se trata de un número muy pequeño de clientes, su impacto en los resultados generales del análisis es mínimo.



Mantener el **último país de residencia** es la opción más adecuada, ya que refleja la situación más actual del cliente. Además, dado que el salario no cambió en ninguno de estos casos, es probable que los cambios de country\_id sean errores en la captura de datos, lo que refuerza la decisión de optar por el dato más reciente.

### 1.1.5. Deceased

La variable deceased es una indicación de si el cliente ha fallecido o no, con los valores 'N' para los clientes vivos y 'S' para los fallecidos. La presencia de esta variable puede ser relevante desde una perspectiva comercial, ya que los clientes fallecidos naturalmente dejarían de interactuar con la plataforma. Sin embargo, durante el análisis se observaron varios factores que llevaron a reconsiderar su uso en los modelos.

**Variabilidad y Distribución de deceased**

En este dataset, la gran mayoría de los clientes están marcados como **vivos** ('N'), con solo **55 clientes fallecidos** ('S'). Además, se detectaron **74 casos mixtos (particiones)**, en los cuales el estado del cliente cambiaba entre 'N' y 'S' en diferentes registros. Esto no tiene sentido desde una perspectiva lógica, ya que una vez fallecido, un cliente no puede revertir su estado. Este hecho sugiere errores en la captura o en el registro de la variable.

**Análisis de Clientes Mixtos y Constantes**

* **Clientes constantes en deceased**: La mayoría de los clientes mantienen su estado de fallecimiento constante a lo largo del tiempo. Esto incluye tanto a los 55 clientes fallecidos como a los demás clientes vivos.
* **Clientes mixtos en deceased**: Los casos mixtos sugieren errores en la captura de los datos o falta de actualización en algunos registros. Dado que el estado fallecido es irreversible, estos casos fueron corregidos asignando el último estado registrado para cada cliente.

**Eliminación de deceased en los Modelos Predictivos**

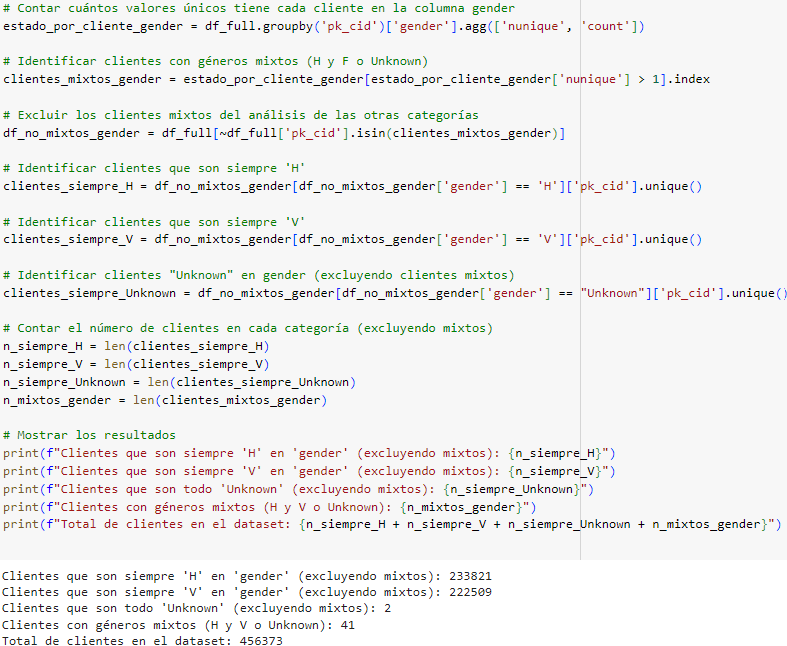
Después de realizar este análisis, se concluyó que la variable deceased **no tiene suficiente capacidad explicativa** para ser utilizada en los modelos de segmentación. Además, se identificó una **alta correlación** entre esta variable y otras como la **edad**, lo que aumenta el riesgo de multicolinealidad si se incluye en los modelos predictivos. Dado que la mayoría de los clientes fallecidos pertenecen a grupos de mayor edad, esta variable no añade valor adicional y podría introducir redundancia.

### 1.1.6. Gender

La variable gender en este dataset cuenta con tres posibles valores: **H** para hombres, **V** para mujeres, y **Unknown** para los casos donde el género no está especificado. El análisis de esta variable es importante para evaluar la segmentación de la clientela en términos de género y para identificar cualquier inconsistencia en los datos.

**Distribución General de gender**

El análisis inicial mostró que la distribución por género es bastante equilibrada, con un **51.24% de hombres** y un **48.76% de mujeres**. Los valores **Unknown** son extremadamente raros, con solo **2 clientes** en esta categoría, lo cual representa un porcentaje insignificante.

****

**Casos Mixtos en gender**

Al revisar los datos, se detectaron **41 casos mixtos**, es decir, clientes que en diferentes registros aparecían con distintos géneros. Esta situación, aunque rara, es un indicativo de posibles errores en la captura de datos. La presencia de géneros mixtos puede deberse a varias razones, como:

* **Errores en el registro** donde el género fue introducido incorrectamente en algún momento.
* **Cambios reales en el género**, aunque dado el bajo número de casos, esta explicación parece poco probable.

Para resolver estos casos, se decidió asignar el **último género registrado** para cada cliente, bajo la premisa de que este es el dato más actualizado y, por tanto, el más fiable.

**Imputación de Unknown**

En los pocos casos donde el género era **Unknown**, se decidió imputar este valor con **Hombre**, ya que ambos clientes superaban los **70 años** y la proporción de hombres es la mayor.

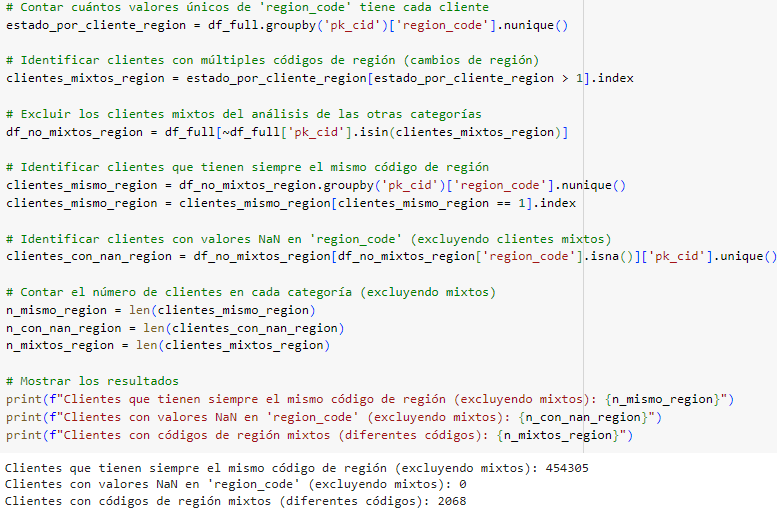
### 1.1.7. Region Code

La variable region code en el conjunto de datos indica la provincia de residencia de los clientes. Al igual que con otras variables, es crucial asegurarse de que esta información esté bien representada y limpia, ya que impactará en el análisis de segmentación y en el entendimiento del comportamiento del cliente según su ubicación geográfica. A continuación, se detallan los pasos realizados para limpiar y tratar esta variable.

**Inicialización y Manejo de Valores Nulos**

1. **Asignación de Unknown**:
   * Se identificaron y asignaron los valores nulos en region\_code como 'Unknown'.
   * Esto se logró mediante el uso de cat.add\_categories() para añadir 'Unknown' a las categorías existentes y luego reemplazar los valores nulos usando fillna().
2. **Conteo de Clientes**:

* Se contabilizaron los clientes por region\_code para identificar cuántos clientes tenían códigos de región constantes, cuántos presentaban valores nulos, y cuántos tenían códigos mixtos (cambios en la región).
* Se encontraron **454,305 clientes con el mismo código de región**, **0 clientes con valores nulos** en region\_code (excluyendo mixtos), y **2,068 clientes con códigos de región mixtos**. Este último grupo indica que algunos clientes han cambiado su región de residencia en diferentes momentos.



**Sustitución de Valores**

1. **Sustitución de Unknown**:

* Para los casos de region\_code que tenían el valor 'Unknown' y cuyo country\_id era 'ES', se decidió sustituir por **28**, que corresponde a Madrid, ya que esta es la provincia más poblada y representativa de España. Para aquellos que no pertenecen a España, se sustituyó por **00**, indicando que no son clientes locales.
* Esta decisión se basa en el principio de que mantener la información más relevante y actualizada proporciona una visión más precisa del cliente.

1. **Mapeo de Códigos a Nombres**:

* Se realizó un mapeo adicional para convertir los códigos de región a sus nombres correspondientes. Esto facilita la interpretación y el análisis posterior, ya que trabajar con nombres en lugar de códigos es generalmente más comprensible.

1. **Asignación de 'Extranjero'**:

* En caso de que el region\_code siguiera siendo nulo después de las imputaciones, se rellenó con el valor **'Extranjero'** para indicar que el cliente reside fuera de España. Esto es importante para clasificar adecuadamente a los clientes que no están localizados en las provincias españolas.

**Visualización de la Distribución Regional**

1. **Distribución de Usuarios en España**:

* Se creó un gráfico interactivo de dispersión en un mapa para visualizar la distribución de usuarios en España. Los puntos en el mapa representaban el número de clientes por cada provincia, ayudando a identificar áreas con mayor concentración de clientes.
* Este análisis geográfico es útil para la toma de decisiones estratégicas en marketing y segmentación.



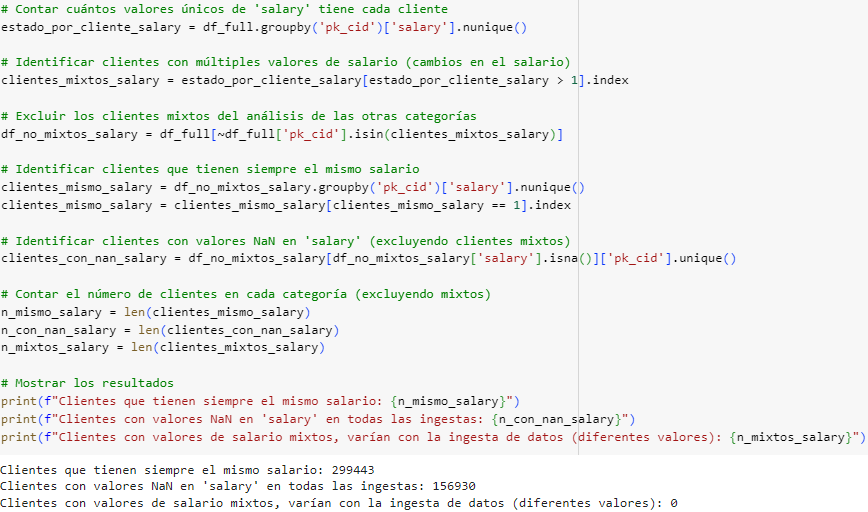
### 1.1.7. Salary

La variable salary representa los ingresos brutos de la unidad familiar de cada cliente y es fundamental para evaluar el potencial financiero y el comportamiento de consumo de los clientes. A continuación, se detallan los pasos realizados para limpiar y preprocesar esta variable.

**Identificación de Valores Nulos y Mixtos**

1. **Conteo de Valores Únicos**:

* Se realizó un análisis para determinar cuántos valores únicos de salary tenía cada cliente. Esto permitió identificar a los clientes con múltiples valores de salario (indicando cambios en sus ingresos) y aquellos que tenían valores nulos.
* Se encontraron **299,443 clientes con el mismo salario**, **156,930 clientes con valores nulos** y **0 clientes con valores de salario mixtos**. Este último resultado sugiere que la mayoría de los clientes tienen un único valor de salario registrado.



1. **Descripción Estadística**:

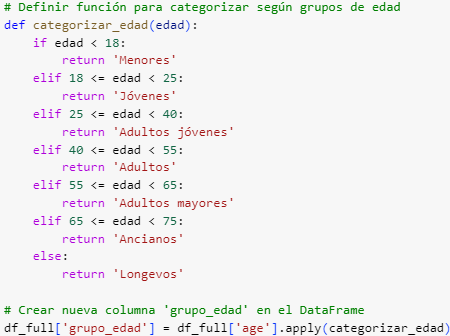
Se generó un resumen estadístico de la variable salary, que reveló que el salario medio era de **115,816.72**, con un rango amplio que va desde **1,202.73** hasta **28,894,395.51**. Este rango amplio puede ser indicativo de outliers o errores en la captura de datos, lo que requiere atención adicional.

1. **Imputación de Valores Nulos**

Para abordar la problemática de los valores nulos, se optó por aplicar una estrategia de imputación basada en el **salario mediano**. La elección de la mediana es estratégica, ya que proporciona una medida más robusta en presencia de valores atípicos.

**Creación de Grupos de Edad**

Como los valores nulos están presentes en varios grupos de edad y países, se decidió segmentar a los clientes en grupos de edad definidos previamente, a saber:



La imputación de salarios nulos se realizó en dos pasos clave:

* **Calcular la Mediana**: Se calculó la mediana de salary para cada combinación de país, grupo de edad, segmento y código de región en el DataFrame sin duplicados. Este paso es crucial para garantizar que la mediana sea representativa y no se vea afectada por valores atípicos.
* **Imputar los Valores Nulos**: Se aplicó una función que, para cada fila del DataFrame, si el salario era nulo, se reemplazaba con la mediana correspondiente calculada en el paso anterior. Este método asegura que las imputaciones sean específicas y relevantes al contexto del cliente.

**Comprobación de Resultados**

A pesar de los esfuerzos por imputar los valores nulos, se constató que todavía quedaban algunos casos con salary como NaN. Para abordar estos restantes **1,856** casos, se decidió aplicar otra capa de imputación usando la mediana de salary agrupando por **grupo de edad** y **país**. Este enfoque fue seleccionado para maximizar la relevancia de los salarios imputados.

Se creó un nuevo DataFrame con las medianas correspondientes y se utilizó para reemplazar los valores nulos restantes. Tras la imputación, se realizó una verificación final para asegurar que no quedaran valores nulos en salary.

Este riguroso proceso de imputación no solo proporciona un enfoque sistemático para manejar los valores nulos en salary, sino que también permite asegurar que el análisis posterior se base en datos más completos y representativos. A través de la creación de grupos de edad y la utilización de la mediana como método de imputación, se busca construir un modelo más preciso y útil para la toma de decisiones en easyMoney.

### 1.1.8. Resumen de Imputaciones

**1. Columnas Creadas**

* **Mes\_partition y Mes\_nombre\_partition**: Se crearon estas columnas para identificar el mes de pk\_partition. Esto puede ser relevante en el futuro si se detectan estacionalidades en el comportamiento de los clientes a lo largo de los meses.
* **Grupo\_edad**: Se segmentaron las edades en los siguientes grupos, lo cual es útil para un análisis más enfocado:
* **Menores**: Personas menores de dieciocho años.
* **Jóvenes**: Personas mayores de dieciocho años y menores de veinticinco.
* **Adultos jóvenes**: Personas mayores de veinticinco y menores de cuarenta años.
* **Adultos**: Personas mayores de cuarenta y menores de cincuenta y cinco.
* **Adultos mayores**: Personas de cincuenta y cinco a sesenta y cinco años.
* **Ancianos**: Personas de sesenta y cinco a setenta y cinco años.
* **Longevos**: Personas de setenta y cinco años en adelante.

**2. Asignaciones de Valores en Variables**

**Edad (Age)**:

* **Variaciones + 3 años**: Se sustituyeron todos los valores de Age por el último valor conocido. Esto se justifica en que, se considera que los errores de tipificación se corrigen con el tiempo; así, una entrada más reciente es más probable que sea correcta.
* **Variación Negativa**: Se sustituyeron las edades con variaciones negativas por el último valor registrado para cada usuario, asegurando que se mantenga la consistencia en los datos.

**Country\_id Mixto**: Para los casos de country\_id mixto (25 clientes), se asignó el último país registrado para su pk\_cid. Esto se debe a que se presume que la variabilidad en estos datos es un error en la recopilación de datos. Suponiendo que los cambios de residencia fueran reales, el salario debería verse afectado, lo que no se observa en estos casos. Por ello, se mantuvo la información más actualizada.

**Deceased**: La variable deceased presenta una variabilidad despreciable. Por lo tanto, no se considera útil para la construcción de modelos, ya que no aporta capacidad predictiva. Además, dada su alta correlación con otras variables, como el grupo de edad, es prudente eliminarla para evitar problemas de multicolinealidad en los modelos.

**Gender**:

* **Unknown Constante**: Se imputaron los valores Unknown y se cambiaron por "H" (Hombre) para los dos casos que superan los 70 años, considerando que es poco probable que estas personas se identifiquen como género fluido. La categoría masculina presenta la mayor proporción, lo que refuerza esta decisión.
* **Unknown Mixto**: Para los registros con gender mixto, se imputó el último género conocido.
* **Géneros Mixtos**: Se modificaron los registros de género mixto tomando el último valor registrado, ya que se considera que la última entrada es la más actualizada y, por ende, la más fiable.

**Region\_code**:

* **Region Unknown**: Los registros con region\_code como "Unknown" se sustituyeron por "28" en el caso de España (correspondiente a Madrid) y "00" para clientes fuera de España. Esto establece un valor más coherente para el análisis posterior.
* Posteriormente, se reemplazaron los códigos de provincia por el nombre correspondiente, y se etiquetaron como "Extranjero" aquellos registros que no pertenecían a España.

**Imputación de Salario:**

* **Imputación de Nulos**: Se decidió aplicar el salario mediano, segmentado por grupo de edad, país de residencia, región de residencia y actividad comercial. Esta imputación se llevó a cabo de forma iterativa; en caso de que no se encontrara un salario mediano aplicable, se iba reduciendo la restricción de las variables en el siguiente orden:
  + Primero se utilizaron **grupo\_edad** y **country\_id**.
  + Si aún quedaban casos con salarios nulos, se procedió a utilizar únicamente **grupo\_edad**.

Es importante destacar que todas las imputaciones de salario se realizaron sobre el conjunto sin duplicados para garantizar que la mediana no se viera afectada por valores atípicos, lo que podría distorsionar los resultados finales.

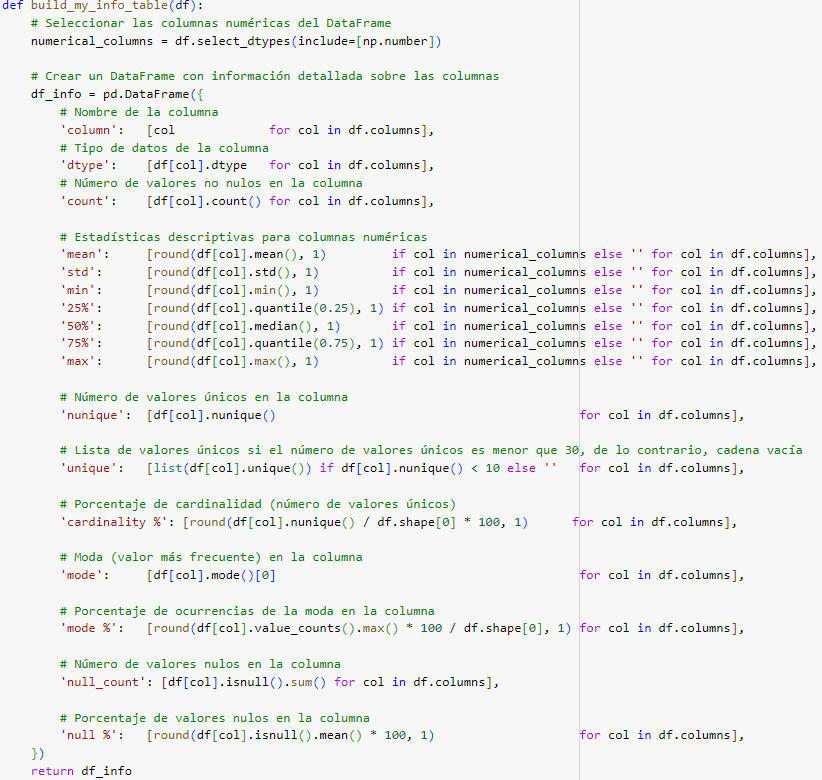
## Product Database

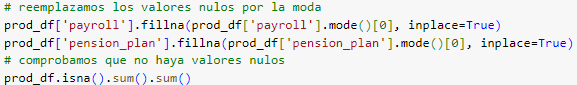
En este segmento del análisis, nos enfocamos en las variables relacionadas con los productos ofrecidos por easyMoney. A continuación, se presentan las variables y su respectiva explicación:

| **Variable** | **Explicación** |
| --- | --- |
| credit\_card | Tarjetas de crédito |
| debit\_card | Tarjetas de débito |
| em\_account\_p | Cuenta easyMoney+ |
| em\_account\_pp | Cuenta easyMoney++ |
| em\_acount | Cuenta easyMoney |
| emc\_account | Cuenta easyMoney Crypto |
| funds | Fondos de inversión |
| loans | Préstamos |
| long\_term\_deposit | Depósitos a largo plazo |
| mortgage | Hipotecas |
| payroll | Domiciliaciones |
| payroll\_account | Cuenta bonificada por domiciliaciones |
| pension\_plan | Plan de pensiones |
| securities | Valores |
| short\_term\_deposit | Depósitos a corto plazo |

### 1.2.1. Análisis Preliminar

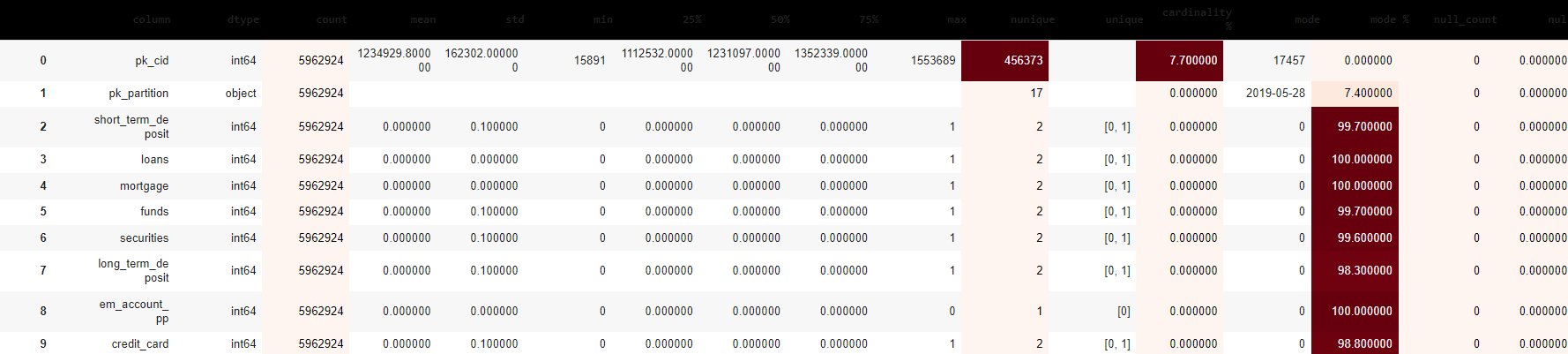
Para iniciar el análisis, cargamos el dataset de productos desde AWS S3. Realizamos un análisis preliminar del DataFrame para observar la distribución de los datos y los valores nulos. Los resultados mostraron que hay **61 nulos** en las columnas **payroll** y **pension\_plan** dentro de un dataset que contiene más de 5 millones de filas. Se procederá a imputar estos valores nulos utilizando la moda. Posteriormente, confirmamos que no haya valores nulos en el DataFrame:





**Tabla de Información General**

Eliminamos columnas que no aportan información *'Unnamed: 0'* y creamos una tabla de información detallada:



Esto nos permite extraer las siguientes conclusiones de la tabla:

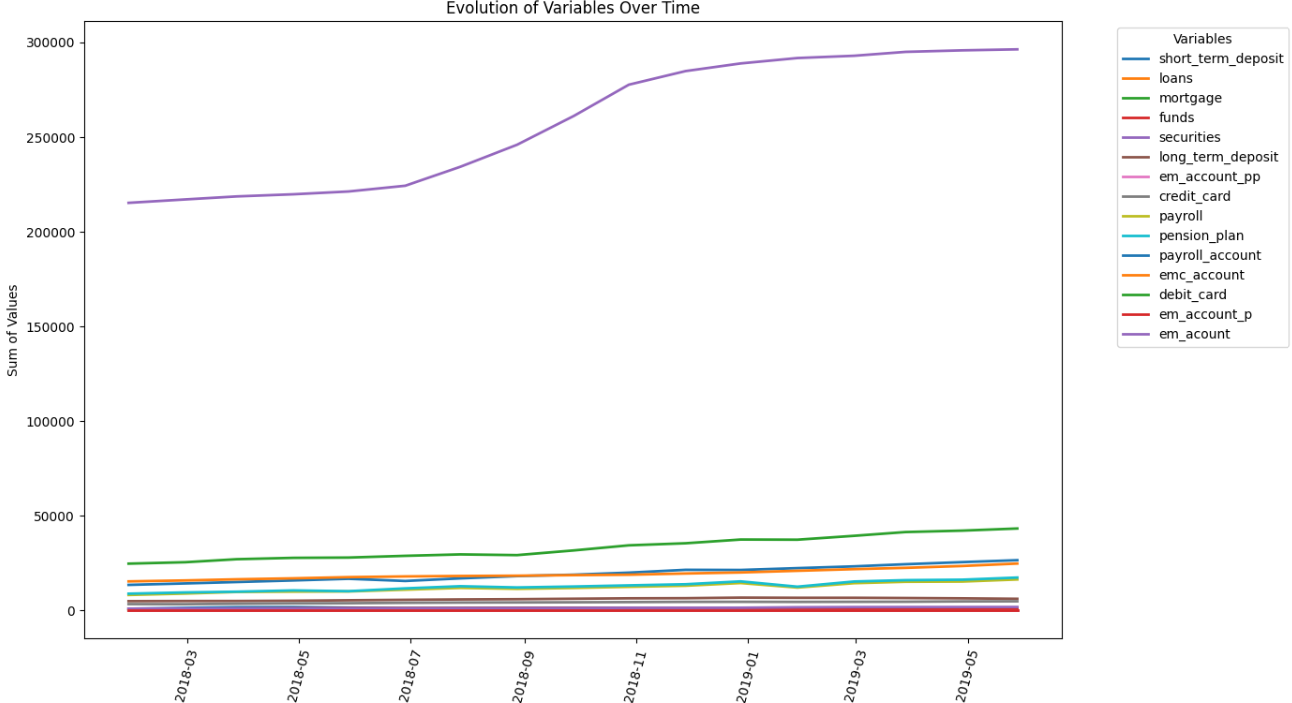
* Se han corregido todos los valores nulos observados.
* A excepción de la cuenta easyMoney, el resto de productos tienen una tasa de contratación/uso menor al 10% sobre el total de particiones.
* Algunos casos concretos, como **loans**, **mortgage**, **em\_account\_pp** y **em\_account\_p**, presentan una tasa de contratación del 0%.

**Modificación de Tipos de Datos**

Para optimizar el uso de memoria, modificamos el tipo de dato de las columnas relacionadas con los productos financieros.

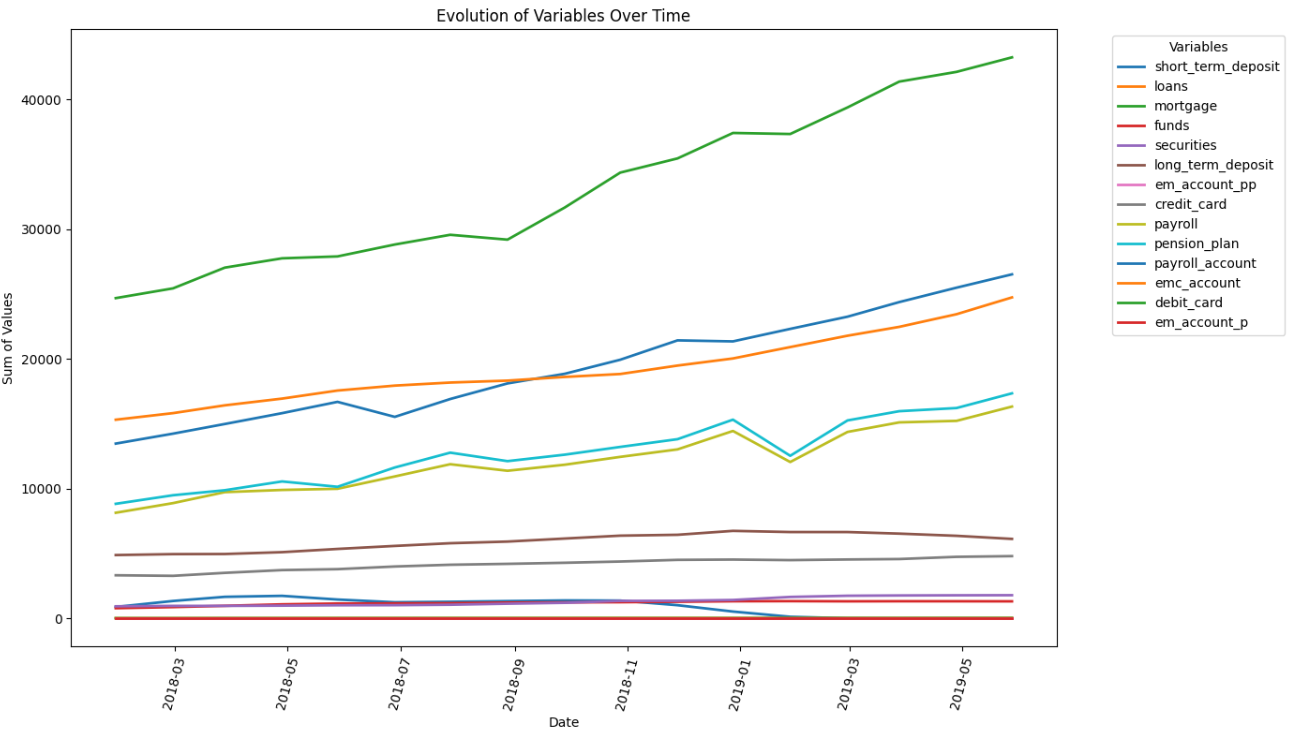
### 1.2.2. Comportamiento de los Productos a lo Largo de las Particiones

Agrupamos por **pk\_partition** y sumamos las demás columnas para observar la evolución de los productos a lo largo del tiempo:



El gráfico resultante revela que **em\_account** es el producto más utilizado, mientras que otros productos como **loans** y **mortgage** tienen una adopción considerablemente menor. Al eliminar **em\_account** del análisis para observar el comportamiento de los demás productos, encontramos que:

* Productos como **tarjetas de débito** y **cuentas de nómina** presentan un crecimiento constante.
* Los productos como **fondos**, **depósitos a largo plazo** e **hipotecas** tienen un uso limitado y constante.



El gráfico muestra la evolución del uso de varios productos financieros a lo largo del tiempo:

* Productos más populares: Las **tarjetas de débito**, **las cuentas de nómina** y los **préstamos** son los más utilizados por los clientes, mostrando un crecimiento constante.
* Uso moderado: Las **tarjetas de crédito** y la **nómina** tienen un uso estable, con un crecimiento más leve.
* Bajo uso: Productos como **valores, fondos, depósitos a largo plazo e hipotecas** tienen un uso limitado y constante, indicando menor adopción.
* Adopción mínima: Productos como **em\_account\_pp, em\_account\_p, y emc\_account** tienen una adopción casi inexistente.

En resumen, mientras algunos productos financieros son ampliamente utilizados, otros presentan oportunidades significativas para aumentar su adopción entre los clientes.

### 1.2.3. Análisis de Clientes Según Tipos de Producto

Al analizar el comportamiento de los productos según tipos de clientes, se observa que existen clientes que repiten en cada partición, así como otros que presentan un comportamiento variable.

La distribución de productos para cada cliente se evalúa y se generan gráficos que muestran el porcentaje de clientes que siempre utilizan un producto, aquellos que no lo usan y los que tienen un comportamiento mixto.

La imagen obtenida presenta un análisis visual del comportamiento de los productos ofrecidos por easyMoney. En el primer gráfico se observa la distribución porcentual de los clientes categorizados según su uso de los productos: aquellos que siempre utilizan el producto, aquellos que nunca lo utilizan (Siempre 0) y aquellos que presentan un comportamiento variable. Esta categorización es crucial para entender las dinámicas de uso y aceptación de los diferentes productos en el portafolio de la empresa.

En general, la mayoría de los productos financieros analizados muestran una notable cantidad de clientes que no los utilizan "Siempre 0". Este hallazgo indica que hay un área significativa de mejora para la empresa, ya que existe una oportunidad para aumentar la adopción de estos productos entre los clientes. La empresa puede investigar las razones detrás de esta falta de uso: ¿son los productos poco conocidos? ¿Hay barreras de acceso? ¿Los clientes no ven el valor en los mismos? Estas preguntas pueden ayudar a identificar estrategias de marketing y educación que fomenten la utilización de productos financieros.

Por otro lado, productos como **em\_account** han demostrado ser ampliamente utilizados y mantienen un uso constante. Esto se debe a que fue el producto original con el que comenzó la empresa, antes de pivotar hacia otros productos financieros.

**Análisis Detallado del Comportamiento por Tipo de Cliente**

El resto de los gráficos analizan más a fondo el comportamiento de los diferentes tipos de clientes a lo largo de las particiones temporales. Este análisis revela patrones específicos que son importantes para entender cómo los productos están siendo utilizados en el tiempo:

* **Clientes que nunca usan el producto**: El comportamiento de estos clientes es notablemente consistente a través de los productos. Antes de julio de 2018, el número de clientes que no utilizaban los productos se mantenía relativamente estable. Sin embargo, a partir de esta fecha, se ha observado un aumento en la inclusión de nuevos clientes en esta categoría. Este cambio puede indicar un desinterés creciente en los productos por parte de algunos usuarios existentes o la incapacidad de la empresa para atraer a nuevos clientes hacia la utilización de productos que previamente no usaban.
* **Clientes que siempre usan el producto**: Estos clientes presentan una tendencia estable, aunque hay una ligera inclinación ascendente en algunos productos después de 2018. Este crecimiento sugiere que, con el tiempo, algunos clientes han comenzado a adoptar productos que anteriormente no utilizaban de manera consistente.
* **Clientes con comportamiento variable**: Este grupo es especialmente interesante, ya que muestra fluctuaciones en su uso de productos. Inicialmente, muchos de estos clientes no utilizaban los productos, pero a partir de julio de 2018, se ha notado una mejora en la aceptación. Sin embargo, algunos productos como **em\_account**, **long\_term\_deposit**, **loans** y **short\_term\_deposit** han visto una reducción en el interés por parte de los clientes. Esto puede indicar que, aunque la aceptación de productos está aumentando, ciertos productos están perdiendo su atractivo o que no cumplen con las expectativas de los usuarios. La empresa debe investigar más a fondo estos casos para entender las razones detrás de esta disminución.

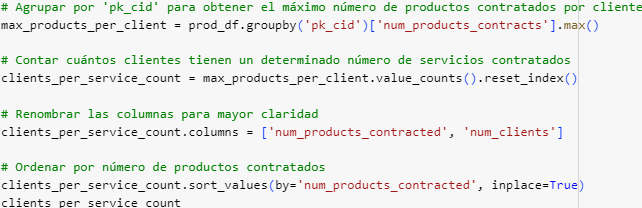
### 1.2.4. Suma de Productos Contratados

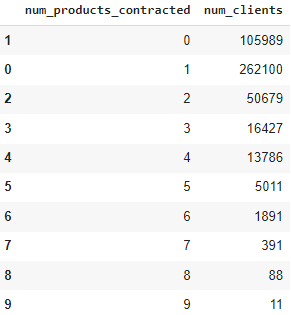
Para entender mejor la interacción de los clientes con los productos financieros ofrecidos por EasyMoney, hemos creado una nueva columna en el DataFrame prod\_df, denominada num\_products\_contracts, que representa la suma de productos contratados por cada registro. Este paso nos permite tener una visión más clara del comportamiento de los clientes en relación a la diversidad de productos que utilizan.

Se observa que la mayoría de los clientes (aproximadamente el 57%) han contratado al menos un producto, mientras que el 23% no tiene ningún producto contratado. Esta información es crucial, ya que sugiere que una parte significativa de la base de clientes todavía no ha explorado los productos disponibles, lo que podría ser una oportunidad para aumentar la adopción de productos.

**Agrupando los Clientes por Productos Contratados**

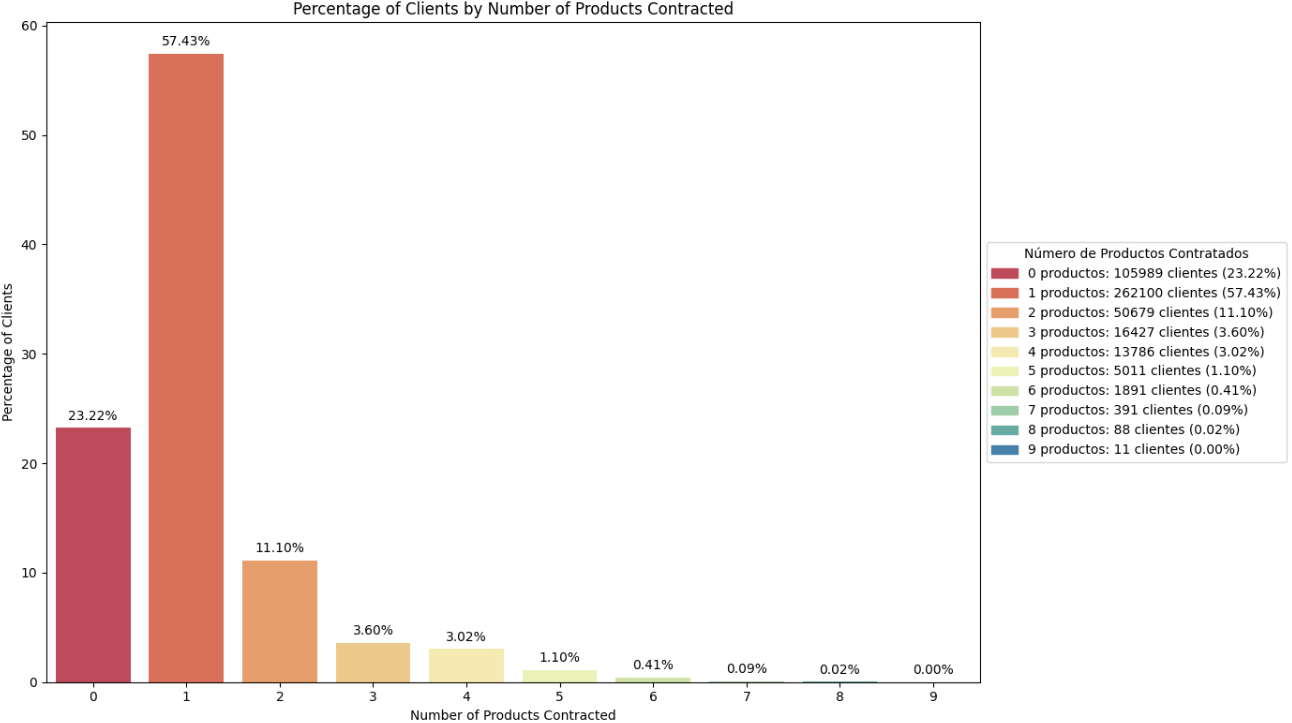
A continuación, agrupamos a los clientes por el número máximo de productos contratados a lo largo de su historia. Esto nos permitirá identificar el número total de clientes que poseen un determinado número de productos:





**Porcentaje de Clientes por Número de Productos Contratados**

Para profundizar en el análisis, calculamos el porcentaje de clientes según el número de productos contratados. A continuación, se presenta un gráfico de barras que ilustra la distribución porcentual de los clientes en función del número de productos contratados:



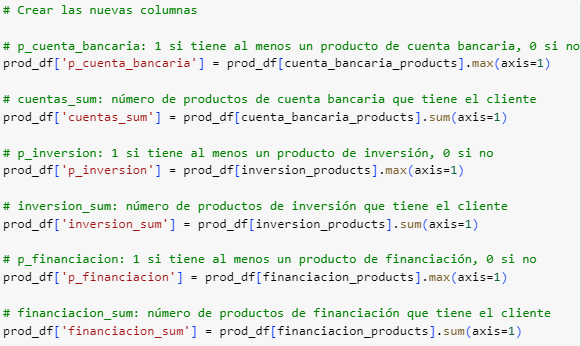
Analizando el gráfico resultante, se pueden extraer las siguientes conclusiones:

* **Más del 57% de los Clientes Tienen al Menos un Producto Contratado**: Este hallazgo indica que una parte significativa de la clientela ha comenzado a explorar los productos de easyMoney. Sin embargo, también implica que existe un porcentaje considerable de clientes (23%) que no ha contratado ningún producto, lo que podría reflejar desinterés, falta de conocimiento o barreras de acceso.
* **Uso de Productos por Parte de los Clientes**: El gráfico muestra que la mayoría de los clientes solo tiene uno o dos productos contratados. Esto sugiere que, si bien hay una adopción inicial, es necesario trabajar en estrategias que fomenten la contratación de más productos por parte de los clientes existentes.
* **Oportunidades para la Empresa**: La gran cantidad de clientes que no han contratado ningún producto presenta una clara oportunidad para easyMoney. Invertir en campañas de marketing y educación puede ser clave para aumentar la adopción de productos.

### 1.2.5. Creación de Variables por Segmentación de Productos

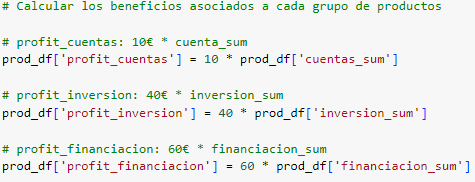
Para entender mejor el rendimiento de los productos ofrecidos por easyMoney, se sugiere agrupar los productos por familias, según su tipología y los ingresos que generan. Se crean las siguientes variables:

1. **p\_cuenta\_bancaria**: Esta variable se establece en 1 si el cliente tiene contratado en esa pk\_partition cualquiera de los siguientes productos: em\_account\_pp, payroll, payroll\_account, emc\_account, debit\_card, em\_account\_p, em\_acount. En caso contrario, se asigna 0.
2. **cuentas\_sum**: Esta columna sumará el número de productos de cuentas bancarias contratados por el usuario en ese momento.
3. **p\_inversion**: Al igual que la variable anterior, se asigna un 1 si el cliente tiene al menos un producto de inversión en su pk\_partition, y 0 en caso contrario. Los productos considerados son: short\_term\_deposit, funds, securities, long\_term\_deposit, pension\_plan.
4. **inversion\_sum**: Esta variable contará el número de productos de inversión que tiene el usuario.
5. **p\_financiacion**: Esta columna recogerá un 1 si el cliente tiene contratado en esa pk\_partition cualquiera de los siguientes productos: loans, mortgage, credit\_card, y 0 si no.
6. **financiacion\_sum**: Sumará el número de productos de financiación contratados por el usuario.



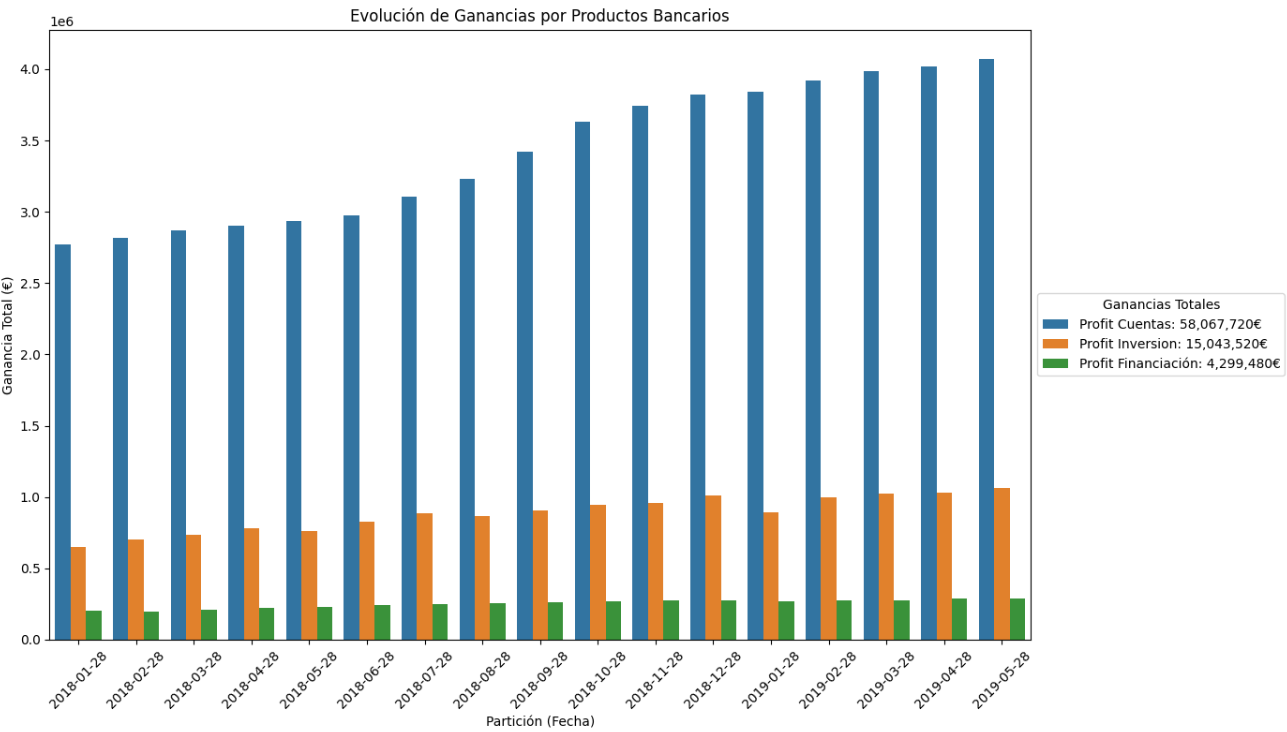
Una vez definidas las variables por tipología de productos, se procede a calcular los beneficios asociados a cada grupo. Los beneficios se estiman de la siguiente manera:

* **Profit Cuentas**: Se asigna un beneficio de 10€ por cada cuenta bancaria contratada.
* **Profit Inversión**: Se asigna un beneficio de 40€ por cada producto de inversión.
* **Profit Financiación**: Se asigna un beneficio de 60€ por cada producto de financiación.



**Análisis de la Evolución de Ganancias**

Para realizar un análisis más detallado sobre la evolución de las ganancias asociadas a los productos, se agrupan los datos por pk\_partition y se suman las ganancias de cada tipo de producto. Luego, creamos un gráfico que ilustre la evolución de las ganancias a lo largo del tiempo:



La gráfica presentada muestra la evolución de las ganancias totales generadas por los diferentes grupos de productos bancarios a lo largo del tiempo, desde enero de 2018 hasta marzo de 2019.

1. **Ganancias Totales**: Las ganancias totales, muestran un crecimiento constante y significativo a lo largo del tiempo, lo que indica una tendencia positiva en la adquisición y uso de productos bancarios por parte de los clientes. Este incremento sugiere que la empresa ha logrado mejorar su oferta o fortalecer su relación con los clientes, alentando a más personas a contratar productos.
2. **Ganancias por Productos**: La gráfica incluye diferentes colores para representar las ganancias generadas por cuentas bancarias (azul), inversiones (naranja) y financiación (verde).

* **Cuentas Bancarias**: Con ganancias totales de 58,067,720€, este grupo representa la mayor parte de los ingresos, lo que implica que los clientes están optando mayoritariamente por este tipo de productos.
* **Inversiones**: Las ganancias de este segmento alcanzan 15,043,520€. Aunque son significativamente menores en comparación con las cuentas bancarias, el hecho de que haya ingresos procedentes de inversiones demuestra un interés por parte de los clientes en diversificar su patrimonio, lo que podría ser un área para explorar más a fondo y potenciar.
* **Financiación**: Con 4,299,480€ en ganancias, el segmento de financiación muestra un interés moderado de los clientes en productos como préstamos y tarjetas de crédito. A pesar de ser el grupo con menores ingresos, es crucial para la salud financiera de la empresa y puede ser un área con potencial de crecimiento si se implementan estrategias adecuadas para incentivar su uso.

1. **Conclusiones**: La evolución de las ganancias indica que la empresa está en una trayectoria positiva, con un aumento notable en las ganancias totales. Sin embargo, se observa una dependencia elevada en productos de cuentas bancarias, lo que sugiere que la empresa debe seguir diversificando su oferta de productos para reducir el riesgo asociado a una única fuente de ingresos. Además, hay oportunidades para mejorar la adopción y uso de productos de inversión y financiación, lo que podría resultar en un aumento sustancial de los ingresos en el futuro.

### 1.2.6. Resumen

1. **Tratamiento de Valores Nulos:** Se identificaron 61 valores nulos en dos columnas clave, **pension\_plan** y **payroll**. Aunque representan una pequeña fracción del total de datos, estos valores pueden afectar el análisis, se decidió reemplazar los valores nulos con la moda, dado que es una técnica efectiva para columnas categóricas en datasets de gran tamaño. Esto asegura que los registros estén completos sin introducir sesgos significativos:
2. **Identificación de Duplicados**: Una validación esencial del preprocesamiento fue la búsqueda de duplicados, los cuales pueden inflar resultados o distorsionar el análisis. No se encontraron duplicados, lo que indica que cada registro en el dataset es único y representa a un cliente o una partición temporal distinta.
3. **Popularidad de los Productos** El producto más utilizado a lo largo del tiempo por los clientes es **em\_acount** (cuenta bancaria electrónica). Este comportamiento refleja una preferencia clara por parte de los usuarios hacia productos que permiten la gestión básica de sus finanzas de manera digital. Este hallazgo tiene implicaciones significativas para el diseño de futuras estrategias de marketing y desarrollo de productos.
4. **Análisis del Comportamiento de los Clientes en las Particiones** Se observó que hay clientes que mantienen su actividad en diferentes particiones del dataset. Estos clientes presentan distintos patrones de comportamiento:

* Algunos mantienen valores constantes (0 o 1) en todas las particiones, lo que sugiere un comportamiento consistente en la contratación o no de productos.
* Otros muestran variabilidad en su comportamiento, contratando o no productos en diferentes momentos.

1. **Agrupación de Clientes por Cantidad de Productos Contratados**: Uno de los puntos clave del análisis fue la creación de una nueva variable: **num\_products\_contracts**, que representa el número de productos contratados por cada cliente, con un rango que va de 0 a 9 productos contratados.

Los resultados muestran que:

* El **57%** de los clientes tienen al menos un producto contratado.
* El **23%** no tiene ningún producto contratado.
* El **20%** restante tiene más de un producto contratado.

1. **Evolución de Ganancias por Segmento de Productos**: A partir del análisis financiero, se observó un comportamiento diferencial en las ganancias generadas por cada segmento de productos financieros. Los resultados indican que:

* **Cuentas**: Constituyen el producto más rentable, acumulando **58.067.720€**, y mostrando un crecimiento continuo a lo largo del periodo analizado.
* **Inversiones**: Han generado **15.043.520€**, consolidándose como el segundo segmento más rentable.
* **Financiación**: Aunque su aportación es menor, ha sumado **4.299.480€** a las ganancias totales.

Este análisis sugiere que las cuentas bancarias son el producto financiero más valioso para la empresa, seguido por las inversiones. Las estrategias futuras podrían centrarse en potenciar estos segmentos, sin descuidar el desarrollo de productos de financiación.

## 1.3. Actividad Comercial

La siguiente sección está dedicada al análisis de la **actividad comercial** de los clientes en la aplicación easyMoney. El conjunto de datos correspondiente contiene información relevante sobre la interacción de los clientes con la plataforma, los canales de captación, fechas de contratación y otros detalles relacionados con su actividad comercial.

### 1.3.1. Descripción del dataset

El dataset proporcionado en esta sección consta de las siguientes variables:

* **active\_customer**: Indicador de actividad del cliente en la aplicación.
* **entry\_channel**: Canal de captación del cliente.
* **entry\_date**: Fecha en la que el cliente realizó la primera contratación a través de easyMoney.
* **segment**: Segmento comercial al que pertenece el cliente.
* **pk\_cid**: Identificador único del cliente.
* **pk\_partition**: Fecha en la que los datos fueron ingeridos.

### 1.3.2. Preprocesamiento de datos

El primer paso en el preprocesamiento de los datos fue eliminar las columnas irrelevantes, como **Unnamed: 0**, que no aportaban información útil.

Se identificaron errores en las fechas de la columna **entry\_date**, como la presencia de días no válidos (29 de febrero en años no bisiestos). Para corregir estos errores, se sustituyeron manualmente las fechas erróneas por las correctas:

A continuación, se generó una tabla descriptiva que resume las principales características de cada variable. Se incluyeron estadísticas como el número de valores únicos, moda, porcentajes de valores nulos, y estadísticas descriptivas para las variables numéricas.

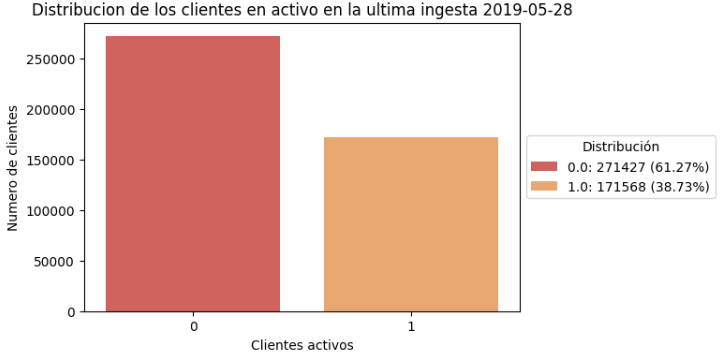
****

En un primer analisis, podemos comprobar bastantes datos a simple vista:

* El número total de clientes, valores únicos de "pk\_cid" = 456373.
* Comprobamos que hay 17 fechas de “pk\_partition” correspondientes a 17 ingestas de datos.
* Hay dos columnas con valores nulos “entry\_channel” y “segment”.
* No tenemos filas duplicadas.

### 1.3.3. Clientes Activos e Inactivos

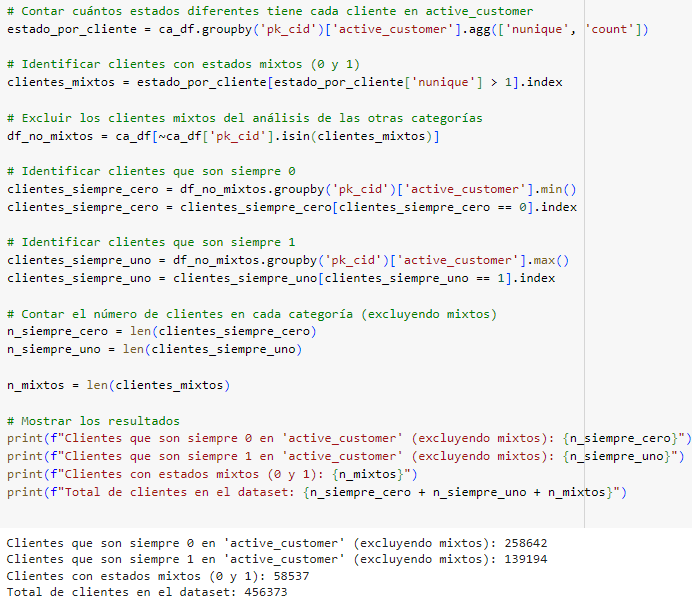
Para analizar la distribución de clientes activos e inactivos, se filtraron los datos de la última ingesta disponible, correspondiente al 28 de mayo de 2019. Los resultados mostraron una proporción significativa de clientes inactivos en la plataforma, con un **61% de inactividad** en ese periodo.



Este resultado refleja una problemática importante que deberá abordarse en la estrategia de retención de clientes, ya que un alto porcentaje de clientes no está utilizando activamente la plataforma.

### 1.3.4. Clientes duplicados por actividad

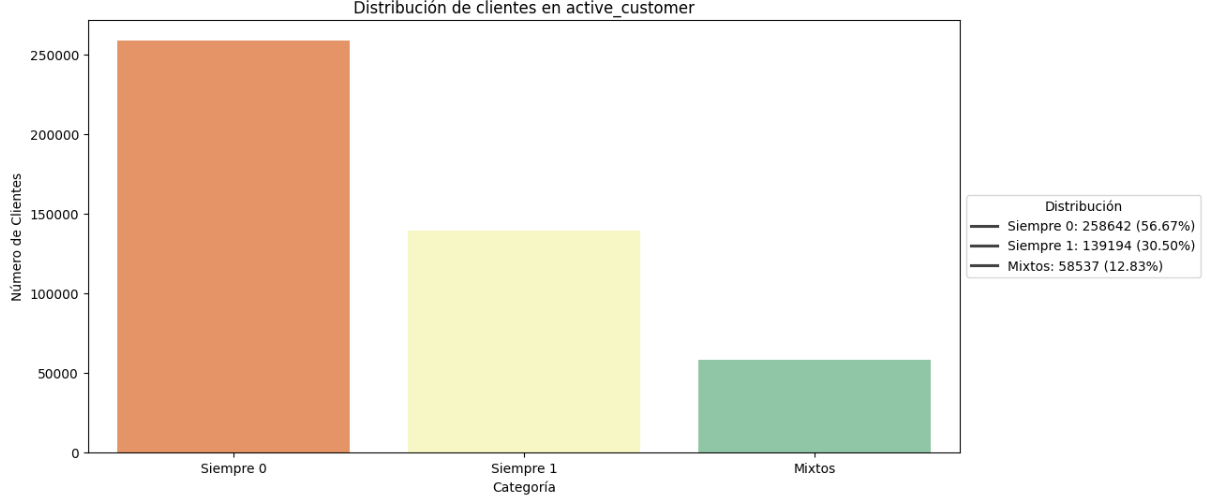
El análisis de la columna **active\_customer** mostró que algunos clientes han cambiado su estado de actividad a lo largo del tiempo. Al agrupar los clientes por su estado de actividad, se identificó una discrepancia en el número de clientes únicos, lo que indica que hay clientes que presentan más de un estado de actividad en diferentes ingestas:



Esta observación refuerza la importancia de analizar el comportamiento de los clientes a lo largo del tiempo para obtener una visión más precisa de su ciclo de vida dentro de la plataforma.

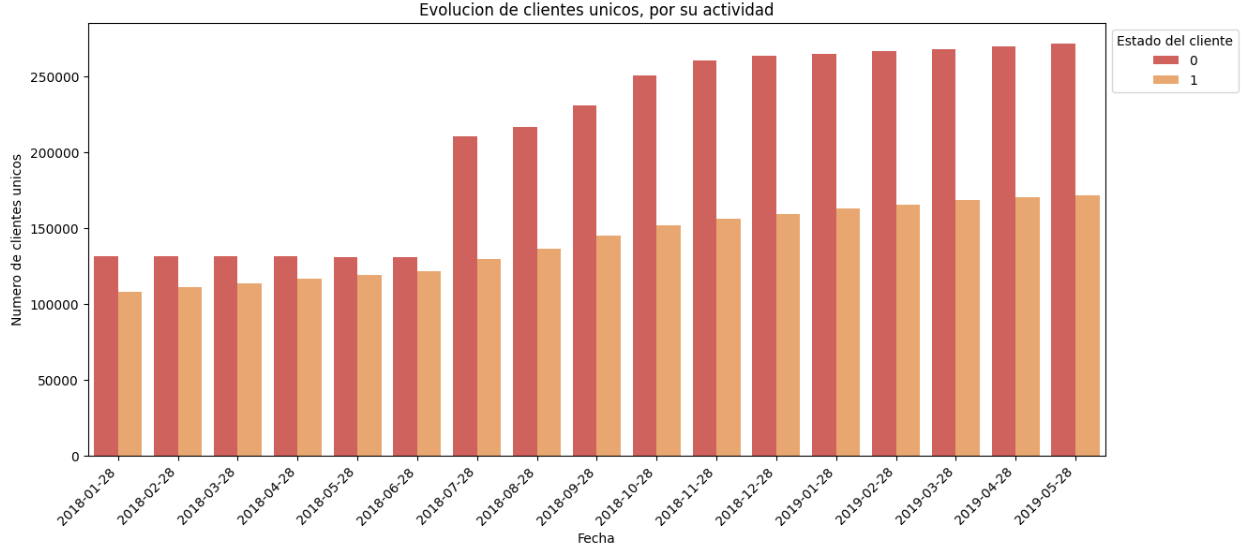
**Gráfica de Distribución**

La mayoría de los clientes (56.67%) se encuentran inactivos durante todo su ciclo de vida, lo cual sugiere un problema en la retención de clientes. Solo un pequeño porcentaje (12.83%) muestra un comportamiento variable en su estado de actividad.



**Evolución Temporal de la Actividad**

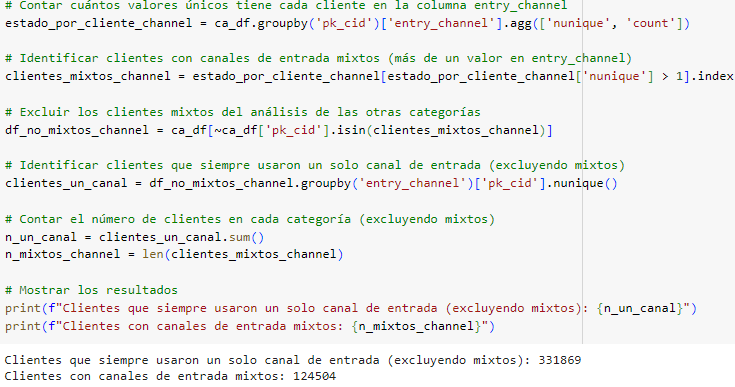
Además, se analizó la evolución temporal de la actividad de los clientes a lo largo del tiempo. Los resultados indican que, aunque la base de clientes ha crecido, una proporción significativa de nuevos clientes no se mantienen activos. Se observó un aumento en el número de clientes inactivos en julio de 2018, lo que puede requerir una estrategia de reactivación para evitar la pérdida de clientes.



### 1.3.5. Canal de Entrada (entry\_channel)

Se exploraron los canales de entrada de los clientes para identificar patrones en la captación de usuarios. Tras el análisis de las categorías, se observó que un porcentaje significativo de los registros presentaba información mixta o faltante. Se tomó la decisión de:

1. **Reemplazar valores 'Unknown'**: Se reemplazó el valor Unknown por el último canal conocido para cada cliente.
2. **Reemplazar clientes con canales mixtos**: Los clientes que habían utilizado más de un canal de entrada fueron reasignados al canal más reciente registrado.



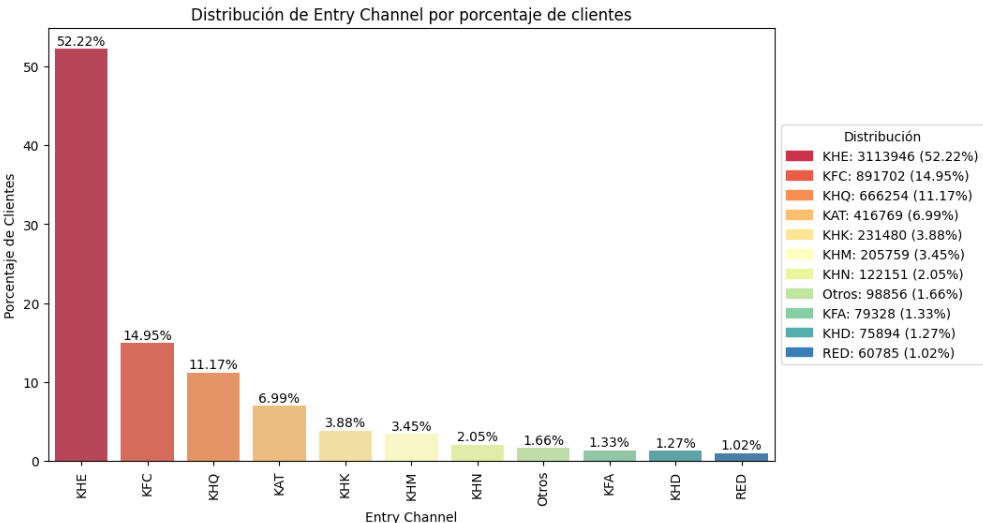
**Distribución de Canales de Entrada**

Una vez depurada la base de datos, se observó que la mayor parte de los clientes ingresó a través de los siguientes canales:

* **KHE**: 52.22% de los clientes.
* **KFC**: 14.95% de los clientes.
* **KHQ**: 11.17% de los clientes.
* **KAT**: 6.99% de los clientes.

Los canales que representaban menos del 1% de los clientes fueron agrupados bajo la categoría "Otros", lo cual proporciona una visión más clara de los principales canales de captación.

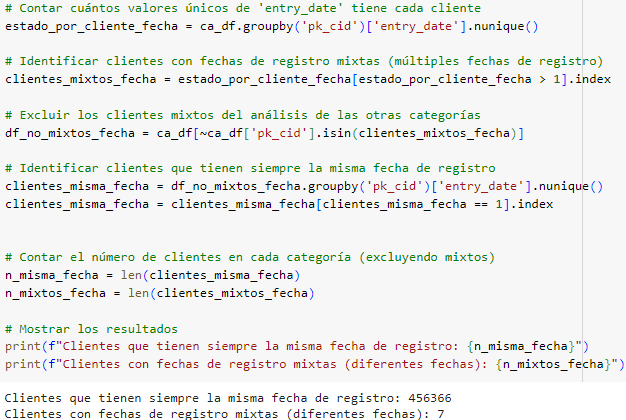
La siguiente gráfica muestra la distribución de clientes por los principales canales de entrada:



### 1.3.6. Fecha de Contratación (entry\_date)

**Evaluación de Fechas de Contratación**

Se realizó un análisis exhaustivo de las fechas de contratación de los clientes para identificar la consistencia de los datos en la columna entry\_date. Se comenzó por contar los valores únicos de entry\_date para cada cliente, con el fin de determinar si había clientes con múltiples fechas de registro. A continuación, se identificaron **7 clientes** que presentaban fechas de registro mixtas. Sin embargo, la gran mayoría de los clientes, **456,366**, tenía siempre la misma fecha de registro.



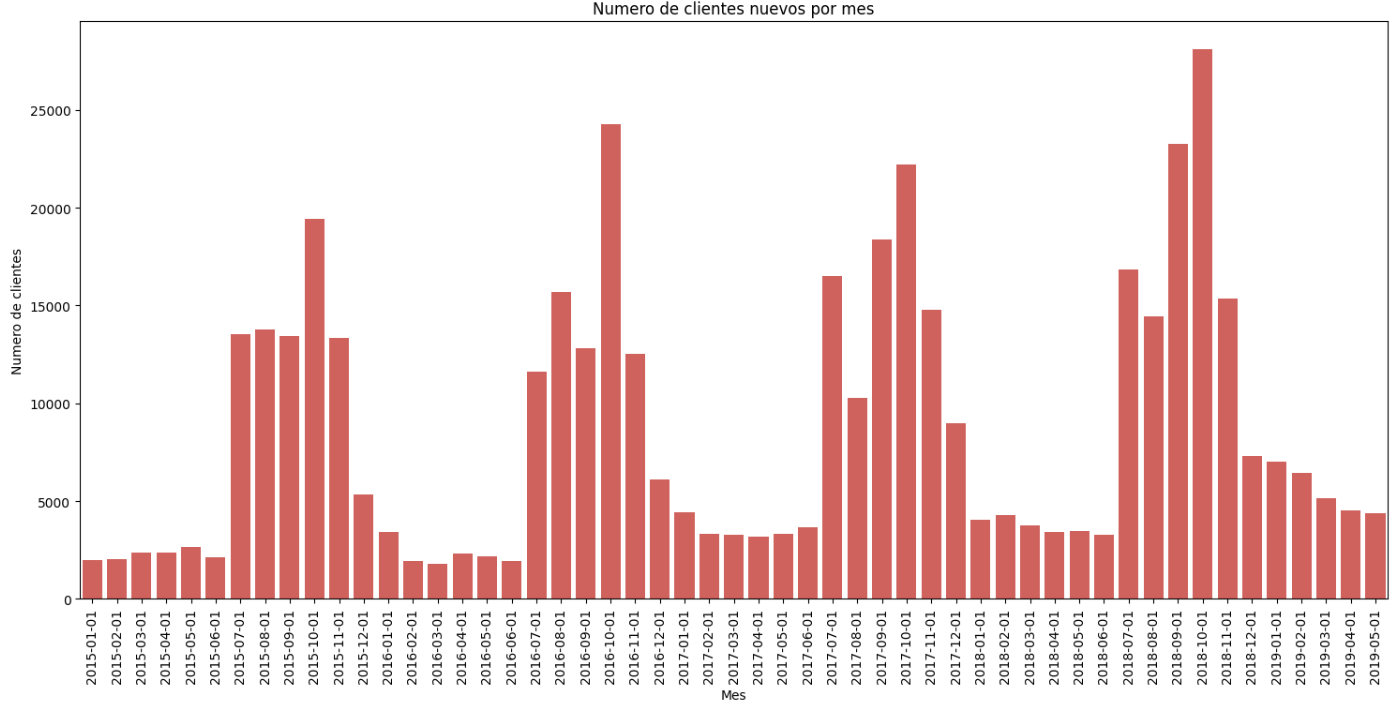
**Corrección de Fechas de Registro**

Se observó que las inconsistencias en entry\_date podrían indicar errores en la carga de datos o actualizaciones incorrectas. Para corregir estos problemas, se decidió establecer la fecha de registro como el valor más temprano registrado para cada cliente. Este enfoque asegura que se mantenga la calidad y la integridad de los datos.

El proceso implicó ordenar los datos por pk\_cid y pk\_partition, y luego aplicar la fecha de la primera partición registrada a todos los registros correspondientes a cada cliente. Tras este ajuste, se verificó que **456,373** clientes ahora tenían la misma fecha de registro, eliminando así cualquier inconsistencia en los datos.

**Comportamiento de Clientes Según Fecha de Contratación**

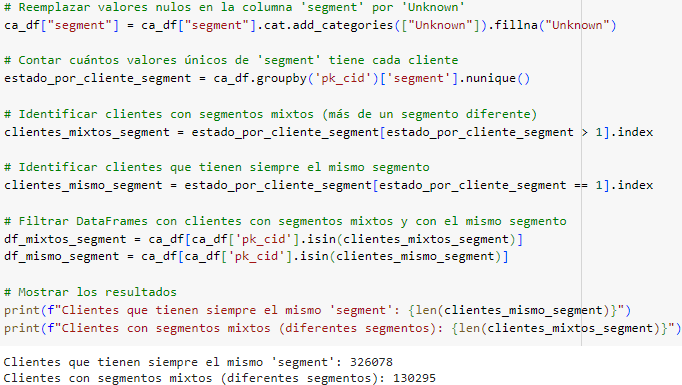
Se creó un dataframe que contenía únicamente los clientes con fechas de registro consistentes. Posteriormente, se agruparon las fechas por mes para analizar la actividad de adquisición de nuevos clientes a lo largo del tiempo. El gráfico resultante mostró ciclos claros en la actividad de adquisición, con picos notables en la captación de nuevos clientes. Se identificaron períodos de alta actividad entre **mediados de julio hasta mediados de agosto** y también **a inicios de enero** de cada año.



### 1.3.7. Segmentación (Segment)

**Análisis General de Segmentos**

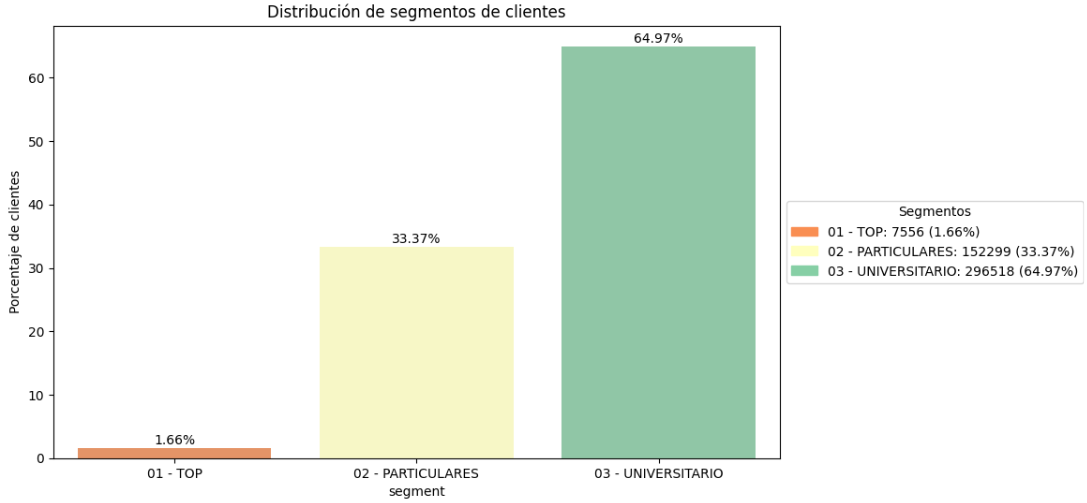
En el siguiente paso, se abordó la segmentación de clientes, comenzando por reemplazar los valores nulos en la columna segment con "Unknown". Se contó el número de valores únicos en segment para cada cliente, lo que reveló que **326,078** clientes tenían siempre el mismo segmento, mientras que **130,295** presentaban segmentos mixtos.



Para manejar las inconsistencias, se definió una función que reemplazó los valores 'Unknown' por el segmento más reciente registrado. Después de aplicar este proceso, se confirmó que ya no había clientes con segmentos mixtos, y **456,373** clientes ahora tenían un único segmento asignado.

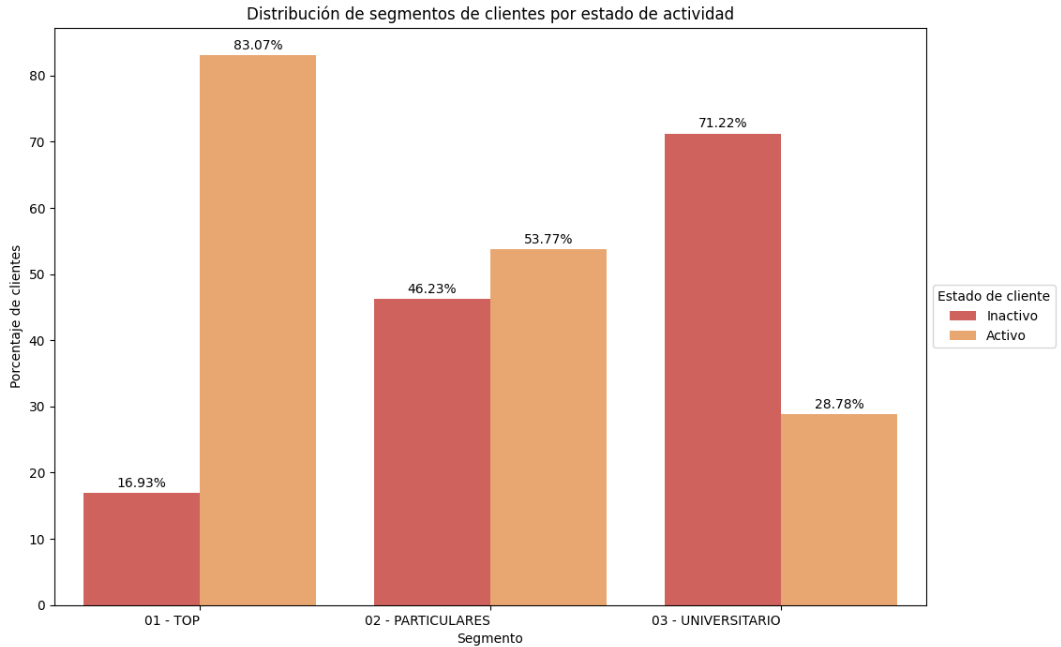
**Distribución de Segmentos**

La distribución final de los segmentos mostró que la mayoría de los clientes pertenecen al segmento **03 - UNIVERSITARIO** (87.24%), seguido de **02 - PARTICULARES** (12.33%) y **01 - TOP** (0.02%). Esto indica una gran concentración de clientes en el segmento universitario, lo que puede sugerir la necesidad de estrategias de marketing más específicas para los otros segmentos.



**Análisis de Clientes Activos e Inactivos por Segmento**

Se realizó un análisis adicional para observar la proporción de clientes activos e inactivos en cada segmento. Los resultados revelaron que los clientes en el segmento **TOP** eran los más activos, mientras que los **universitarios** mostraban una menor actividad en comparación con los demás segmentos.

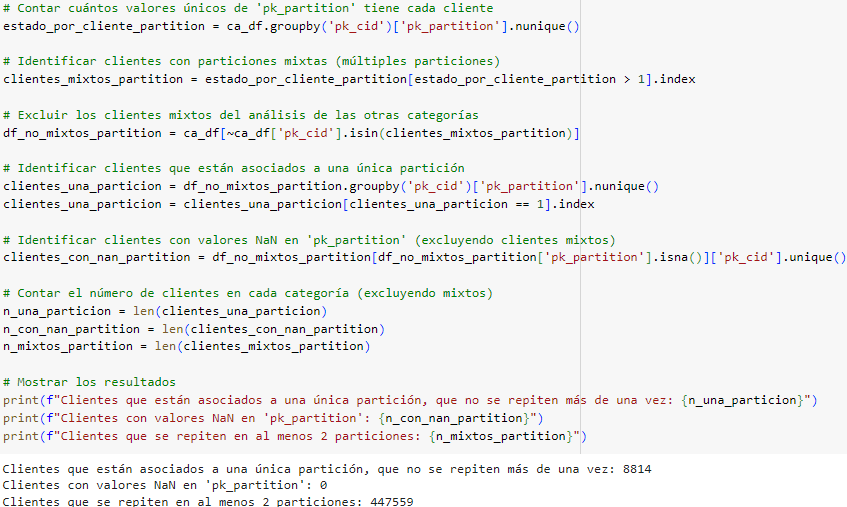


### 1.3.8. Particiones (pk\_partition)

Se realizó un análisis para determinar la presencia de clientes en distintas particiones (pk\_partition). En primer lugar, se contó el número de particiones únicas asociadas a cada cliente. Los resultados mostraron que:

* **8,814** clientes están asociados a una única partición, es decir, no se repiten en más de una.
* **447,559** clientes están presentes en al menos dos particiones, lo que sugiere que son usuarios recurrentes.
* No se encontraron clientes con valores NaN en pk\_partition.

Estos hallazgos resaltan que una gran mayoría de los clientes (aproximadamente el 98%) tiene actividad en múltiples particiones, lo cual es un indicador positivo de continuidad en el uso de la aplicación.

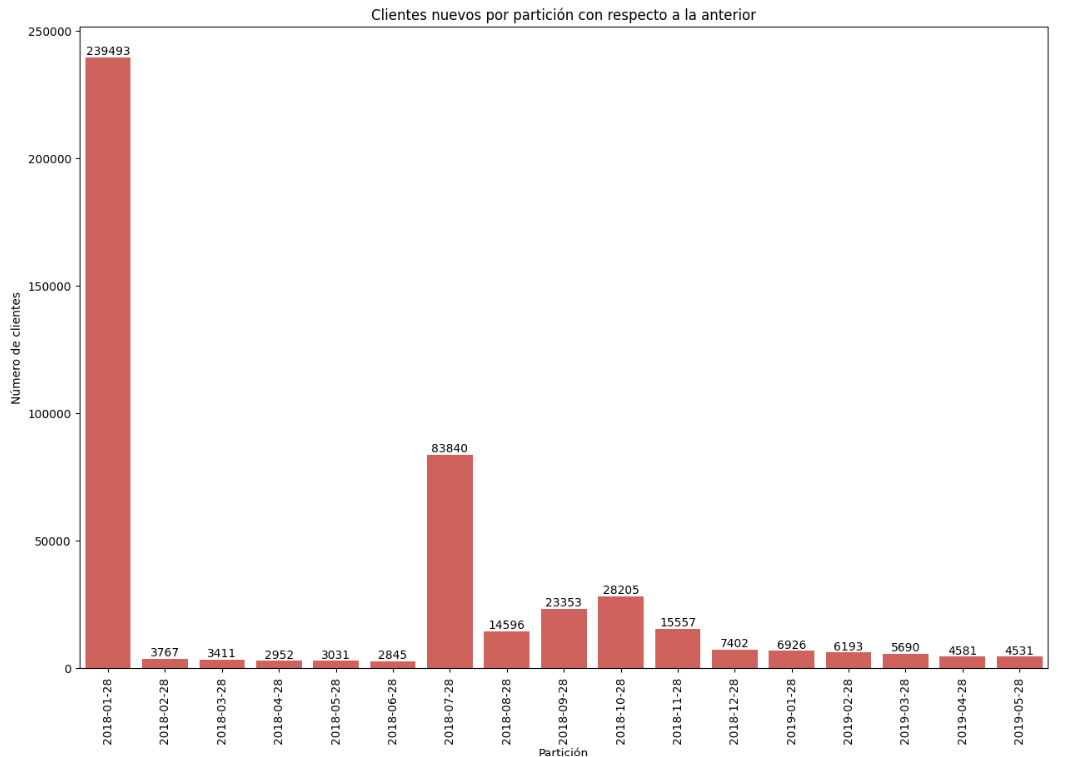


**Identificación de Clientes Nuevos en Cada Partición**

A continuación, se procedió a analizar la adquisición de nuevos clientes en cada partición en comparación con la anterior. Para ello, se ordenaron los datos por pk\_partition y pk\_cid. Se creó un conjunto que almacenaba todos los clientes vistos hasta el momento y, al iterar sobre las particiones, se identificaron los nuevos clientes al restar los clientes actuales de los acumulados.

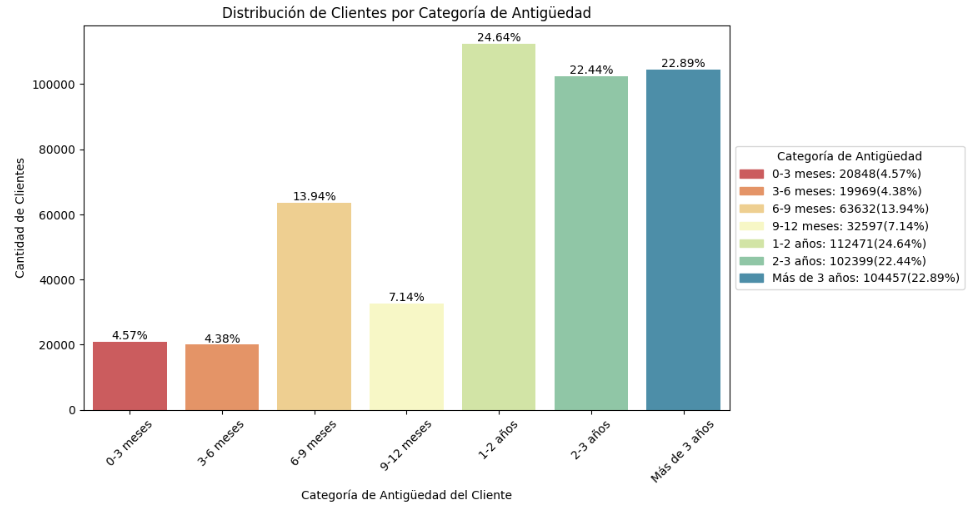
La suma total de nuevos clientes por partición coincide con el número total de clientes únicos en el dataset, **456,373**.

El gráfico resultante muestra picos de adquisición en varias particiones, lo que podría estar vinculado a campañas o eventos específicos. Hacia el final del período analizado, se observa una tendencia a la baja en la adquisición de nuevos clientes, sugiriendo la necesidad de nuevas estrategias para atraer a clientes o una posible saturación del mercado.



**Distribución de Clientes según Antigüedad**

Para complementar el análisis, se evaluó la antigüedad de los clientes en la plataforma. Se calculó la antigüedad en días a partir de la última fecha de partición (pk\_partition) y la fecha de registro (entry\_date), y se categorizó en intervalos de tiempo. Las categorías definidas fueron:



Los resultados mostraron que más del **70%** de los clientes tienen más de un año de antigüedad, lo cual es un indicativo positivo en términos de retención y lealtad. Sin embargo, se identificó una caída notable en la categoría de 9-12 meses en comparación con la de 6-9 meses, lo que sugiere que se debe prestar atención a este grupo para mejorar la retención. Además, la mayor concentración de clientes se encontró en el rango de 1 a 3 años, lo que representa una oportunidad para implementar estrategias de fidelización dirigidas a este segmento.

### 1.3.8. Resumen

1. **Comportamiento de Clientes**: Se clasificaron a los clientes según su actividad:

* **Siempre Activos**: 139,194 clientes.
* **Nunca Activos**: 258,642 clientes.
* **Actividad Variable**: 58,537 clientes. Este análisis revela que el 12% de los clientes cambian su estado de actividad, indicando una necesidad de estrategias para mejorar la retención.

1. **Análisis de entry\_channel**: Se identificaron 124,504 clientes con canales de entrada mixtos (27% del total). Se reemplazaron los valores Unknown con el último canal conocido. Los principales canales son KHE (52.22%) y KFC (14.94%).
2. **Análisis de entry\_date**: Se ajustaron las fechas de registro a la más temprana para los clientes con registros mixtos. Se observó un patrón de adquisición de clientes con picos que podrían estar relacionados con campañas específicas y una estabilización hacia el final del período.
3. **Segmentación de Clientes**: La mayoría de los clientes pertenecen al segmento 03 - UNIVERSITARIO, seguido de 02 - PARTICULARES y 01 - TOP. Se limpiaron los segmentos mixtos y se asignó el segmento más común a los clientes con información faltante.
4. **Clientes por pk\_partition**: Se identificaron 8,814 clientes que solo están presentes en una partición. En total, 447.559 clientes se repiten en más de una partición. El análisis de nuevos clientes en cada partición reveló picos en la adquisición, lo que sugiere que podrían ser necesarios nuevos enfoques para atraer clientes.
5. **Distribución de Clientes según Antigüedad**: Más del 70% de los clientes tienen más de un año de antigüedad. Se identificó una caída en la retención entre los 9-12 meses, lo que sugiere áreas de mejora en la fidelización.

# Tarea 3 - Segmentación de Clientes

Este apartado tiene como objetivo segmentar la base de clientes de *Easy Money* en siete u ocho grupos diferenciados para personalizar sus estrategias comerciales y de marketing. A partir de la segmentación, se busca proporcionar insights que ayuden a la compañía a optimizar la relación con sus clientes y aumentar la efectividad de sus campañas publicitarias. Para ello, se han empleado técnicas avanzadas de análisis de datos, reducción de dimensionalidad y clustering.

## 3.1. Análisis de Variables

Para lograr una segmentación efectiva, se partió de la identificación y selección de las variables que proporcionaran la mayor relevancia en términos de comportamiento de los clientes. Entre las variables seleccionadas se encuentran:

* **Variables demográficas:** Edad, género, nivel educativo, y lugar de residencia. Estas variables son esenciales para diferenciar clientes por etapas de vida, necesidades y preferencias.
* **Comportamiento de compra:** Frecuencia y monto de transacciones, tipo de productos contratados (préstamos, inversión, servicios de pago, etc.). Estas características proporcionan información clave sobre los hábitos de consumo financiero de los clientes.
* **Interacción con productos fintech:** Frecuencia en el uso de productos digitales de la plataforma, como préstamos, créditos rápidos o soluciones de inversión.

La selección de estas variables es crucial ya que permiten una segmentación basada en características tanto estáticas (edad, género) como dinámicas (comportamiento de compra, uso de productos). Este enfoque híbrido garantiza que los segmentos resultantes sean significativos para la estrategia comercial.

Analizando más en detalle la aportación que puede ofrecer cada una de las variables

* **age (Edad):** Es fundamental para entender la etapa de vida en la que se encuentra el cliente, lo que afecta sus necesidades financieras.
* **country\_id (País de residencia):** Podría ser útil para identificar variaciones geográficas en el comportamiento de los clientes.
* **region\_code (Provincia):** Para segmentar a nivel más granular dentro de un país, especialmente si hay diferencias regionales significativas.
* **gender (Sexo):** Aunque menos relevante en algunos casos, podría influir en la preferencia por ciertos productos financieros.
* **salary (Ingresos brutos):** Indica la capacidad de compra e inversión del cliente, crucial para segmentar en términos de valor potencial.
* **active\_customer (Actividad del cliente):** Indicador de si el cliente está activamente utilizando los productos, importante para identificar clientes leales o en riesgo de abandono.
* **entry\_channel (Canal de captación):** Puede influir en el comportamiento inicial y la adopción de productos del cliente.
* **credit\_card, debit\_card (Tarjetas de crédito/débito**): Reflejan el uso de productos financieros cotidianos y la disposición del cliente a utilizar crédito.
* **em\_account\_p, em\_account\_pp, em\_acount, emc\_account (Cuentas de easyMoney):** Estas variables muestran el nivel de compromiso del cliente con los diferentes productos de la empresa, siendo clave para entender la amplitud de su relación con la empresa.
* **loans, mortgage (Préstamos e Hipotecas):** Indican el nivel de deuda y el compromiso financiero a largo plazo del cliente.
* **funds, securities (Fondos y Valores):** Reflejan la predisposición del cliente a invertir y su perfil de riesgo.
* **short\_term\_deposit, long\_term\_deposit (Depósitos a corto y largo plazo**): Muestran las preferencias del cliente por la seguridad y el rendimiento financiero a corto o largo plazo.
* **payroll, payroll\_account (Domiciliaciones y cuentas bonificadas):** Indican el nivel de vinculación del cliente con la empresa a través de la domiciliación de su salario y otros ingresos recurrentes.
* **pension\_plan (Plan de pensiones):** Refleja la planificación a largo plazo del cliente y su preocupación por la jubilación.
* **categoria\_antiguedad:** Muestra la retención del cliente a largo plazo.

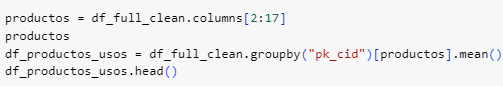
## 3.2. Preprocesamiento

El conjunto de datos utilizado para la segmentación proviene de diversas fuentes, incluyendo información transaccional, características demográficas y el uso de productos financieros. Para asegurar la calidad de los datos, se llevó a cabo un proceso de preprocesamiento exhaustivo.

* **Codificación de variables categóricas:** Todas las variables categóricas, como el canal de entrada y el grupo de edad, se transformaron a formato numérico mediante Frecuency Encoding. Este paso permitió que las categorías fueran correctamente interpretadas por los algoritmos de machine learning sin introducir relaciones artificiales entre las diferentes categorías.
* **Análisis de correlación:** Una vez que las variables categóricas fueron transformadas a formato numérico, se realizó un análisis de correlación para identificar variables altamente correlacionadas. Aunque algunas variables mostraron una correlación superior a 0.9, se decidió mantenerlas en el modelo debido a la posibilidad de que proporcionen insights distintos sobre los clientes.
* **Estandarización de variables numéricas:** Se aplicó el método RobustScaler para estandarizar las variables numéricas, lo cual fue fundamental para asegurar que todas las variables tuvieran el mismo peso durante el proceso de clustering. Este enfoque es especialmente útil en presencia de outliers, ya que el escalado se basa en la mediana y los cuartiles, protegiendo los resultados de distorsiones.

### 3.2.1. Agrupación por Productos Usados

El primer paso fue **agrupar a los clientes según los productos que tienen contratados** a lo largo de diferentes particiones temporales, calculando la media de cada producto por cliente. Esto permite evaluar el grado de uso de cada producto por parte de los clientes de forma precisa.



Este paso permite obtener un perfil detallado del uso de productos financieros por cliente, permitiendo que posteriormente los segmentos puedan diferenciarse en función de la cantidad y tipo de productos contratados. La media de cada producto por cliente refleja la intensidad de uso en cada categoría de productos.

### 3.2.2. Número Máximo de Productos Contratados por Cliente

Luego, se calculó el **número máximo de productos contratados por cada cliente hasta la fecha**. Esto proporciona una métrica clara del compromiso del cliente con la plataforma y el uso de sus servicios.



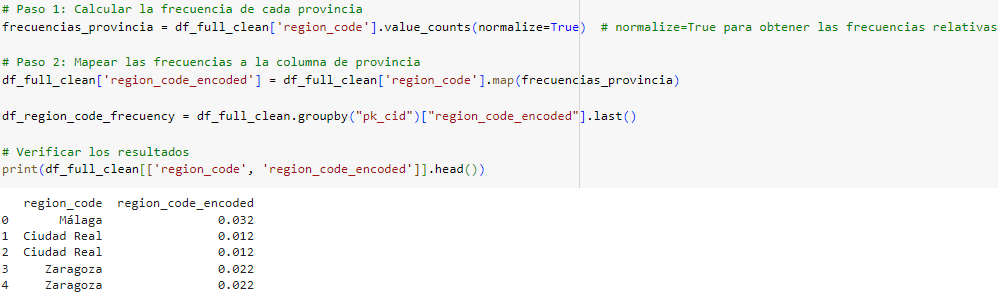
Este cálculo permite identificar a los clientes con un mayor nivel de vinculación con la empresa, ya que aquellos que han contratado más productos pueden considerarse más fieles o comprometidos. Esto es fundamental para detectar segmentos de alto valor o potencial.

### 3.2.3. Tratamiento de la Variable Region Code

Para la columna **region\_code**, se barajaron distintas opciones, incluyendo **One-Hot Encoding** y **Label Encoding**, pero finalmente se decidió utilizar **Frequency Encoding**. Este enfoque se eligió para evitar la alta dimensionalidad y relaciones ordinales incorrectas, asegurando un mejor rendimiento del modelo de K-Means, ya que permite representar la distribución de clientes por región de manera efectiva, sin inflar el número de columnas como sucedería con One-Hot Encoding. Esto reduce la dimensionalidad del dataset y asegura que el modelo capture la densidad relativa de clientes en cada provincia.

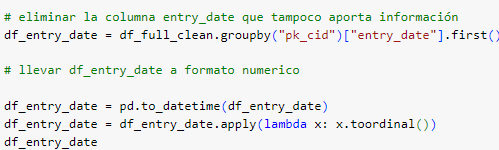
Pasos seguidos:

1. Calcular la frecuencia de cada provincia en el dataset.
2. **Mapear las frecuencias** calculadas a cada cliente en función de su provincia de residencia.



### 3.2.4. Transformación de la Fecha de Ingreso

Se transformó la columna **entry\_date** a un formato numérico, ya que, si bien la fecha original no aporta información directa, al convertirla en su representación ordinal (número de días desde una fecha base), se puede utilizar en los modelos de machine learning. La transformación de fechas a un formato numérico permite incluir el factor tiempo en el análisis. En este caso, la fecha de ingreso de un cliente puede influir en su nivel de compromiso con la empresa o en el uso de productos financieros, por lo que es importante incluirla de una forma que sea compatible con los modelos de machine learning.



### 3.2.5. Eliminación de Columnas Irrelevantes

Algunas columnas fueron eliminadas por no aportar valor suficiente para el análisis de segmentación. Entre ellas, las columnas **country\_id**, **deceased** y **em\_account\_pp**, las cuales no brindaban información relevante o habían sido tratadas previamente en la tarea 1. En este caso, estas variables no aportaban información adicional para la segmentación y, por tanto, su eliminación simplifica el análisis sin perder valor.

### 3.2.6. Cálculo del Grado Medio de Actividad

Para la variable **active\_customer**, se agrupó por clientes y se calculó la media, lo que indica el grado promedio de actividad de los clientes en la plataforma. Este indicador refleja cuán activos son los clientes en el uso de los productos y servicios de *Easy Money*, proporcionando una métrica importante para diferenciar entre clientes que son usuarios frecuentes y aquellos que pueden estar en riesgo de abandono.



### 3.2.7. Limpieza de Variables Finales

Para el resto de las variables que ya venían preprocesadas en tareas anteriores, se seleccionaron los valores más recientes (de la última partición) para mantener una única fila por cliente. Las variables seleccionadas fueron **gender, salary, mes\_partition, grupo\_edad, age, entry\_channel, segment, categoria\_antiguedad**.

Este último paso asegura que se mantenga la información más actualizada de cada cliente, necesaria para la segmentación. Mantener una sola fila por cliente facilita la implementación del modelo de clustering y asegura que se estén utilizando datos consistentes y recientes.

## 3.3. Matriz de Correlación y Variables Altamente Correlacionadas

Una vez que todas las variables fueron convertidas a un formato numérico, se calculó la **matriz de correlación**, la cual muestra las relaciones lineales entre las variables. Para visualizar la matriz, se generó un **heatmap** utilizando la biblioteca Seaborn, lo cual permite identificar fácilmente las variables que están altamente correlacionadas (aquellas con un coeficiente de correlación mayor a 0.9).



Se fijó un **umbral de 0.9** para considerar variables como altamente correlacionadas. En este caso, las parejas de variables que superaban este umbral fueron:

* 'payroll' y 'pension\_plan'
* 'segment\_02 - PARTICULARES' y 'segment\_03 - UNIVERSITARIO'

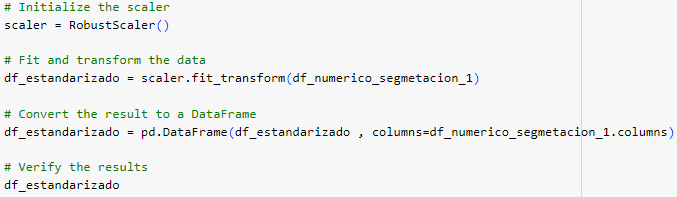
A pesar de estar correlacionadas, ambas parejas de variables fueron conservadas en el análisis, ya que representan aspectos distintos del comportamiento de los clientes que podrían ofrecer insights valiosos.



## 3.4. Estandarización de las Variables Numéricas

Una vez realizadas las transformaciones y el análisis de correlación, se procedió a la **estandarización de las variables numéricas**. Este proceso es necesario para que todas las variables tengan el mismo rango y no influyan desproporcionadamente en el modelo de clustering. Para ello, se utilizó **RobustScaler** de Scikit-learn. **RobustScaler** es particularmente útil cuando hay outliers en los datos, ya que utiliza la mediana y los cuartiles para escalar las variables, lo que lo hace más robusto ante la existencia de valores extremos.

La estandarización es un paso crucial para algoritmos como **K-Means**, que dependen de las distancias entre los puntos. Sin estandarización, las variables con valores absolutos más grandes (como salarios o montos financieros) podrían dominar el análisis y hacer que el clustering se sesgue hacia ellas.



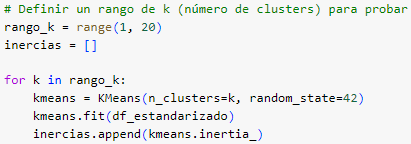
## 3.5. Ejecución del Clustering

En esta sección, se evaluaron tres enfoques distintos para el clustering de los datos de clientes de *Easy Money*: clustering con el dataset completo, clustering con reducción de dimensionalidad mediante PCA y clustering con ingeniería de características. Cada uno de estos enfoques fue diseñado para evaluar la calidad de la segmentación obtenida en términos de inercia y la puntuación de silueta, con el objetivo de determinar cuál enfoque proporciona clusters más claros y útiles para el análisis comercial.

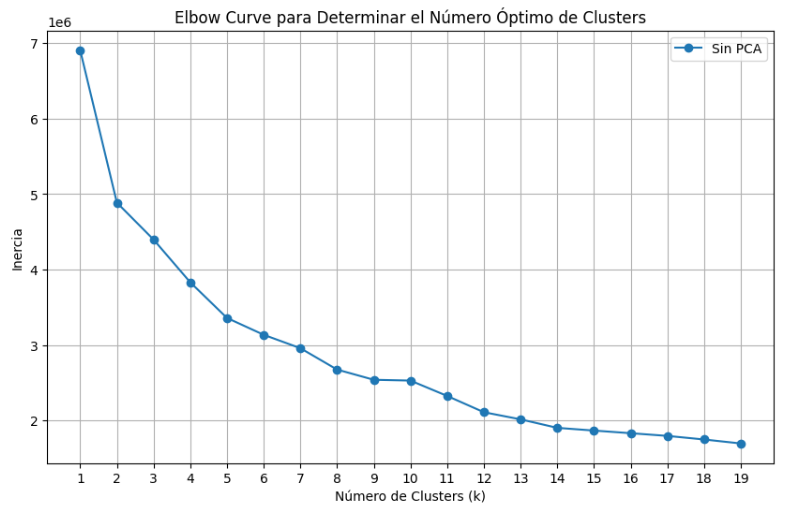
### 3.5.1. Clustering con el Dataset Completo

Este enfoque aplicó el algoritmo de **K-Means** directamente sobre el dataset estandarizado completo (sin reducción de dimensionalidad ni transformación adicional de las variables).

* **Definición de un rango de clusters:** Se probaron diferentes valores de **k** (número de clusters), variando entre 1 y 20, para encontrar el número óptimo de clusters. Para cada iteración, se calculó la **inercia** óptima para segmentar los datos.



* **Curva del Codo:** La gráfica de la **curva del codo** mostró que la inercia disminuía rápidamente al incrementar el número de clusters hasta **k = 5**. A partir de ese punto, la inercia se estabilizó, lo que indica que 5 clusters podrían ser suficientes para capturar la estructura de los datos.



Basado en el análisis, se seleccionó **5 clusters** como el número óptimo en este enfoque. Este resultado sugiere que el dataset completo estandarizado puede ser segmentado de manera eficiente en 5 grupos sin necesidad de reducción de dimensionalidad ni ingeniería de características adicionales.

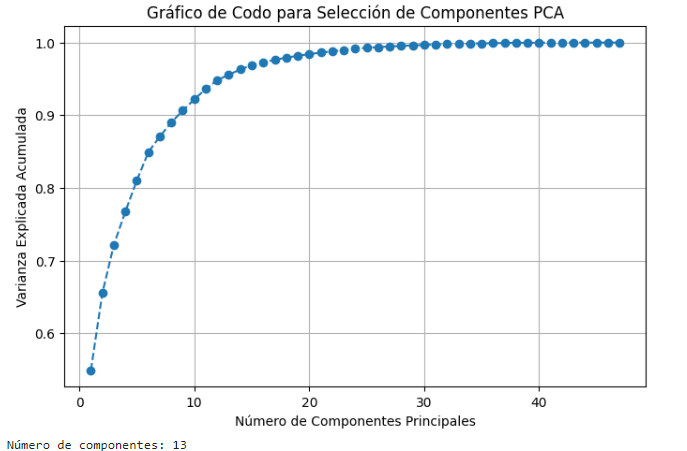
### 3.5.2. Clustering con Reducción de Dimensionalidad mediante PCA

En este enfoque, se utilizó **Análisis de Componentes Principales (PCA)** para reducir la dimensionalidad del dataset antes de aplicar el clustering. Este método busca simplificar los datos al identificar los componentes principales que capturan la mayor parte de la varianza.



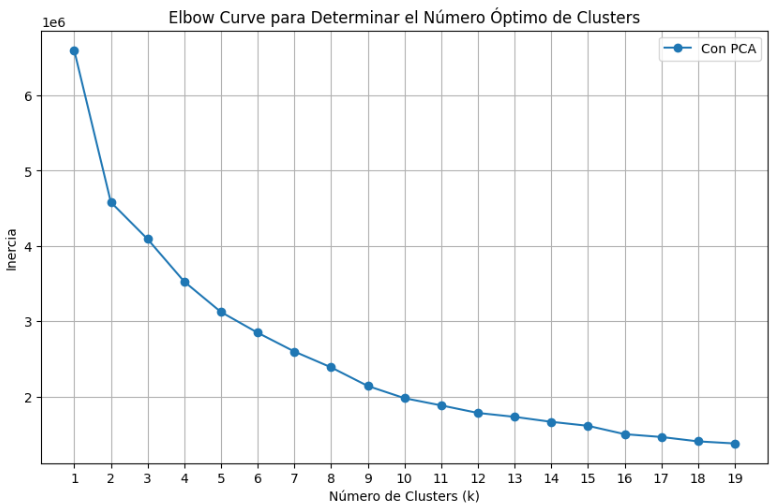


* **Selección del número de componentes:** Se seleccionaron suficientes componentes principales para retener el **95% de la varianza** de los datos originales, lo que redujo el número de variables de 47 a **13 componentes** principales.





* **Curva del Codo para PCA:** Al igual que en el enfoque anterior, se aplicó el algoritmo de **K-Means** sobre el dataset reducido a 13 componentes principales, probando un rango de valores para **k** entre 1 y 20. El gráfico del codo mostró una reducción en la inercia similar a la observada sin PCA, pero con clusters más compactos, lo que sugiere que los grupos resultantes son más coherentes.



Aunque los resultados son similares al clustering sin PCA, la reducción de dimensionalidad permitió que los clusters fueran más compactos y eficientes. El análisis sugiere que **4 o 5 clusters** es el número óptimo, con una notable reducción en la inercia comparado con el dataset completo.

### 3.5.3. Clustering con Ingeniería de Características

En este enfoque, se realizó **ingeniería de características** para crear nuevas variables a partir de las originales, con el objetivo de mejorar la segmentación de clientes. El proceso consistió en agrupar productos financieros, categorizar canales de entrada, clasificar los salarios y eliminar variables redundantes, entre otros pasos. Tras realizar estos ajustes, se procedió al análisis de correlación, estandarización y aplicación del clustering.

Agrupación de Productos Financieros

Se crearon nuevas columnas que agrupan los productos en tres categorías clave: **cuentas**, **ahorro/inversión** y **financiación**. Estas agrupaciones permitieron simplificar la representación de los productos financieros sin perder información relevante.

* **Cuentas:** Se agruparon productos como cuentas de depósito a corto y largo plazo, cuentas de nómina y cuentas de ahorro.
* **Ahorro/inversión:** Incluye productos de fondos, valores y planes de pensiones.
* **Financiación:** Se agrupan productos como préstamos, tarjetas de crédito, hipotecas y tarjetas de débito.

La agrupación de productos simplifica el modelo y permite capturar de manera más general el comportamiento financiero de los clientes sin perder detalle relevante.

Categorías de Canales de Entrada

Se decidió agrupar los canales de entrada en **principales** y **secundarios**, lo que permite reducir la dimensionalidad y hacer más manejable el análisis.

Los canales principales seleccionados fueron: 'KHE', 'KHQ', 'KFC'. Todos los demás canales se clasificaron como secundarios. Agrupar los canales de entrada ayuda a reducir el número de categorías mientras se preserva la información sobre el origen de los clientes, facilitando la identificación de patrones de comportamiento en función del canal de ingreso.

Eliminación de Variables Redundantes

Se eliminaron varias variables que podrían introducir redundancias o sesgos en el modelo. Las variables eliminadas incluyen: **segment, grupo\_edad, gender, region\_code\_encoded**

Eliminar variables redundantes es una práctica común para evitar la multicolinealidad y mejorar la eficiencia del modelo de clustering. Estas variables no aportaban nueva información significativa tras la creación de nuevas características.

Codificación de la Antigüedad

La variable **categoria\_antiguedad** se codificó en función de la frecuencia de cada categoría en el dataset. Codificar las categorías de antigüedad mediante **frequency encoding** permite capturar la variabilidad en la distribución de los clientes sin inflar el número de columnas, lo que mantiene el modelo más compacto.

Clasificación de Salarios

La columna de **salary** se transformó en una variable categórica, dividiendo los salarios en tres rangos: **Ingreso Bajo**, **Ingreso Medio** e **Ingreso Alto**, según los percentiles 33% y 66%.

Clasificar los salarios en categorías simplifica el análisis y permite evaluar cómo los ingresos influyen en la segmentación de clientes, además de evitar que el salario tenga un impacto desproporcionado en el clustering.

Conversión de Variables Categóricas a Numéricas

Las variables categóricas restantes se convirtieron a formato numérico mediante **One-Hot Encoding** para que pudieran ser procesadas por el algoritmo de clustering.

El **One-Hot Encoding** asegura que las variables categóricas sean tratadas correctamente por el algoritmo de machine learning, sin introducir relaciones artificiales entre categorías.

Análisis de Correlación y Estandarización

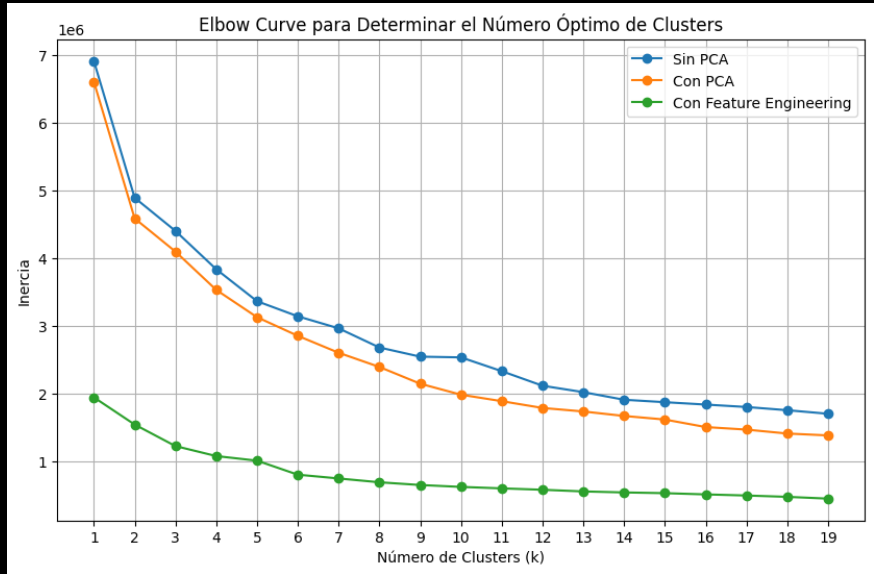
Se calculó la **matriz de correlación** para identificar variables altamente correlacionadas, que podrían ser eliminadas si presentaban redundancia. En este caso, no se identificaron variables altamente correlacionadas (con un umbral mayor a 0.9).

En esta sección se aplicó el algoritmo **K-Means** utilizando los tres enfoques previamente definidos: el **dataset original**, el dataset con **reducción de dimensionalidad mediante PCA**, y el dataset con **ingeniería de características**. El objetivo fue identificar el número óptimo de clusters que proporcionara una segmentación clara y manejable.

### 3.5.4. Definición del Rango de Clusters

Para cada uno de los tres enfoques, se probó un rango de **k** entre 1 y 20, calculando la **inercia** para cada valor de clusters. A continuación, se presenta un análisis comparativo basado en las inercias obtenidas y representadas en la **curva del codo**.

La **curva del codo** mostró el punto en el que agregar más clusters deja de reducir significativamente la inercia, indicando que los clusters adicionales no aportan información relevante. A continuación, se observa la visualización de la curva del codo para los tres enfoques:



* **Clustering sin PCA:** Proporciona una segmentación razonable con 5 clusters, pero los resultados son menos compactos en comparación con los otros enfoques.
* **Clustering con PCA:** Mejora la compactación de los clusters sin perder demasiada información, mostrando que 4 o 5 clusters son suficientes para capturar la estructura de los datos.
* **Clustering con Feature Engineering:** Este enfoque muestra la mayor reducción de inercia, lo que indica que las nuevas variables son más efectivas para segmentar a los clientes, con 5 a 7 clusters recomendados.

En conclusión, la **ingeniería de características** resultó ser el enfoque más efectivo para agrupar a los clientes, seguido por el uso de **PCA**. Ambos enfoques permitieron una segmentación más compacta y coherente, facilitando la interpretación y la aplicación de los resultados en estrategias comerciales.

## 3.6. Interpretación del Clustering

Una vez completado el proceso de clustering, es necesario realizar una interpretación y un análisis de los resultados obtenidos para seleccionar el número adecuado de clusters y comprender las características principales de cada grupo. El análisis incluyó la comparación de resultados con **4, 5 y 6 clusters**, y se tomó la decisión final de trabajar con **6 clusters**, que ofrecían el equilibrio adecuado entre detalle y manejabilidad.

### 3.6.1. Selección del Número de Clusters

El análisis de la **curva del codo** indicó que los valores óptimos para el número de clusters estaban entre 5 y 6, pero la elección final de **6 clusters** se basó en la observación de la distribución y el tamaño de los grupos.

* **Aplicación de K-Means con diferentes clusters:** Se probaron **4, 5 y 6 clusters** y se analizó la distribución de los clientes en cada uno.
* **Distribución de clientes:** El número de clientes por cluster fue evaluado para ver si cada cluster era lo suficientemente grande y balanceado para generar insights significativos. Al trabajar con 6 clusters, se observó una distribución equilibrada que facilitaba tanto la interpretación como la implementación de estrategias comerciales.



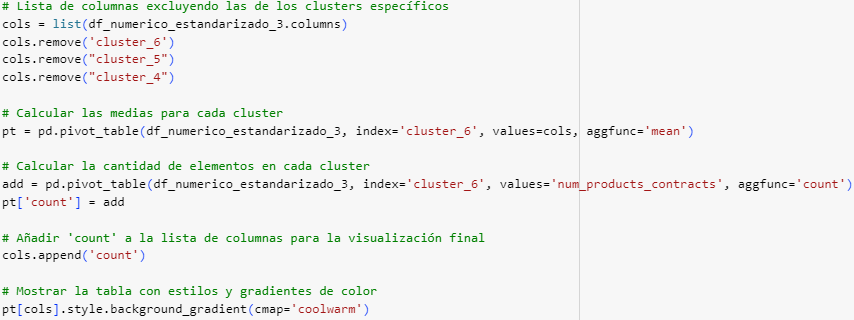
* **Decisión de trabajar con 6 clusters:** Aunque la Directora General (Carol) había propuesto dividir los clientes en 7 u 8 grupos, se optó por trabajar con **6 clusters** porque la adición de más clusters no mejoraba significativamente la reducción de la inercia. Además, la distribución equilibrada en 6 grupos proporcionaba una base sólida para la segmentación, con clusters manejables y diferenciados.

Se seleccionó el número de **6 clusters** para obtener un balance adecuado entre la utilidad de la segmentación y la posible implementación. Este enfoque permite que los grupos sean lo suficientemente grandes para generar insights valiosos sin perder especificidad.

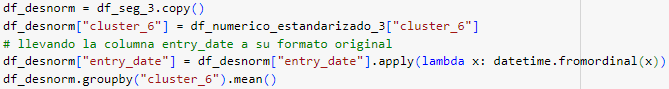
### 3.6.2. Resumen de Clusters y Características Medias

Se calcularon las **medias de las características** por cluster, lo que permitió identificar los perfiles de cada grupo, así como la cantidad de clientes en cada cluster.

* **Cálculo de las medias por cluster:** Se generó un resumen de las principales características de los clientes agrupados en cada cluster, incluyendo la cantidad de productos contratados, la actividad de los clientes y otros factores relevantes.



* **Desnormalización de las variables:** Para una mejor interpretación de los resultados, algunas variables, como **entry\_date**, fueron desnormalizadas y transformadas a su formato original.



### 3.6.3. Interpretación de los Clusters

A continuación, se describen los 6 clusters, sus características principales y las estrategias recomendadas para cada grupo:

Cluster 0: Jóvenes de Baja Actividad

* Edad media: 25.87 años.
* Productos contratados: 1.093 productos.
* Clientes activos: 28.4%.
* Perfil: Jóvenes con baja actividad y bajo compromiso con la entidad. Son clientes de bajo valor con poca participación en productos de financiación.
* Estrategias: Incentivar el uso de los servicios mediante programas de recompensas y ofertas personalizadas.

Cluster 1: Nuevos Clientes de Bajo Compromiso

* Edad media: 35.2 años.
* Productos contratados: 0.485 productos.
* Clientes activos: 16.6%.
* Perfil: Clientes de mediana edad, que ingresaron recientemente con baja actividad. La mayoría ingresó por canales secundarios.
* Estrategias: Focalizar en campañas de enganche personalizadas y mejorar la experiencia digital.

Cluster 2: Jóvenes Inactivos Sin Productos

* Edad media: 28.88 años.
* Productos contratados: 0.039 productos.
* Clientes activos: 3.7%.
* Perfil: Jóvenes con muy poca actividad y prácticamente ningún producto contratado. Están en riesgo de abandono.
* Estrategias: Ofrecer incentivos para reactivar la actividad, como descuentos y promociones.

Cluster 3: Jóvenes Activos en Crecimiento

* Edad media: 23.27 años.
* Productos contratados: 1.08 productos.
* Clientes activos: 40.6%.
* Perfil: El grupo más joven, con un alto nivel de actividad y potencial de crecimiento en el uso de productos financieros.
* Estrategias: Desarrollar programas educativos financieros y promociones adaptadas a sus necesidades.

Cluster 4: Clientes Premium Altamente Comprometidos

* Edad media: 37.41 años.
* Productos contratados: 3.871 productos.
* Clientes activos: 94.5%.
* Perfil: Clientes con un alto compromiso y gran valor. Son usuarios intensivos de productos financieros complejos.
* Estrategias: Mantener la lealtad mediante programas de fidelización y asesoría personalizada.

Cluster 5: Seniors Comprometidos con Potencial

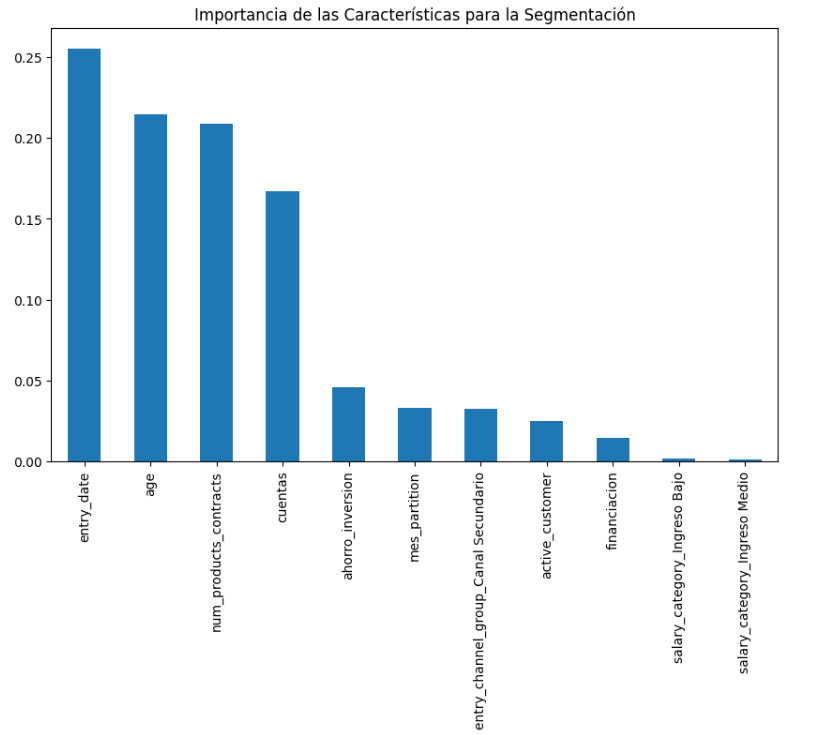
* Edad media: 49.63 años.
* Productos contratados: 1.401 productos.
* Clientes activos: 75.4%.
* Perfil: Clientes de mayor edad, con alta actividad y un nivel moderado de contratación de productos. Tienen potencial para adoptar productos más complejos.
* Estrategias: Ofrecer productos financieros adaptados y organizar eventos que fortalezcan la relación con la entidad.

### 3.6.4. Análisis de la Importancia de Características

Para identificar las características más importantes que diferencian los clusters, se utilizó un modelo de **Random Forest**, que permitió evaluar el peso relativo de cada variable en la segmentación.

Las variables con mayor peso relativo en la segmentación de clientes son:

* **entry\_date** y **age** resultaron ser las variables más relevantes para la segmentación, lo que indica que el comportamiento de los clientes varía en función de cuándo ingresaron al sistema y su edad.
* **num\_products\_contracts** (número de productos contratados) también fue una variable importante, reflejando el nivel de compromiso de los clientes con la entidad.
* Variables como **cuentas**, **ahorro\_inversion** y **active\_customer** también aportaron información relevante, aunque en menor medida.



# Tarea 5 – Seguimiento de Campaña

El éxito de una campaña de marketing digital depende de la capacidad de medir, interpretar y actuar sobre los datos de desempeño obtenidos. En el caso de EasyMoney, la campaña de marketing por correo electrónico dirigida a 10,000 clientes segmentados tiene como objetivo impulsar la adopción de un producto financiero clave: la tarjeta de crédito **EasyMoney**. Dado que la segmentación se ha realizado de manera cuidadosa, agrupando a los clientes en cinco clústeres distintos basados en características demográficas, socioeconómicas y de comportamiento, la medición del éxito de la campaña es fundamental para garantizar su efectividad.

Este plan de medición tiene el propósito de definir las métricas clave (KPIs) que permitirán evaluar de manera rigurosa la efectividad de la campaña. Los KPIs proporcionarán una visión cuantificable del rendimiento de la campaña, permitiendo ajustar las estrategias en función de los resultados obtenidos. Los indicadores seleccionados miden aspectos cruciales como la interacción con el contenido del correo, la conversión de clientes potenciales y el retorno económico, asegurando que la campaña no solo logre un impacto inmediato, sino que también fomente una relación a largo plazo con los clientes de EasyMoney.

## 5.1. KPIs (Key Performance Indicators)

### 5.1.1. Tasa de Apertura del Correo Electrónico (Open Rate)

La tasa de apertura mide el porcentaje de correos electrónicos que fueron abiertos en relación con el total de correos enviados. Es un indicador temprano de la efectividad del asunto y la relevancia del mensaje para el cliente.

Este KPI es esencial para evaluar si los correos electrónicos están captando la atención de los clientes en los diferentes clústeres. Una tasa de apertura alta indicará que los mensajes son percibidos como relevantes y atractivos, lo cual es clave para inducir las siguientes interacciones.

### 5.1.2. Tasa de Clics (Click-Through Rate, CTR)

El CTR mide el porcentaje de clientes que hicieron clic en algún enlace dentro del correo electrónico en relación con el total de correos abiertos. Este indicador proporciona una visión clara sobre el nivel de interés generado por el contenido del correo y si el llamado a la acción (CTA) fue efectivo.

El CTR es especialmente relevante cuando se trata de la venta de productos financieros, ya que el objetivo es motivar a los clientes a actuar, ya sea visitando la página de la tarjeta de crédito o iniciando el proceso de solicitud.

### 5.1.3. Tasa de Conversión (Conversion Rate)

La tasa de conversión mide el porcentaje de clientes que completaron la acción deseada, en este caso, solicitar la tarjeta de crédito, en relación con el total de correos abiertos. Este KPI es uno de los más críticos, ya que indica el éxito real de la campaña en términos de generación de ventas.

Es importante medir esta métrica por separado para cada clúster, ya que la efectividad de los mensajes puede variar según las características demográficas y socioeconómicas de cada grupo.

### 5.1.4. Valor de Vida del Cliente (Customer Lifetime Value, CLV)

El CLV mide el valor económico total que un cliente aportará a la empresa a lo largo de su relación con la misma. Este KPI no solo se centra en la conversión inmediata, sino también en el impacto a largo plazo que tendrá la adquisición de nuevos clientes de tarjetas de crédito.

El CLV ayuda a determinar si el costo de adquirir un cliente nuevo es justificable con respecto al valor total que ese cliente aportará. Es un KPI clave en una campaña de productos financieros, ya que se debe analizar el impacto a largo plazo más allá de la compra inicial.

\* La tasa de Churn, también llamado tasa de cancelación, mide la rapidez con la que una empresa pierde clientes dentro de un plazo determinado. En el contexto de las aplicaciones móviles, representa el porcentaje de usuarios que han dejado de usar tu aplicación, ya sea que la desinstalen o cancelen su suscripción.

### 5.1.5. Retorno sobre la Inversión (ROI)

El ROI mide la rentabilidad de la campaña en términos del retorno generado por cada unidad monetaria invertida en la misma. En el caso de EasyMoney, es esencial conocer si la inversión en la campaña está generando un retorno positivo.

El ROI es un indicador clave para evaluar el éxito global de la campaña, especialmente cuando se utiliza para la venta de productos financieros, donde el costo de adquisición de clientes puede ser alto.

### 5.1.6. Retención de Clientes

La retención de clientes mide el porcentaje de clientes que, después de haber solicitado la tarjeta de crédito, siguen activos y utilizando los productos de EasyMoney a lo largo del tiempo. La retención es un KPI importante, ya que refleja la lealtad y el valor a largo plazo del cliente.

Este indicador es especialmente relevante para productos financieros como las tarjetas de crédito, donde el éxito no solo se mide por la adquisición, sino también por la utilización continua.

### 5.1.7. Tasa de Adopción del Producto

Este KPI mide la tasa a la cual los clientes que han recibido la tarjeta de crédito comienzan a utilizarla para realizar compras o acceder a los beneficios. Es crucial para medir no solo la conversión inicial, sino también el uso activo del producto.

Este KPI es vital para asegurar que la campaña no solo genere solicitudes de tarjetas, sino que también promueva su uso continuo.

### 5.2 Estrategia de Medición Ampliada

La estrategia de medición automatizada debe permitir un seguimiento detallado de las interacciones con los correos electrónicos y sus efectos en los comportamientos de los clientes. La implementación de herramientas adecuadas para el análisis en tiempo real es fundamental para hacer ajustes oportunos y asegurar el máximo rendimiento de la campaña.

### 5.2.1 Pre-lanzamiento

Configuración de herramientas de análisis y segmentación:

Integración CRM y Email Marketing: Es esencial que las plataformas de email marketing se sincronicen con un sistema de CRM (Customer Relationship Management) para permitir un seguimiento del comportamiento del cliente, como aperturas, clics y conversiones, en un solo sistema. Este paso garantiza que se pueda trazar un perfil detallado de cada clúster.

Definición de umbrales de éxito inicial: Se deberán establecer valores umbral para cada KPI antes de iniciar la campaña. Por ejemplo, una tasa de apertura mínima esperada para ciertos clústeres o un CTR promedio que defina éxito o necesidad de ajuste.

Segmentación dinámica: Basada en los clústeres predefinidos, el CRM debe estar configurado para ajustar automáticamente la segmentación si se detecta un cambio en las interacciones con el email, permitiendo la entrega de contenidos más personalizados en tiempo real.

### 5.2.2 Durante la campaña

Monitoreo diario de KPIs clave por clúster:

Visualización de datos en tiempo real: Es esencial utilizar paneles de control que permitan la visualización en tiempo real de KPIs como tasa de apertura, CTR, conversión y probabilidad de compra para cada clúster. Esto permitirá identificar tendencias emergentes.

Ajustes en tiempo real:

A/B Testing continuo: Las pruebas A/B deben ejecutarse durante toda la campaña, modificando el asunto del correo, las imágenes o el contenido del cuerpo según la interacción de cada clúster. El sistema debe poder adaptar el contenido en base a los resultados de rendimiento observados.

Personalización y segmentación: Si un clúster específico muestra una tasa de apertura baja o un CTR reducido, se puede ajustar en tiempo real el mensaje o las ofertas mostradas para mejorar el rendimiento.

Alertas automáticas:

Configuración de alertas automáticas cuando ciertos KPIs caen por debajo de los valores predeterminados o si alguno de los clústeres muestra un comportamiento anómalo, como una caída drástica en la tasa de conversión o un aumento inesperado de cancelaciones de suscripción.

### 5.2.3 Post-campaña

Evaluación exhaustiva de datos:

Análisis por clúster: Se debe realizar un análisis detallado de los resultados por clúster, evaluando no solo el rendimiento en términos de KPIs tradicionales como CTR y tasa de conversión, sino también métricas más profundas como valor del ciclo de vida del cliente (CLV) y retención.

Segmentación de ingresos y CLV: Es importante medir el retorno sobre la inversión (ROI) para cada clúster, calculando cuánto ha generado cada uno en términos de ingresos directos e indirectos. Además, se deben analizar las diferencias en el CLV por clúster para detectar cuáles podrían ser los grupos más rentables a largo plazo.

Evaluación del impacto a largo plazo:

Retención a 3 y 6 meses: Para evaluar el impacto duradero de la campaña, es clave analizar la retención de clientes a los 3 y 6 meses después de la campaña. ¿Se han mantenido más activos aquellos que recibieron la oferta? ¿Cómo varía la retención entre clústeres?

Lecciones aprendidas y optimización futura:

Optimización para campañas futuras: A partir de los resultados, se identificarán las estrategias más efectivas para cada clúster. Se propondrán mejoras para la próxima campaña, ajustando las estrategias de personalización, el uso de canales adicionales (como redes sociales o notificaciones push) y cualquier otra acción que mejore los resultados.

### 5.2.4 Adicional: Métricas Avanzadas

Análisis predictivo:

Utilizar técnicas de machine learning para predecir el comportamiento futuro de los clústeres basándose en los datos recopilados durante la campaña. Esto puede incluir modelos de regresión o clasificación para identificar los clientes con mayor probabilidad de conversión o abandono.

### 5.3. KPI Adaptados a los Clústeres

Clúster 0: "Urban Explorers"

* Enfoque en KPI: Tasa de apertura y conversión.
* Justificación: Este clúster tiene una probabilidad de compra media, y su poder adquisitivo es bajo. El objetivo será atraerlos con beneficios inmediatos, como descuentos locales.

Clúster 1: "City High-Flyers"

* Enfoque en KPI: Tasa de clics e ingreso promedio por cliente.
* Justificación: Dada la alta probabilidad de compra y los ingresos elevados, el foco será maximizar los ingresos generados por este segmento.

Clúster 2: "Balanced Lifestyles"

* Enfoque en KPI: Tasa de conversión y retención.
* Justificación: Este clúster tiene una alta probabilidad de compra, pero el objetivo a largo plazo será la retención, enfocándose en el valor a largo plazo.

Clúster 3: "Cultural Trendsetters"

* Enfoque en KPI: Tasa de apertura y retención.
* Justificación: Los clientes de este clúster son culturalmente sofisticados, y la campaña debe enfocarse en mantener su interés con beneficios que aseguren su lealtad.

Clúster 4: "Elite Professionals"

* Enfoque en KPI: ROI y CLV.
* Justificación: Debido a su perfil de altos ingresos y gran probabilidad de compra, el objetivo será maximizar el valor a largo plazo de este segmento.