Análisis comportamiento clientes “ABC Private Limited”

# Resumen

La empresa ABC Private Limited desea comprender el comportamiento de compra de sus clientes (en concreto, el importe de la compra) en relación con varios productos de diferentes categorías. Nos han enviado las compras de varios clientes de productos seleccionados de un mes.

Ahora, quieren construir un modelo para predecir el monto de compra de los clientes frente a varios productos, lo que les ayudará a crear una oferta personalizada para los clientes frente a diferentes productos.

# Datos

La empresa nos ha enviado un fichero con las siguientes columnas:

* User\_ID: ID del cliente asignado por la empresa
* Product\_ID: referencia del producto
* Gender: género. Hay M y H.
* Age: edad. Variable de intervalos
* Occupation: ya se convirtió de valor categórico a columna numérica. Hay 20 profesiones distintas en el estudio.
* City\_Category: hay 3 tipos de ciudad. Los que viven en la ciudad A, los que viven en la ciudad B y los que viven en la ciudad C
* Stay in Current City Years: número de años que llevan viviendo en la ciudad.
* Marital\_Status: 0 soltero y 1 casado
* Product Category 1: ya se convirtió de valor categórico a columna numérica.
* Product Category 2: ya se convirtió de valor categórico a columna numérica.
* Product Category 3: ya se convirtió de valor categórico a columna numérica.
* Purchase: importe de la compra realizada.

# Limpieza de los datos

Una vez revisadas las columnas del fichero que nos ha mandado el cliente, se han observado los siguientes puntos y se han realizado las modificaciones pertinentes:

* No hay valores duplicados.
* Crear columna Gender e indicar si son Mujer o Hombre.
* Crear columna Marital Status para cambiar 0 por soltero y 1 por casado.
* Hay celdas vacías en las columnas Product Category 2 y 3, pero no ha sido posible eliminarlas.

# Análisis descriptivo

Una vez visto las columnas que teníamos y haber limpiado los datos, procedemos a hacer una un análisis descriptivo de las variables numéricas, de las categóricas y de las temporales.

Análisis descriptivo numérico

En nuestro conjunto de datos tenemos 2 variables numéricas:

* **Años en la ciudad:**

1. Media (1,86):

La media es 1,86 años, lo que indica que, en promedio, los clientes en este caso llevan aproximadamente 1,86 años viviendo en su ciudad actual.

2. Error típico (0,0017):

El error típico es bastante bajo, lo que sugiere que la estimación de la media es bastante precisa y no está muy influenciada por variaciones grandes.

3. Mediana (2):

La mediana es 2, lo que significa que el valor central de la distribución es 2 años. Es decir, el 50% de los clientes viven en su ciudad actual por más de 2 años, y el 50% lo hacen por menos.

4. Moda (1):

La moda es 1, lo que indica que la mayoría de los clientes llevan 1 año viviendo en su ciudad actual. Este es el valor más frecuente.

5. Desviación estándar (1,29):

Una desviación estándar de 1,29 sugiere que hay una dispersión moderada de los datos alrededor de la media, lo que implica que hay una variabilidad significativa en los años que los clientes han estado en su ciudad.

6. Varianza de la muestra (1,66):

Esta cifra es simplemente el cuadrado de la desviación estándar, y nos da una idea de la dispersión de los datos. Una varianza de 1,66 indica que los datos no están completamente concentrados alrededor de la media, sino que hay una dispersión moderada.

7. Curtosis (-1,07):

La curtosis negativa indica que la distribución de los datos es más plana que una distribución normal (una distribución con curtosis 0). Esto significa que los valores extremos (muy bajos o muy altos) son menos frecuentes en esta variable, lo cual es coherente con la moda y la mediana que sugieren una concentración en los valores cercanos a 1 o 2 años.

8. Coeficiente de asimetría (0,32):

El coeficiente de asimetría positivo sugiere que la distribución tiene una ligera cola hacia la derecha (algunos valores son más altos que el resto, pero no de manera extrema). Esto indica que, aunque la mayoría de los clientes viven en su ciudad por 1 o 2 años, hay algunos que llevan más tiempo.

9. Rango (4):

El rango de 4 años (desde 0 hasta 4) indica que la variable está limitada en un intervalo bastante pequeño, lo cual significa que la mayoría de los clientes tienen un tiempo de residencia en la ciudad de entre 0 y 4 años.

10. Mínimo (0) y Máximo (4):

El mínimo es 0, lo que sugiere que hay clientes que probablemente se han mudado a la ciudad recientemente. El máximo es 4, lo que indica que no hay clientes que hayan estado en la ciudad más tiempo que esto.

11. Suma (1,022,256) y Cuenta (550,068):

La suma de los valores es 1,022,256 y la cuenta de los datos es 550,068, lo que nos dice que el análisis está basado en una muestra grande de datos.

En resumen, la mayoría de los clientes parecen haber vivido en su ciudad por poco tiempo, con una moda de 1 año. La distribución es relativamente dispersa, pero hay una ligera tendencia hacia los valores más bajos. Además, la presencia de una pequeña asimetría positiva indica que un pequeño grupo de clientes ha estado en la ciudad por más tiempo, aunque no en grandes cantidades.

* **Compras:**

1. Media (9263,97):

La media de las compras es 9263,97, lo que indica que el gasto promedio de los clientes en el análisis es aproximadamente 9,264 unidades monetarias. Es un valor representativo del comportamiento general, pero puede no reflejar perfectamente el comportamiento de todos los clientes debido a la dispersión de los datos.

2. Error típico (6,77):

El error típico de 6,77 sugiere que la estimación de la media es bastante precisa, lo que indica que la variación en los datos alrededor de la media no es demasiado grande. En este caso, podemos confiar en que la media es una buena representación del gasto promedio.

3. Mediana (8047):

La mediana de 8047 muestra que el 50% de los clientes gastan menos de 8047, y el otro 50% gasta más. Dado que la mediana es más baja que la media, esto indica que hay algunos clientes que están gastando significativamente más, lo que nos puede sugerir que la distribución de los datos está sesgada hacia valores más altos.

4. Moda (7011):

La moda es 7011, lo que significa que la cantidad más frecuente de compra es 7011 unidades monetarias. Este es el valor que más veces se repite en el conjunto de datos, lo que refuerza la idea de que muchos clientes tienden a gastar alrededor de este monto.

5. Desviación estándar (5023,07):

Una desviación estándar de 5023,07 sugiere que hay una gran variabilidad en los montos de compra entre los clientes. Los valores individuales de las compras se dispersan bastante alrededor de la media, lo que indica una diferencia significativa entre el comportamiento de compra de los clientes.

6. Varianza de la muestra (25231185,95):

La varianza, que es el cuadrado de la desviación estándar, nos muestra cuán dispersos están los datos. Una varianza tan alta (25231185,95) indica una gran dispersión, lo que confirma que los clientes tienen comportamientos de compra muy diversos.

7. Curtosis (-0,34):

La curtosis negativa de -0,34 sugiere que la distribución de las compras es relativamente plana en comparación con una distribución normal. Esto significa que no hay una gran cantidad de valores extremos, es decir, no hay muchas compras muy altas o muy bajas que se aparten significativamente del resto. Sin embargo, hay una ligera cola hacia la derecha, lo que implica que hay algunos valores de compra muy altos (pero no en exceso).

8. Coeficiente de asimetría (0,60):

Un coeficiente de asimetría positivo de 0,60 indica que la distribución de las compras está sesgada hacia la derecha. Esto significa que, aunque la mayoría de los clientes gastan menos, hay algunos clientes que gastan cantidades mucho mayores, lo que arrastra la media hacia valores más altos.

9. Rango (23949):

El rango de 23949 entre el valor mínimo (12) y el máximo (23961) indica una gran variabilidad en las compras. Algunos clientes gastan muy poco, mientras que otros gastan considerablemente más, lo que se refleja en una gran diferencia entre el valor mínimo y el valor máximo.

10. Mínimo (12) y Máximo (23961):

El valor mínimo de 12 unidades monetarias y el valor máximo de 23961 unidades monetarias indican que las compras oscilan entre estas dos cifras. Esto también refuerza la idea de que la distribución es bastante dispersa.

11. Suma (5,095,812,742) y Cuenta (550,068):

La suma total de las compras es 5,095,812,742 unidades monetarias, y el número de clientes (o registros) es 550,068. Esto confirma que la muestra es bastante grande, lo que nos da más confianza en la representatividad de los resultados.

En resumen, la variable "Purchase" muestra un comportamiento de compra altamente disperso, con una media significativamente más alta que la mediana, lo que indica que algunos clientes están realizando compras mucho más grandes que la mayoría. La distribución tiene una ligera asimetría hacia la derecha, lo que sugiere que aunque la mayoría de los clientes realizan compras moderadas, existe un pequeño grupo que realiza compras mucho mayores. La variabilidad (desviación estándar y varianza) es alta, lo que también refleja que no hay un comportamiento de compra uniforme entre los clientes.

Análisis descriptivo categórico

De todas las columnas a analizar, tenemos las siguientes variables categóricas:

* Gender: El 75,31% de los clientes son hombres, mientras que solo el 24,69% son mujeres. La mayoría de las compras en este negocio las realizan hombres.
* Age: La mayoría de los clientes pertenecen al grupo de 26-35 años (39,92%), seguido por 36-45 años (20%) y 18-25 años (18,12%), lo que indica que el público principal de la empresa está entre los 18 y 45 años. La participación disminuye en edades mayores, con solo 15.31% de clientes entre 46 y 55 años y 3.91% en mayores de 55. Los menores de 18 representan apenas 2,75%, mostrando que no son un segmento fuerte. Esto sugiere que la empresa podría enfocarse en estrategias para atraer a clientes mayores de 45 años o consolidar su éxito con adultos jóvenes.
* Occupation: Las ocupaciones más comunes entre los clientes son las categorías 4 (13,15%), 0 (12,66%) y 7 (10,75%), mientras que algunas, como la 8 (0,28%) y la 9 (1,14%), tienen muy baja representación. La distribución es bastante dispersa, lo que indica que la empresa atiende a una variedad de sectores laborales, aunque algunos predominan más que otros. Para un análisis más profundo, sería clave identificar si las ocupaciones más representadas también son las que más gastan y explorar oportunidades para atraer clientes de categorías menos presentes.
* City Category: La mayoría de los clientes provienen de ciudades categoría B (42,03%), seguidas por C (31,12%) y A (26,85%), lo que indica una distribución relativamente equilibrada, aunque con mayor presencia en ciudades tipo B. Esto sugiere que estas ciudades pueden ser clave para la estrategia comercial de la empresa. Sin embargo, sería útil analizar si el gasto promedio varía entre categorías y explorar oportunidades de crecimiento en las ciudades tipo A, que tienen la menor participación.
* Marital Status: La mayoría de los clientes son solteros (59,03%), mientras que los casados representan el 40,97%. Esto indica que la empresa tiene una mayor proporción de clientes sin pareja, lo que podría influir en sus patrones de compra y en las estrategias de marketing. Sería interesante analizar si hay diferencias en el gasto entre ambos grupos y si ciertos productos son más populares entre solteros o casados, lo que permitiría diseñar campañas más personalizadas.
* Product Category 1: Las categorías más frecuentes son la 5 (27,44%), la 1 (25,52%) y la 8 (20,71%), representando en conjunto más del 70% de las observaciones. Esto indica que estos productos son los más comúnmente adquiridos. Por otro lado, las categorías menos frecuentes incluyen la 9 (0,07%), la 17 (0,11%) y la 14 (0,28%), lo que sugiere que estos productos tienen una presencia muy reducida en las compras. La distribución no es uniforme, con una clara concentración en unas pocas categorías, mientras que muchas otras tienen una participación mínima en la muestra.
* Product Category 2: Las categorías más frecuentes son la 8 (17,03%), la 14 (14,64%) y la 2 (13,07%), concentrando una parte significativa del total. Otras categorías con presencia relevante incluyen la 16 (11,49%) y la 15 (10,06%). En contraste, las categorías menos representadas son la 7 (0,17%), la 3 (0,77%) y la 18 (0,74%), indicando una baja frecuencia en comparación con el resto. La distribución está relativamente dispersa, aunque con ciertas categorías dominando en mayor medida las observaciones.
* Product Category 3: El análisis de la variable Product\_Category\_3 muestra que las categorías más frecuentes son la 16 (19,56%), la 15 (16,79%) y la 14 (11,05%), representando una gran proporción del total. También destacan las categorías 5 (9,99%) y 17 (10,01%), con una presencia relevante. En el otro extremo, las categorías con menor frecuencia son la 3 (0,37%), la 4 (1,12%) y la 10 (1,03%), indicando que son productos menos comunes en el conjunto de datos. La distribución muestra una concentración en unas pocas categorías con una mayor presencia en comparación con otras de menor impacto.

Análisis categórico temporal

En nuestro estudio no tenemos ninguna variable que sea del tipo temporal.

# Análisis Compras

Análisis Años en la Ciudad vs Compras

En este gráfico se puede apreciar la relación entre los años de permanencia en la ciudad y el monto total de compras realizadas. Se observa que las personas que han vivido en la ciudad por 1 año tienen el mayor volumen de compras (1.792.872.533), representando una proporción significativa del total. En contraste, quienes llevan 0 años (recién llegados) tienen el menor monto de compras (682.979.229). A partir de 2 años en adelante, las cantidades de compra son relativamente similares, con valores entre 785.884.390 y 949.173.931.

Análisis Género vs Compras

El análisis de estas variables muestra que los hombres representan la mayor parte del total de compras, con un 76,72%, mientras que las mujeres contribuyen con un 23,28%. Esta diferencia significativa sugiere que la clientela masculina tiene un mayor gasto en la empresa minorista "ABC Private Limited".

Análisis Edad vs Compras

El análisis de la relación entre la edad y las compras revela que el grupo de 26 a 35 años es el que más contribuye a las compras totales, representando la mayor suma de compras con 2.031.770.578,00. Le sigue el grupo de 36 a 45 años con 1.026.569.884,00, mostrando una tendencia en la que los consumidores en edad laboral activa realizan la mayor parte de las compras. Los grupos más jóvenes (0-17 y 18-25) y los mayores de 46 años tienen montos de compra significativamente menores, con una caída progresiva a medida que aumenta la edad.

Análisis Ocupación vs Compras

El análisis de la ocupación en relación con las compras muestra una distribución variada del gasto según la categoría ocupacional. Se observa que las ocupaciones con los montos de compra más altos son la categoría 4 (666.244.484) y la categoría 0 (635.406.958), seguidas de la categoría 7 (557.371.587). Por otro lado, las ocupaciones con menor volumen de compras son la categoría 8 (14.737.388) y la categoría 9 (54.340.046).

Análisis Categoría Ciudad vs Compras

El gráfico muestra que la mayor proporción de compras proviene de la categoría de ciudad "B", con un 41,52% del total, seguida por la categoría "C" con un 32,65%, y finalmente la categoría "A" con un 25,83%. Esto indica que las ciudades de tipo "B" son las que más contribuyen al volumen de compras, mientras que las de tipo "A" tienen la menor participación.

Análisis Estado civil vs Compras

El análisis de la relación entre el estado civil y las compras muestra que las personas solteras generan un mayor volumen de compras (3.008.927.447) en comparación con las casadas (2.086.885.295). Esto representa aproximadamente el 59% del total de compras realizadas por solteros y el 41% por casados.

Análisis Categoría Producto 1 vs Compras

El análisis de la relación entre la categoría de productos 1 y las compras muestra que la categoría con mayor impacto en las ventas es la número 1, acumulando 1.910.013.754 en compras, lo que representa una parte significativa del total. Le siguen las categorías 5 y 8, con 941.835.229 y 854.318.799 respectivamente. Estas tres categorías juntas representan una gran parte del total de compras, indicando una fuerte preferencia por estos productos. En contraste, hay varias categorías con ventas significativamente menores, como las categorías 19 y 20, que muestran cifras considerablemente más bajas.

Análisis Categoría Producto 2 vs Compras

El análisis de la variable categoría de producto 2 vs compras muestra que hay una gran variabilidad en las cantidades de compra según la categoría del producto. Algunas categorías, como la 2 (670M) y la 8 (658M), tienen importes de compra significativamente altos, lo que indica que estos productos son altamente demandados. También destacan las categorías 14, 15 y 16, con compras superiores a 390M cada una. Sin embargo, hay categorías con compras bastante bajas, como la 7 (4M) y la 18 (25M), lo que sugiere que estos productos tienen menor popularidad o menor disponibilidad. Un aspecto interesante es que hay una gran cantidad de compras clasificadas como "(en blanco)" (1.305M).

Análisis Categoría Producto 3 vs Compras

El análisis de la categoría de producto 3 frente a las compras muestra que una gran parte de los datos está en blanco (3.150.994.343,00), lo que sugiere que muchos productos no tienen una categoría 3 asignada. Dentro de las categorías definidas, las más destacadas en términos de compras son las categorías **16 (391.040.983,00), 15 (345.662.769,00) y 5 (201.858.094,00)**. Otras categorías con compras significativas incluyen la **14, 17 y 8**. En contraste, categorías como **3, 4 y 10** presentan volúmenes de compras menores.

# Dashboard

Después de hacer los análisis correspondientes, los KPI’s claves son:

* Total de compras
* Compras según categoría del producto
* Compras en función de los años en la ciudad
* Compras según la ocupación
* Compras en función del tipo de ciudad

Los segmentadores que usaremos serán:

* Género
* Edad
* Estado civil

Los gráficos que componen nuestro dashboard son:

* Gráfico según la compra por categoría de producto
* Gráfico de las compras según los años que lleven en la ciudad
* Gráfico de las compras dependiendo de la ocupación
* Gráfico de las compras en función del tipo de ciudad

# Conclusiones

Después de realizar los análisis correspondientes de los datos de ABC Private Limited hemos llegado a las siguientes conclusiones:

* Los clientes entre 26-35 años son los que más compran, seguidos por los de 36-45 años, lo que indica que la mayor parte del consumo proviene de adultos jóvenes y de mediana edad.
* Se observan diferencias significativas en el gasto según la ocupación, con ciertos grupos profesionales comprando mucho más que otros, lo que podría orientar estrategias de segmentación y marketing.
* La mayor parte de las compras provienen de la categoría de ciudad B (41.52%), seguida de la C y A, lo que sugiere que las estrategias comerciales deben priorizar estos segmentos.
* Los solteros gastan más que los casados, posiblemente debido a un mayor poder adquisitivo disponible o patrones de consumo distintos.
* En Categoría 1, el producto 1 es el más comprado, con una diferencia significativa.
* En Categoría 2, hay una gran cantidad de compras en la categoría (en blanco), lo que sugiere que algunos productos no tienen esta clasificación.
* En Categoría 3, más del 60% de las compras también están en la categoría (en blanco), lo que podría indicar datos faltantes o una menor segmentación en esta dimensión.

Algunas recomendaciones para ABC Private Limited podrían ser personalizar las ofertas (dado que ciertos segmentos de edad y ocupación compran más, se pueden desarrollar campañas específicas), optimizar el inventario (priorizar productos más comprados y mejorar la categorización de productos faltantes) o expandirse en ciudades clave (invertir en estrategias para capturar más mercado en las ciudades con mayor volumen de compras).