# 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.

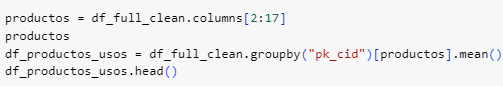
## 3.2. Preprocesamiento

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.



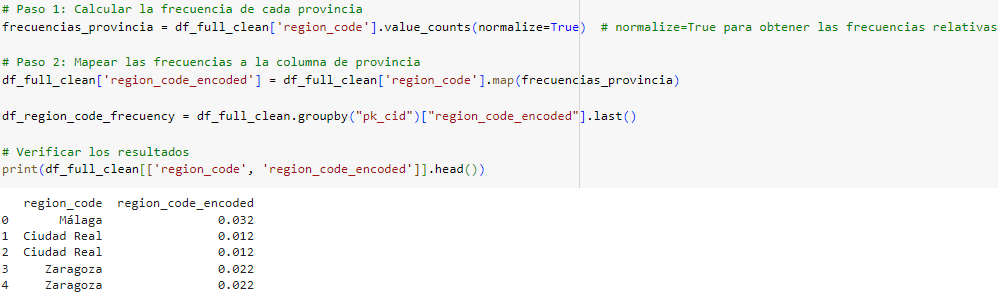
Este cálculo permite identificar a los clientes con un mayor nivel de vinculación con la empresa, ya que aquellos que han contratado más productos pueden considerarse más fieles o comprometidos. Esto es fundamental para detectar segmentos de alto valor o potencial.

### 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.

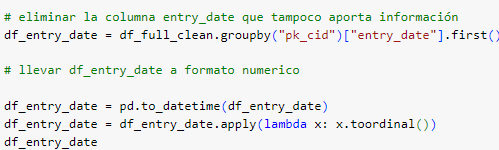
Pasos seguidos:

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'**

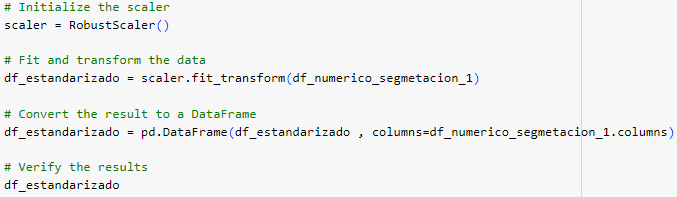
A pesar de estar correlacionadas, ambas parejas de variables fueron conservadas en el análisis, ya que representan aspectos distintos del comportamiento de los clientes que podrían ofrecer insights valiosos.



## 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.



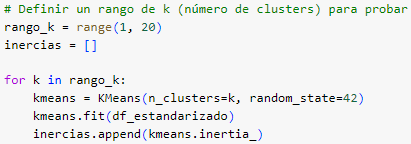
## 3.5. Ejecución del Clustering

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.

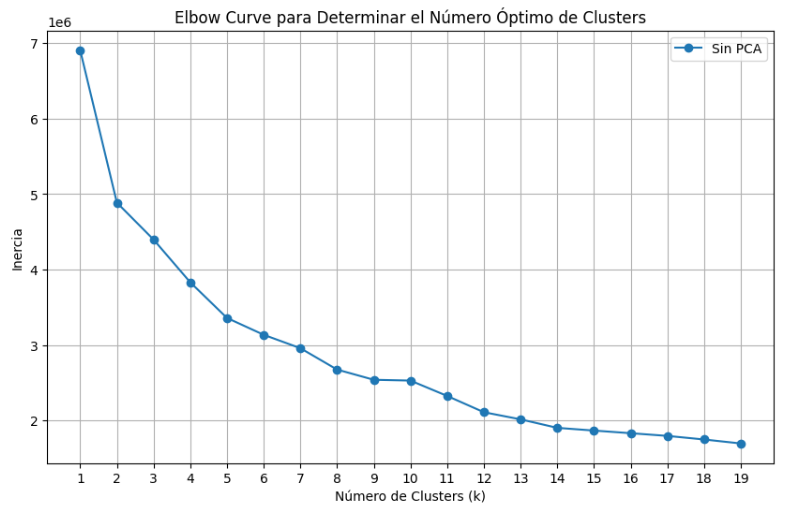
### 3.5.1. Clustering con el Dataset Completo

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.

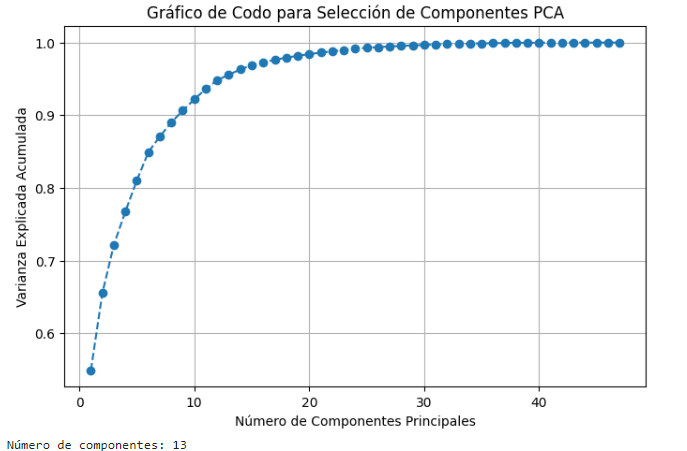
### 3.5.2. Clustering con Reducción de Dimensionalidad mediante PCA

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.



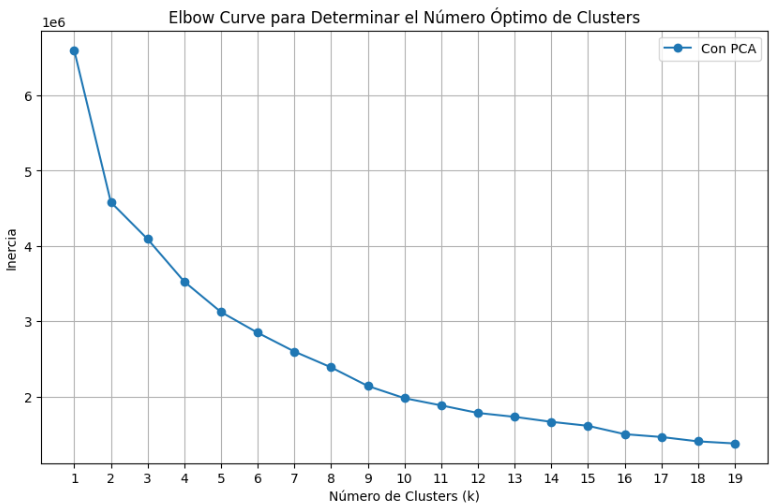


* **Selección del número de componentes:** Se seleccionaron suficientes componentes principales para retener el **95% de la varianza** de los datos originales, lo que redujo el número de variables de 47 a **13 componentes** principales.





* **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**

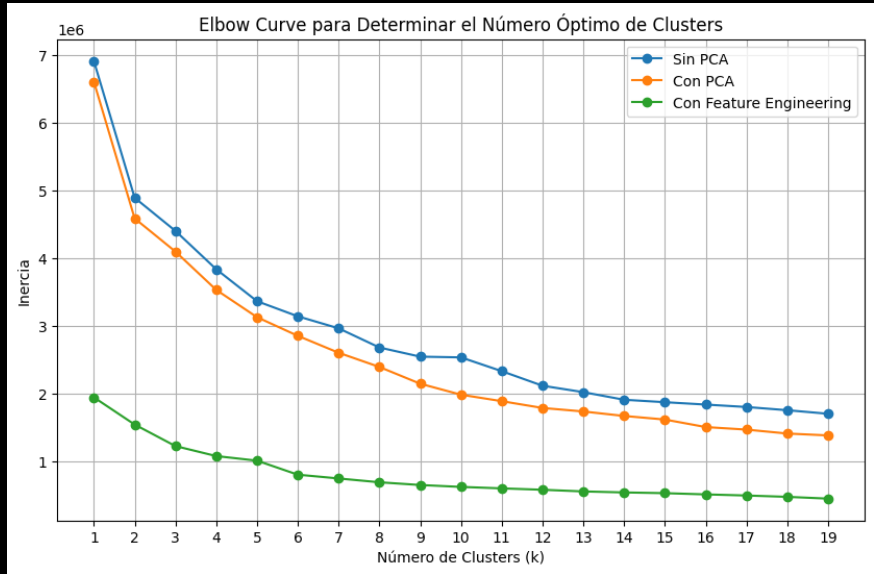
Se calculó la **matriz de correlación** para identificar variables altamente correlacionadas, que podrían ser eliminadas si presentaban redundancia. En este caso, no se identificaron variables altamente correlacionadas (con un umbral mayor a 0.9).

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 de un 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.



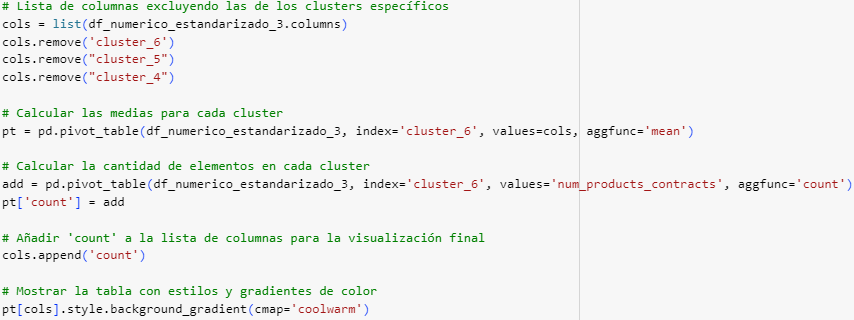
* **Decisión de trabajar con 6 clusters:** Aunque la Directora General (Carol) había propuesto dividir los clientes en 7 u 8 grupos, se optó por trabajar con **6 clusters** porque la adición de más clusters no mejoraba significativamente la reducción de la inercia. Además, la distribución equilibrada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

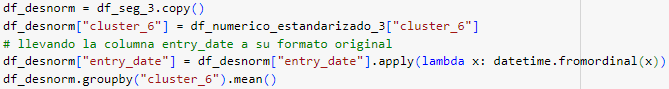
### 3.6.2. Resumen de Clusters y Características Medias

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 las 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.

