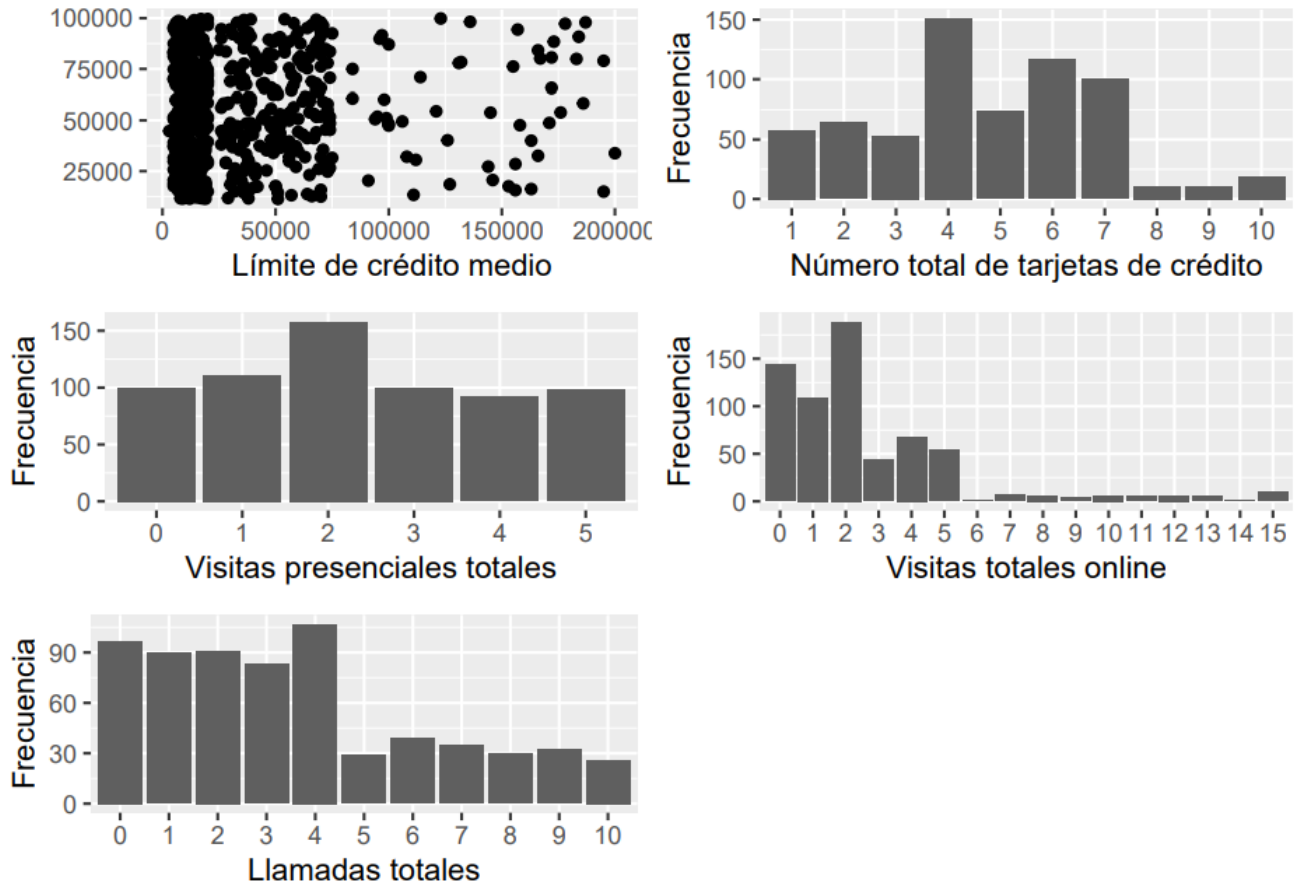
INFORME AGRUPACIÓN ESTADÍSTICA

1.- Preprocesamiento de los datos

El dataset original cuenta con 659 filas y 7 columnas de las cuales la primera y la segundo son variables índice. En principio el dataset no tiene valores faltantes como NA's. Si exploramos los datos, nos encontramos con el individuo 661 el cual cuenta con varios valores -99, los cuales se pueden interpretar como valores faltantes así que eliminamos esa instancia. También comprobamos si existen otros individuos con valores '-99' pero no es así.

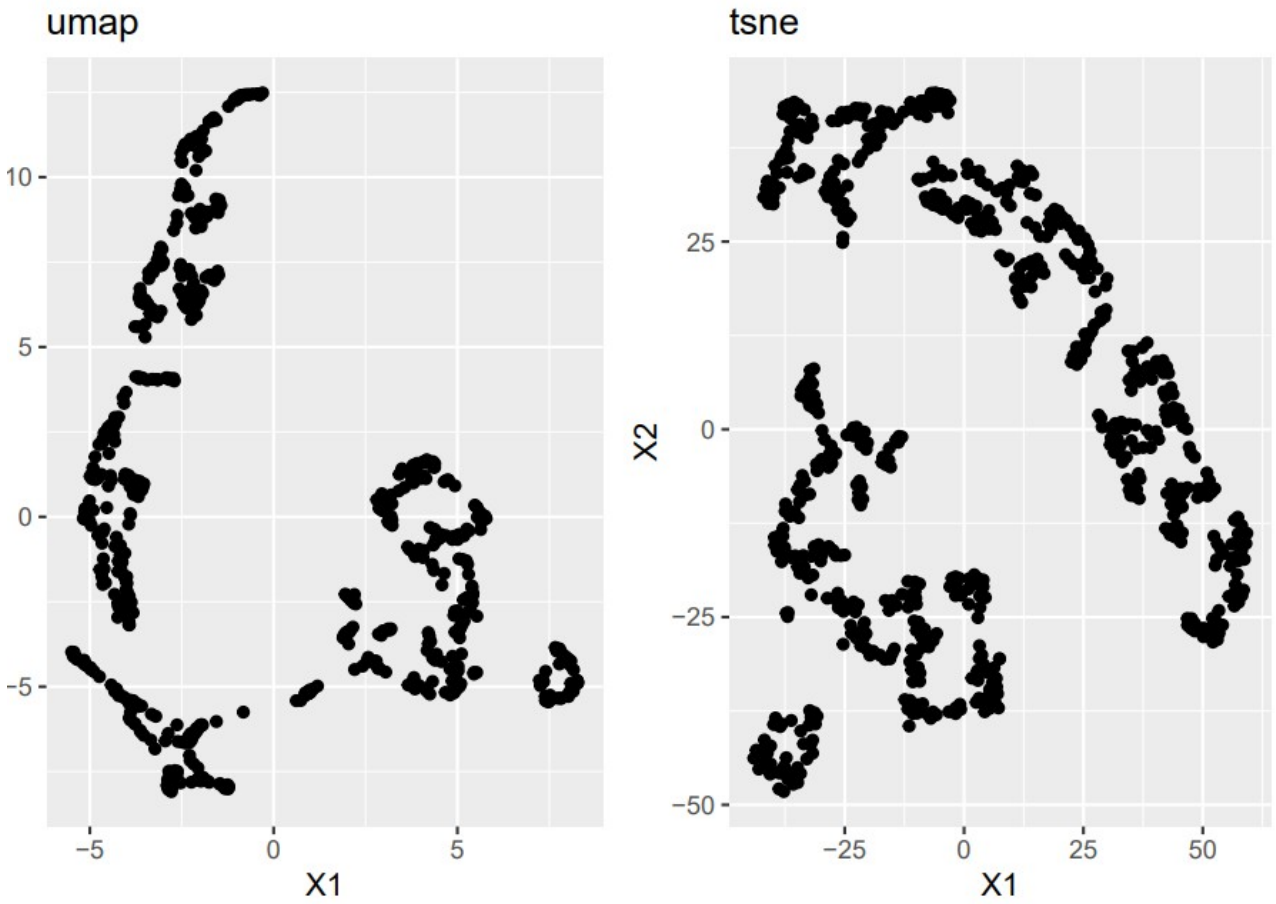
Calculamos los posibles outliers mediante la distancia de Mahalanobis, que es una función apropiada para detectar outliers de conjuntos de datos multivariantes. Los individuos con una distancia mayor a una en concreto, dada por un test y mediante la correción por Bonferroni serán los valores atípicos. Los individuos 66, 661 y 662 son los que superan este umbral.

Visualizando los datos en diagramas de barras nos damos cuenta de que hay 3 valores muy altos en diferentes variables en comparación con el resto de datos, las cuales se corresponden con estos 3 individuos previamente localizados. Eliminamos estas 3 instancias y con esto ya tenemos el conjunto de datos limpio y listo para analizar.



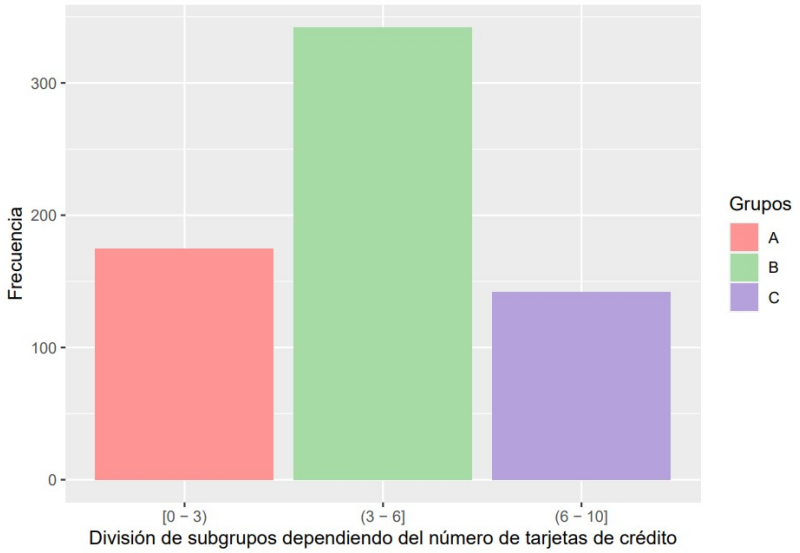
Conjunto de datos limpio

2.- Reducción de la dimensionalidad.
Para reducir la dimensión del conjunto de datos y poder visualizarlo en una gráfica en dos dimensiones hacemos uso de las técnicas Tsne y Umap.Tsne tiende más a preservar la topología local de los datos mientras que Umap tiende más a preservar la topología global.



Podríamos segmentar relativamente fácil tres grupos en t-SNE, mientras que no están facil de segmentar a simple vista la representación de UMAP. (Esto se da por lo explicado anteriormente, t-SNE tiende a enfatizar la separación entre los puntos en el espacio de baja dimensión, mientras que UMAP tiende a enfatizar la preservación de la estructura de los vecinos más cercanos)

3.- Agrupamiento de variables por frecuencia.
La variable continua crédito límite medio la discretizamos en varios rangos, bajo (3-10k), bajo/medio (10k-18k), medio/alto (18k-48k), alto (48k-200k) para facilitar la diferenciación de grupos.
Las demás variables al ser ya categóricas, las agrupamos también en un rango de valores.

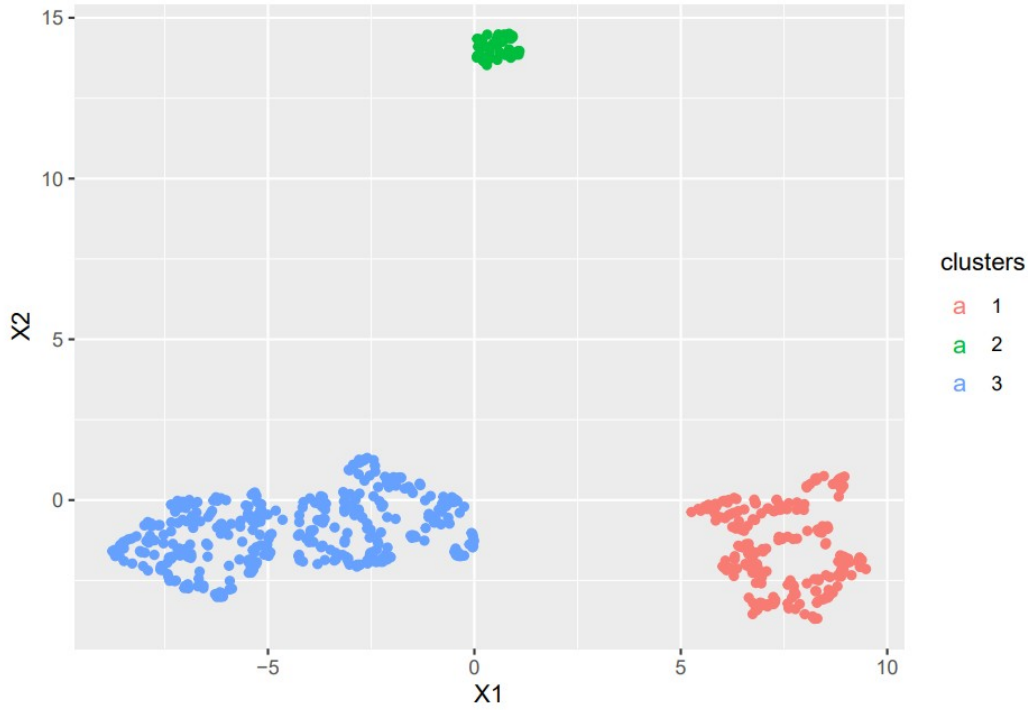


4.- Clusterizacion

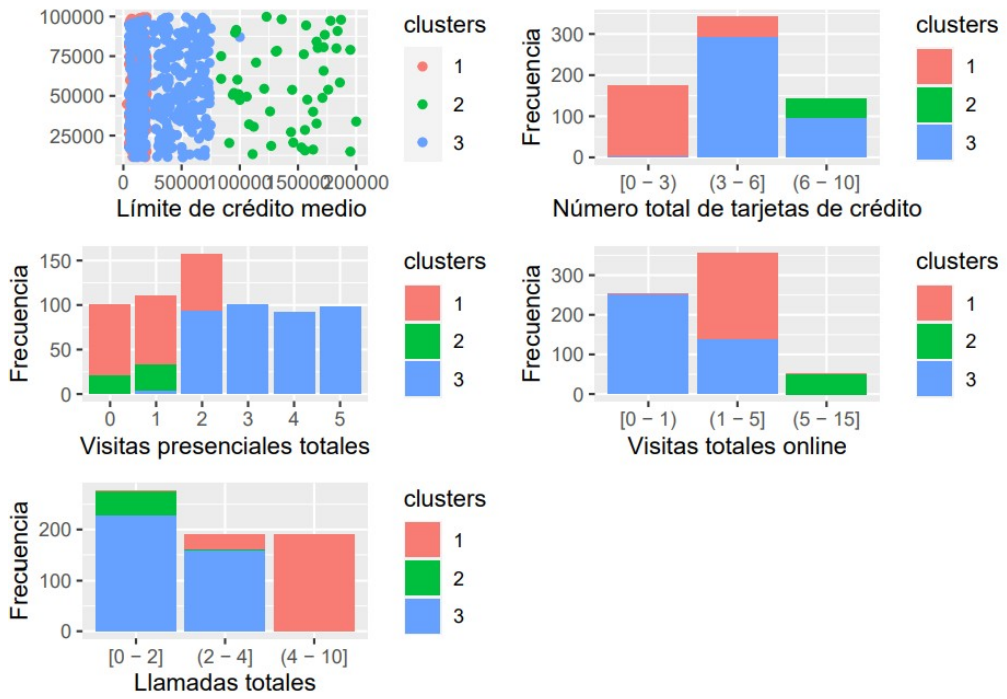
Mediante la librería cluster, vamos a realizar un análisis de clustering mediante el algoritmo k-medias. Para determinar el número de clusters, nos hacemos uso del método del codo (que nos refleja un valor de 3 clusters) y del método de Silhoutte (cuyo valor más óptimo también es de k=3)

La visualización de los datos escalados en UMAP, y coloreados según el cluster al que pertenezcan es la siguiente:

Podemos observar que con los datos escalados, es mucho más obvio que los datos se dividen en 3 grupos.



Ahora, ¿cómo se refleja la pertenencia de cada individuo a un cluster en las variables del conjunto de datos?



Podemos observar que existen tres grupos bien diferenciados, el rojo, verde y azul.
El grupo representado con valor rojo son los individuos con menor límite de crédito medio (de 0a 18000), son los que menos tarjetas de crédito poseen y los que menos visitas presenciales hacen, estan en la media en cuanto a visitas online (entre 1 y 5 veces) y son los que más llamadas hacen.

El grupo azul son los individuos cuyo límite de crédito medio está entre los 0 y los 50000, tienen entre 4 y 7 tarjetas de crédito, son los que menos visitas online y menos llamadas hacen los que más acuden presencialmente al banco.

El grupo verde, por último son los que más crédito límite medio tiene (entre 100000 y 200000), los que más tarjetas de crédito tienen y los que más visitas online hacen (esto último con mucha diferencia) y los que menos visitas presenciales y llamadas hacen.

Resumen final justificado para directivos no técnicos

Con este estudio, hemos utilizado técnicas estadísticas para agrupar a nuestros clientes en diferentes grupos según su uso de las herramientas del banco y los datos que poseemos
. Después de analizar los datos, hemos llegado a la conclusión de que existen 3 grupos bien diferenciados. Cada grupo presenta características únicas en cuanto a su comportamiento y necesidades bancarias.
Al identificar a nuestros clientes según estos grupos, podremos diseñar estrategias personalizadas para cada uno de ellos con el fin de mejorar su experiencia y satisfacción con nuestros servicios bancarios.

Una interpretación de estos datos podría ser:

Grupo rojo: "Jóvenes emprendedores". Este grupo podría ser representativo de jóvenes que están empezando en su vida financiera y tienen un límite de crédito bajo, pero están activamente explorando sus opciones y contactando al banco con frecuencia para obtener asesoramiento y tomar decisiones informadas.

Grupo azul: "Familias de clase media". Este grupo podría representar a personas con un límite de crédito promedio que tienen un número moderado de tarjetas de crédito y visitan la sucursal bancaria con frecuencia. Podrían ser familias que buscan construir su patrimonio y mantener sus finanzas en orden.

Grupo verde: "Inversionistas de alto nivel". Este grupo podría ser representativo de personas que tienen un límite de crédito alto y utilizan una variedad de herramientas financieras, como múltiples tarjetas de crédito y banca en línea. Podrían ser inversores experimentados que buscan maximizar su patrimonio y obtener altas ganancias a largo plazo.

Otra interpretación sería dividirlos en “Usuarios de llamadas”, “Clientes Presenciales” y “Clientes Digitales”.