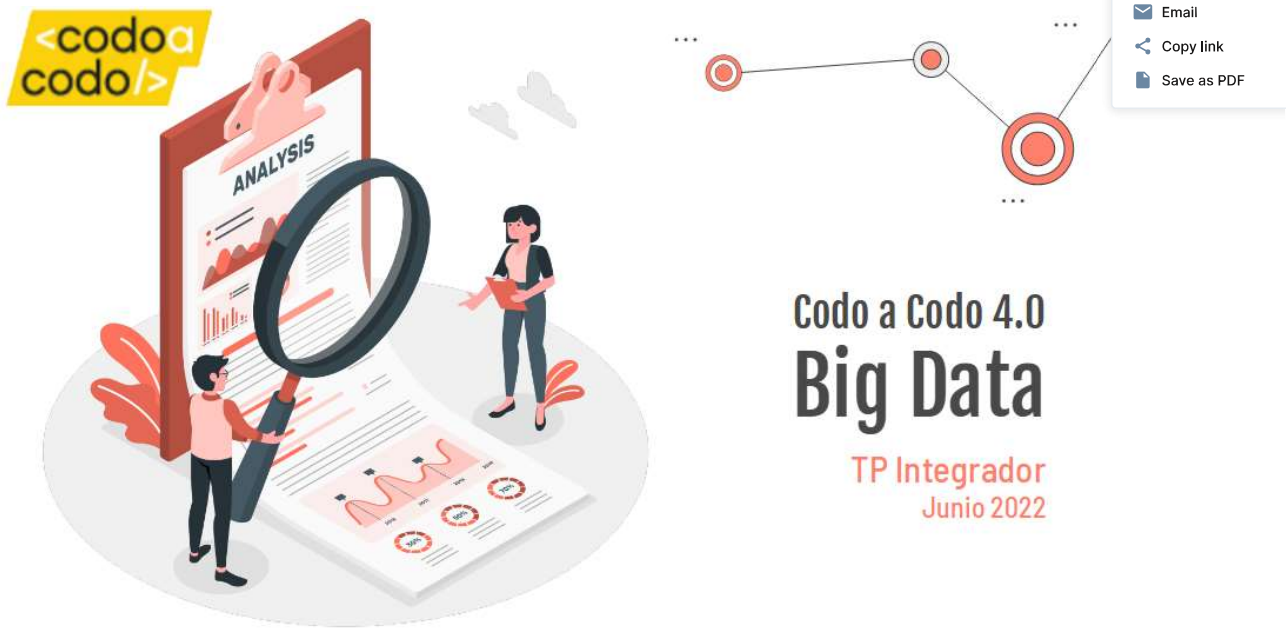
TP Integrador - CaC Big Data



Paso 1: Alcances del proyecto y obtener datos

Alcances del proyecto

En este proyecto vamos a integrar tres sets de datos con lista de vendedores, lista de artículos y registro de operaciones de un mes, que nos ayudarán a responder a las preguntas:

- ¿Cuál es el artículo más vendido? (unidades)
- ¿Qué artículo es el que más ingresos nos proporcionó?
- ¿A qué vendedor debe otorgarse el bono por "Mejor vendedor del mes"?
- ¿Hay grandes variaciones en ventas a lo largo del mes?

Para este proyecto utilizaremos herramientas de Pandas para análisis exploratorio, Numpy para el análisis de ciertas columnas y Matplotlib/Seaborn para visualización de resultados.

Descripción y obtención de los datos

Fuentes de datos

- articles.db: BD con datos de los artículos.
- sellers.xlsx: datos de los vendedores.
- orders.csv: registro de las ventas de un mes.

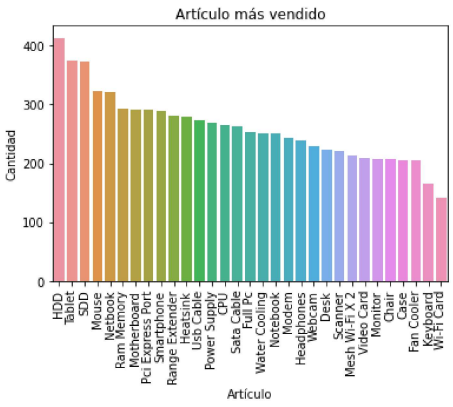
1. ¿Cuál es el artículo más vendido? (en unidades)

```
# RESOLUCIÓN ANALÍTICA
df2 = my_df.groupby('article_name').sum()
por_cant = df2.sort_values('quantity', ascending=False)
print(por_cant['quantity'].head(1))
print(df2.head())
```

```
article_name
HDD    413
Name: quantity, dtype: int64
              week  quantity  total_amount
article_name
CPU             66       266      37138.92
Case            54       206       7807.40
Chair           56       207      69477.48
Desk            60       223      29012.30
Fan Cooler      64       205        871.25
```

```
# RESOLUCIÓN GRÁFICA
sns.barplot(x=df2.index,y=df2['quantity'],data=df2, saturation=.8, order=df2.sort_values('quantity', ascending=False).index).set_title("Artículo más vendid
sns.set( rc = {'figure.figsize' :(20, 10),  'axes.labelsize': 12 },style='whitegrid',font_scale =1.9)
plt.xticks(rotation=90)
plt.xlabel('Artículo')
```

```
plt.ylabel('Cantidad')
plt.show()
```



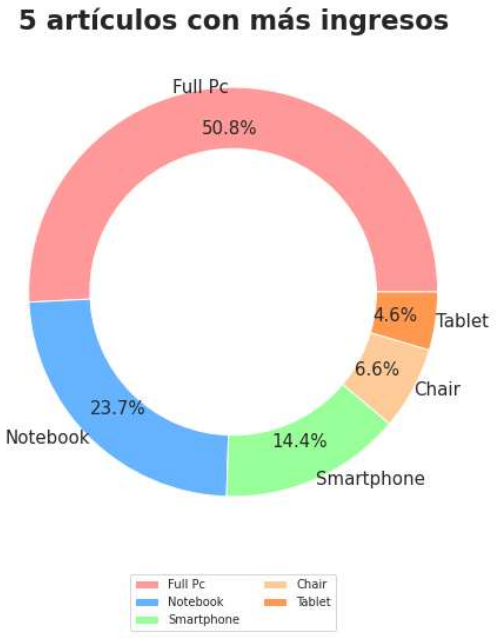
El artículo más vendido en cantidades, es el HDD, seguido por Tablet y SDD

2. ¿Qué artículo es el que más ingresos nos proporcionó?

```
# RESOLUCIÓN ANALÍTICA
df3=(my_df.groupby('article_name').sum()).sort_values('total_amount', ascending=False).head(5)
print(df3['total_amount'])
```

```
article_name
Full Pc       538335.93
Notebook      251000.00
Smartphone    152250.00
Chair          69477.48
Tablet         48620.00
Name: total_amount, dtype: float64
```

```
# RESOLUCIÓN GRÁFICA
fig = plt.figure()
fig.set_figheight(8)
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#ff9950']
textprops = {"fontsize":15}
plt.pie(x=df3['total_amount'],labels=df3.index,textprops =textprops, colors=colors, autopct='%.1f%%',pctdistance=0.8, labeldistance=1)
plt.legend(fontsize=10, loc='upper center', bbox_to_anchor=(0.5, -0.04), ncol=2)
plt.title('5 artículos con más ingresos',fontweight='bold')
centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.show()
```



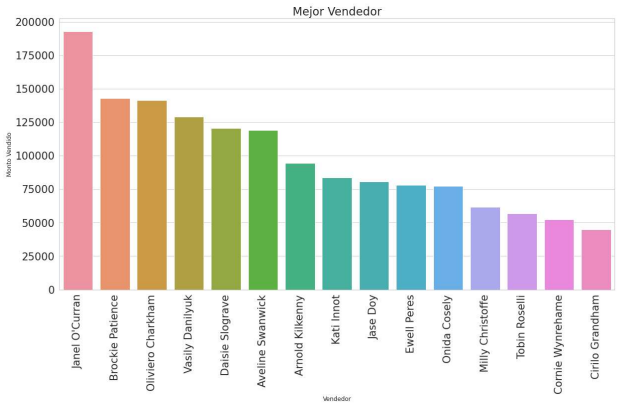
El artículo con mayores ingresos generados es Full PC, seguido de Notebook y Smartphone. Vemos que este top 3 no guarda relación con los artículos más vendidos en cantidades, debido a que influye el precio unitario de cada artículo.

3. ¿A qué vendedor debe otorgarse el bono por "Mejor vendedor del mes"?

```
# RESOLUCIÓN ANALÍTICA
df4 =(my_df.groupby('seller_name').sum()).sort_values('total_amount', ascending=False)
print('Respuesta:', df4.head(1) )
print()
print(df4[['quantity']+['total_amount']])
```

```
Respuesta:                week  quantity  total_amount
seller_name
Janel O'Curran   174       703     192832.47

                  quantity  total_amount
seller_name
Janel O'Curran          703     192832.47
Brockie Patience        441     142709.88
Oliviero Charkham       555     141329.76
Vasily Danilyuk         521     129157.55
Daisie Slograve         554     120520.11
Aveline Swanwick        629     118874.33
Arnold Kilkenny         583      94552.04
Kati Innot              512      83704.62
Jase Doy                582      80628.31
Ewell Peres             496      78144.32
Onida Cosely            535      77373.37
Milly Christoffe        442      61733.69
Tobin Roselli           519      56984.42
Cornie Wynrehame        523      52253.57
Cirilo Grandham         470      45009.40
```

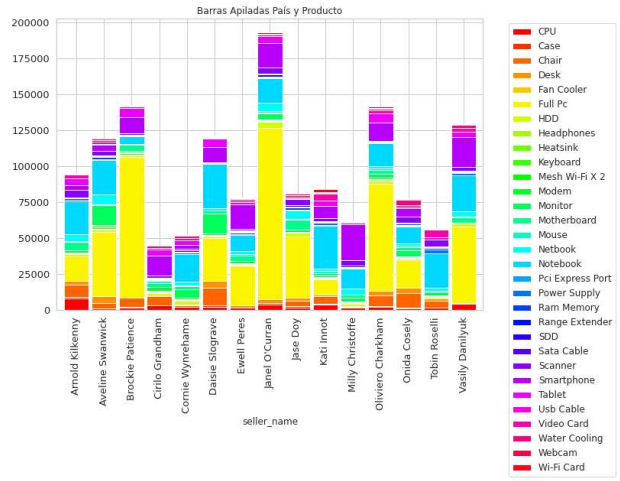
```
# RESOLUCIÓN GRÁFICA2
sns.barplot(x=df4.index,y=df4['total_amount'],data=df4, saturation=.8, order=df4.sort_values('total_amount', ascending=False).index).set_title("Mejor Vende
sns.set( rc = {'figure.figsize' :(20, 10),  'axes.labelsize': 12 },style='whitegrid',font_scale =1.9)
plt.xticks(rotation=90)
plt.xlabel('Vendedor')
plt.ylabel('Monto Vendido')
plt.show()
```



El vendedor que genera mayores ingresos y que además también vende la mayor cantidad de artículos es Janel O'Curran.

Veamos la forma en que cada vendedor distribuye sus ventas.

```
cross3.plot(kind = 'bar',
            stacked = True,
            title = 'Barras Apiladas País y Producto',
            mark_right = True,
            cmap='hsv',
            fontsize=13,
            figsize= (10,7),
            width=0.9,
            style='whitegrid').set_facecolor('w')
plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
plt.rc('legend', fontsize=10)
```



Janel O'Curran es el vendedor con mayores ventas, y en parte esto viene impulsado porque se concentra en vender los productos que están en el top3 de ingresos, es decir, Full Pc, Notebook y SmartPhones. Hay otros vendedores, como Tobin Roselli, cuya cartera de ventas, no se orienta a los productos más vendidos; quizás pudiera intentar cambiar su mix de productos, y ofrecer a los clientes los productos que generan más ingresos para la empresa.

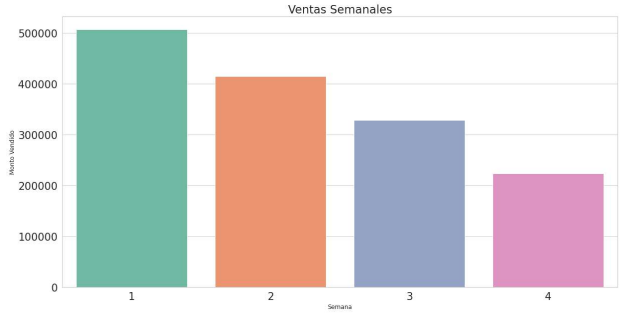
4. ¿Hay grandes variaciones en ventas a lo largo del mes?

Si es así, ¿en qué momento debería lanzar una campaña de promociones?

```
# RESOLUCIÓN ANALÍTICA
df5 = (my_df.groupby('week').sum()).sort_values('total_amount',ascending=False)
print(df5)
```

```
      quantity  total_amount
week
1         2449     507458.81
2         2444     415364.44
3         2114     329140.03
4         1058     223844.56
```

```
# RESOLUCIÓN GRÁFICA 2
sns.barplot(x=df5.index,y=df5['total_amount'],data=df5, palette="Set2",saturation=.8, order=df5.sort_values('total_amount', ascending=False).index).set_tit
sns.set( rc = {'figure.figsize' :(20, 10),  'axes.labelsize': 12 },style='whitegrid',font_scale =1.9)
plt.xticks(rotation=0)
plt.xlabel('Semana')
plt.ylabel('Monto Vendido')
plt.show()
```



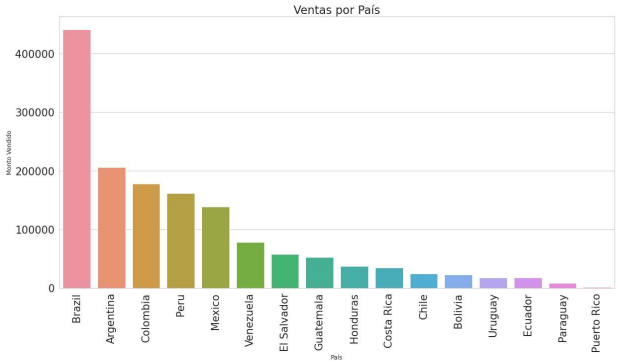
Las personas realizan sus compras principalmente en la semana 1 del mes, y luego van disminuyendo paulatinamente, siendo la cuarta y última semana, la que presenta menores ventas.

5. ¿Cuál es el país con mayor venta (total_amount?

```
# RESOLUCIÓN ANALÍTICA
df6 = (my_df.groupby('country_name').sum()).sort_values('total_amount',ascending=False)
print('Respuesta:', df6.head(1))
print()
print(df6[['quantity']+['total_amount']].head())
```

```
Respuesta:               week  quantity  total_amount
country_name
Brazil         717      2515     441271.85

              quantity  total_amount
country_name
Brazil            2515     441271.85
Argentina          947     205832.78
Colombia           881     177514.29
Peru              1027     161421.12
Mexico             846     138619.99
```

```
#RESOLUCIÓN GRÁFICA
sns.barplot(x=df6.index,y=df6['total_amount'],data=df6, saturation=.8, order=df6.sort_values('total_amount', ascending=False).index).set_title("Ventas por
sns.set( rc = {'figure.figsize' :(12, 10),  'axes.labelsize': 12 },style='whitegrid',font_scale =1)
plt.xticks(rotation=90)
plt.xlabel('País')
plt.ylabel('Monto Vendido')
plt.show()
```

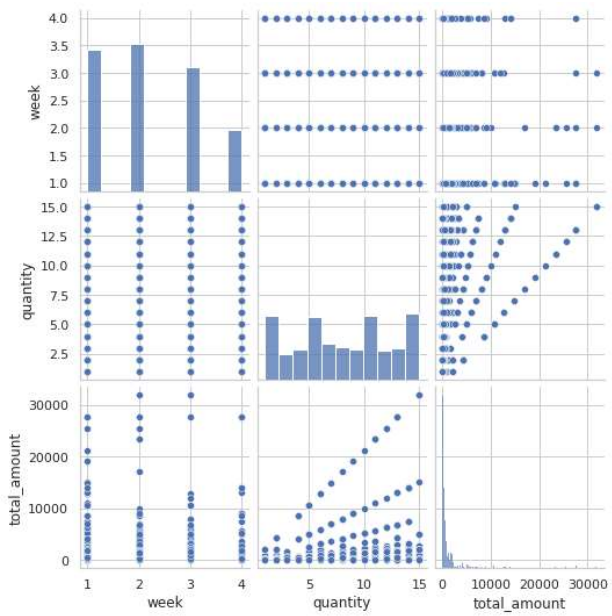


El país con mayor venta es Brazil, tanto a nivel de ventas en valor monetario, como en cantidades vendidas.

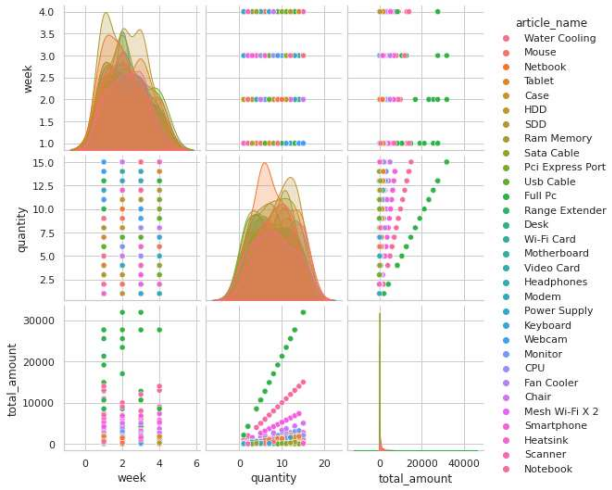
6. ¿Las variables quantity y total_amount guardan alguna relación?

```
# RESOLUCIÓN

# Para encontrar relación hacemos un pairplot
# Las variables que pudieran relacionarse son:week, quantity y total_amount
sns.pairplot(my_df)
# En el gráfico se puede ver que hay una relación entre quantity y total_amount
```

```
<seaborn.axisgrid.PairGrid at 0x7f05092e6400>
```




```
# Para encontrar mejor relación, distinguimos por article_name
# Esto es lógico porque total_amount se obtuvo el cálculo de quantity*unit_price
# La relación entre quantity y total_amount se da en diferente proporción para cada artículo

sns.pairplot(my_df, hue='article_name')
plt.rc('legend', fontsize=12)
plt.show()
```



Hay relación entre las variables cantidad y total_amount, las cuales tienen una correlación positiva, es decir, al aumentar la cantidad vendida, también aumenta el monto total (total_amount). Esto es evidente que ocurriera, debido a que la variable total_amount la construimos partiendo de la multiplicación de cantidad*precio_unitario. Adicionalmente, esa relación se puede distinguir por producto, esto porque cada producto tiene su precio.

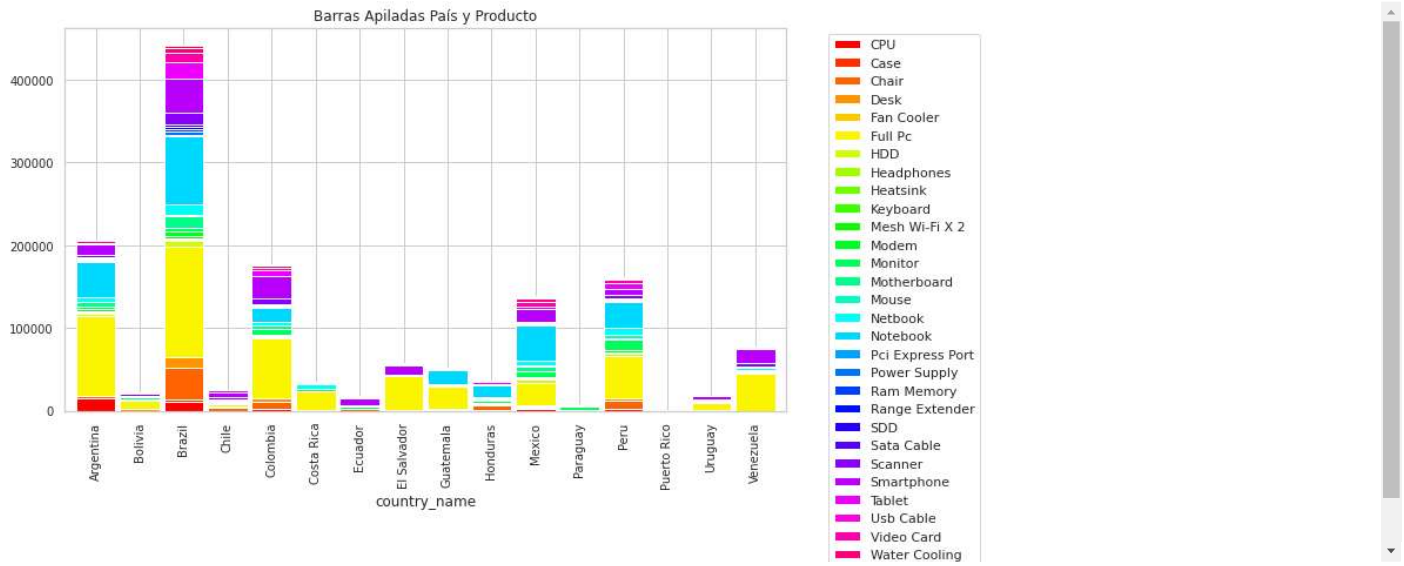
7. ¿En qué productos se basa Brasil para ser el país con el mayor ingreso en este análisis?

```
# RESOLUCIÓN ANALÍTICA
#Creamos una tabla cruzada para ver los productos vendidos en cada país
cross2=pd.crosstab(my_df.country_name, my_df.article_name, my_df.total_amount, aggfunc=np.sum)
print(cross2.head())
```

```
Bolivia            NaN     NaN       NaN   2081.6         NaN   10639.05
Brazil        11448.84  2690.9  37591.68  12619.7      246.50  134052.03
Chile              NaN     NaN   4363.32   1040.8         NaN    2127.81
Colombia       2373.54   644.3   8391.00   3772.9       29.75   72345.54

article_name      HDD  Headphones  Heatsink  Keyboard  ...     SDD  \
country_name                                           ...
Argentina     3714.16       442.7       NaN     700.6  ...  1606.0
Bolivia           NaN         NaN       NaN     293.8  ...     NaN
```

```
Brazil        6499.78      1700.9    1030.0     836.2  ...  2574.0
Chile         1147.02         NaN     120.0      67.8  ...     NaN
Colombia      1857.08      1281.5     260.0     632.8  ...  1628.0

article_name  Sata Cable  Scanner  Smartphone   Tablet  Usb Cable  Video Card  \
country_name
Argentina          81.32   2220.0     13125.0      NaN      94.40      1972.5
Bolivia              NaN   2775.0         NaN      NaN      61.95      1446.5
Brazil            121.98  14430.0     41475.0  20280.0     188.80     10520.0
Chile              23.54   2775.0      6300.0   1430.0        NaN         NaN
Colombia            4.28   7215.0     27300.0   6500.0     150.45      2630.0

article_name  Water Cooling   Webcam  Wi-Fi Card
country_name
Argentina            1687.5   200.70      715.32
Bolivia                 NaN      NaN         NaN
Brazil               5535.0  1043.64     2205.57
Chile                   NaN   280.98         NaN
Colombia             2160.0  1043.64     1490.25

[5 rows x 31 columns]
```

```
# RESOLUCIÓN GRÁFICA
cross2.plot(kind = 'bar',
            stacked = True,
            title = 'Barras Apiladas País y Producto',
            mark_right = True,
            cmap='hsv',
            fontsize=10,
            width=0.9,
            figsize= (11,6),
            style='whitegrid').set_facecolor('w')
sns.set( rc = {'figure.figsize' :(10, 10),  'axes.labelsize': 12 },style='whitegrid',font_scale =1)
plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
plt.rc('legend', fontsize=12)
```

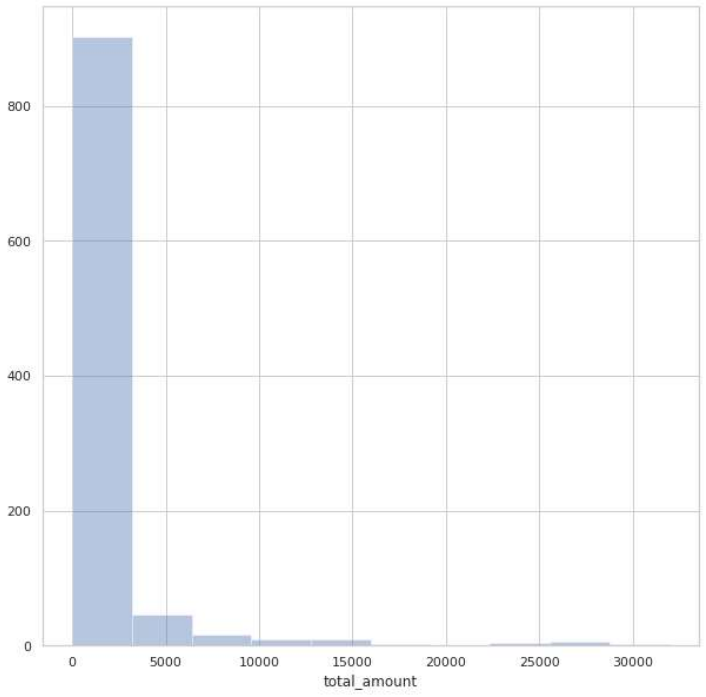


Brasil es el país con más ventas, pero adicionalmente al ver a detalle los productos que vende para entender el éxito que ha logrado, vemos que sus ventas se concentran en los mismos productos que vimos anteriormente con mayores ventas: Full Pc, Notebook y SmartPhone; es decir, se enfocan en los productos que generan más ingresos.

8. Análisis del Total_amount

```
sns.distplot(my_df.total_amount,kde=False, bins=10)
plt.show()
```

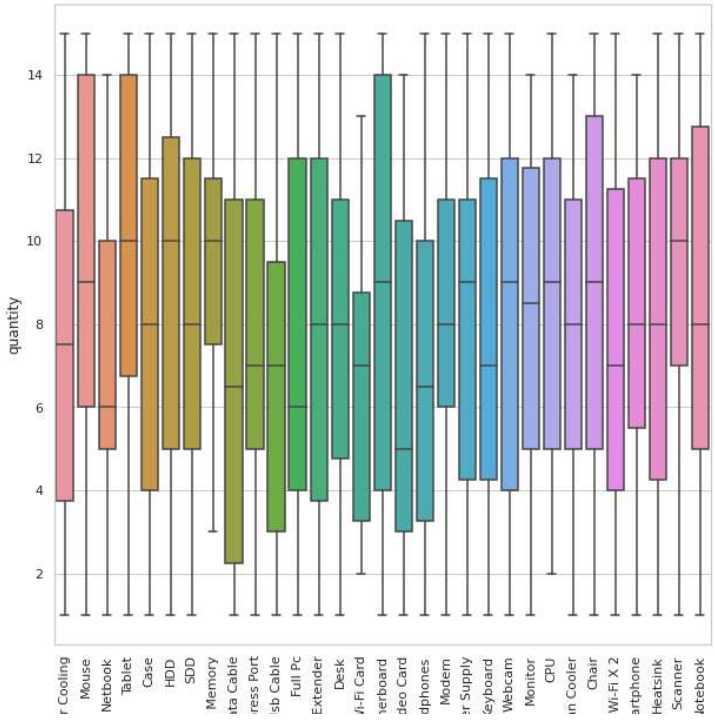
```
/shared-libs/python3.9/py/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Plea
  warnings.warn(msg, FutureWarning)
```



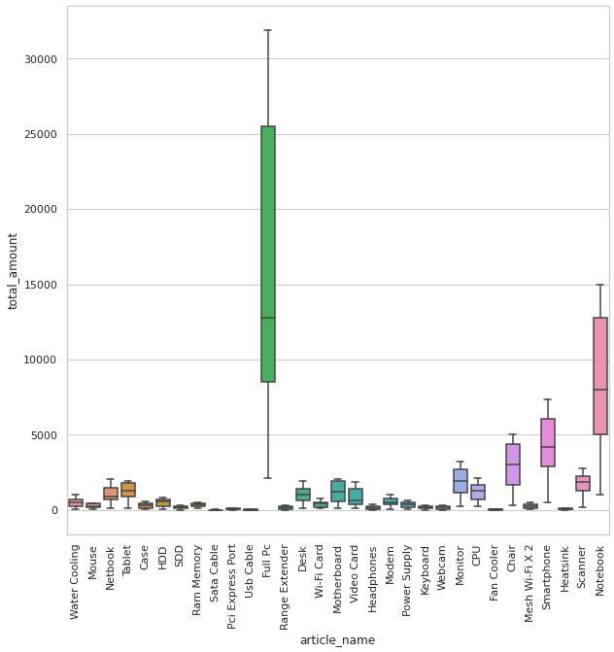
El total_amount se concentra en productos con ingresos entre 0-5000, mientras que en mucha menor proporción se encuentra en un total_amount mayor a 5000.

9. Full_Pc

```
sns.boxplot(x='article_name', y='quantity', data=my_df)
plt.xticks(rotation=90)
plt.show()
```




```
sns.boxplot(x='article_name', y='total_amount', data=my_df)
plt.xticks(rotation=90)
plt.show()
```



Cuando hacemos el boxplot del producto por quantity y total_amount, vemos que el artículo "Full Pc" (el que genera mayores ingresos) para quantity tiene una mediana baja respecto al resto de artículos, mientras que para total_amount, la mediana es de aproximadamente 13.000$, siendo el producto con mayor valor, aportando así mayores ingresos. Es decir, vendiendo pocas cantidades, se tienen mayores ingresos.

Conclusiones y propuestas

En esta empresa hay distintos tipos de productos, y cada uno de ellos tiene una venta en cantidad y en monto. Hay productos que tienen una alta rotación en sus ventas, pero el monto es bajo; mientras que hay otros productos que tienen montos más elevados, lo cual genera mayores ingresos. (En este caso, haría falta sumar al análisis la tabla de costos, para conocer realmente el ingreso neto por productos, y conocer la estrategia a seguir).

El top3 de productos más vendidos en cantidad son:

- HDD
- Tablet
- SDD

Sin embargo, si hacemos el top3 pero con los artículos con mayores ingresos, el ranking es el siguiente:

- Full Pc
- Notebook
- SmartPhone

Como vemos estos top3 son muy diferentes, no hay coincidencia entre los productos; ya que aunque HDD, Tablet y SDD se venden con mayor frecuencia, su precio unitario es bajo y genera bajos ingresos a la empresa.

El país con mayor ventas es Brasil, y esto lo logra debido a que sigue una estrategia en donde vende mayor cantidad de productos, pero adicionalmente se concentra en los 3 productos que generan mayores ingresos, es decir, Full Pc, Notebook y SmartPhone. Si vendieran la misma cantidad, pero de productos con bajos ingresos, pudiera ocurrir que no sea el n° 1 en ingresos.

Al buscar relación entre las variables evaluadas en este análisis, vemos que hay una correlación positiva entre quantity y total_amount, y esto ocurre debido a que total_amount fue un dato que obtuvimos de nuestro dataset original, cuando multiplicamos "quantity*unit_price". Y dicha relación se puede ver reflejada graficamente, en donde para cada producto esa relación es única.

La propuesta sería, incluir una tabla con los costos de cada producto, para obtener en ingreso neto, y conocer qué productos son los que generan mayor rentabilidad a la empresa.

Otra propuesta sería evaluar las ventas logradas por el mejor vendedor (Janel O'Curran) para poder determinar si ese modelo sería aplicable a otros vendedores en diferentes paises, y así lograr mayores ingresos a nivel global. Claro, allí se deben evaluar otros factores, como los niveles de inventarios de cada país, la oferta y demanda, satisfacción de los empleados, etc.