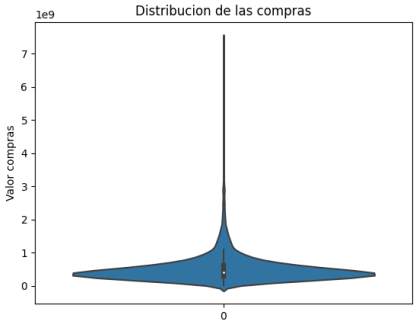
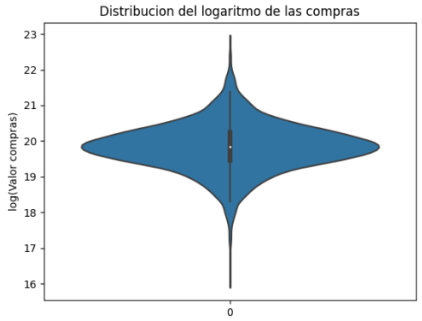
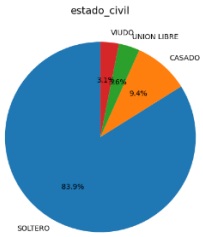
Punto 3: ejercicio de segmentación

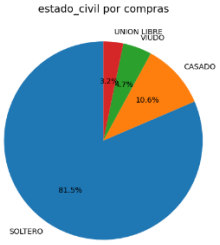
Para este punto se pide realizar un análisis de segmentación con el archivo de Excel enviado. Antes de realizar este análisis, se debe realizar un análisis exploratorio de datos.

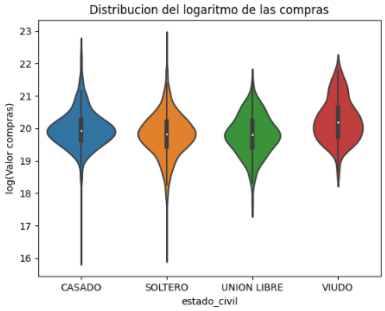
1. **Campos**: el archivo tiene 6 campos ID, COMPRAS ACUMULADAS, MESES CUMPLIENDO VENTAS, ESTADO CIVIL, SEXO, Ubicación geográfica. En ninguno hay nulos. Se cambian a minúsculas, se reemplazan espacios por guion bajo y se eliminan las tildes. Adicionalmente, se propone la creación del logaritmo natural de las compras acumulada para estandarizar los valores.
2. **Visualizaciones**:

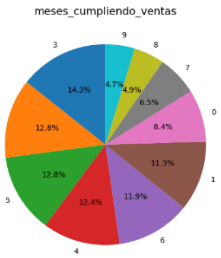


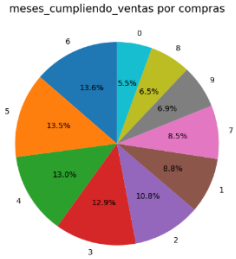


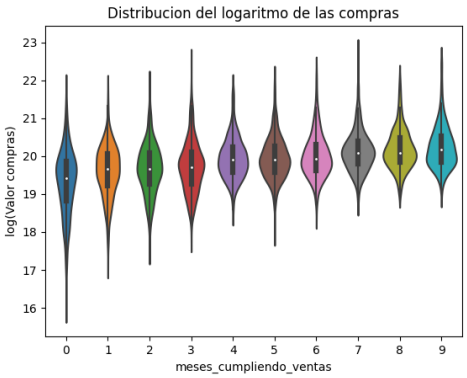


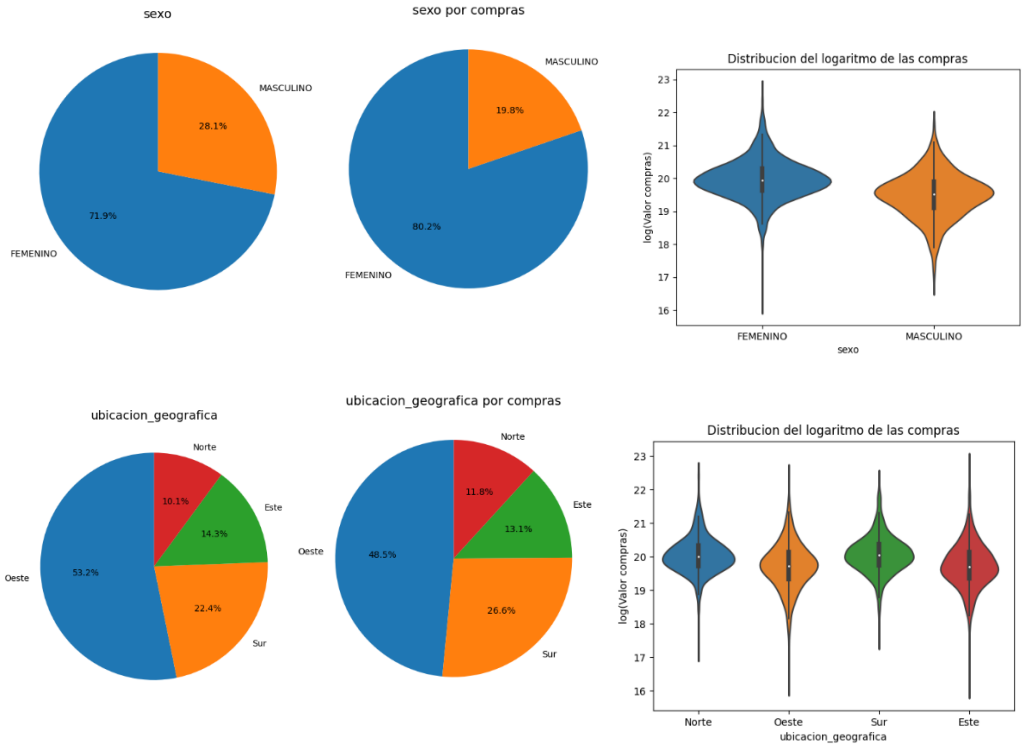








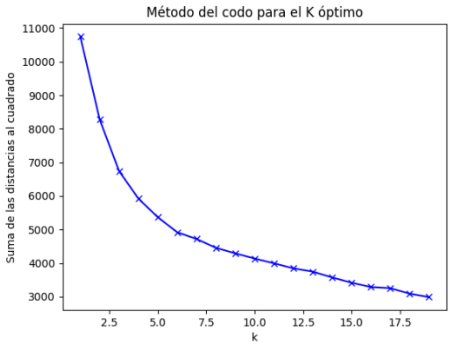




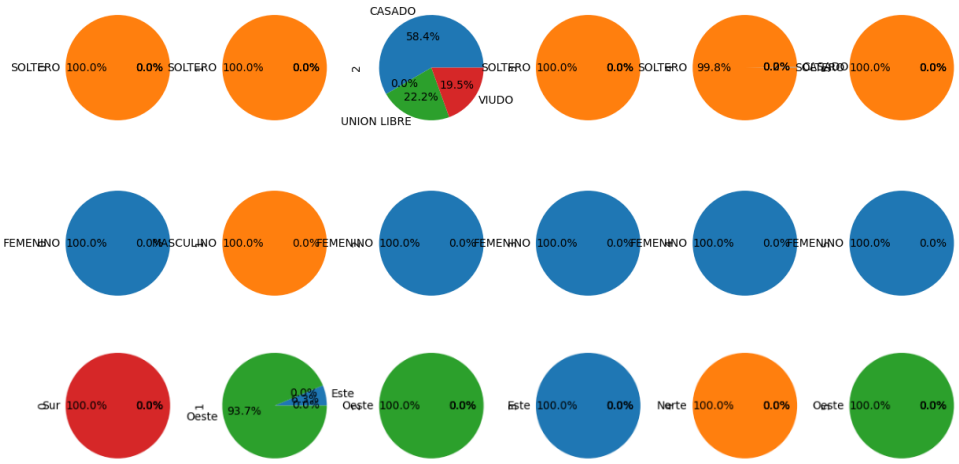
Se puede observar que la mayoría de los valores de compra están distribuidos hasta los 200 millones, con algunos casos atípicos llegando hasta los 700 millones. Por estado civil no se distingue ninguna diferencia del valor medio de compra entre los diferentes estados, sin embargo, se observa que *SOLTERO* es el de mayor peso, con más del 83%. De igual forma, los valores de compra se concentran los meses de cumplimiento 4, 5 y 6; además de que parece notarse incrementos en las medias de estos valores en la medida que tiene mayores meses de cumplimiento. El sexo *FEMENINO* domina al sexo *MASCULINO* en las compras, tanto en totales con una participación por encima del 80%, como en medias. Finalmente, a pesar de que la ubicación *OESTE y SUR* componen casi el 75% de las compras, estas ubicaciones parecen no tener mucha diferencia en sus medias.

3. **Variables nuevas y propuestas:** adicional a la nueva variable del logaritmo de las compras, se propone crear variables *dummies* de las variables categóricas de estado civil, ubicación, sexo y meses de cumplimiento. En caso de tener acceso a la información detrás también propondría incluir otras variables numéricas como la meta de venta y su diferencia con el real, cantidad de productos vendidos (si aplica), precios promedio de venta, precios promedios de ticket, entre otros. La razón de estas variables es que los algoritmos de segmentación se comportan mejor en la medida que se incluyen variables numéricas al modelo. Al final se estandariza la información utilizando un transformador de Mínimos y Máximos.
4. **Método de K-means**: se propone utilizar el algoritmo de *k-medias* para segmentar la información previamente mencionada y procesada. Utilizando el método del codo se observa que hay una clara diferencia a partir de 6 grupos, sin embargo, el algoritmo sigue optimizando para valores más grandes. Sin embargo, como ejercicio inicial se propone

tomar una cantidad de segmentos pequeña, específicamente 6, ya que la creación de un gran número de segmentos resulta impráctica para el manejo por parte de la empresa.



5. **Variables importantes:** se realiza un algoritmo de clasificación de árboles para identificar las variables de importancia y se identifica que estas son el *SEXO, UBICACIÓN GEOGRÁFICA y ESTADO CIVIL*.



Se puede observar que en el grupo 1 se clasificaron todos los hombres, mientras que solo en el grupo 2 existen diferentes estados civiles. En el grupo 0 se clasificaron las ventas en zona sur, en el 3 zona este y zona norte en el 4.

6. **Conclusiones**: después de realizar la segmentación se concluye que se debe realizar la segmentación con las siguientes dos recomendaciones: primero, se debería balancear el conjunto de datos ya que la distribución de estas variables importantes está sesgada para alguna de las categorías; segundo, se deben incluir nuevas variables para complementar la segmentación y que el algoritmo pueda trabajar mejor.