



Reto: Preguntas Iniciales

Por Jacobo Hirsch Rodríguez - A00829679, Eryk Elizondo González - A01284899, Cleber Gerardo Pérez Galicia - A01236390, Juan Pablo Bernal Lafarga - A01742342.

Carga de Datos y Limpieza

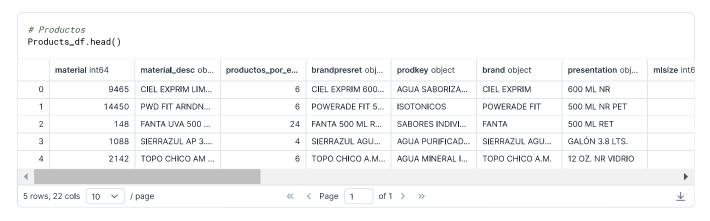
```
# Cargar librerias
import pandas as pd
import numpy as np
import statistics as st

# Cargar datos
Sales_df = pd.read_csv('ventas.csv')
Customers_df = pd.read_csv('customers_sampled.csv')
Products_df = pd.read_csv('20230223_productos.csv')

# Limpieza
Sales_df.columns = Sales_df.columns.str.lower()
Customers_df.columns = Customers_df.columns.str.lower()
Products_df.columns = Products_df.columns.str.lower()
Sales_df['calmonth_convert'] = pd.to_datetime(Sales_df['calmonth'], format='%Y%m')
```

| | <pre>Ventas ales_df.head()</pre> | | | | | | | | | | |
|-----|----------------------------------|----------------|----------------|-----------------|--------------------|--|--|--|--|--|--|
| | customerid int64 | material int64 | calmonth int64 | uni_box float64 | calmonth_convert (| | | | | | |
| 0 | 499920078 | 9151 | 201909 | 0.4364 | 2019-09-01 00:0 | | | | | | |
| 1 | 499920078 | 2287 | 201909 | 3.1701 | 2019-09-01 00:0 | | | | | | |
| 2 | 499920078 | 4526 | 201909 | 0.2818 | 2019-09-01 00:0 | | | | | | |
| 3 | 499920078 | 14050 | 201909 | 0.2642 | 2019-09-01 00:0 | | | | | | |
| 4 | 499920078 | 1333 | 201909 | 2.1134 | 2019-09-01 00:0 | | | | | | |
| ows | 5 cols 10 v / p | age | « | < Page 1 of 1 | > » | | | | | | |

| | <pre># Clientes Customers_df.head()</pre> | | | | | | | | | | | | | | |
|--------|---|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|-----------|--|--|--|--|--|--|--|
| | customerid int64 | pc_agr_300m flo | pc_comercial_30 | pc_generales_30 | pc_habitacional | pc_habitacional | pc_industrial_30 | pc_minero | | | | | | | |
| 0 | 499920078 | 0 | 0 | 6.11 | 48.4 | 37.71 | 6.28 | | | | | | | | |
| 1 | 499920499 | 0 | 0 | 0 | 89.38 | 6.39 | 4.23 | | | | | | | | |
| 2 | 499921473 | 0 | 1.17 | 15.51 | 66.28 | 16.65 | 0.39 | | | | | | | | |
| 3 | 499921557 | 0 | 0 | 81.14 | 16.57 | 1.99 | 0 | | | | | | | | |
| 4 | 499921908 | 0 | 0 | 0 | 100 | 0 | 0 | | | | | | | | |
| 4 | ▼ | | | | | | | | | | | | | | |
| 5 rows | 5 rows, 207 cols 10 v / page | | | | | | | | | | | | | | |



```
# Ventas 2019
Sales_2019 = Sales_df[Sales_df['calmonth'] < 202001]

# Ventas 2020
Sales_2020 = Sales_df[Sales_df['calmonth'].astype(str).str[:4] == '2020']

# Ventas 2021
Sales_2021 = Sales_df[Sales_df['calmonth'].astype(str).str[:4] == '2021']

# Ventas 2022
Sales_2022 = Sales_df[Sales_df['calmonth'].astype(str).str[:4] == '2022']</pre>
```

```
print(f'Hubo {len(Sales_2019)} ventas en 2019')
print(f'Hubo {len(Sales_2020)} ventas en 2020')
print(f'Hubo {len(Sales_2021)} ventas en 2021')
print(f'Hubo {len(Sales_2022)} ventas en 2022')

Hubo 207318 ventas en 2019
Hubo 633643 ventas en 2020
Hubo 714856 ventas en 2021
Hubo 791293 ventas en 2022
```

1. ¿Qué meses tienen más venta?

```
print('El mes con mayor ventas en 2019 fue:', st.mode(Sales_2019['calmonth']), 'con',
len(Sales_2019[Sales_2019['calmonth']==st.mode(Sales_2019['calmonth'])]), 'ventas')
print('El mes con mayor ventas en 2020 fue:', st.mode(Sales_2020['calmonth']), 'ventas')
len(Sales_2020[Sales_2020['calmonth']==st.mode(Sales_2020['calmonth'])), 'ventas')
print('El mes con mayor ventas en 2021 fue:', st.mode(Sales_2021['calmonth']), 'ventas')
len(Sales_2021[Sales_2021['calmonth']==st.mode(Sales_2021['calmonth']), 'ventas')
print('El mes con mayor ventas en 2022 fue:', st.mode(Sales_2022['calmonth']), 'ventas')

El mes con mayor ventas en 2019 fue: 201910 con 53016 ventas
El mes con mayor ventas en 2020 fue: 202010 con 56308 ventas
El mes con mayor ventas en 2021 fue: 202112 con 62598 ventas
El mes con mayor ventas en 2022 fue: 202208 con 68708 ventas
```

En el siguiente conteo del número de ventas por mes se considerará el periodo 2020-2022, pues los registros del año 2019 cubren únicamente las ventas desde agosto hasta diciembre, quedando con 7 meses de registros faltantes.

```
Sales_Dict = {2020: Sales_2020, 2021: Sales_2021, 2022: Sales_2022}
Months = [[202001,202101,202201],[202002,202102,202202],[202003,202103,202203],[202004,202104,202204],
 [202005, 202105, 202205], [202006, 202106, 202206], [202007, 202107, 202207], [202008, 202108, 202208], [202009, 202109, 202209], [202008, 202108, 202208], [202008, 202108, 202208], [202008, 202108, 202208], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108], [202008, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 202108, 2021
[202010, 202110, 202210], [202011, 202111, 202211], [202012, 202112, 202212]]
Months_Dict = ['Enero', 'Febrero', 'Marzo', 'Abril', 'Mayo', 'Junio', 'Julio', 'Agosto', 'Septiembre', 'Octubre',
 'Noviembre', 'Diciembre']
total_len = 0
print('(------PERIODO DE VENTAS 2020 A 2022-----)')
for i in range(12):
          total_len = 0
          for year, month in zip(Sales_Dict.keys(), Months[i]):
                    total_len += len(Sales_Dict[year][Sales_Dict[year]['calmonth'] == month])
          print(f"Total de ventas en {Months_Dict[i]}: {total_len}")
(------PERIODO DE VENTAS 2020 A 2022-----)
Total de ventas en Enero: 161854
Total de ventas en Febrero: 161694
Total de ventas en Marzo: 176898
Total de ventas en Abril: 174084
Total de ventas en Mayo: 180480
Total de ventas en Junio: 182138
Total de ventas en Julio: 181511
Total de ventas en Agosto: 184465
Total de ventas en Septiembre: 184046
Total de ventas en Octubre: 185164
Total de ventas en Noviembre: 182705
Total de ventas en Diciembre: 184753
```

Los meses con mayor número de ventas en el periodo 2020-2022 han sido: Octubre, seguido de Diciembre y Agosto.

2. ¿En qué meses se vende menos?

```
print('El mes con menor ventas en 2019 fue:', Sales_2019['calmonth'].value_counts().idxmin(),
'con',len(Sales_2019[Sales_2019['calmonth'] ==Sales_2019['calmonth'].value_counts().idxmin()],
'ren',len(Sales_2020[Sales_2020['calmonth'] ==Sales_2020['calmonth'].value_counts().idxmin(),
'con',len(Sales_2020[Sales_2020['calmonth'] ==Sales_2020['calmonth'].value_counts().idxmin()],
'ventas')
print('El mes con menor ventas en 2021 fue:', Sales_2021['calmonth'].value_counts().idxmin(),
'con',len(Sales_2021[Sales_2021['calmonth'] ==Sales_2021['calmonth'].value_counts().idxmin()],
'ventas')
print('El mes con menor ventas en 2022 fue:', Sales_2022['calmonth'].value_counts().idxmin(),
'con',len(Sales_2022[Sales_2022['calmonth'] ==Sales_2022['calmonth'].value_counts().idxmin()],
'ventas')

El mes con menor ventas en 2019 fue: 201911 con 50742 ventas
El mes con menor ventas en 2020 fue: 202004 con 48097 ventas
El mes con menor ventas en 2021 fue: 202101 con 52246 ventas
El mes con menor ventas en 2021 fue: 202101 con 52246 ventas
El mes con menor ventas en 2022 fue: 202201 con 59312 ventas
```

Además, gracias a la tabla anterior podemos ver que los meses con peores ventas en el periodo 2020-2022 fueron: 1ro Febrero, 2do Enero y 3ro Abril.

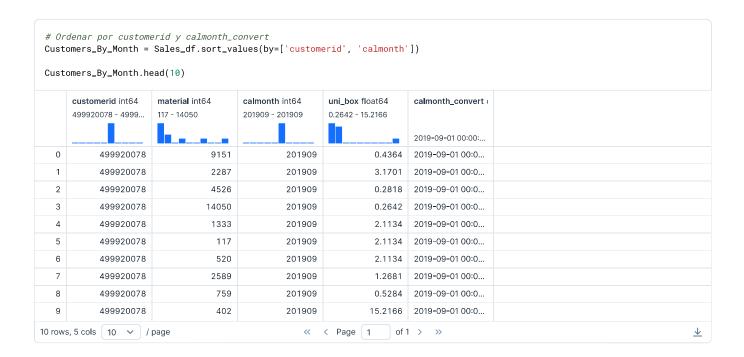
3. ¿Cuántas veces un cliente compra un producto que no compraba antes?

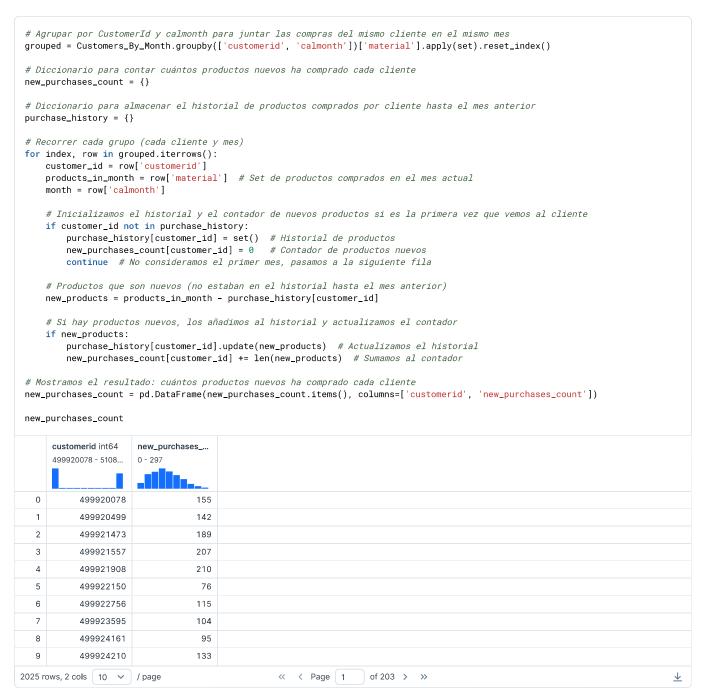
Se obtienen cuantos clientes únicos hay.

```
num_customers = Sales_df['customerid'].nunique()

# Mostramos el resultado
print(num_customers)
```

Se organiza el datafreame por id de cliente y se ordena con fecha de menor a mayor.



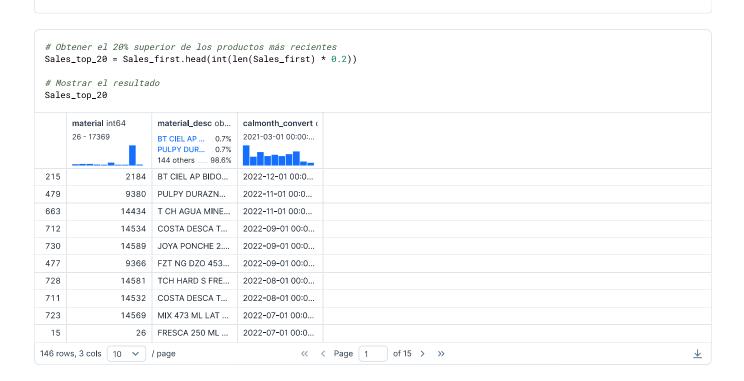


La función de arriba regresa un diccionario con la cantidad de veces que los diferentes clientes (las llaves del diccionario) compraron un producto nuevo, la función considera que se compra un producto nuevo y lo analiza mes a mes.

4. ¿Qué productos de lanzamiento tiene la base de datos? [definir por equipo el término producto de lanzamiento]

Definimos un producto de lanzamiento aquel culla primera venta sea dentro de los últimos 6 meses de la base de datos.

```
# Primera venta para cada material
Sales_first = (
     pd.merge(Sales_df, Products_df, on='material', how='inner') # Merge con los datos de los productos
     .groupby('material', as_index=False) # Agrupar por material
     .agg({'material_desc': 'first', 'calmonth_convert': 'min'}) # Obtener la primera fecha de venta
     . sort\_values(by='calmonth\_convert', \ ascending=False) \ \# \ \textit{Ordenar por fecha m\'as reciente}
)
 # Obtener la fecha más reciente y restarle 6 meses
fecha_limite = Sales_df.agg({'calmonth_convert': 'max'})['calmonth_convert'] - pd.DateOffset(months=6)
 # Filtrar las ventas que ocurrieron después de la fecha límite
Sales_launch = Sales_first[Sales_first['calmonth_convert'] > fecha_limite]
 # Mostrar el resultado
Sales_launch
      material int64
                         material_desc ob...
                                           calmonth convert (
      26 - 14595
                                           2022-07-01 00:00:...
                         BT CIEL AP ... 5.3%
                         PULPY DUR... 5.3%
                         17 others
                                    89.5%
215
                  2184
                         BT CIEL AP BIDO...
                                           2022-12-01 00:0...
                         PULPY DURAZN...
                                           2022-11-01 00:0...
 479
                  9380
 663
                 14434
                         T CH AGUA MINE...
                                           2022-11-01 00:0
 712
                 14534
                         COSTA DESCA T...
                                           2022-09-01 00:0...
                         JOYA PONCHE 2....
 730
                 14589
                                           2022-09-01 00:0...
                  9366 FZT NG DZO 453...
                                           2022-09-01 00:0...
477
 728
                 14581
                         TCH HARD S FRE...
                                            2022-08-01 00:0...
 711
                         COSTA DESCA T...
                                            2022-08-01 00:0...
                 14532
 723
                 14569
                         MIX 473 ML LAT ...
                                            2022-07-01 00:0...
  15
                    26 FRESCA 250 ML ...
                                            2022-07-01 00:0...
                                                                            of 2 > >>
19 rows, 3 cols 10 v / page
                                                        << < Page 1</pre>
                                                                                                                                              \underline{\downarrow}
```



```
# Obtener la lista de tipos de productos únicos
Products_list = pd.DataFrame(Products_df['producttype'].unique(), columns=['producttype'])
# Mostrar el número de tipos de productos
print(f"Hay {len(Products_list)} tipos de productos.")
# Mostrar los tipos de productos
print("Los tipos de productos son:")
print(Products_list)
Hay 25 tipos de productos.
Los tipos de productos son:
              producttype
          AGUA SABORIZADA
             ISOTONICOS
1
          SABORES REGULAR
2
          AGUA PURIFICADA
3
             AGUA MINERAL
5
            COLAS REGULAR
6
      LECHE UHT SABORIZADA
7
      BEBIDAS REFRESCANTES
      BEBIDAS ENERGETICAS
8
         JUGOS Y NECTARES
9
10
          BEBIDAS DE SOYA
               NARANJADAS
12
            SABORES LIGHT
13 BEBIDA CON ELECTROLITOS
14
             COLAS LIGHT
15
              CAFE MOLIDO
                   MIXTOS
16
17
                     TE
18
               CAFE GRANO
19
       BEBIDAS INFANTILES
           AGUA FUNCIONAL
20
21 LECHE UHT ESPECIALIZADA
22
        BEBIDA ALCOHOLICA
23
        LECHE UHT REGULAR
24
            YOGURT BATIDO
```

```
# Obtener y mostrar el número de clientes únicos
print(f"Existen {Customers_df['customerid'].nunique()} clientes.")
# Separar los valores en 'sub_canal_comercial', explotar los valores y obtener los comercios únicos
Commerce_list = pd.DataFrame(Customers_df['sub_canal_comercial'].str.split(' / ').explode().unique(),
columns=['sub_canal_comercial'])
# Mostrar la cantidad y los tipos de comercios
print(f"Hay {len(Commerce_list)} tipos de comercios.")
print("Los comercios son:")
print(Commerce_list)
Existen 2041 clientes.
Hay 21 tipos de comercios.
Los comercios son:
                           sub_canal_comercial
                                 Estanquillos
1
                                      kioscos
2
                                    Abarrotes
3
                                    Almacenes
4
                                      Bodegas
5
                                     Víveres
6
                             Frutas y Verduras
7
                              Hogar con Venta
8
                                   Carnicería
9
                                     Pollería
10
                                   Pescadería
                             Cerveza y Licores
11
12
                                  Tortillería
13
                        Farmacia Independiente
14
                          Mayorista Abarrotero
15
                                   Panadería
16
                                   Pastelería
17
                                    Minisuper
18
19 Tiendas de Alimentos Especializados Orgánicos
20
                  TDC/Proximidad Independiente
```