Proyecto Final Data Science I

SuperStore 2019 – 2022

By: Sofia Paula Franke Pavòn

Coderhouse

20/07/2024

DataSet

La información registrada en el dataset proviene de un Hipermercado, el período que cubre es desde el 2019 hasta 2022.

El dataset se obtuvo de la web Kaggle
link: https://www.kaggle.com/datasets/timchant/supstore-dataset-2019-2022?resource=download

Contiene un total de 19 columnas y 9.993 filas, y su información es cualitativa y cuantitativa. Es bastante completo permitiendo generar una interesante investigación económica, de consumo y marketing. La base de datos está guardado en csv, pero a traves de Colab estará en formato ".ipynb".

Diccionario del dataset

- Irder_id (ID del pedido), Order_date (fecha del pedido), Ship_date (fecha de envío) : detalles de identificación y sincronización de cada transacción.
- Customer (Cliente), Segment (Segmento) : Información sobre el comprador y el segmento de mercado (por ejemplo, Consumidor, Oficina Central).
- Manufactory (Manufactura), Product_name (Nombre del producto), Category (Categoría), Subcategory (Subcategoría) : Detalles sobre el origen, tipo y clasificación del producto
- Region (Región), City (Ciudad), State (Estado), Country (País), Zip (Código Postal) : Información geográfica sobre la ubicación de la transacción.
- Sales (Ventas), Quantity (cantidad), Discount (descuento), Profit (beneficio), Profit_margin (margen de beneficio) : cifras financieras que proporcionan información sobre el impacto económico de la transacción.

Objetivo

El dataset con el que trabajaré arroja un conjunto de datos apreciables para el análisis de las ventas de un hipermercado durante el período 2019 hasta 2022. Aquí se podrá valorar las categorías y productos más requeridos, pero en contra partida visualizaremos aquellos que fueron menos solicitados por el cliente para proyectar como reforzar esas ventas y promover mayor circulación. También tendremos en cuenta los países que más consumen, y como redoblar la apuesta para motivar aquellos países que aún no predominan en este registro. Trataremos de comprender si los beneficios tuvieron influencia para que genere mas ventas de productos específicos y entender si debemos cambiar de estrategia en otros productos que no fueron tanto.

Hipótesis

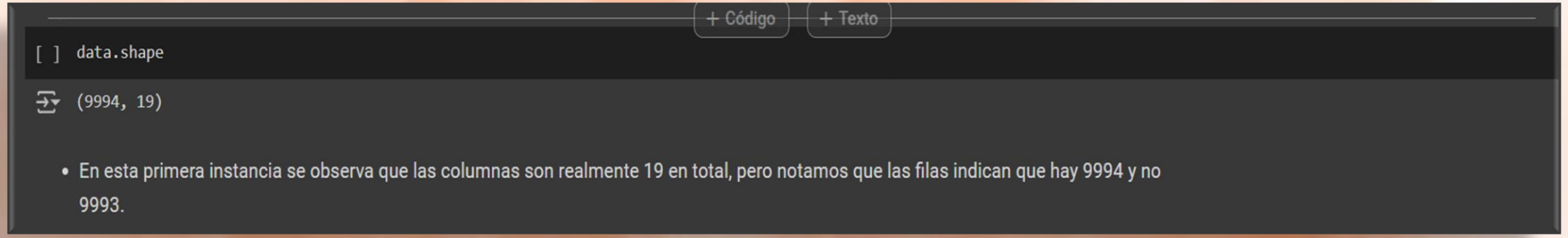
- Se considera que a los largo de este periodo de tiempo hubo un gran volumen de ventas en las categorias de oficina y tecnología por arriba del resto.
- Con buenos márgenes de beneficio, y las ciudades con mayores compras no superior a 6 ciudades.
- El volumen de ventas superior a los 700, y ventas totales mayores a 25.000mil.
- Y pronosticaré que las ventas dentro de tecnología seguirá siendo la venta con más fuerza en el mercado, pero podría darse que los productos de mueble podría ser otro fuerte en la ventas y se realizan estrategia de ventas para la comodidad y practicidad.

CONOCIMIENTO DEL DATASET

	order_id	order_date	ship_date	customer	manufactory	product_name	segment	category	subcategory	region	zip	city	state	country	discount	profit	quantity	sales	profit_margin
0	US-2020-103800	1/3/2019	1/7/2019	Darren Powers	Message Book	Message Book, Wirebound, Four 5 1/2" X 4" Form...	Consumer	Office Supplies	Paper	Central	77095	Houston	Texas	United States	0.2	5.5512	2	16.448	0.3375
1	US-2020-112326	1/4/2019	1/8/2019	Phillina Ober	GBC	GBC Standard Plastic Binding Systems Combs	Home Office	Office Supplies	Binders	Central	60540	Naperville	Illinois	United States	0.8	-5.4870	2	3.540	-1.5500
2	US-2020-112326	1/4/2019	1/8/2019	Phillina Ober	Avery	Avery 508	Home Office	Office Supplies	Labels	Central	60540	Naperville	Illinois	United States	0.2	4.2717	3	11.784	0.3625
3	US-2020-112326	1/4/2019	1/8/2019	Phillina Ober	SAFCO	SAFCO Boltless Steel Shelving	Home Office	Office Supplies	Storage	Central	60540	Naperville	Illinois	United States	0.2	-64.7748	3	272.736	-0.2375
4	US-2020-141817	1/5/2019	1/12/2019	Mick Brown	Avery	Avery Hi-Liter EverBold Pen Style Fluorescent ...	Consumer	Office Supplies	Art	East	19143	Philadelphia	Pennsylvania	United States	0.2	4.8840	3	19.536	0.2500
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	US-2023-126221	12/30/2022	1/5/2023	Chuck Clark	Eureka	Eureka The Boss Plus 12-Amp Hard Box Upright V...	Home Office	Office Supplies	Appliances	Central	47201	Columbus	Indiana	United States	0.0	56.5110	2	209.300	0.2700
9990	US-2023-143259	12/30/2022	1/3/2023	Patrick O'Donnell	Other	Bush Westfield Collection Bookcases, Fully Ass...	Consumer	Furniture	Bookcases	East	10009	New York City	New York	United States	0.2	12.1176	4	323.136	0.0375
9991	US-2023-143259	12/30/2022	1/3/2023	Patrick O'Donnell	Wilson Jones	Wilson Jones Legal Size Ring Binders	Consumer	Office Supplies	Binders	East	10009	New York City	New York	United States	0.2	19.7910	3	52.776	0.3750
9992	US-2023-143259	12/30/2022	1/3/2023	Patrick O'Donnell	Other	Gear Head AU3700S Headset	Consumer	Technology	Phones	East	10009	New York City	New York	United States	0.0	2.7279	7	90.930	0.0300
9993	US-2023-156720	12/30/2022	1/3/2023	Jill Matthias	Other	Bagged Rubber Bands	Consumer	Office Supplies	Fasteners	West	80538	Loveland	Colorado	United States	0.2	-0.6048	3	3.024	-0.2000

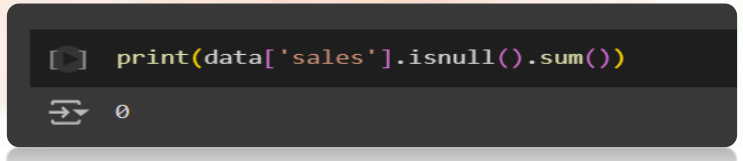
Aquí tenemos una primera visualización de la estructura de los datos, tanto columnas (las 19) como filas (las primeras 5 filas y las 5 filas ultimas). A primera vista se observa que contiene datos de tipo numéricos como alfanuméricos, fechas, entre otros.

Conocimiento exacto del las filas y columnas



Conocimiento de las columnas totales

No hay faltantes de valores



No contiene datos Nulos

```
[ ]  data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   order_id        9994 non-null   object
 1   order_date      9994 non-null   object
 2   ship_date       9994 non-null   object
 3   customer        9994 non-null   object
 4   manufactory     9994 non-null   object
 5   product_name    9994 non-null   object
 6   segment         9994 non-null   object
 7   category        9994 non-null   object
 8   subcategory     9994 non-null   object
 9   region          9994 non-null   object
 10  zip             9994 non-null   int64
 11  city            9994 non-null   object
 12  state           9994 non-null   object
 13  country         9994 non-null   object
 14  discount        9994 non-null   float64
 15  profit          9994 non-null   float64
 16  quantity        9994 non-null   int64
 17  sales           9994 non-null   float64
 18  profit_margin   9994 non-null   float64
dtypes: float64(4), int64(2), object(13)
memory usage: 1.4+ MB
```

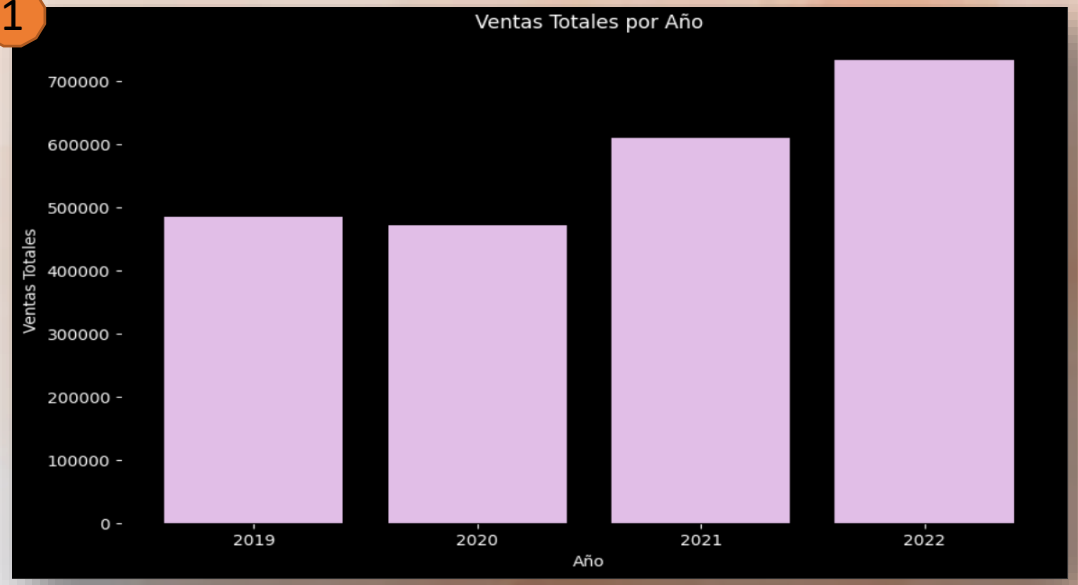
Descripción general del dataset

```
[ ]  data.describe()
```

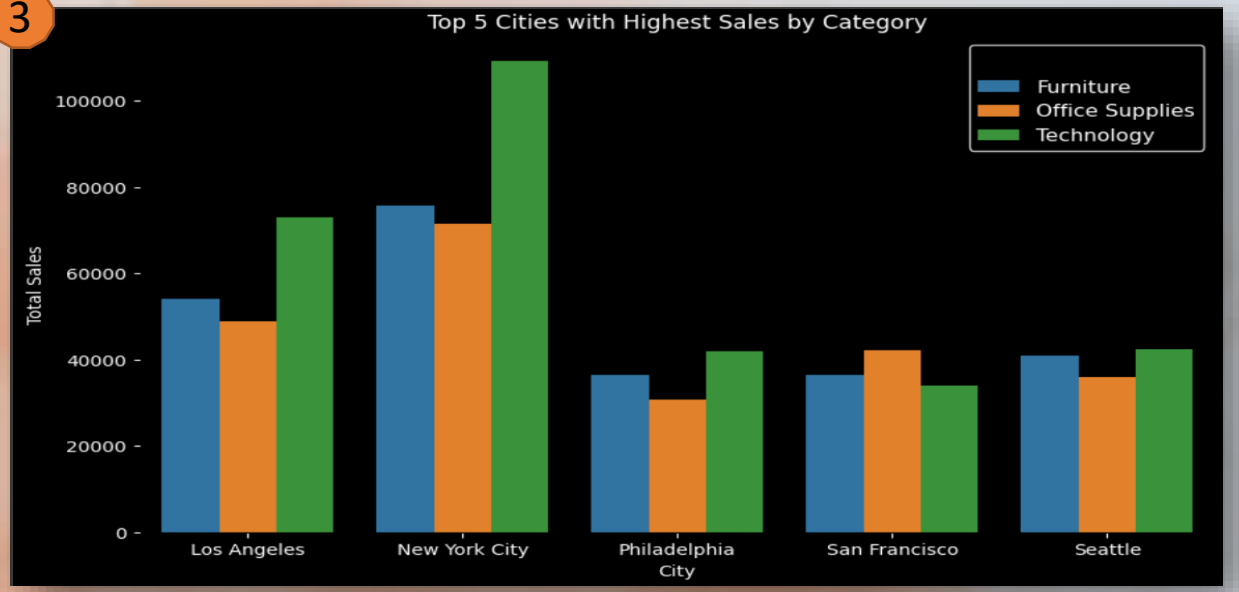
	zip	discount	profit	quantity	sales	profit_margin
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	55190.371723	0.156203	28.656896	3.789574	229.858001	0.120314
std	32063.705315	0.206452	234.260108	2.225110	623.245101	0.466754
min	1040.000000	0.000000	-6599.978000	1.000000	0.444000	-2.750000
25%	23223.000000	0.000000	1.728750	2.000000	17.280000	0.075000
50%	56430.500000	0.200000	8.666500	3.000000	54.490000	0.270000
75%	90008.000000	0.200000	29.364000	5.000000	209.940000	0.362500
max	99301.000000	0.800000	8399.976000	14.000000	22638.480000	0.500000

Visualizaciones Gráficas de interés

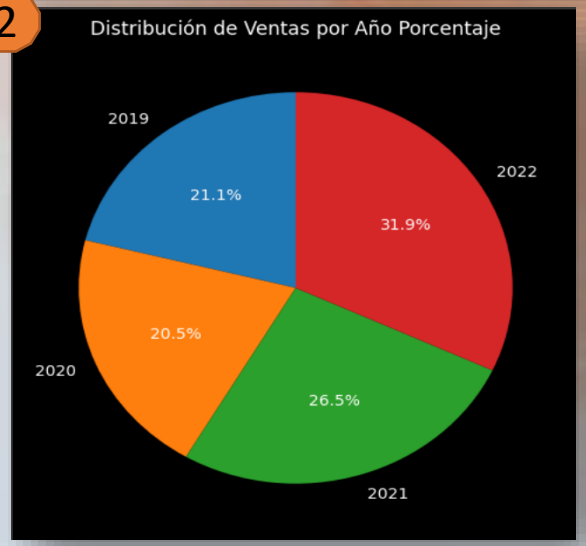
1



3



2



1) Histograma de ventas segun los años registrados en el dataset

2) Gráfico de torta con los porcentajes de venta en los periodos de tiempo 2019 al 2022 de las ventas

3) Grafico de barras que indican las mayores compras según la Categoría, Ciudad y ventas

Gráfico de Histograma Top 10 de las ciudades que realizaron más compras

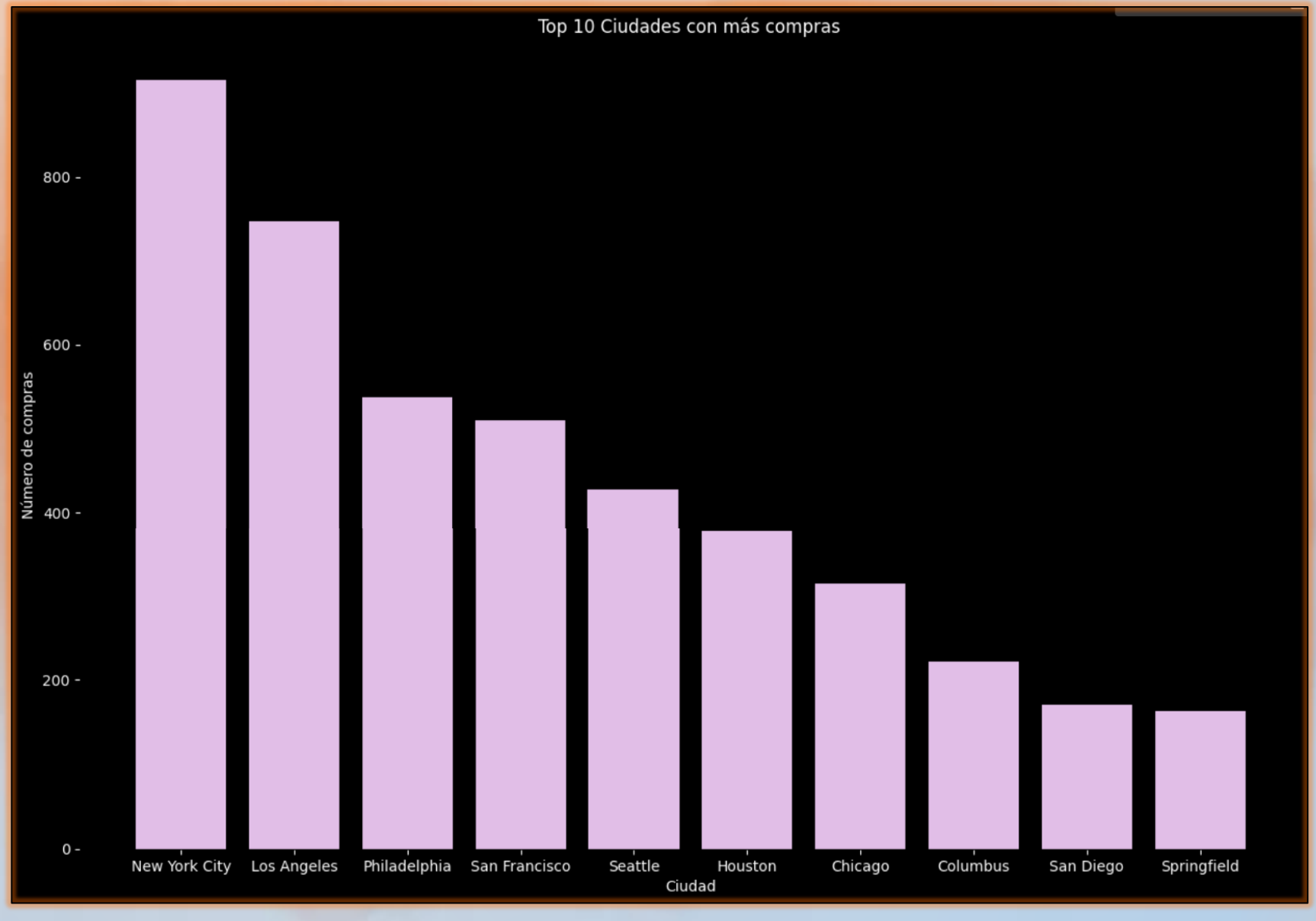
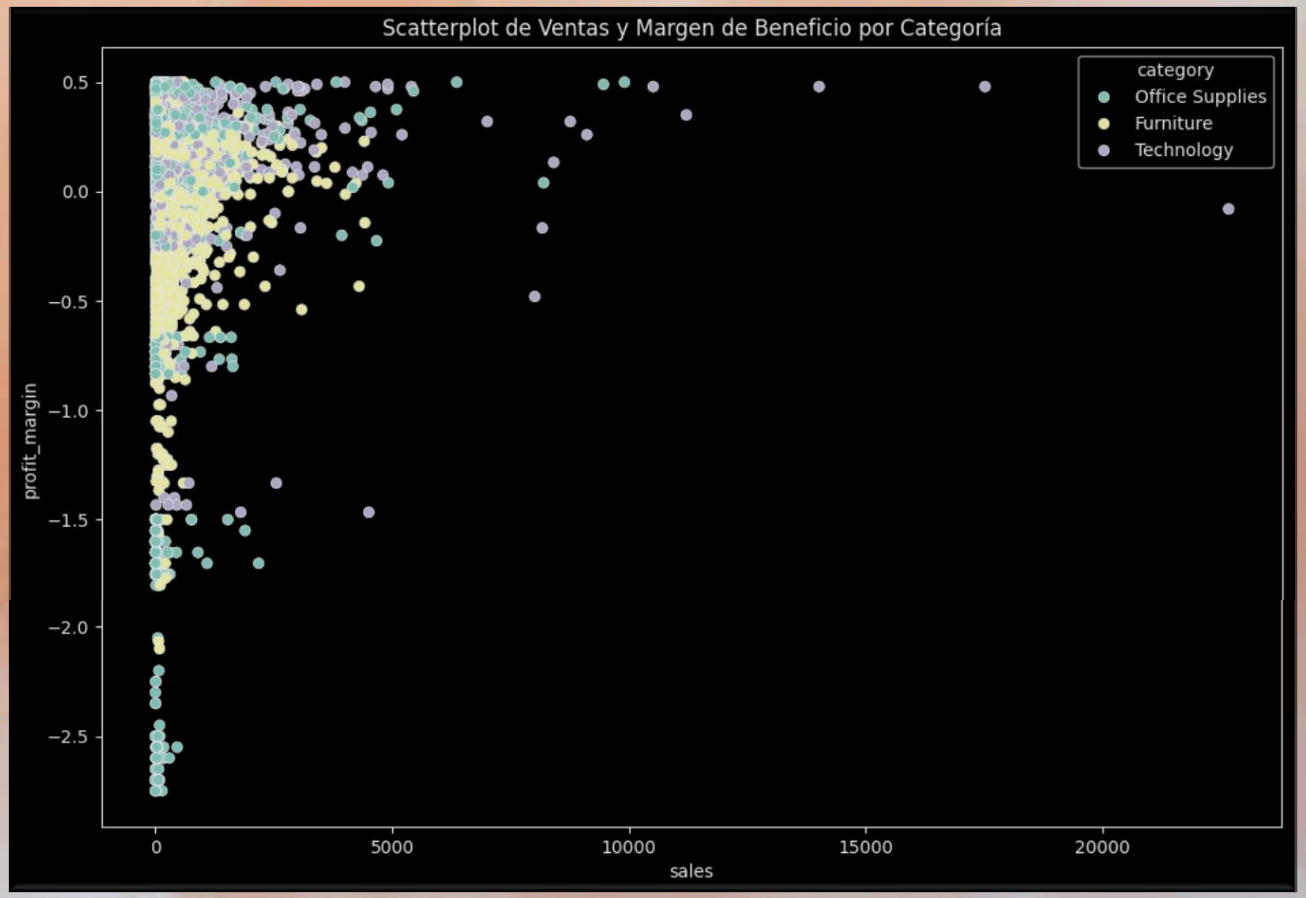
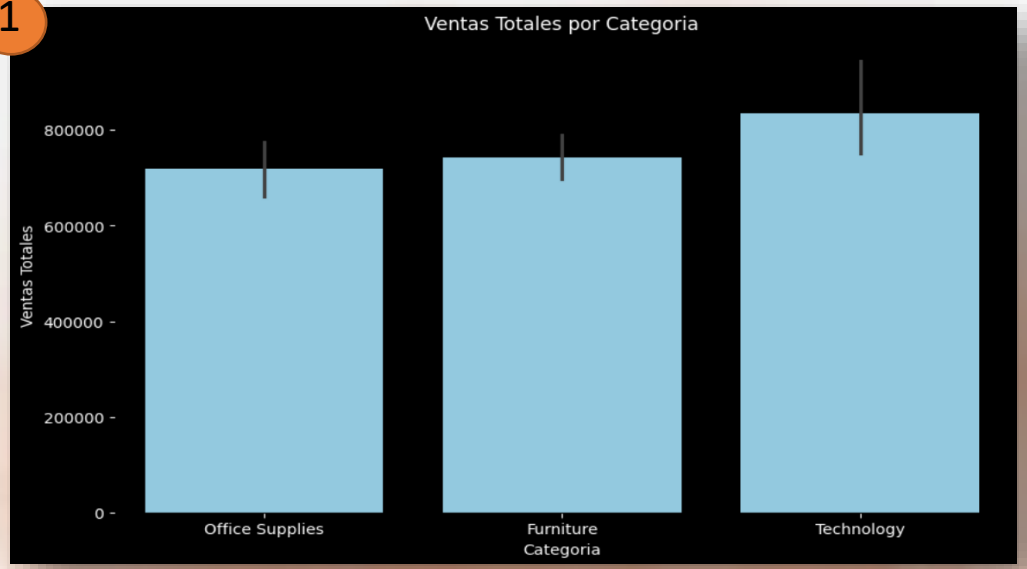


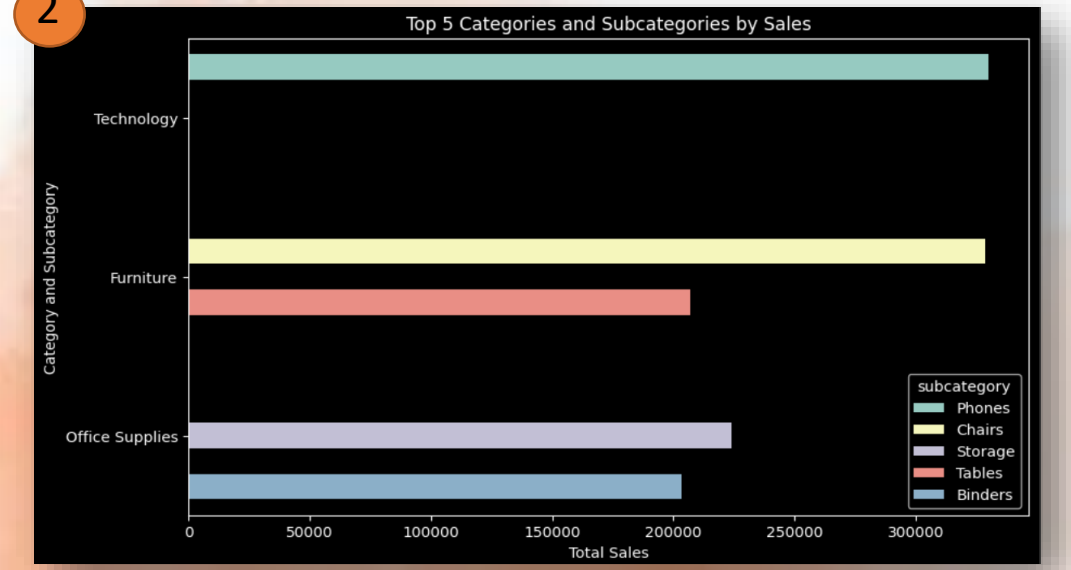
Gráfico Scatterplot de las ventas con mejores márgenes de beneficios según la categoría



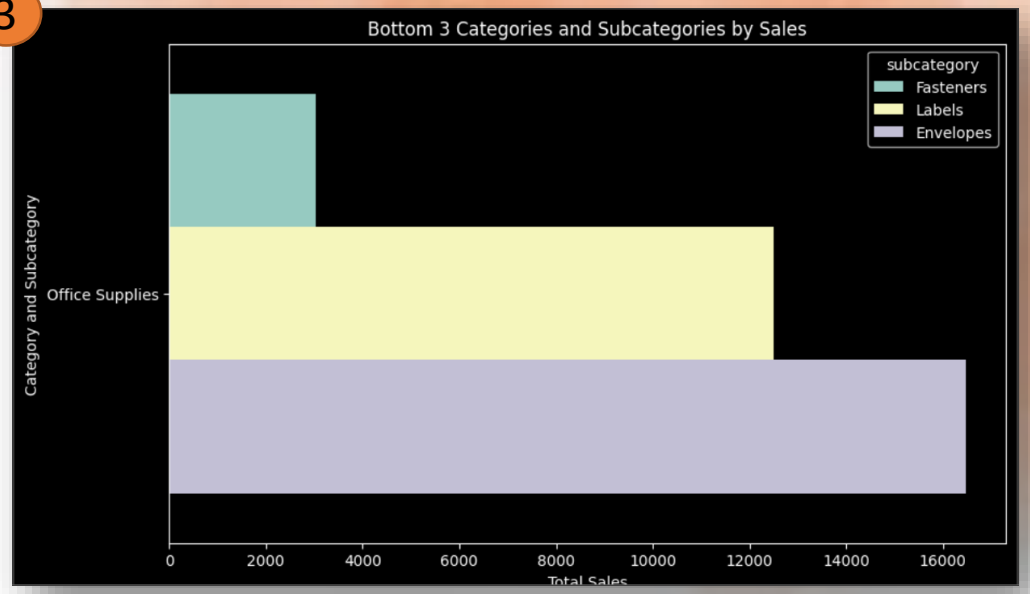
1



2



3

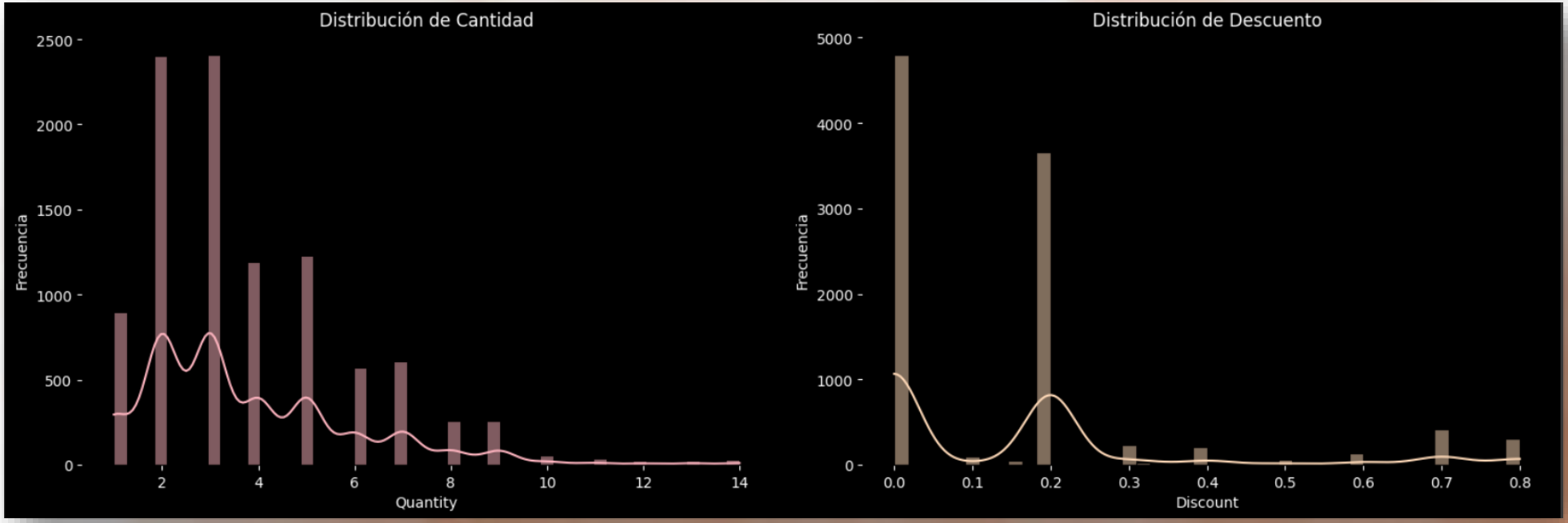


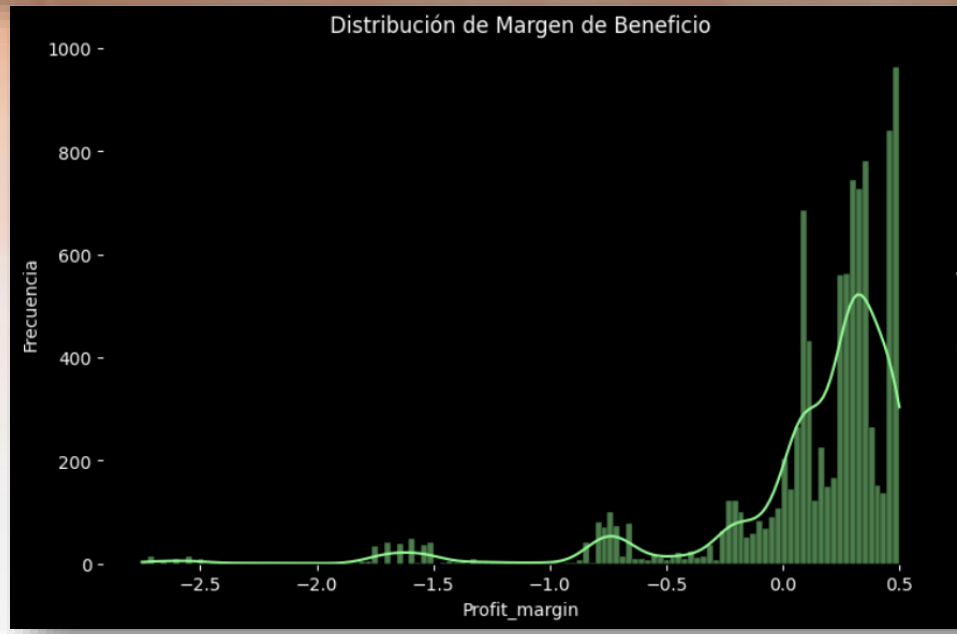
1) Grafico Barplot Ventas por Categoría

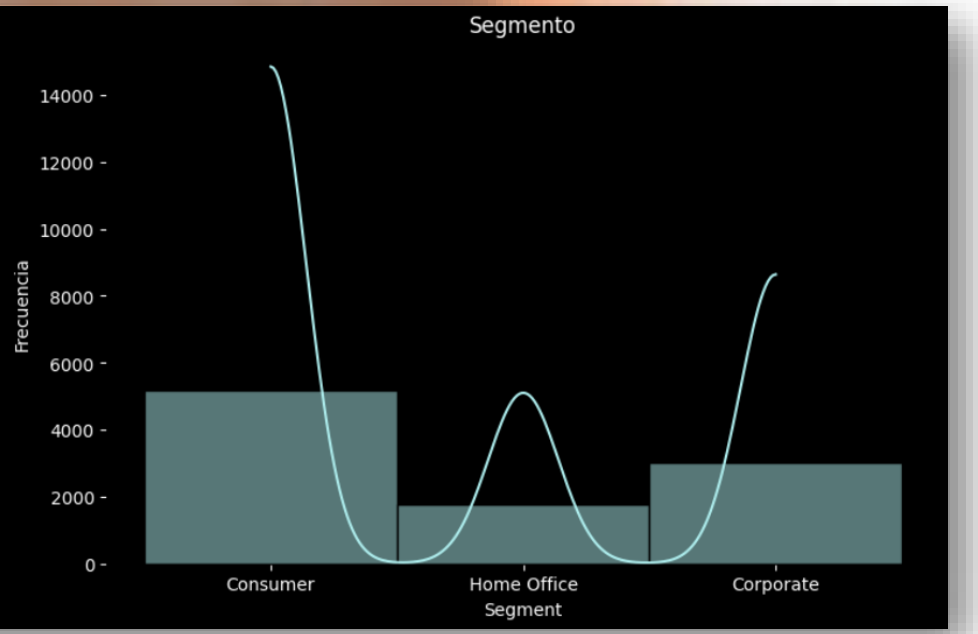
2) Gráfico de barras con un Top 5 de las Categorías y Subcategorías más vendidas

3) Gráfico de Barras Top 3 de las Categorías y Subcategorías menos vendidas.

Mas columnas de interés







Evaluación para saber cuantas variables tengo en las columnas que son categóricas, antes de proceder a trabajar en ML (Machine Learning)

Variables registradas en category (categoria)

```
[ ]  # Indicación de las variables que hoy en la columna Category

     data['category'].unique()
```

```
array(['Office Supplies', 'Furniture', 'Technology'], dtype=object)
```

Variables registradas en Subcategory (subcategoria)

```
[ ]  data['subcategory'].unique()
```

```
array(['Paper', 'Binders', 'Labels', 'Storage', 'Art', 'Chairs',
       'Fasteners', 'Phones', 'Furnishings', 'Accessories', 'Bookcases',
       'Envelopes', 'Appliances', 'Tables', 'Supplies', 'Machines',
       'Copiers'], dtype=object)
```

Variables registradas en City (ciudad)

```
data['city'].unique()
```

```
array(['Houston', 'Naperville', 'Philadelphia', 'Athens', 'Los Angeles',
       'Henderson', 'Huntsville', 'Laredo', 'Springfield', 'Dover',
       'Mount Pleasant', 'Newark', 'San Francisco', 'Bossier City',
       'Roswell', 'Scottsdale', 'Jonesboro', 'Smyrna', 'Westland',
       'Miami', 'Las Vegas', 'Lafayette', 'Alexandria', 'Rapid City',
       'San Diego', 'New York City', 'Detroit', 'Mission Viejo',
       'Green Bay', 'Saint Petersburg', 'Seattle', 'Escondido',
       'Romeoville', 'Chesapeake', 'Linden', 'North Las Vegas',
       'Columbia', 'Concord', 'Dallas', 'Chicago', 'Arlington', 'Lubbock',
       'Richmond', 'Woodstock', 'Moreno Valley', 'El Paso', 'Medford',
       'Columbus', 'Elmhurst', 'Wilmington', 'Margate', 'Yonkers',
       'Des Moines', 'Royal Oak', 'Denver', 'Roseville', 'Jacksonville',
       'Logan', 'Huntington Beach', 'Tampa', 'Raleigh', 'Lakeville',
       'Jackson', 'Burbank', 'Lakeland', 'Asheville', 'Hamilton',
       'Knoxville', 'Portage', 'Tucson', 'Greensboro', 'Delray Beach',
       'Fresno', 'Pomona', 'Albuquerque', 'Plano', 'Apple Valley',
       'Brownsville', 'Long Beach', 'Revere', 'Vallejo', 'Virginia Beach',
       'Dearborn Heights', 'Decatur', 'Lancaster', 'Mobile', 'Marietta',
       'Glendale', 'Toledo', 'Chandler', 'Great Falls', 'Lewiston',
       'Austin', 'Lodi', 'Redondo Beach', 'Bloomington', 'Baltimore',
       'San Jose', 'Troy', 'San Gabriel', 'Jamestown', 'Memphis',
       'Rochester', 'Lake Charles', 'Louisville', 'Appleton',
       'Middletown', 'San Antonio', 'Freeport', 'Lawrence', 'Kent',
       'Fort Worth', 'Watertown', 'Milwaukee', 'Franklin', 'Oakland',
       'Hialeah', 'West Jordan', 'Eau Claire', 'Cleveland', 'Akron',
       'Midland', 'San Marcos', 'Bellevue', 'Murray', 'Buffalo Grove',
       'Lakewood', 'Little Rock', 'Orem', 'Aurora', 'Peoria', 'Bristol',
       'Harrisonburg', 'Hempstead', 'Mishawaka', 'Lawton', 'Waynesboro',
       'Meriden', 'Pueblo', 'Minneapolis', 'Phoenix', 'Chester', 'Salem',
       'Southaven', 'Cincinnati', 'Deltona', 'Plainfield', 'Palm Coast',
       'El Cajon', 'Buffalo', 'Hackensack', 'Niagara Falls',
       'League City', 'Sioux Falls', 'New Rochelle', 'Riverside', 'Omaha',
       'Atlanta', 'Draper', 'Apopka', 'Charlotte', 'Pleasant Grove',
       'Bangor', 'Texas City', 'Trenton', 'Vacaville', 'Hollywood',
       'Fairfield', 'Hampton', 'North Miami', 'Saint Charles',
       'Grand Rapids', 'Billings', 'Oceanside', 'Owensboro', 'Santa Fe',
       'Fayetteville', 'Bowling Green', 'Santa Clara', 'Tulsa', 'Oswego',
       'Pasco', 'Tyler', 'Macon', 'Greenville', 'Lowell', 'Clifton',
       'Gresham', 'Oxnard', 'Olathe', 'Cary', 'Odessa', 'Tempe',
       'Corpus Christi', 'Chula Vista', 'Garland', 'Boca Raton',
       'Mesquite', 'Clarksville', 'Boynton Beach', 'Reno', 'Evanston',
       'Durham', 'Indianapolis', 'Manteca', 'Pasadena', 'Mount Vernon',
       'Edmonds', 'Everett', 'Parma', 'Beaumont', 'Montgomery',
       'Texarkana', 'Newport News', 'Fort Lauderdale', 'Rock Hill',
       'Rancho Cucamonga', 'Garden City', 'Lorain', 'Cranston',
       'Avondale', 'Mason', 'Orange', 'Portland', 'Medina', 'Irving',
       'Nashville', 'Wausau', 'Redding', 'Madison', 'Reading',
       'Carrollton', 'Johnson City', 'Manhattan', 'Cedar Hill',
       'Moorhead', 'Provo', 'Des Plaines', 'Salt Lake City',
       'Coon Rapids', 'Monroe', 'Bolingbrook', 'Sacramento',
       'Saint Louis', 'Woonsocket', 'Brentwood', 'Utica', 'Tigard',
       'Skokie', 'Orlando', 'Sandy Springs', 'Clinton', 'Oklahoma City',
       'Gilbert', 'Olympia', 'Mesa', 'Caldwell', 'Marion', 'Florence',
       'Grand Prairie', 'Thornton', 'Port Arthur', 'Colorado Springs',
       'Beverly', 'Anaheim', 'Cottage Grove', 'Taylor', 'Providence',
       'Baytown', 'Woodbury', 'Park Ridge', 'Bartlett', 'Bozeman',
       'West Palm Beach', 'Rome', 'Suffolk', 'Kenosha', 'Perth Amboy',
       'Hot Springs', 'Las Cruces', 'Sterling Heights', 'Leominster',
       'Altoona', 'Coppell', 'Bethlehem', 'New Castle', 'Plantation',
       'Chico', 'Lehi', 'Auburn', 'San Bernardino', 'Thousand Oaks',
       'Coral Springs', 'Covington', 'Normal', 'Lansing', 'Spokane',
       'Norwich', 'Norfolk', 'Farmington', 'Santa Maria', 'New Albany',
       'Mcallen', 'Daytona Beach', 'Washington', 'Tinley Park', 'Allen',
       'Cuyahoga Falls', 'Camarillo', 'Wilson', 'Frankfort',
       'Haltom City', 'Wichita', 'Manchester', 'Paterson', 'Pocatello',
       'Layton', 'East Point', 'Carol Stream', 'Vineland', 'Holyoke',
       'Amarillo', 'Bakersfield', 'Port Saint Lucie', 'Highland Park',
       'South Bend', 'Hattiesburg', 'Kirkwood', 'Boise', 'Eagan',
       'Redmond', 'Yucaipa', 'New Bedford', 'Allentown', 'Murrieta',
       'Bedford', 'Holland', 'Charlottesville', 'Tamarac', 'La Quinta',
       'Redlands', 'North Charleston', 'Quincy', 'Lincoln Park',
       'Dubuque', 'Broken Arrow', 'Murfreesboro', 'Rockford', 'Bayonne',
       'Cambridge', 'Rockville', 'Hillsboro', 'Warner Robins',
       'Ann Arbor', 'Santa Barbara', 'Noblesville', 'Orland Park',
       'Sparks', 'Salinas', 'Conway', 'Burlington', 'Lebanon', 'Helena',
       'Rio Rancho', 'Frisco', 'Morristown', 'Lake Elsinore',
       'Pembroke Pines', 'Champaign', 'Dearborn', 'Santa Ana',
       'Tallahassee', 'Temecula', 'Costa Mesa', 'Glenview', 'Lindenhurst',
       'Bullhead City', 'Superior', 'Dublin', 'Visalia', 'Missoula',
       'Gaithersburg', 'Longview', 'Westfield', 'Gulfport',
       'Atlantic City', 'Sierra Vista', 'Chattanooga', 'Belleville',
       'La Crosse', 'Round Rock', 'Milford', 'Andover', 'Redwood City',
       'Harlingen', 'Bryan', 'Malden', 'Littleton', 'Saint Peters',
       'Norman', 'Gastonia', 'Grapevine', 'Jefferson City',
       'San Clemente', 'Hesperia', 'Encinitas', 'Yuma', 'Waterbury',
       'Warwick', 'Passaic', 'Parker', 'Longmont', 'York', 'Broomfield',
       'Pensacola', 'Hendersonville', 'West Allis', 'Kenner', 'Davis',
       'Edinburg', 'Fort Collins', 'Sheboygan', 'Pharr', 'Englewood',
```

Variables en la columna Orden_Fecha (Order_date)

```
[ ]  # Indicación de las variables de la columna order_date

     data['order_date'].unique()
```

```
array(['1/3/2019', '1/4/2019', '1/5/2019', ..., '12/28/2022',
       '12/29/2022', '12/30/2022'], dtype=object)
```

Generar LabelEncoder (para pasar los object a numeros para poder trabajar mejor el ML)

```
        order_id order_date   ship_date       customer    manufactory  \
0  US-2020-103800    1/3/2019    1/7/2019  Darren Powers  Message Book
1  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober           GBC
2  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober         Avery
3  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober         SAFCO
4  US-2020-141817    1/5/2019   1/12/2019     Mick Brown         Avery

                                      product_name      segment  \
0  Message Book, Wirebound, Four 5 1/2" X 4" Form...     Consumer
1         GBC Standard Plastic Binding Systems Combs  Home Office
2                                       Avery 508  Home Office
3                        SAFCO Boltless Steel Shelving  Home Office
4  Avery Hi-Liter EverBold Pen Style Fluorescent ...     Consumer

         category subcategory   region    zip          city          state  \
0  Office Supplies       Paper  Central  77095       Houston          Texas
1  Office Supplies     Binders  Central  60540    Naperville       Illinois
2  Office Supplies      Labels  Central  60540    Naperville       Illinois
3  Office Supplies     Storage  Central  60540    Naperville       Illinois
4  Office Supplies         Art     East  19143  Philadelphia   Pennsylvania

        country  discount    profit  quantity     sales  profit_margin  \
0  United States       0.2    5.5512         2   16.448         0.3375
1  United States       0.8   -5.4870         2    3.540        -1.5500
2  United States       0.2    4.2717         3   11.784         0.3625
3  United States       0.2  -64.7748         3  272.736        -0.2375
4  United States       0.2    4.8840         3   19.536         0.2500
```

```
   subcategory_encoded
0                   12
1                    3
2                   10
3                   14
4                    2
```

-En este caso use el LabelEncoder ya que la cantidad de variables son significativas, generando una sola colunma con numeros que representan a cada subcategoria.

Generar One Hot Encoding de la columna Category

```
        order_id order_date   ship_date       customer    manufactory  \
0  US-2020-103800    1/3/2019    1/7/2019  Darren Powers  Message Book
1  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober           GBC
2  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober         Avery
3  US-2020-112326    1/4/2019    1/8/2019  Phillina Ober         SAFCO
4  US-2020-141817    1/5/2019   1/12/2019     Mick Brown         Avery

                                      product_name      segment  \
0  Message Book, Wirebound, Four 5 1/2" X 4" Form...     Consumer
1         GBC Standard Plastic Binding Systems Combs  Home Office
2                                       Avery 508  Home Office
3                        SAFCO Boltless Steel Shelving  Home Office
4  Avery Hi-Liter EverBold Pen Style Fluorescent ...     Consumer

         category subcategory   region  ...          state        country  \
0  Office Supplies       Paper  Central  ...          Texas  United States
1  Office Supplies     Binders  Central  ...       Illinois  United States
2  Office Supplies      Labels  Central  ...       Illinois  United States
3  Office Supplies     Storage  Central  ...       Illinois  United States
4  Office Supplies         Art     East  ...   Pennsylvania  United States

   discount    profit  quantity     sales  profit_margin  category_Furniture  \
0       0.2    5.5512         2   16.448         0.3375                 0.0
1       0.8   -5.4870         2    3.540        -1.5500                 0.0
2       0.2    4.2717         3   11.784         0.3625                 0.0
3       0.2  -64.7748         3  272.736        -0.2375                 0.0
4       0.2    4.8840         3   19.536         0.2500                 0.0
```

```
   category_Office Supplies  category_Technology
0                       1.0                  0.0
1                       1.0                  0.0
2                       1.0                  0.0
3                       1.0                  0.0
4                       1.0                  0.0

[5 rows x 22 columns]
```

- Aqui se encodio la columna de Categoria usando One Hot Encoding, ya que las variables son solo tres. Efectivamente se generaron columnas de las variables de categoria con los nombres: Category_Furniture, Category_Office y Category_Technology.

LabelEncoder columna City

```
         order_id order_date  ship_date       customer    manufactory  \
0  US-2020-103800   1/3/2019   1/7/2019  Darren Powers  Message Book
1  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober           GBC
2  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober         Avery
3  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober         SAFCO
4  US-2020-141817   1/5/2019  1/12/2019     Mick Brown         Avery

                                      product_name      segment  \
0  Message Book, Wirebound, Four 5 1/2" X 4" Form...     Consumer
1         GBC Standard Plastic Binding Systems Combs  Home Office
2                                      Avery 508  Home Office
3                  SAFCO Boltless Steel Shelving  Home Office
4  Avery Hi-Liter EverBold Pen Style Fluorescent ...     Consumer

          category subcategory   region  ...          city          state  \
0  Office Supplies       Paper  Central  ...       Houston          Texas
1  Office Supplies     Binders  Central  ...    Naperville       Illinois
2  Office Supplies      Labels  Central  ...    Naperville       Illinois
3  Office Supplies     Storage  Central  ...    Naperville       Illinois
4  Office Supplies         Art     East  ...  Philadelphia   Pennsylvania

         country discount    profit  quantity    sales  profit_margin  \
0  United States      0.2    5.5512         2   16.448         0.3375
1  United States      0.8   -5.4870         2    3.540        -1.5500
2  United States      0.2    4.2717         3   11.784         0.3625
3  United States      0.2  -64.7748         3  272.736        -0.2375
4  United States      0.2    4.8840         3   19.536         0.2500

   subcategory_encoded  city_encoded
0                   12           207
1                    3           321
2                   10           321
3                   14           321
4                    2           374

[5 rows x 21 columns]
```

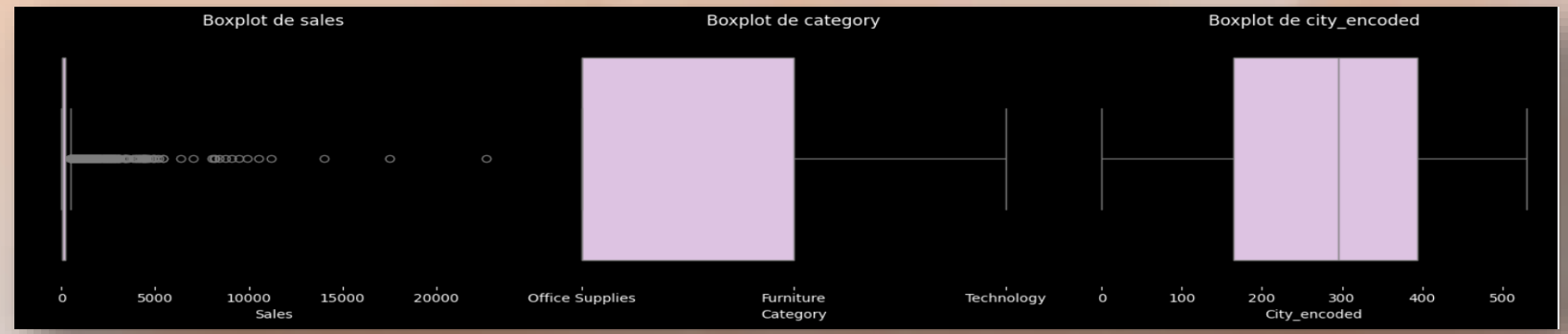
- En el caso de Ciudad, se uso tambien el LabelEncoder, ya que las ciudades registradas superan la cantidad de 300 ciudades. Esto ayudara a trabajar mejor e el Machine Learning (ML)

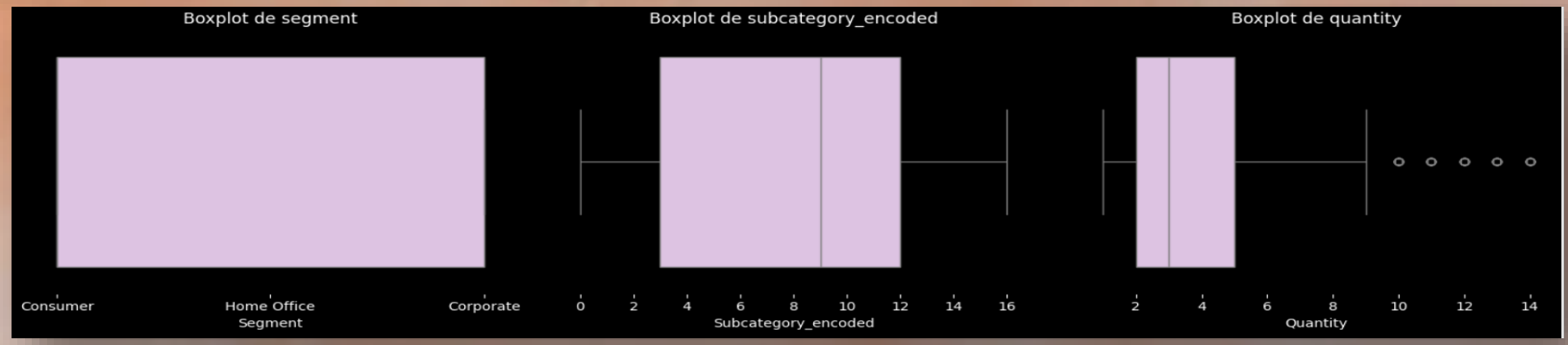
LabelEncoder de la columna Order Date

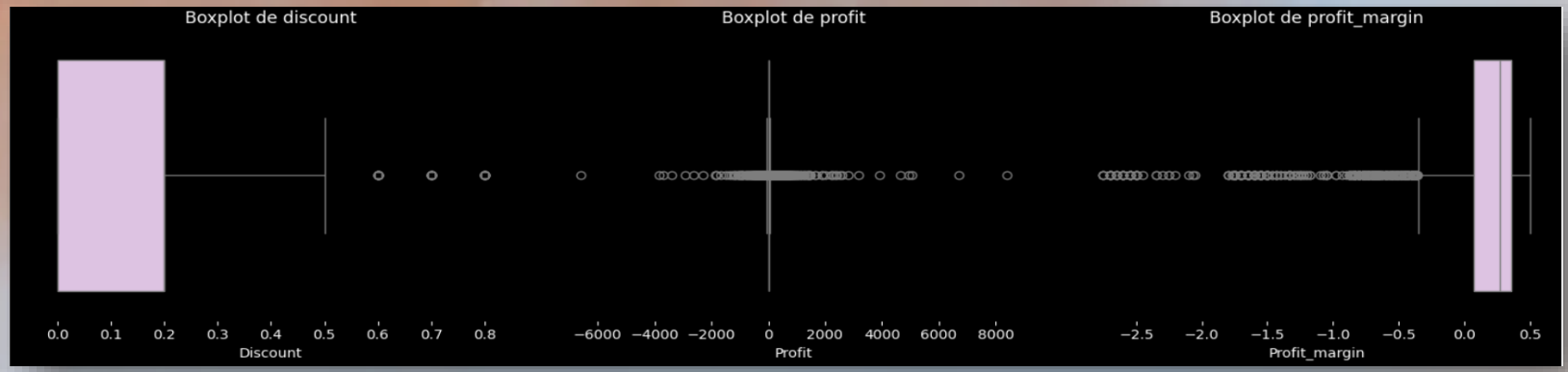
```
         order_id order_date  ship_date       customer    manufactory  \
0  US-2020-103800   1/3/2019   1/7/2019  Darren Powers  Message Book
1  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober           GBC
2  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober         Avery
3  US-2020-112326   1/4/2019   1/8/2019  Phillina Ober         SAFCO
4  US-2020-141817   1/5/2019  1/12/2019     Mick Brown         Avery

                                      product_name      segment  \
0  Message Book, Wirebound, Four 5 1/2" X 4" Form...     Consumer
1         GBC Standard Plastic Binding Systems Combs  Home Office
2                                      Avery 508  Home Office
3                  SAFCO Boltless Steel Shelving  Home Office
4  Avery Hi-Liter EverBold Pen Style Fluorescent ...     Consumer

          category subcategory   region  ...          state        country  \
0  Office Supplies       Paper  Central  ...          Texas  United States
1  Office Supplies     Binders  Central  ...       Illinois  United States
2  Office Supplies      Labels  Central  ...       Illinois  United States
3  Office Supplies     Storage  Central  ...       Illinois  United States
4  Office Supplies         Art     East  ...   Pennsylvania  United States

   discount    profit  quantity    sales  profit_margin  subcategory_encoded  \
0       0.2    5.5512         2   16.448         0.3375                   12
1       0.8   -5.4870         2    3.540        -1.5500                    3
2       0.2    4.2717         3   11.784         0.3625                   10
3       0.2  -64.7748         3  272.736        -0.2375                   14
4       0.2    4.8840         3   19.536         0.2500                    2

   city_encoded  order_date_encoded
0           207                  55
1           321                  66
2           321                  66
3           321                  66
4           374                  69

[5 rows x 22 columns]
```

- Como el dataset informa que las fechas estan como Object, determine realizarle tambien un LabelEncoder. la columna que se le genero se llama Order_date_encoded.

Gráficos para visualizar los outliers de las columnas de interés







IQR (Índice Intercuartil) - lo utilizaré para los gráficos que presentan anomalías en el cuartil, los casos son: Sales, Category, Segment, Discount, Quantity, profit_margin y profit. Reemplazo: mediana Categoría: moda

	zip	discount	profit	quantity	sales	\
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	
mean	55190.371723	0.111756	11.051194	3.642786	88.383819	
std	32063.705315	0.115233	16.834794	1.963765	107.885487	
min	1040.000000	0.000000	-39.637000	1.000000	0.444000	
25%	23223.000000	0.000000	3.210000	2.000000	17.280000	
50%	56430.500000	0.200000	8.666500	3.000000	54.485000	
75%	90008.000000	0.200000	15.252300	5.000000	105.980000	
max	99301.000000	0.500000	70.722000	9.000000	498.260000	

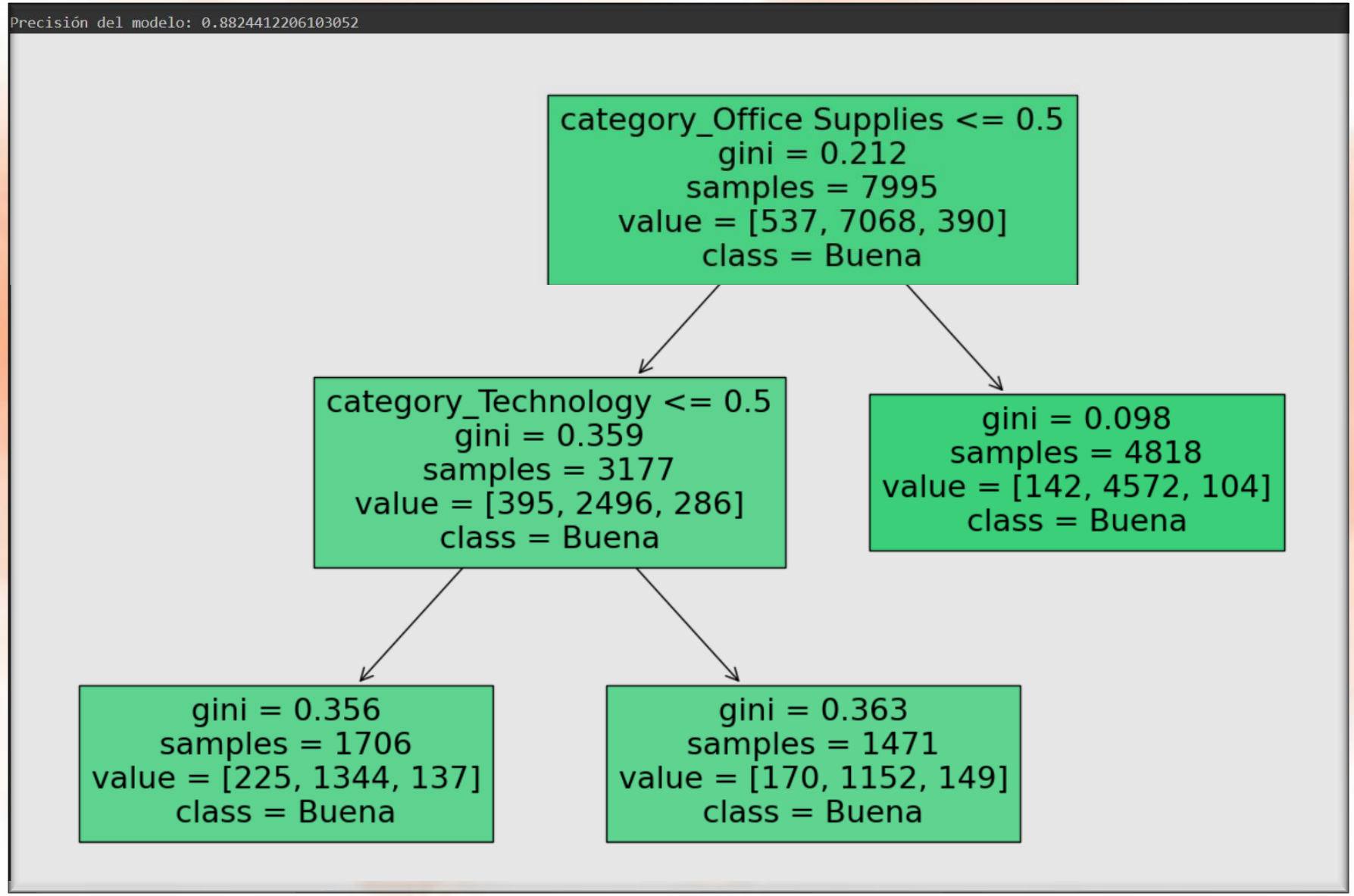
	profit_margin	subcategory_encoded	city_encoded	order_date_encoded
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	0.249004	7.590454	279.957274	626.091955
std	0.182709	5.051429	139.157896	370.432362
min	-0.350000	0.000000	0.000000	0.000000
25%	0.125000	3.000000	164.000000	293.000000
50%	0.270000	9.000000	295.000000	597.000000
75%	0.362500	12.000000	394.000000	968.000000
max	0.500000	16.000000	530.000000	1235.000000

Z-score - Lo utilizare a los gráficos que muestran una distribución normal y le calcularé las desviaciones. En este caso se harán con la columna: City_encoded, y subcategory_encoded. Se reemplaza por la media

	zip	discount	profit	quantity	sales	\
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	
mean	55190.371723	0.111756	11.051194	3.642786	88.383819	
std	32063.705315	0.115233	16.834794	1.963765	107.885487	
min	1040.000000	0.000000	-39.637000	1.000000	0.444000	
25%	23223.000000	0.000000	3.210000	2.000000	17.280000	
50%	56430.500000	0.200000	8.666500	3.000000	54.485000	
75%	90008.000000	0.200000	15.252300	5.000000	105.980000	
max	99301.000000	0.500000	70.722000	9.000000	498.260000	

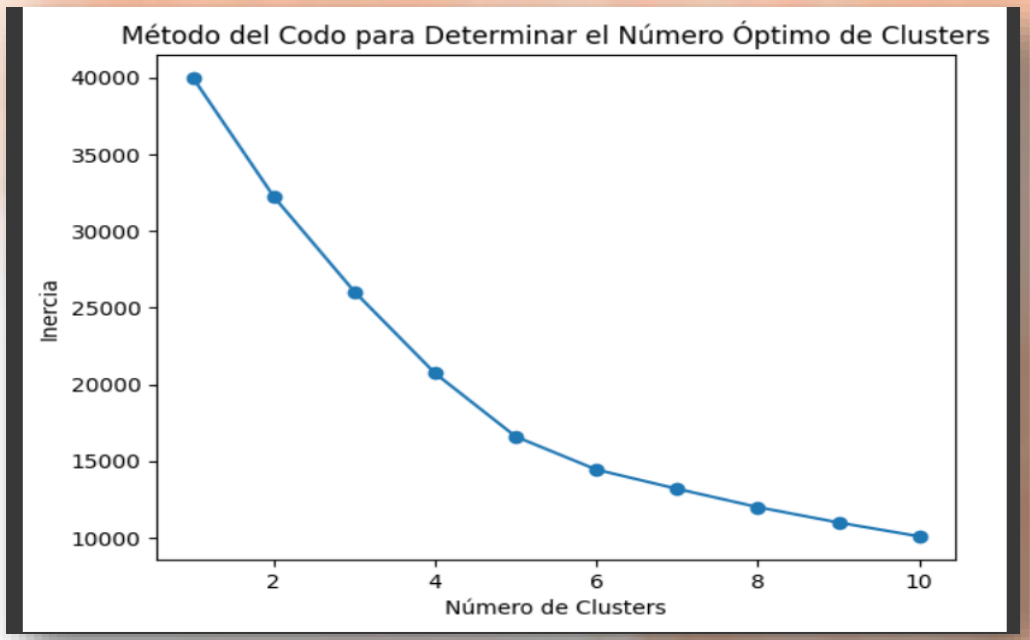
	profit_margin	subcategory_encoded	city_encoded	order_date_encoded
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	0.249004	7.590454	279.957274	626.091955
std	0.182709	5.051429	139.157896	370.432362
min	-0.350000	0.000000	0.000000	0.000000
25%	0.125000	3.000000	164.000000	293.000000
50%	0.270000	9.000000	295.000000	597.000000
75%	0.362500	12.000000	394.000000	968.000000
max	0.500000	16.000000	530.000000	1235.000000

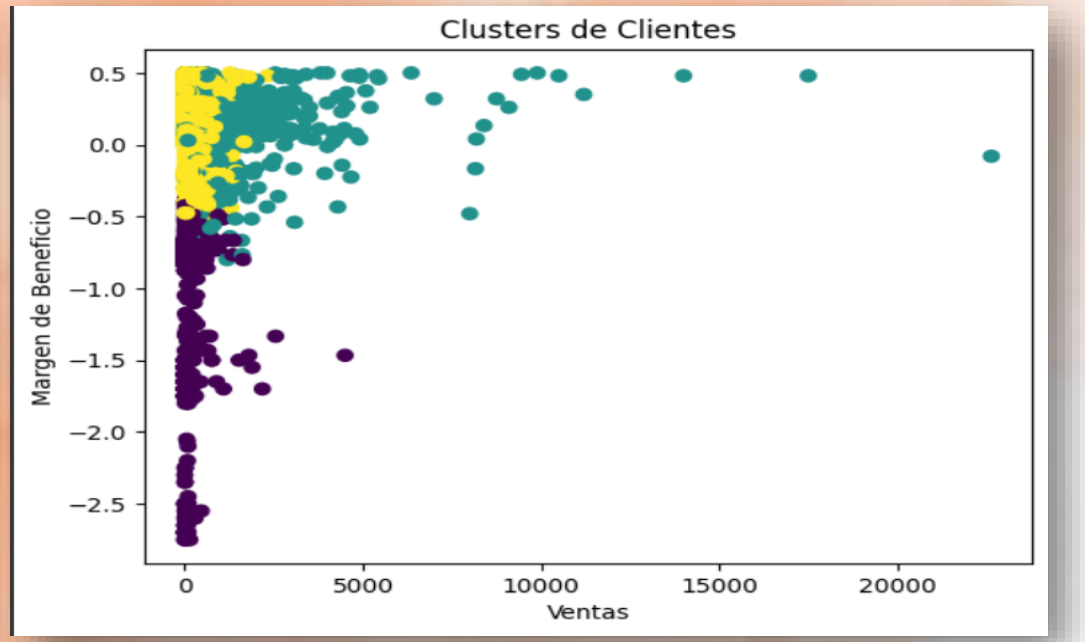
Realización de un Árbol de Decisión con los Sales y Category.



El resultado marca un 88% de certeza en la predicción, siendo las futuras ventas más importantes Office Suplies y Technology. Esto podría también mostrar que se debe reforzar la otra categoría Furniture para lograr mayor ventas y rotación de mercadería. Mejorando el marketing, precios e incluso beneficios.

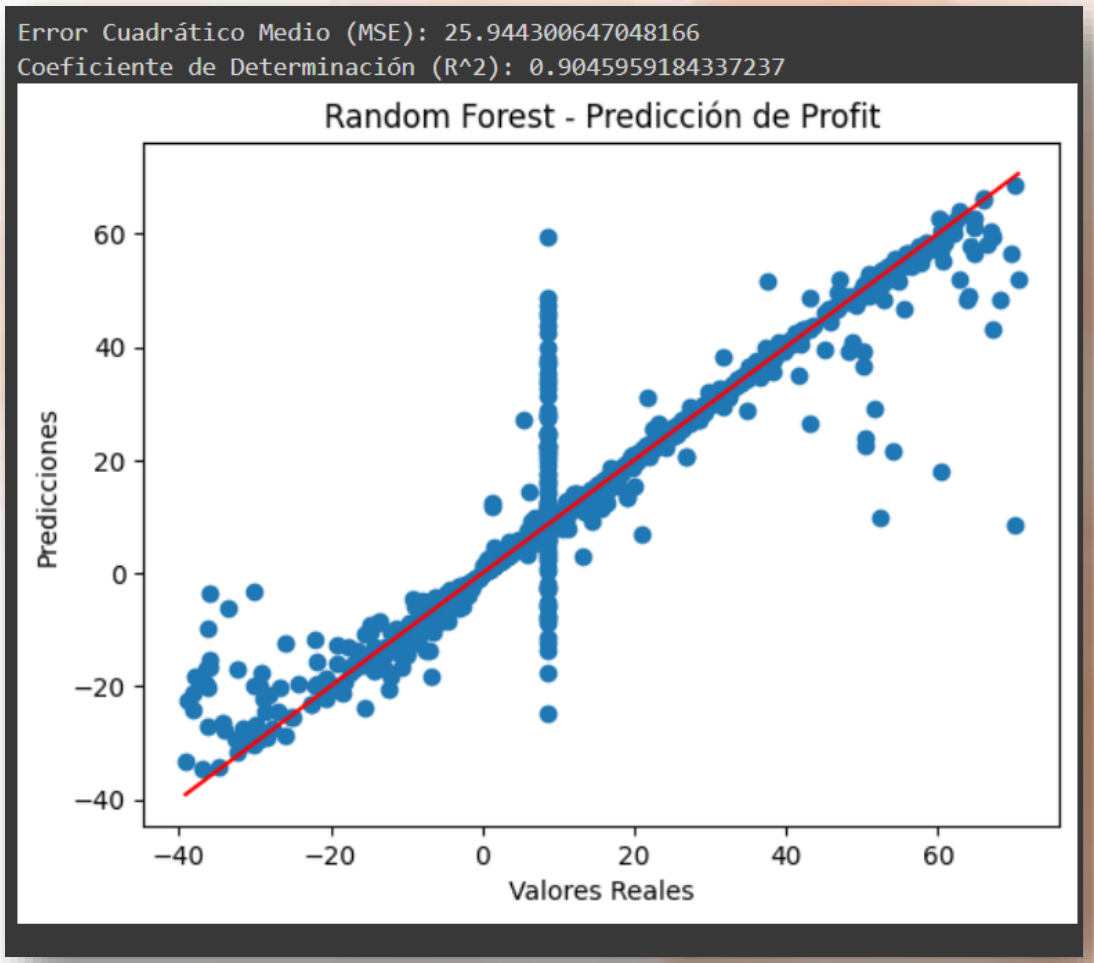
Realización de K-Means (no supervisado) para la predicción de las ventas usando variables de interés Sales, City, Category, quantity, y profit margen registradas.





Las escalas vemos en ambos gráficos que termina siendo ascendente a descendente, mostrado que los productos vendidos se realizaron a mayor margen de beneficio, y los casos aislados fue ventas con poco margen de beneficio según lo registrado en la base de datos. Si se ofrece buenos márgenes de beneficios pueden aún crecer el registro de las ventas y que se sostenga en temporalidad y cantidad de ventas, pero serà necesario seguir mejorando el registro de las mismas. Este resultado es similar al gráfico que se realizó al principio "Gráfico Scatterplot de las ventas con mejores márgenes de beneficios según la categoría".

Algoritmo de Random Forest para ajustar cuadrados y ver predicción



Determiné en tomar el algoritmo de random forest, ya que se ajusta mejor a los datos registrados, teniendo un coeficiente del 90% y un margen de error llegando al 26% (25.94). Caso contrario con otros algoritmos lineales. La recta es ascendente prediciendo un crecimiento favorable en el tiempo según nuestro registro. Contamos con casos aislados que pueden deberse a un conjunto de sucesos como los beneficios adquiridos en las ventas, descuentos cantidades, etc. Si es importante mejorar el registro de los datos para que sea fácil de procesar y analizar, sostener márgenes de beneficios o descuentos a los consumidores y sostener la modalidad de ventas que se realizaron. No marca necesidad de mejores en lo que es marketing, aunque puede avaluarse en nivelar los productos con menor venta y menos beneficio con los productos más demandados.