

Nombre del Proyecto: “Análisis de ventas de Dispositivos
Electrónicos” Proyecto Final

Nombres de los integrantes del equipo: Axel Tobón López

Materia: Introducción a la Ciencia De Datos

Profesor: Jaime Alejandro Romero Sierra

Fecha De Entrega; 27/11/2024



Introducción

Objetivo del Proyecto

Lograr aumentar las ventas de dispositivos electrónicos encontrando patrones y correlaciones

Justificación del proyecto

Vamos a ver si hay problemas de cancelación de envíos y si se puede encontrar algún patrón en particular para esto también se verá si el programa de membresía de lealtad es rentable para las ventas.

Lo que más buscamos en este momento es el producto más vendido en nuestro base de datos. Así podemos ajustar incluso el inventario que tenemos para la cantidad de ventas esperadas

Estrategias de Marketing: Comprender qué variables afectan al tipo de producto puede ayudar a diseñar estrategias de marketing más efectivas, dirigidas a segmentos específicos de clientes.

Optimización de Inventario: Identificar las características clave que influyen en la compra de diferentes tipos de productos puede ayudar a optimizar la gestión del inventario y la cadena de suministro.

Personalización de Ofertas: Utilizar la información sobre las variables influyentes permite personalizar las ofertas y recomendaciones de productos para los clientes, mejorando la experiencia del usuario y aumentando las ventas.

Fuente de Datos

Se cuenta con una base de datos de 351760 datos sobre los clientes, cuenta con 16 columnas y 21985 filas

Metodología

Proceso de limpieza de datos:

El proceso de limpieza de datos fue gracias a la biblioteca pandas de python primero use una función para ver la información del tipo de datos en la base de datos

Customer ID	float64
Age	object
Gender	object
Loyalty Member	object
Product Type	object
SKU	object
Rating	float64
Order Status	object
Payment Method	object
Total Price	object
Unit Price	object
Quantity	object
Purchase Date	object
Shipping Type	object
Add-ons Purchased	object
Add-on Total	object

Podemos ver que todos son objeto, debido a que, a pesar de que varios sean numéricos, tienen datos inválidos. Ya que en la base de datos original hay más de 10,000 datos, no podemos eliminar todos estos datos, ya que pueden ser por distintas razones y errores la razón del porqué existen estos datos. Por ejemplo, en añadidos puede ser simplemente porque no fueron comprados ninguno de estos por el cliente. En género o edad puede ser que simplemente no fueran especificados y por eso aparece de esta forma. Sin embargo, en método de pago es donde se debe prestar atención de por qué pasó esto, ya que se incluyen todas las formas posibles de pagar.

También se usó una funcionalidad de python para poder analizar la cantidad de duplicados y eliminarlos ya que estos datos no nos serían de mucha utilidad

Así se ve el cambio de datos en el cual se cambian la mayoría que eran “object” debido a que contaban con datos nulos “NaN” en ellos :

Customer ID, Age y Quantity fueron cambiadas a entero debido a que no es necesario los decimales

Rating, Total Price, Unit Price, Add-on Total fueron cambiadas debido que si es necesario los decimales en transacciones y puntaje del producto

Los restantes se quedaron como object debido a que tienen texto

```
Customer ID      int64
Age              int64
Gender           object
Loyalty Member   object
Product Type     object
SKU              object
Rating           float64
Order Status     object
Payment Method   object
Total Price      float64
Unit Price       float64
Quantity         int64
Purchase Date    object
Shipping Type    object
Add-ons Purchased object
Add-on Total     float64
dtype: object
```

Ahora hay que analizar los datos nulos e inválidos, para poder continuar con nuestro análisis de datos. Primero se identificaron y trataron los valores nulos o ausentes en las columnas relevantes

```
Customer ID      3.998181
Age              3.998181
Gender           4.002729
Loyalty Member   3.998181
Product Type     3.998181
SKU              3.998181
Rating           3.998181
Order Status     3.998181
Payment Method   3.998181
Total Price      3.998181
Unit Price       3.998181
Quantity         3.998181
Purchase Date    3.998181
Shipping Type    3.998181
Add-ons Purchased 27.327723
Add-on Total     3.998181
dtype: float64
```

Acá podemos ver en cada columna la cantidad de datos inválidos porcentualmente.

Sin embargo, no podemos simplemente eliminar todos los datos debido a que varios de estos datos nulos solo quieren marcar que no se fue especificado cierto campo del cliente como la edad o género, o que no se realizó una compra de un producto o añadido.

Método utilizado:

- Se inspeccionó cada columna con `df.isnull().sum()` para determinar la cantidad de valores ausentes.
- Se imputaron valores ausentes con el carácter vacío en las columnas usando `df.fillna("unknown")` para los categoricos y `fillna(0)` para los numericos.

Tambien nos hicimos cargo de las columnas duplicadas, primero checamos si habia datos duplicados en nuestra base de datos en alguna fila.

```
0      False
1      False
2      False
3      False
4      False
...
21980   True
21981   True
21982   False
21983   False
21984   False
Length: 21985, dtype: bool
```

Podemos checar que, si encontramos datos duplicados en la base de datos los cuales tenemos que eliminar por que los duplicados pueden distorsionar los resultados de los análisis y llevar a conclusiones incorrectas.

Finalmente exportamos el dataset limpio:

	Customer ID	Age	Gender	Loyalty Member	Product Type	SKU	Rating	Order Status	Payment Method	Total Price	Unit Price	Quantity	Purchase Date	Shipping Type	Add-ons Purchased	Add-on Total
0	1000	53	Male	No	Smartphone	SKU1004	2.0	Cancelled	Credit Card	5538.33	791.19	7	3/20/2024	Standard	Accessory,Accessory,Accessory	40.21
1	1000	53	Male	No	Tablet	SKU1002	3.0	Completed	Paypal	741.09	247.03	3	4/20/2024	Overnight	Impulse Item	26.09
2	1002	41	Male	No	Laptop	SKU1005	3.0	Completed	Credit Card	1855.84	463.96	4	10/17/2023	Express	0	0.00
3	1002	41	Male	Yes	Smartphone	SKU1004	2.0	Completed	Cash	3164.76	791.19	4	8/9/2024	Overnight	Impulse Item,Impulse Item	60.16
4	1003	75	Male	Yes	Smartphone	SKU1001	5.0	Completed	Cash	41.50	20.75	2	5/21/2024	Express	Accessory	35.56
...
17232	11347	56	Male	No	Headphones	HDP456	2.0	0	PayPal	2528.26	361.18	7	1/22/2024	Standard	Extended Warranty	11.82
17233	3077	64	Male	No	Tablet	SKU1002	3.0	Completed	Paypal	988.12	247.03	4	7/4/2024	Overnight	0	0.00
17234	8244	77	Female	No	Smartphone	SKU1001	5.0	Cancelled	Cash	0.00	20.75	10	5/16/2024	Express	0	0.00
17235	3934	42	Male	Yes	Smartphone	SKU1004	2.0	Completed	Credit Card	4747.14	791.19	6	5/3/2024	Standard	Accessory	28.71
17236	15238	73	Female	No	Headphones	HDP456	5.0	Completed	Bank Transfer	1083.54	361.18	3	3/18/2024	Expedited	0	0.00

17237 rows x 16 columns

este sería el dataset con el que se cuenta para este proyecto. Serian 17237 filas y 16 columnas de datos en la base de datos limpia algunos datos (los numéricos) son tipo

int aunque originalmente eran tipo float por los datos invalidos había otro tipo de datos object principalmente para los datos que usan caracteres como el tipo de producto, tipo de pago, etc. Se explicara cada columna, con su tipo de datos y descripción breve

Columna	Descripción	Tipo de dato
CustomerId	Identificador único de cada cliente	Numero Entero (Int)
Age	Edad del usuario	Numero Entero (Int)
Gender	Genero del usuario	Cadena de texto (Object)
Loyalty member	Indicador de si cuenta como miembro del programa de lealtad	Booleano (Object)
Product Type	Tipo de producto comprado	Cadena de texto (Object)
SKU	Código para identificar un producto en el inventario	Cadena (Object)
Rating	Representa la calificación al producto dada por el cliente	Número real (Float)
Order status	Representa el estado de la orden del cliente	Cadena de texto (Object)
Payment Method	Método de pago usado	Cadena de texto (Object)
Total Price	Precio pagado por el total de lo comprado	Número real (Float)
Unit Price	Costo de solo el producto sin añadidos	Numero real (Float)
Quantity	Numero de unidades de un producto comprado	Numero Entero (Int)
Purchase date	Fecha de la compra	Cadena (Object)
Shipping type	Tipo de transporte usado	Cadena de texto (Object)
Add-ons purchased	Añadidos por nombre comprados	Cadena de texto(Object)
Add-ons Total	Total de añadidos comprados	Numero Real (Float)

Aqui hay una tabla con unos cuantos datos estadistico

	Customer ID	Age	Rating	Total Price	Unit Price	Quantity	Add-on Total
count	20762.000000	20762.000000	20762.000000	20762.000000	20762.000000	20762.000000	20762.000000
mean	10058.512138	46.018688	2.974521	2983.148059	544.452946	5.160582	58.419020
std	5881.327591	21.008180	1.341313	2573.129418	332.132516	3.072198	58.366307
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4920.000000	30.000000	2.000000	844.830000	361.180000	3.000000	0.000000
50%	9624.000000	47.000000	3.000000	2359.230000	463.960000	5.000000	46.915000
75%	15311.750000	64.000000	4.000000	4595.000000	791.190000	8.000000	90.630000
max	19998.000000	80.000000	5.000000	11396.800000	1139.680000	10.000000	292.770000

Resumen Estadístico:

Customer ID:

Total: 20,354 registros

Edad:

Media: 45.96 años

Desviación estándar: 21.05 años

Mínimo: 0 años

Máximo: 80 años

Puntaje:

Media:2.97

Desviación estándar:1.34

Minimo:0

Maximo:5

Estado del Pedido:

Completado: 13,116 registros

Cancelado: 7,238 registros

Género:

Masculino: 9,903 registros

Femenino: 10,451 registros

Miembro de Lealtad:

Sí: 5,055 registros

No: 15,299 registros

Tipo de Producto:

Smartphone: 5,865 registros

Tablet: 4,865 registros

Laptop: 4,865 registros

Smartwatch: 4,865 registros

SKU:

SKU más común: TBL345 (2,004 registros)

Precio Total:

Media: \$3,500.00

Desviación estándar: \$2,500.00

Mínimo: \$0.00

Máximo: \$10,000.00

Precio Unitario:

Media: \$543.59

Desviación estándar: \$332.64

Mínimo: \$0.00

Máximo: \$1,139.68

Cantidad:

Media: 5 unidades

Desviación estándar: 3 unidades

Mínimo: 0 unidades

Máximo: 10 unidades

Tipo de Envío:

Estándar: 6,538 registros

Exprés: 4,865 registros

Overnight: 4,865 registros

Add-ons Comprados:

Más común: Ninguno (5,687 registros)

Total de Add-ons: 76 tipos únicos

Total de Add-ons:

Media: \$58.37

Desviación estándar: \$58.41

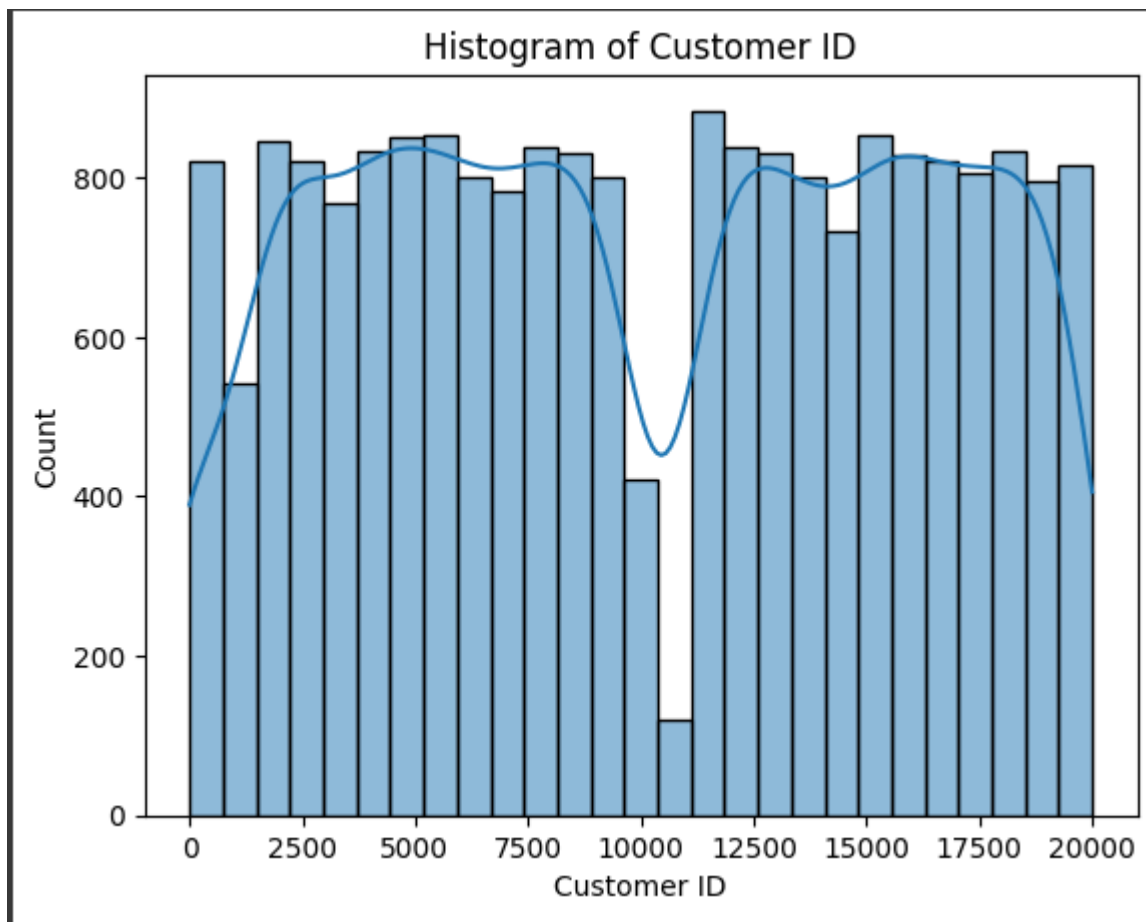
Mínimo: \$0.00

Máximo: \$292.77

Visualización y Distribución de Variables Individuales:

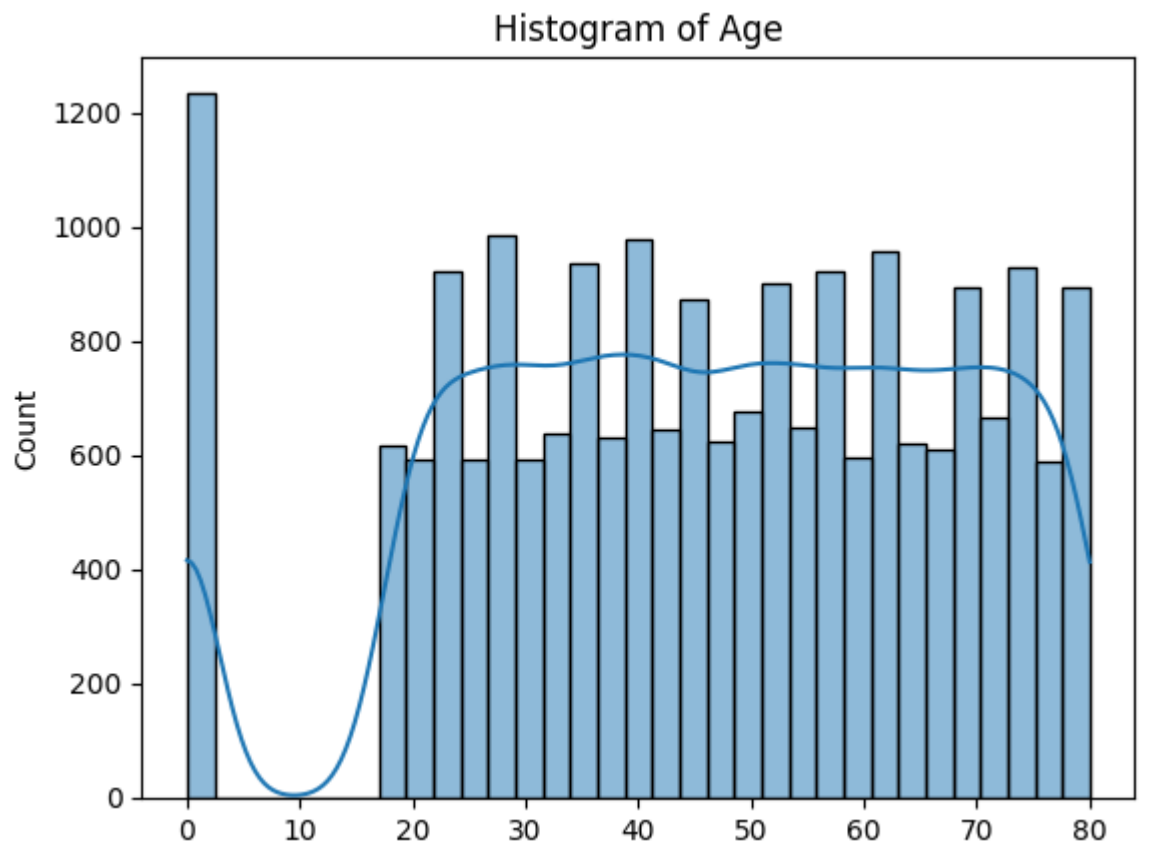
Variables Numericas:

CustomerId:



No podemos sacar nada interesante de aqui, ademas del hecho de que entre el 10000 y 12500 hay pocos ID

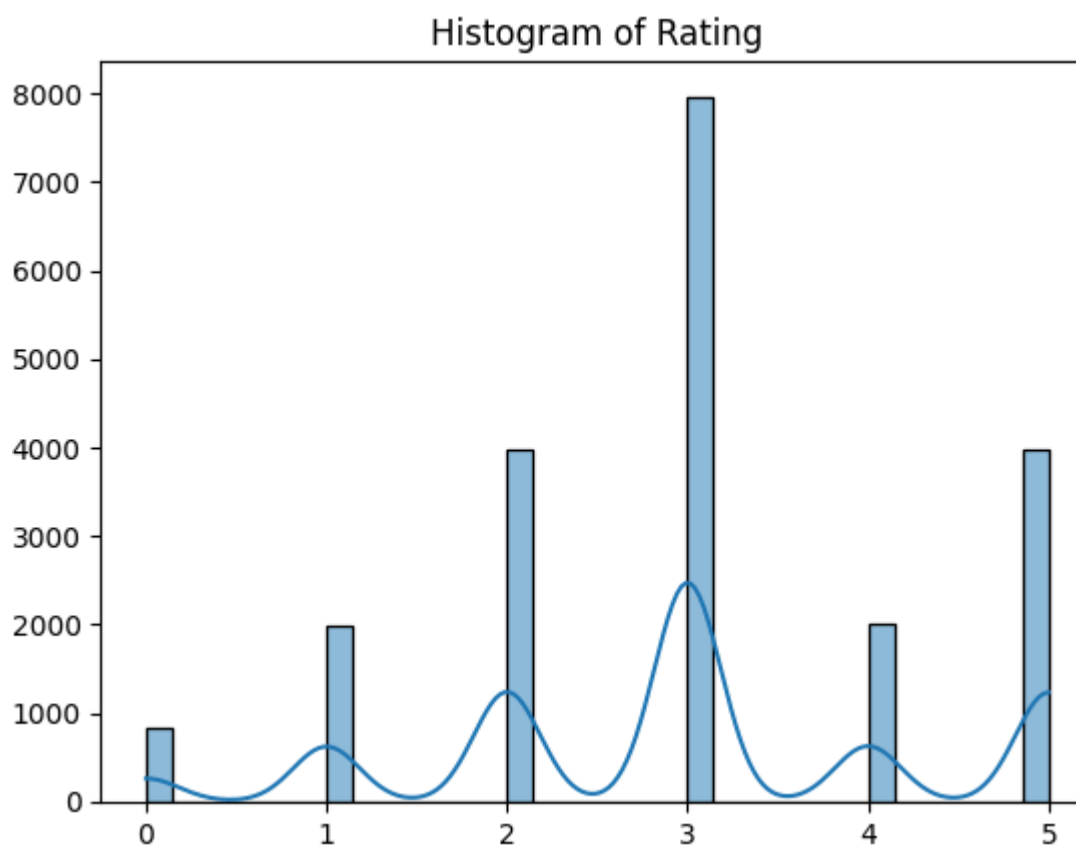
Edad:



En este caso en concreto los que aparecen en 0 es porque eligieron no poner su edad,

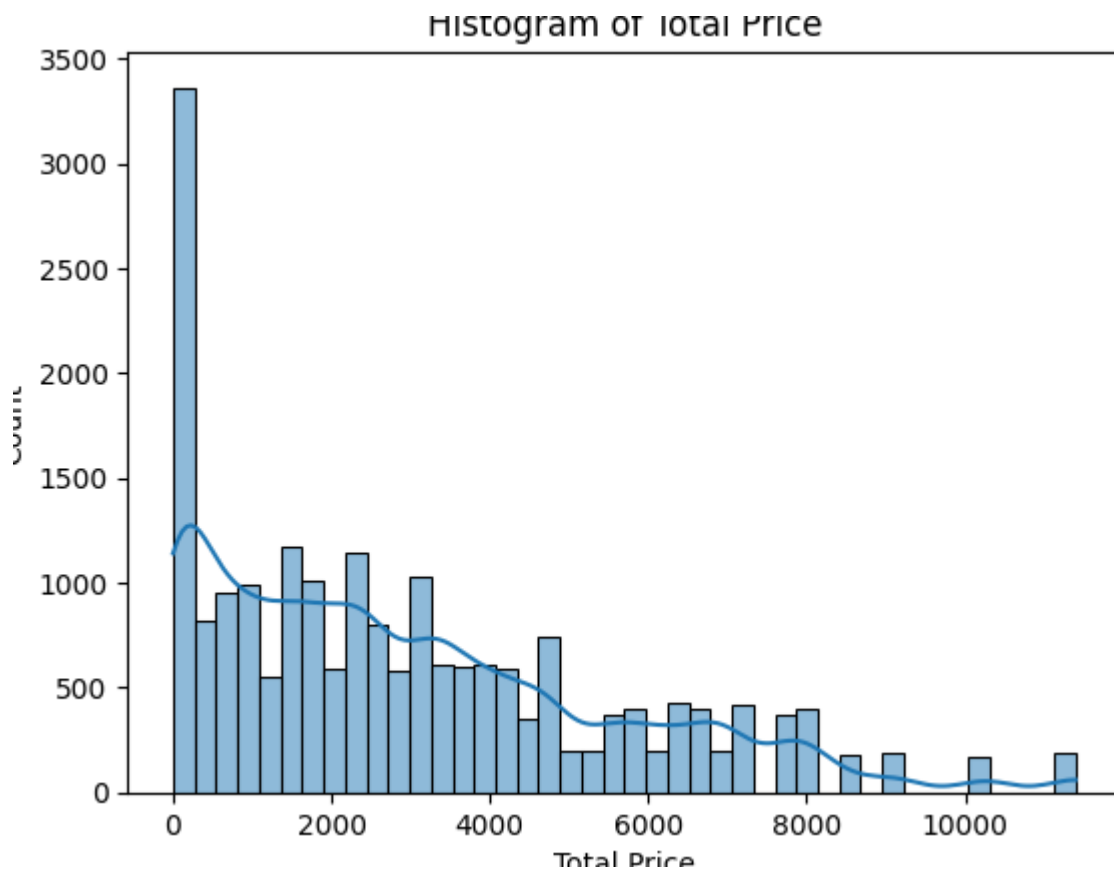
Sobre las demás edades no podemos sacar mucho en concreto ya que está muy parejo, pero podríamos deducir que el promedio es de unos 50 años y hay una cantidad bastante sorprendente a los 80 años

Rating:



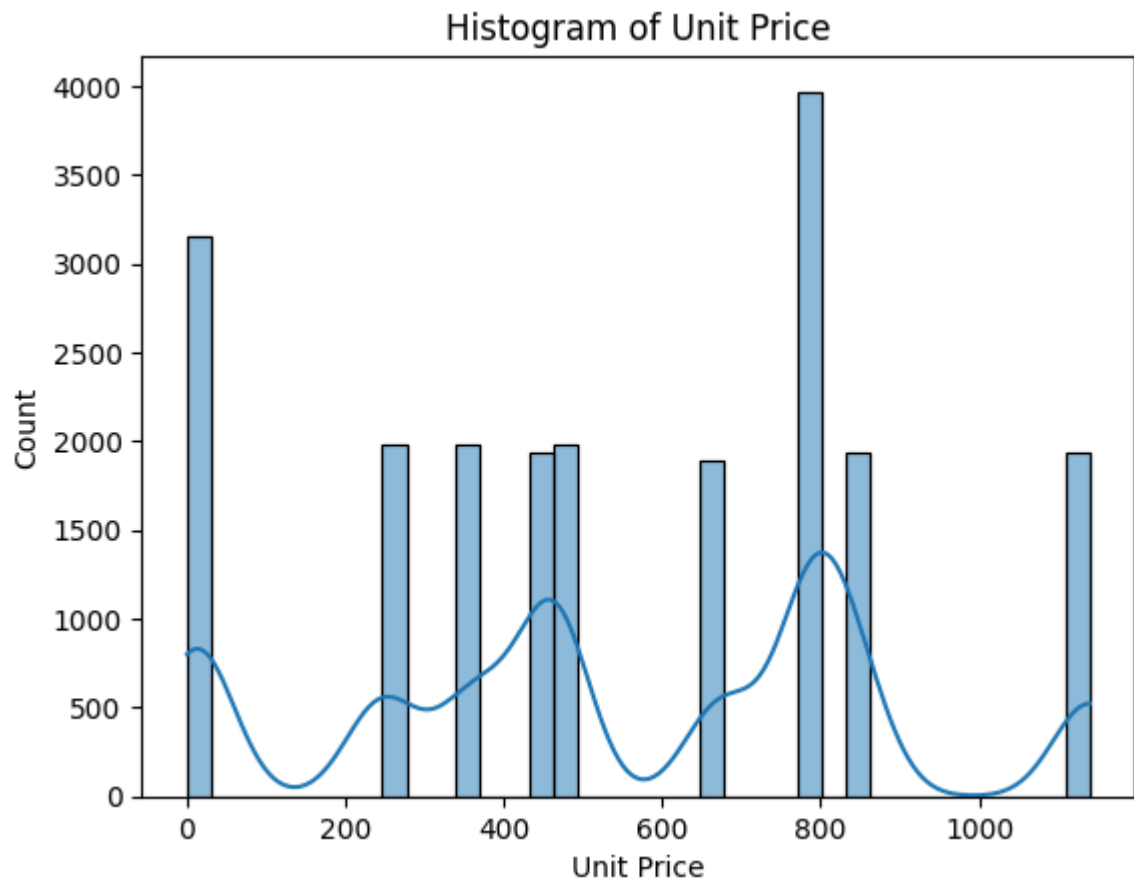
Aquí se ve de una manera un poco más clara que el puntaje promedio es de 3, esto significa que se puede mejorar los productos/añadidos

Precio Total:



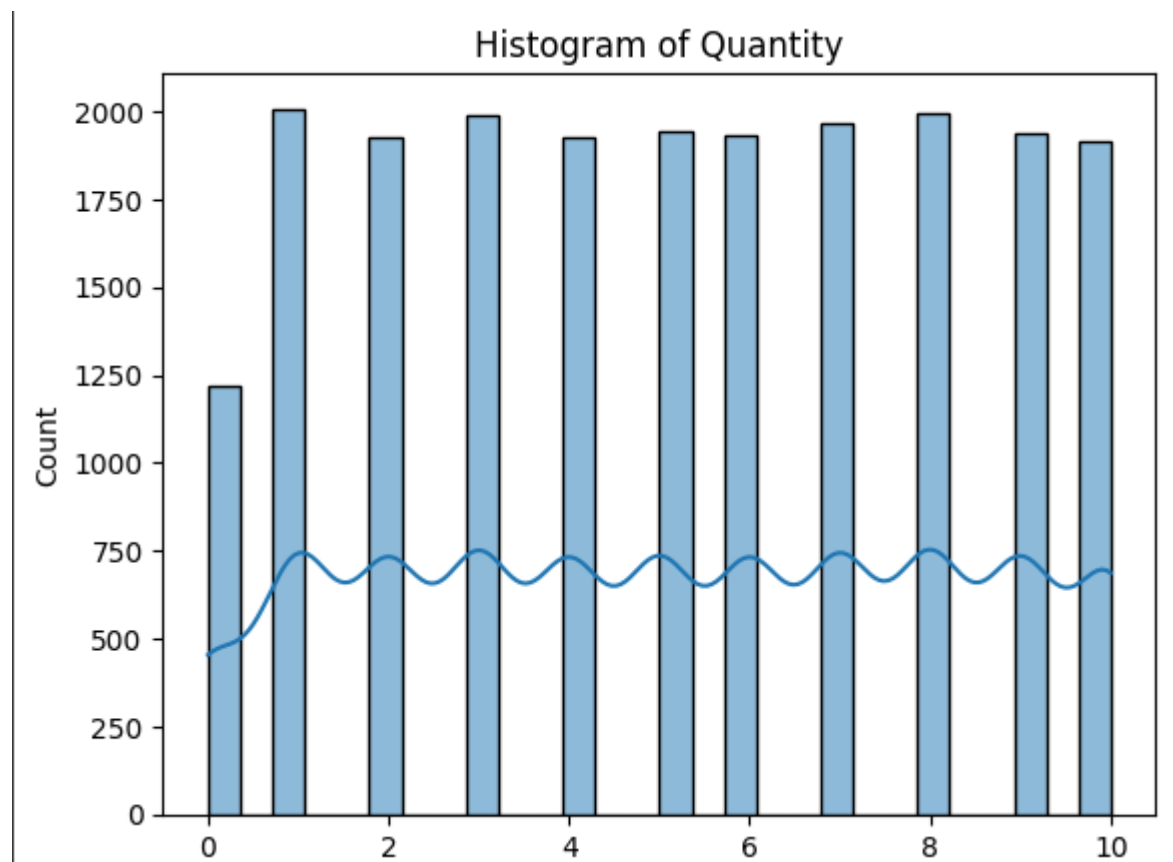
La razón del porque el precio total más común es 0, es probable que sea a un error o a que solo cuenta el valor total de los productos y no los añadidos entonces puede haber personas que compraron solo añadidos. También se puede deber a los pedidos cancelados, que, aunque parece que carecen de importancia podríamos intentar buscar un motivo del porque fueron cancelados o un patrón en los clientes q cancelas sus productos comprados.

Precio por unidad:



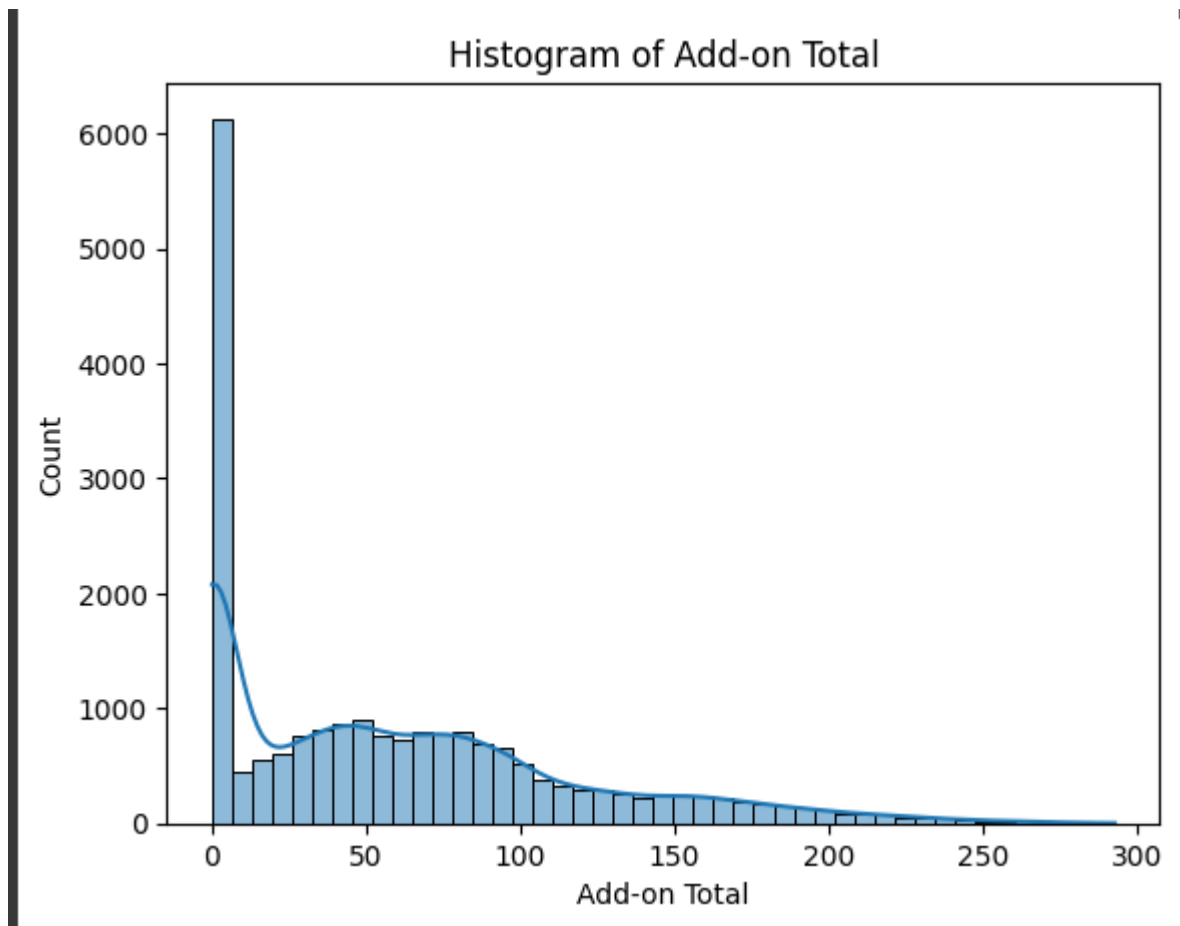
En este caso el precio por unidad más común es aproximadamente de \$800, también encontramos que hay varios 0 suponemos que es porque no se compró un producto si no un añadido

Cantidad:



Aquí se ve una compra de cantidad de productos bastante balanceada entre todas las variables, a excepción de los que no compraron ningún producto, lo cual nos indica que en las graficas de total price y unit price hay un par de errores, debido a que habíamos visto q había bastante cantidad de clientes q aparentaban no haber comprado nada.

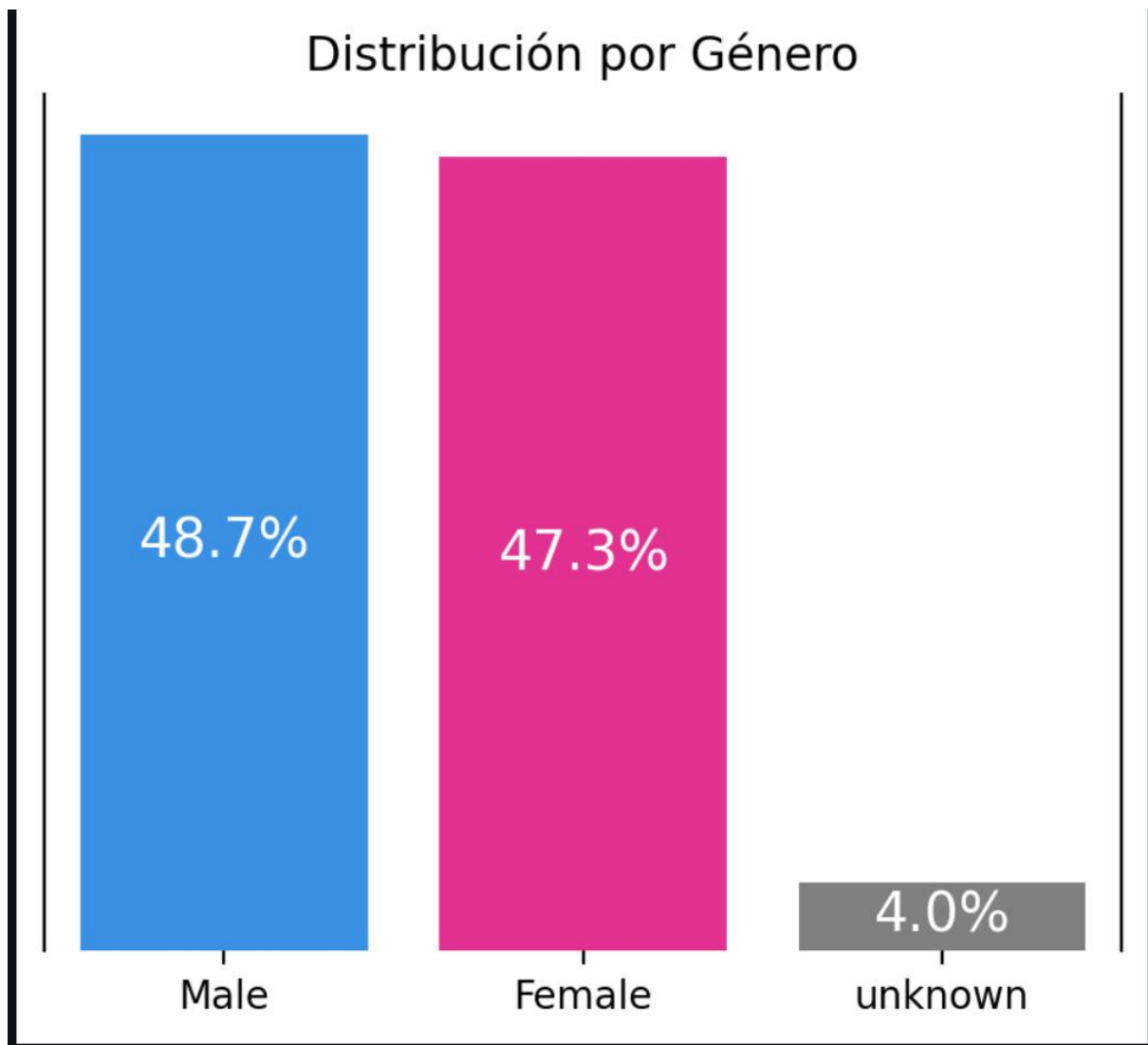
Add-ons Total:



Se puede denotar que la mayoría de clientes no compraron ni un solo add-on o si lo hacían normalmente no gastaban más de \$100, ya que se puede observar que no hay ni un solo dato en \$300.

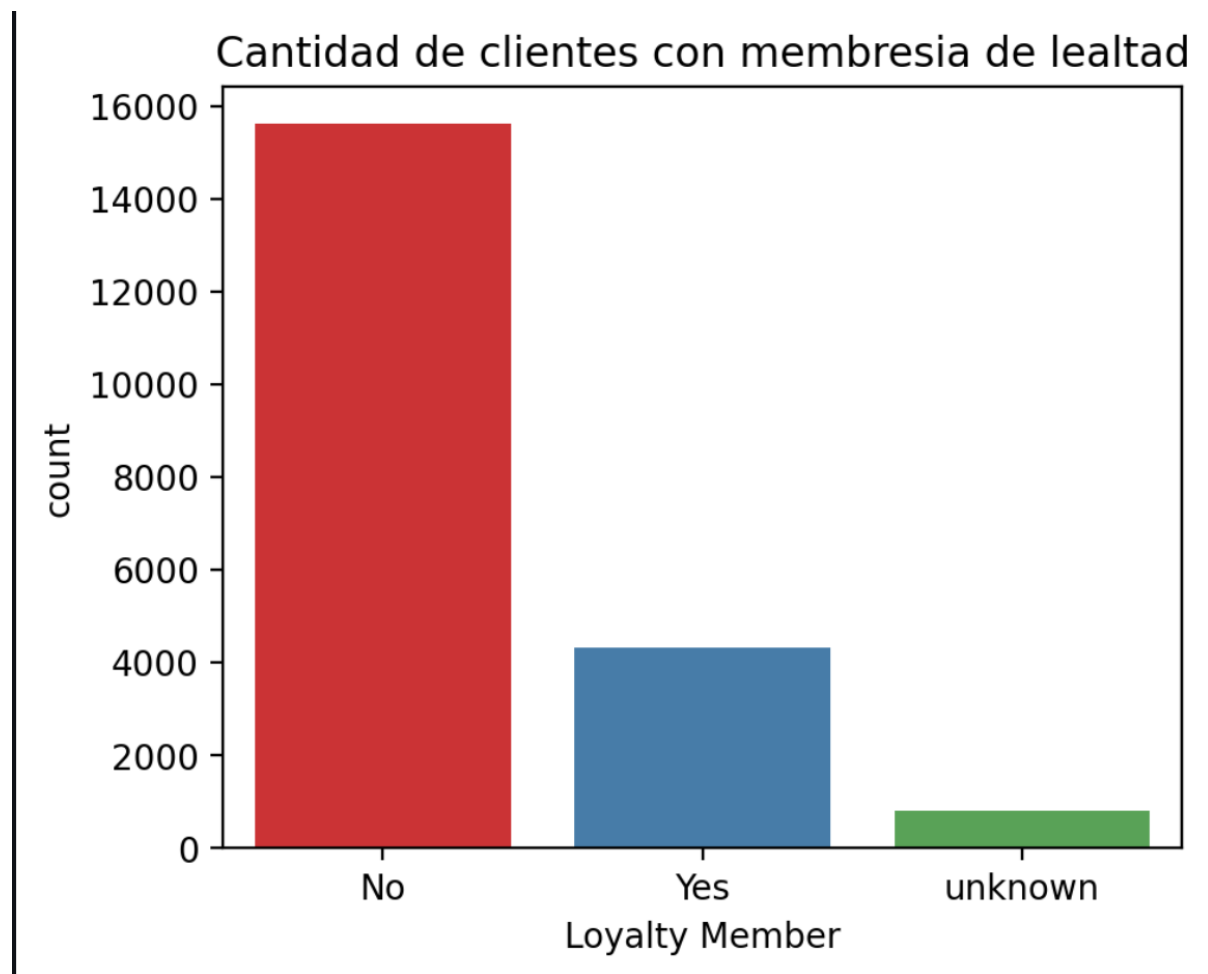
Variables Categóricas:

Genero:



En la variable genero vemos un balance en cuanto a mujeres y hombres entonces podemos deducir que el producto que se vende busca a ambos tipos de publico. El dato desconocido puede ser por que sea no-binario o el usuario no se haya sentido cómodo al compartir dicha informacion

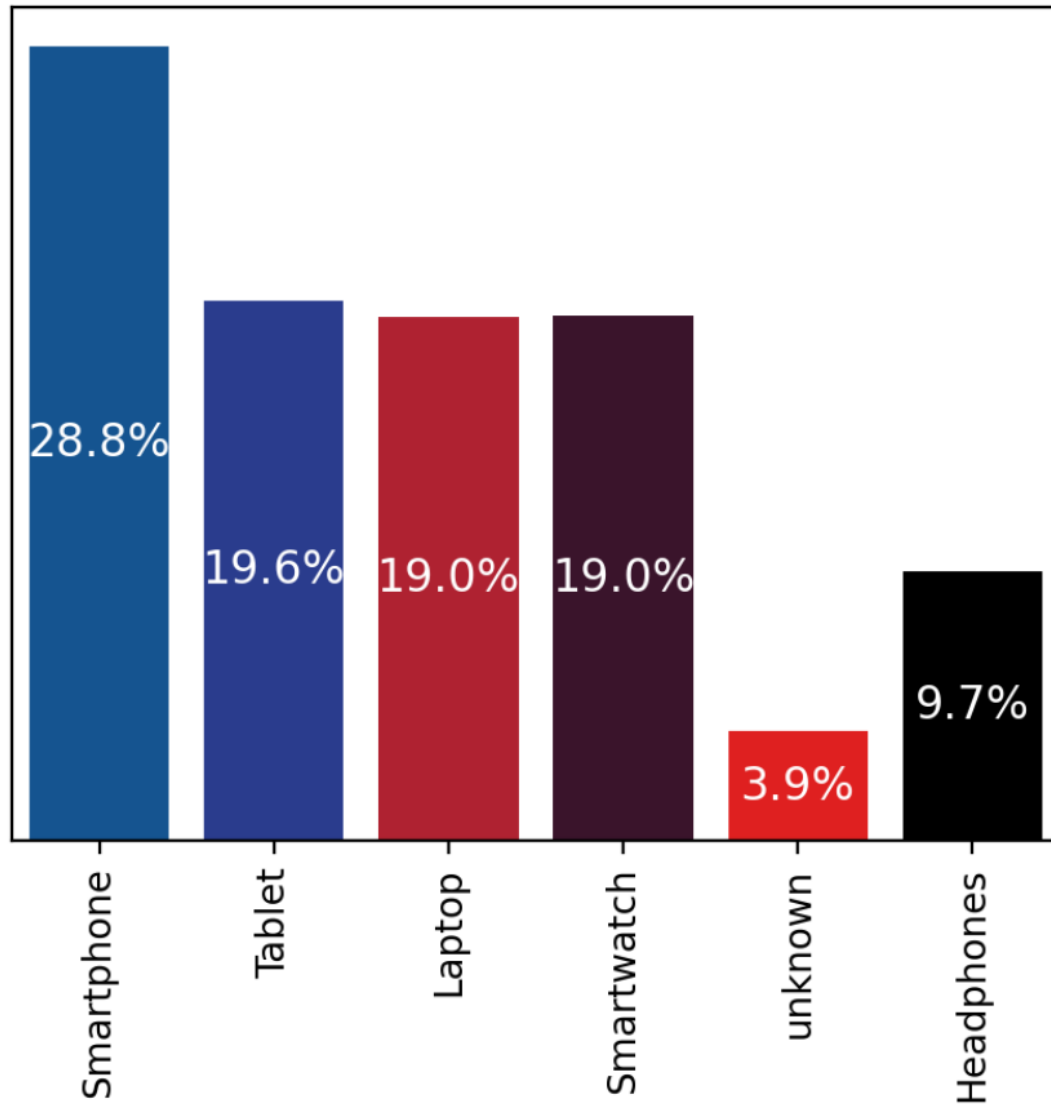
Miembros leales:



Esta grafica es bastante importante, porque va a hacer que nos planteemos si es realmente importante este programa si la mayoría de las clientes no lo quieren, tambien podriamos aumentar los incentivos a que usen este servicio o suscripción con más beneficios, de esta forma podemos ganar más clientes leales

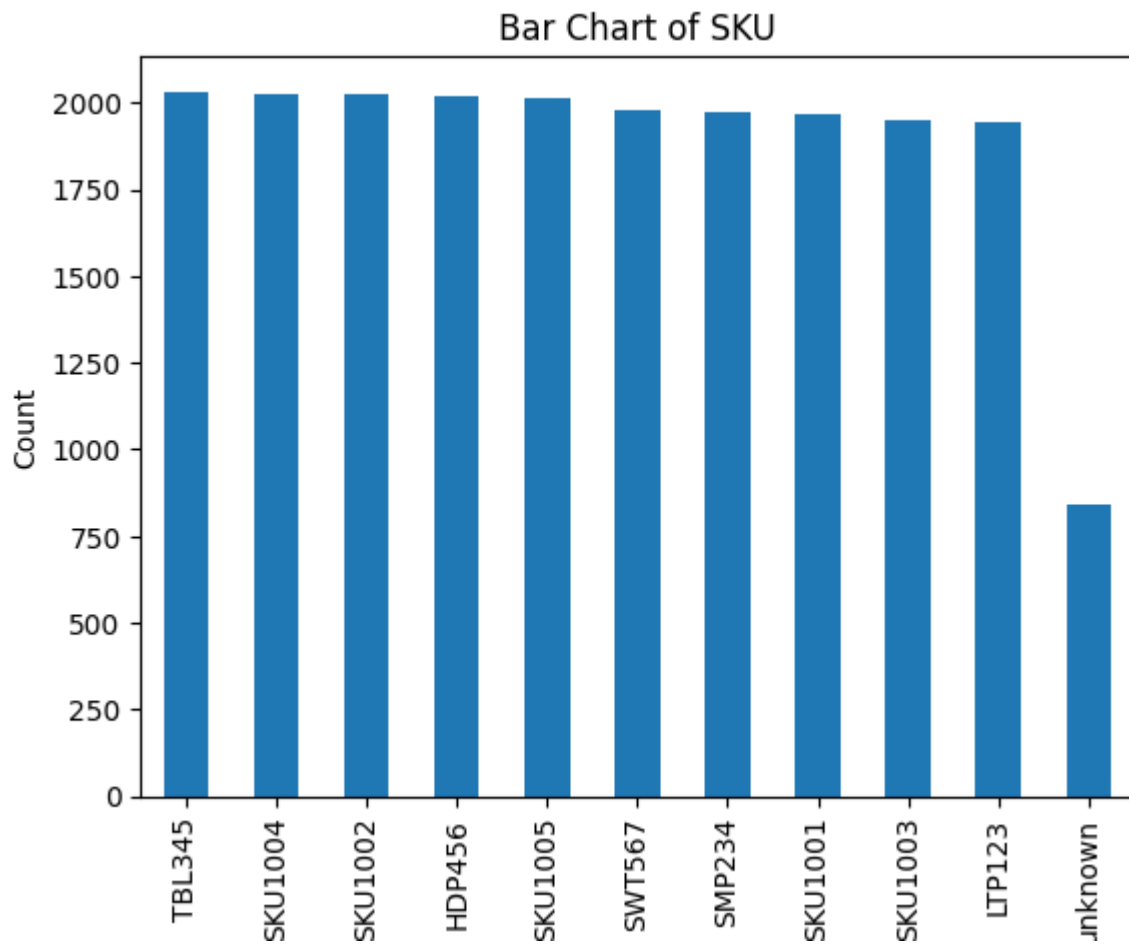
Tipo de producto :

Distribución por Producto



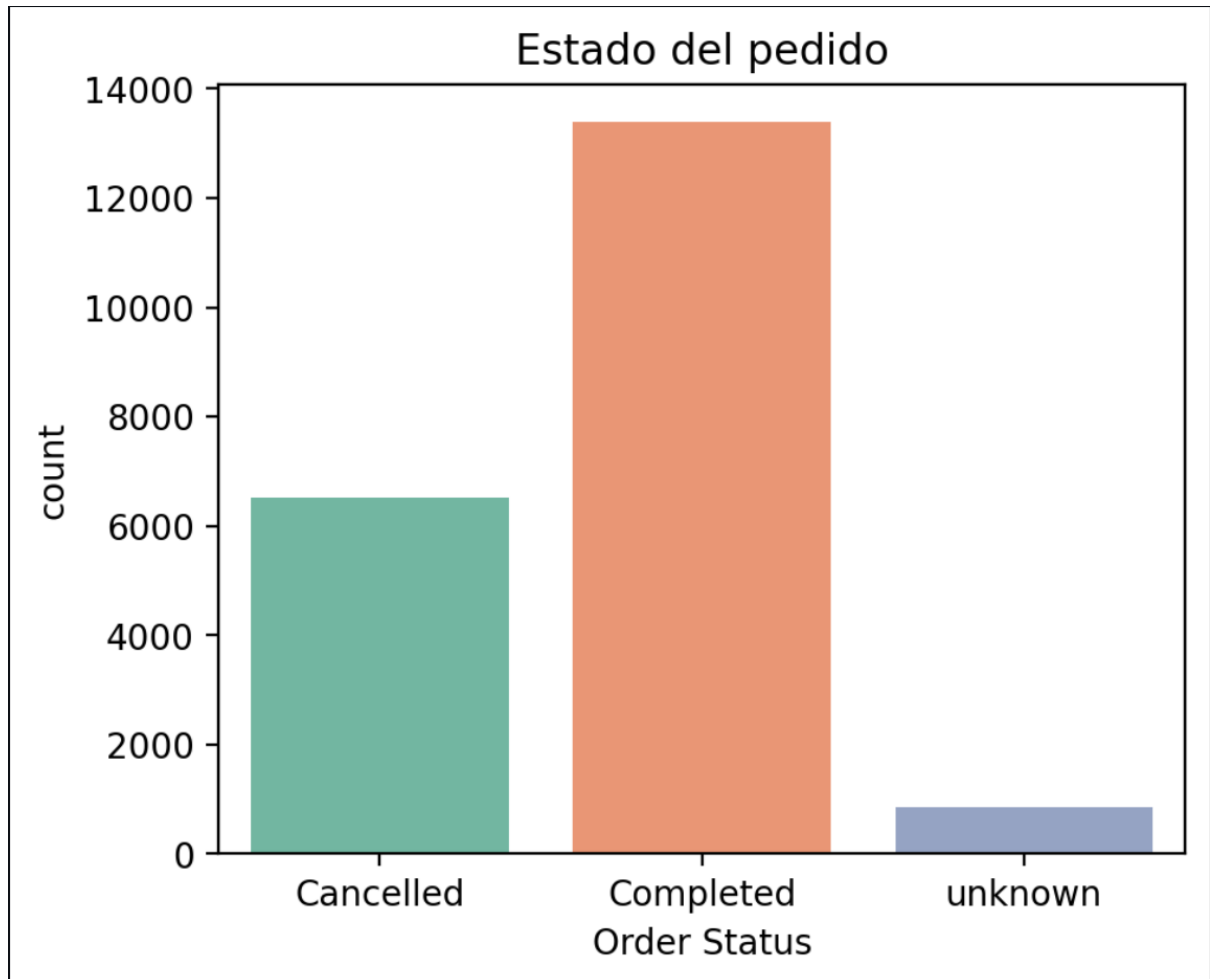
El producto más vendido fue el smartphone con algo de diferencia de los demás lo que nos dice que hay que aumentar en el inventario este producto ya que es el más vendido el menos vendido fueron los headphones hay que reducir la producción de este si sobran en el inventario. Unknown solo son los datos que no compraron ningún producto.

SKU(Stock Keeping Unit):



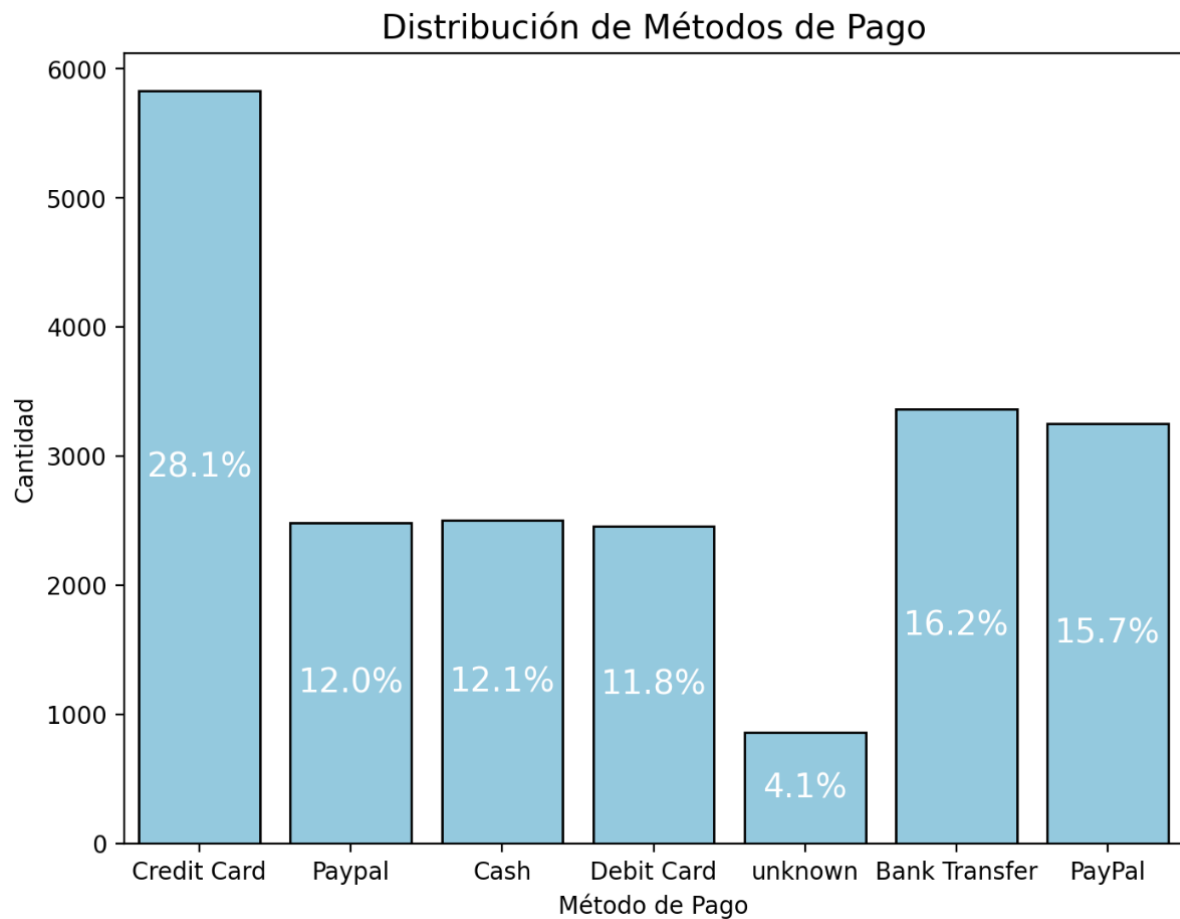
Estos indicadores dicen un producto en específico esta grafica no nos sirve.

Estado de la orden:



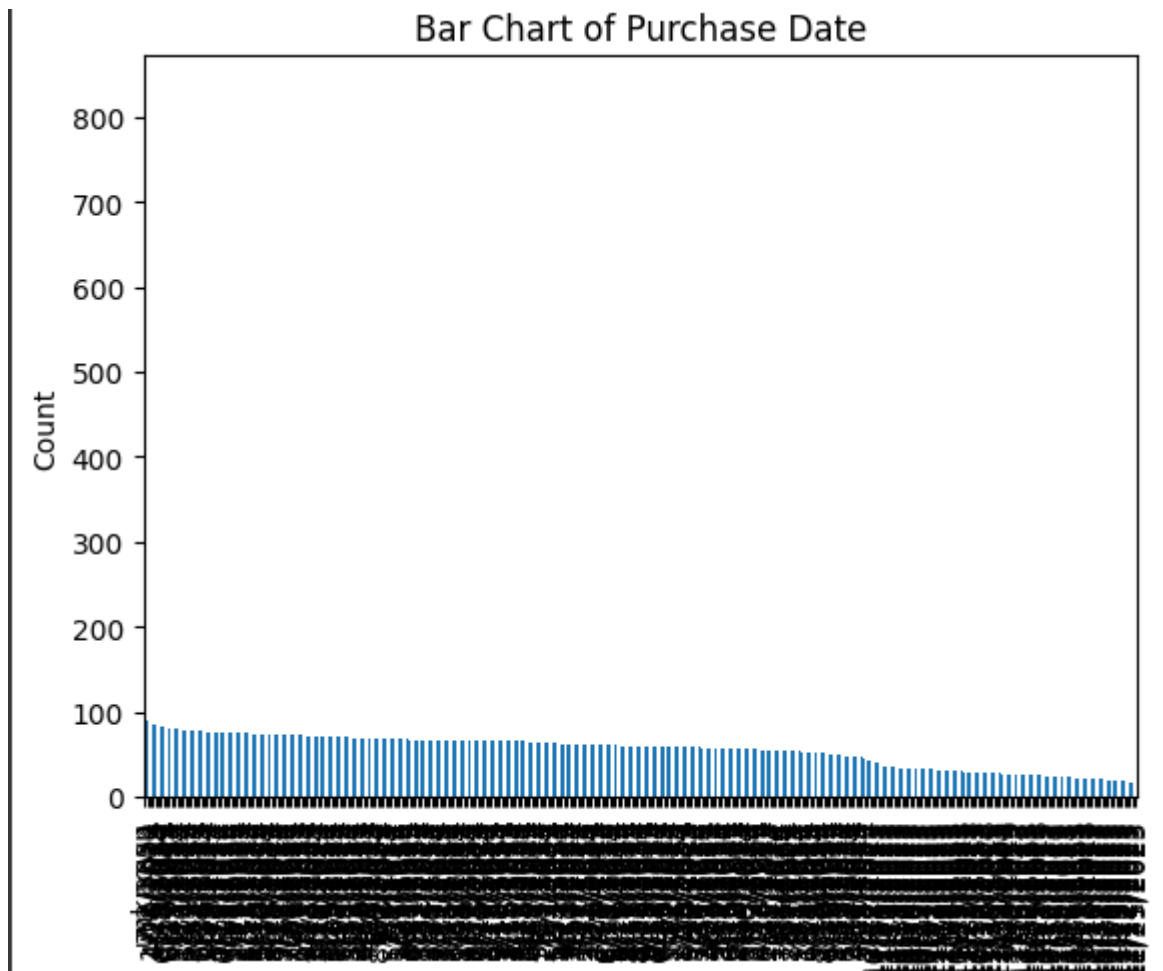
Esta grafica es muy importante para buscar motivos del por qué la orden fue cancelada o completada, principalmente por el tipo de transporte, podremos ver cual es el transporte que menos completa los pedidos

Metodos de pago:



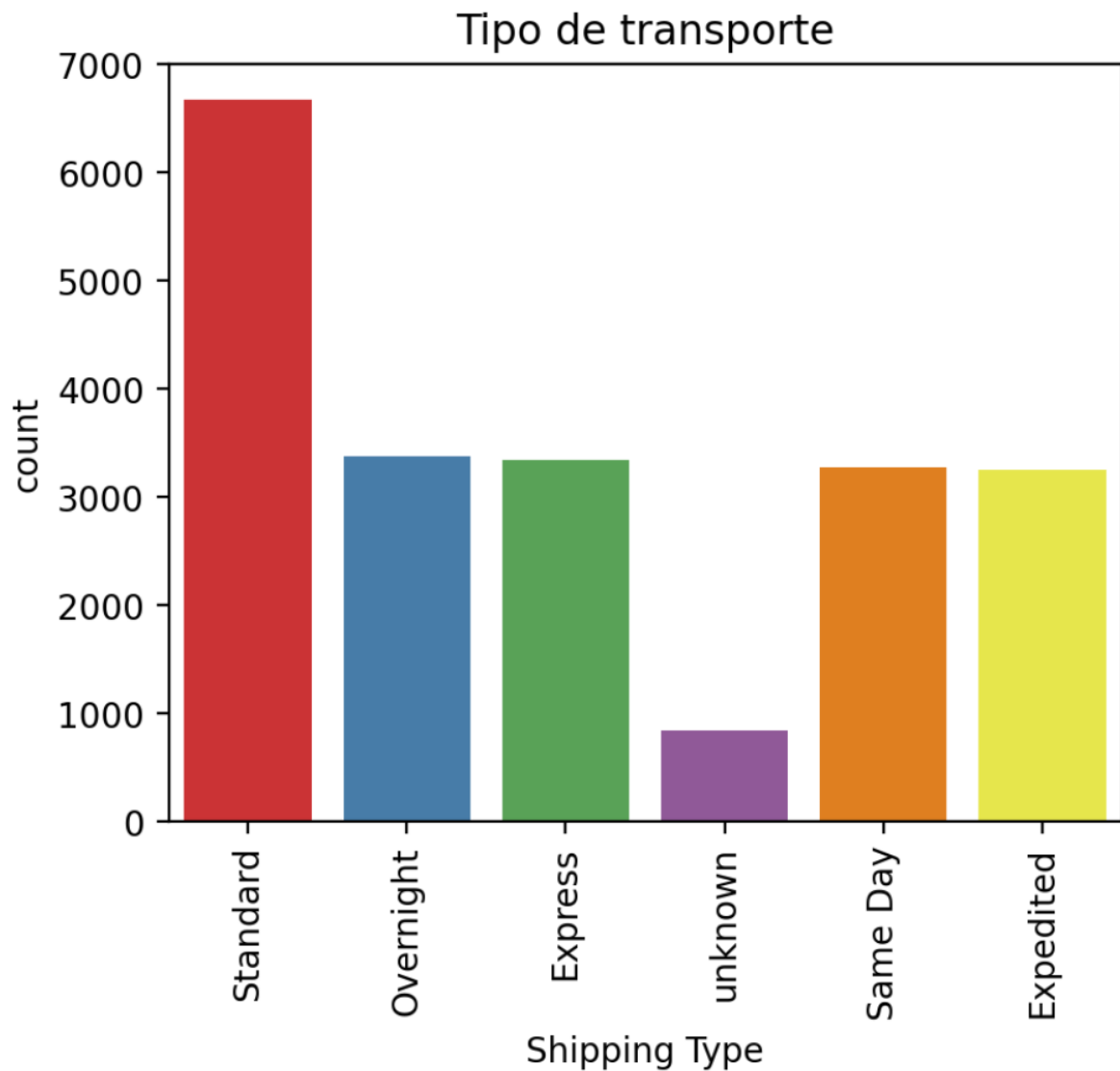
Aquí podemos ver los métodos de pago más comunes con porcentajes, tal vez podríamos usar esta grafica para relacionarlo con otras cuestiones como si con esto podemos predecir que van a comprar y cuanto van a gastar en productos, add-ons e incluso transporte.

Fecha de la compra:



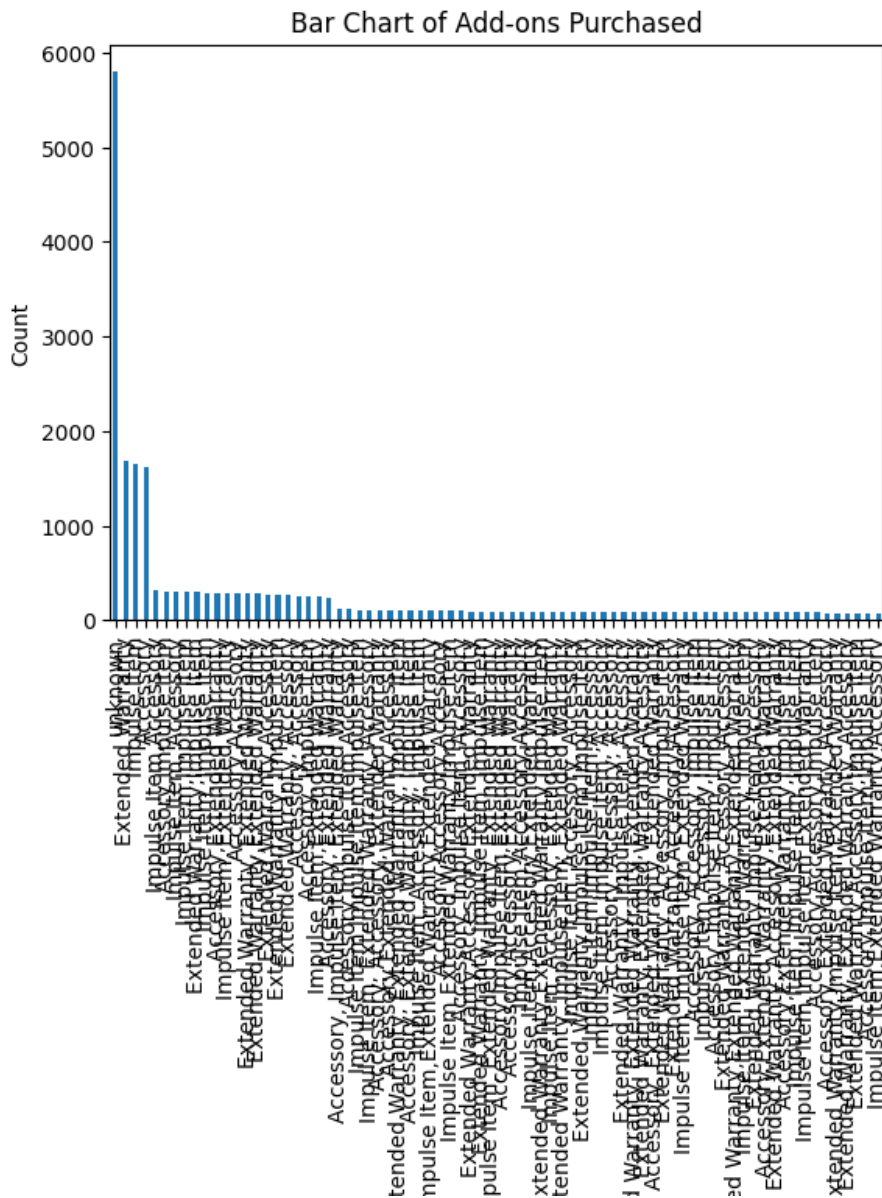
Esta grafica no es de nada útil ya que es imposible analizar

Tipo de transporte:



El tipo de transporte más comun es el estandar probablemente por la sencillez y precio de este hay que hacer análisis para ver si es posible saber si tiene algo q ver los pedidos cancelados con el tipo de transporte

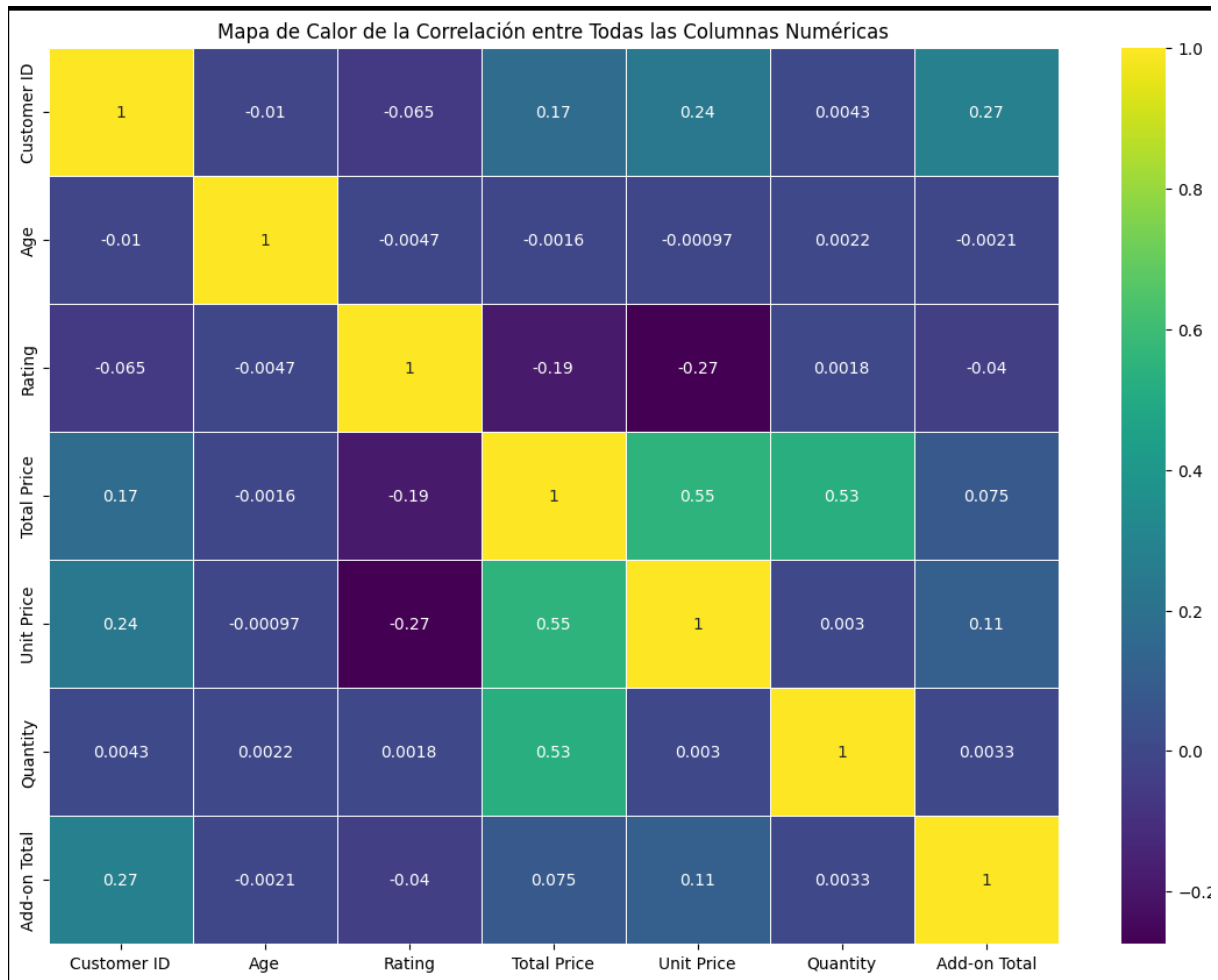
Add on Total:



Aunque esta grafica no nos sea util se puede distinguir una cosa unicamente y es que el dato mas frecuente es “unknown” lo cual probablemente significa que no son muy comprados los add-ons

Correlación entre variables:

Matriz de correlacion:

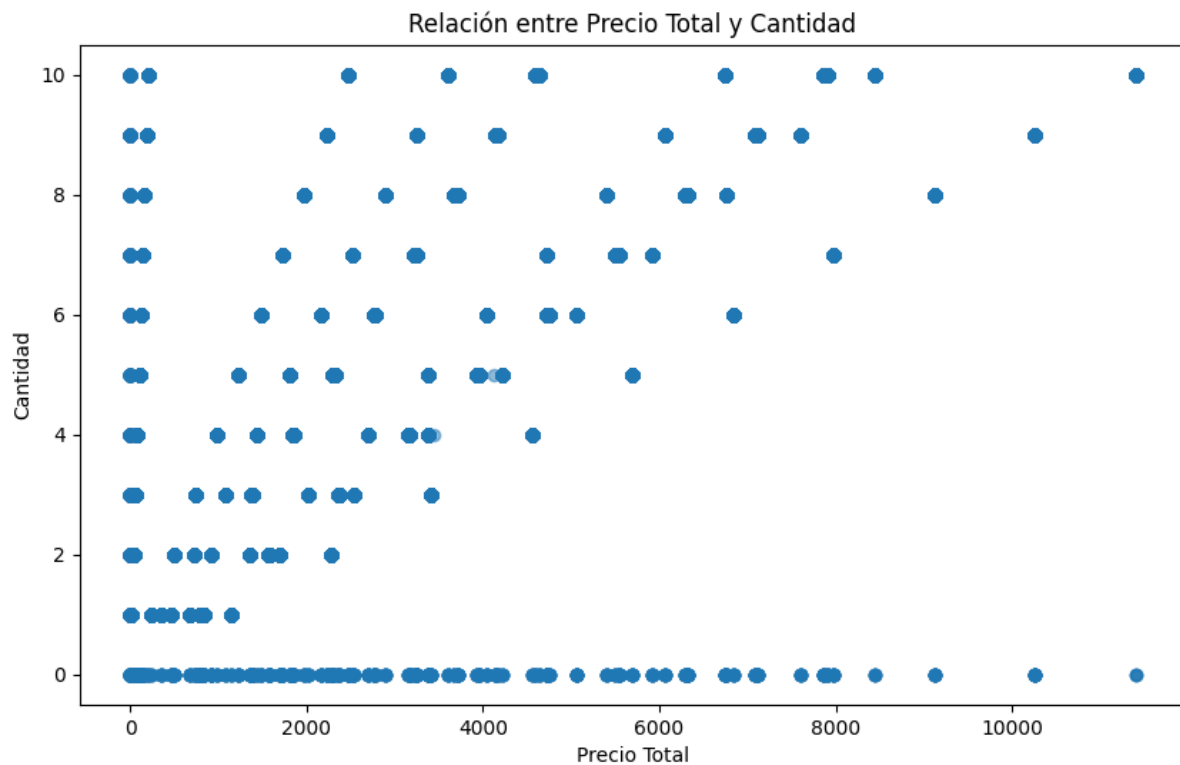


- **Relacion Quantity-Total Price (0.53):**

Estas variables tienen la relación más estrecha que encontramos, lo que nos dice que es probable encontrar una. Se hará una gráfica para averiguar que tan estrechas son las variables.

Gráfica que representa la correlación

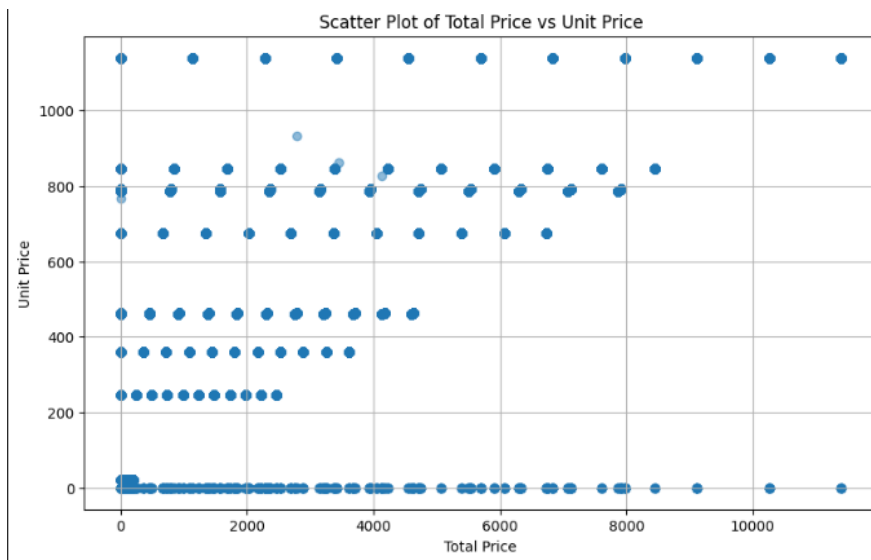
El único patrón destacable es que a medida que aumenta la cantidad de productos comprados, también tiende a aumentar el precio total. A excepción del valor 0 que es probablemente un error o tal vez sea un patrón de que los clientes que ponen más



cantidad de productos cancelan sus pedidos, lo cual explicaría porque el precio total es 0.

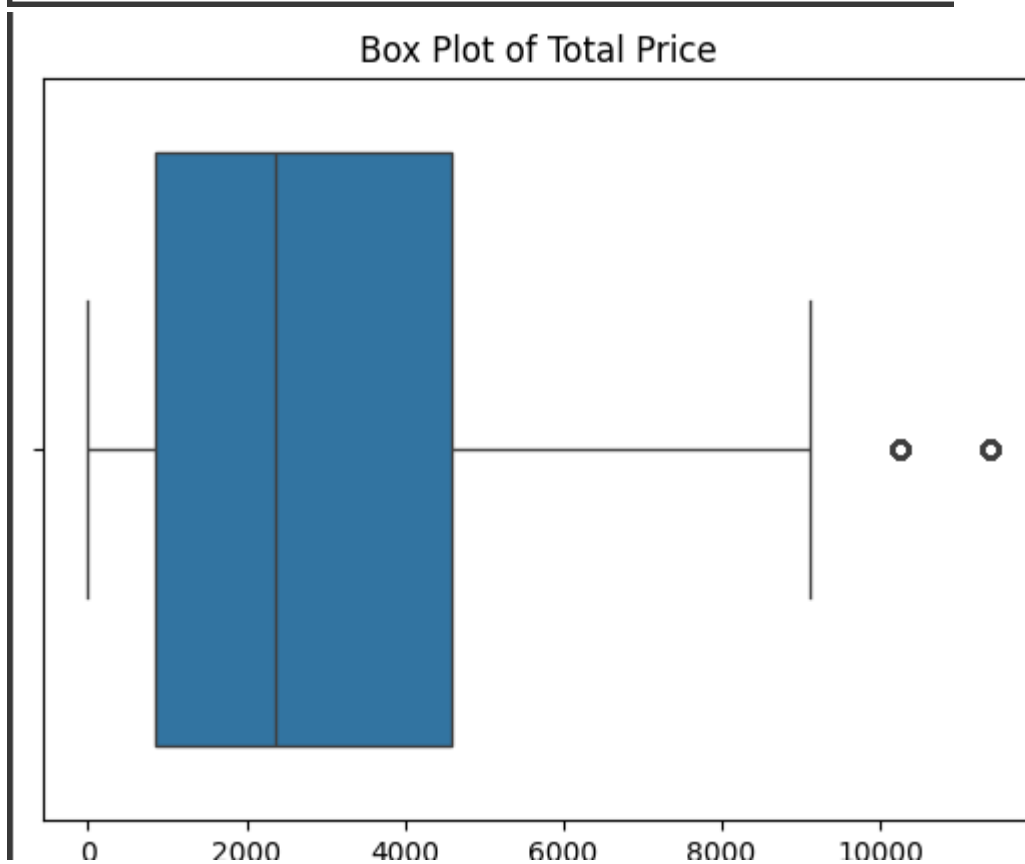
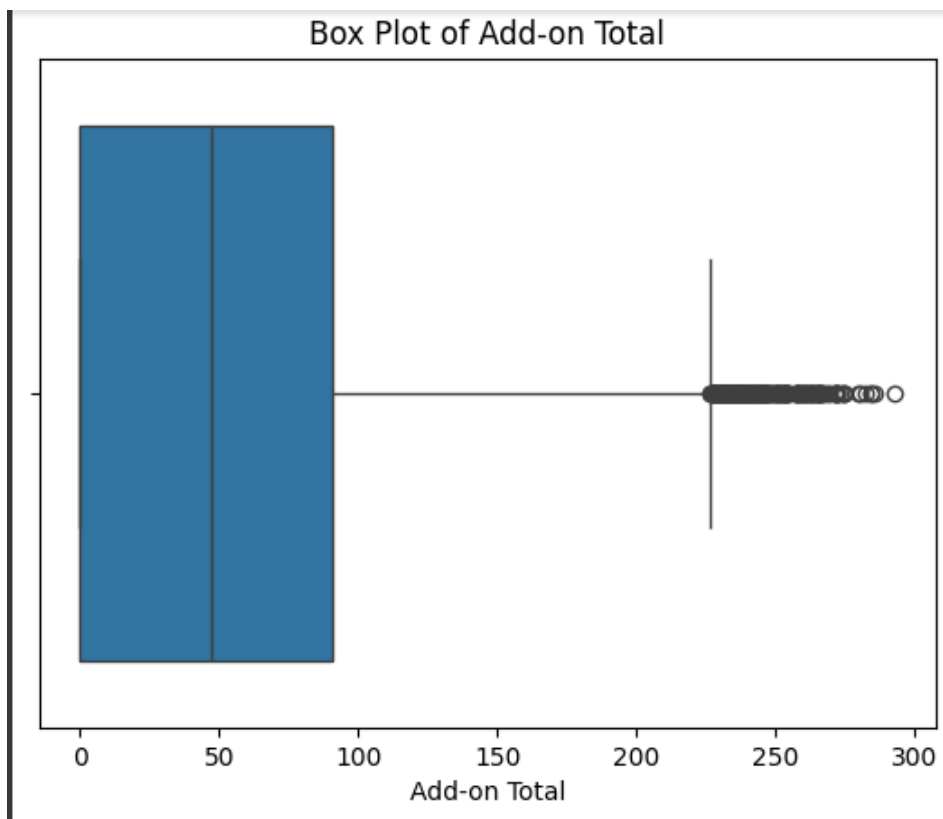
- **Relación Unit Price-Total Price (0.55):**

Este es de la relación más alta de variables que encontramos en el mapa de calor, probablemente sea porque el valor por cada unidad comprada marca el precio total gastado. El hecho de que la relación sea positiva indica que a medida que aumenta el precio por unidad también aumenta el precio total



Análisis de Valores Atípicos (Outliers):

Para encontrar los valores atípicos se usaron box plots para ver de manera visual los datos que salen del rango de este.



Ahora para la eliminacion de estos datos usaremos un proceso llamado detección y eliminación de valores atípicos (outliers) utilizando el rango intercuartílico (IQR)

Q1: El primer cuartil (25%) de la columna Add-on Total.

Q3: El tercer cuartil (75%) de la columna Add-on Total

$IQR = Q3 - Q1$

El IQR es la diferencia entre el tercer cuartil y el primer cuartil.

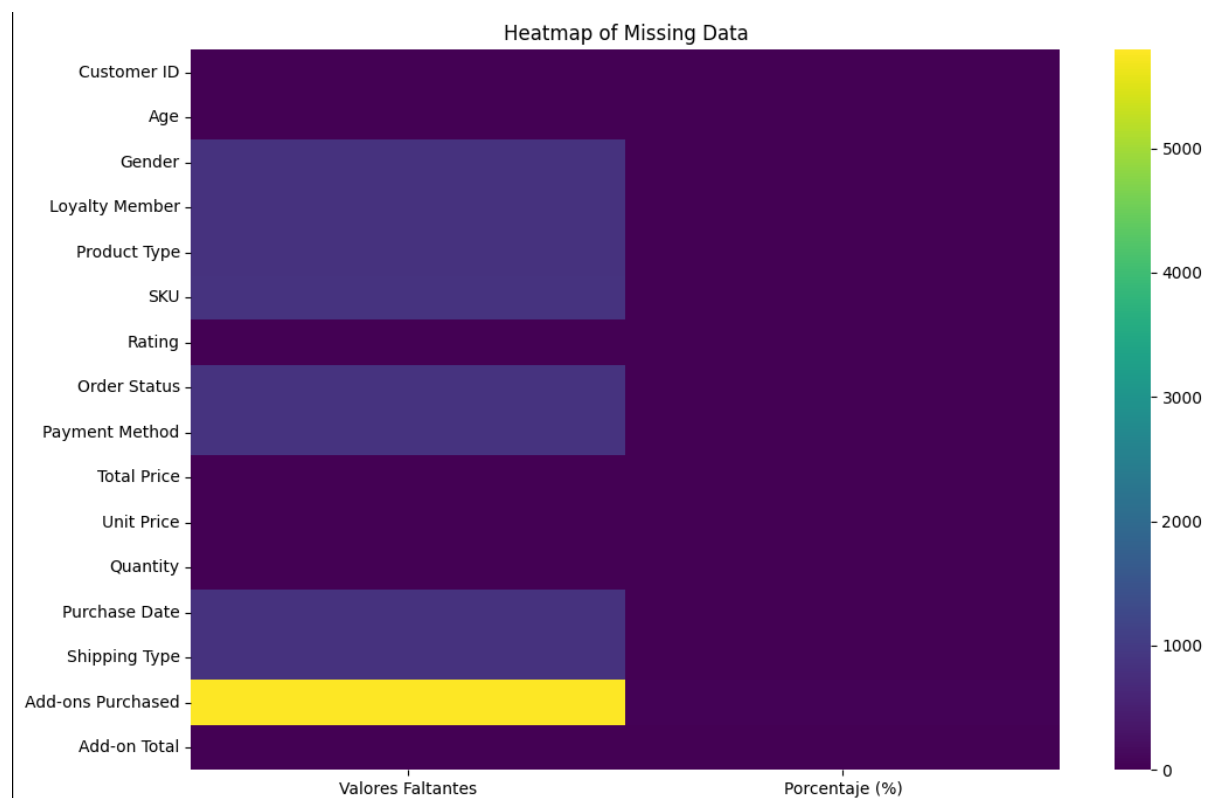
Límite Inferior: $Q1 - 1.5 * IQR$ Límite Superior: $Q3 + 1.5 * IQR$ Estos límites se utilizan para identificar los valores que se consideran atípicos.

```
df = df[(df['Add-on Total'] >= lower_bound) & (df['Add-on Total'] <= upper_bound)]
```

Se eliminan los valores que están por debajo del límite inferior o por encima del límite superior, quedando solo los datos dentro del rango permitido.

Análisis de Valores Faltantes

Mapa de calor datos faltantes



La barra de la derecha muestra la relación entre los colores y la cantidad de datos faltantes

	Column	Missing Values	Percentage
0	Customer ID	0	0.000000
1	Age	0	0.000000
2	Gender	835	4.021771
3	Loyalty Member	822	3.959156
4	Product Type	818	3.939890
5	SKU	839	4.041037
6	Rating	0	0.000000
7	Order Status	844	4.065119
8	Payment Method	859	4.137366
9	Total Price	0	0.000000
10	Unit Price	0	0.000000
11	Quantity	0	0.000000
12	Purchase Date	830	3.997688
13	Shipping Type	837	4.031404
14	Add-ons Purchased	5793	27.901936
15	Add-on Total	0	0.000000

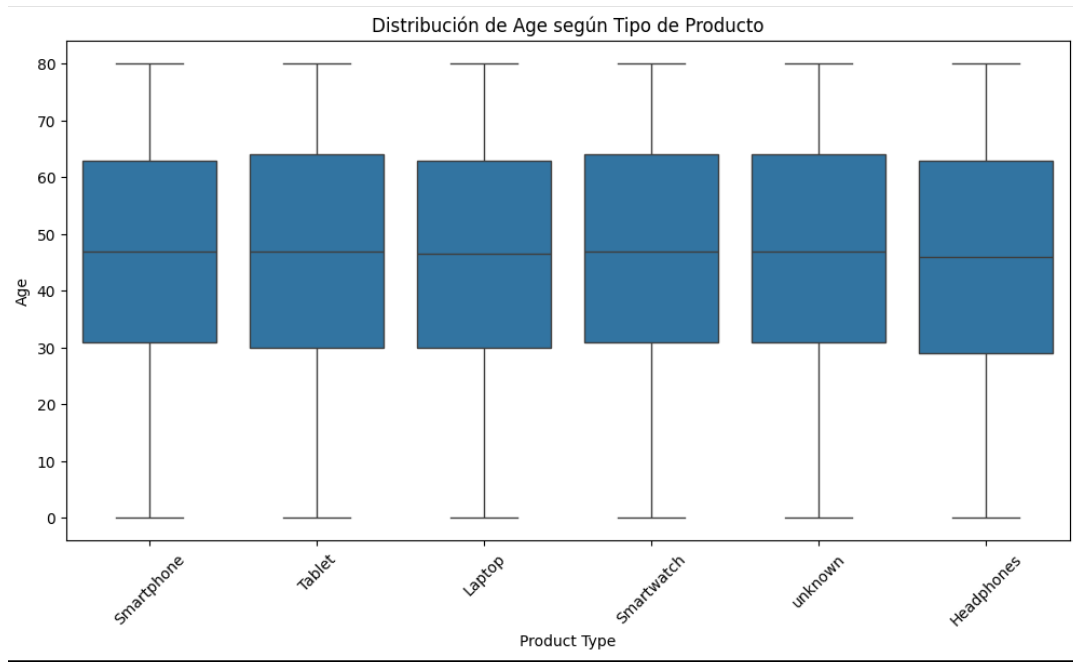
Estrategia de Imputación o Eliminación:

Como fue explicado anteriormente, el metodo o estrategia usada fue imputacion por relleno de valores.

La unica columna peligrosa por los datos faltantes es “add-ons purchased” sin embargo por varias consideraciones se considera lo suficientemente importante para no ser eliminado

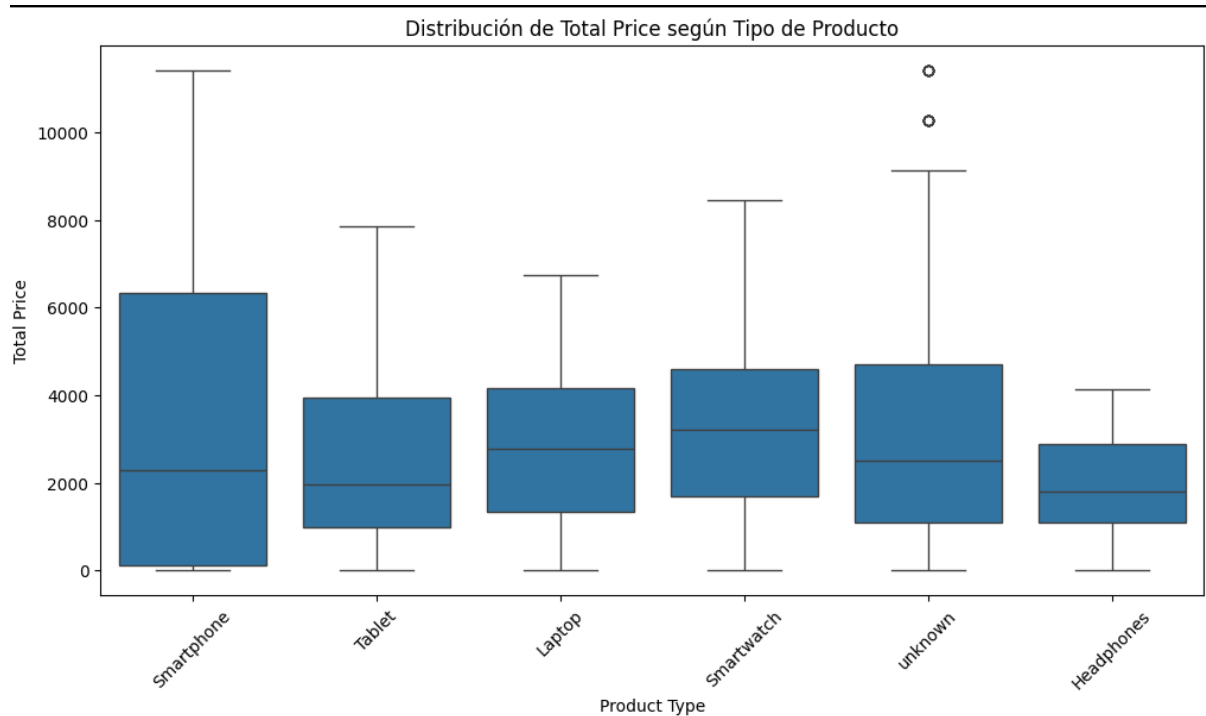
. Relación entre Variables Categóricas y Numéricas

- **Edad segun tipo de producto**



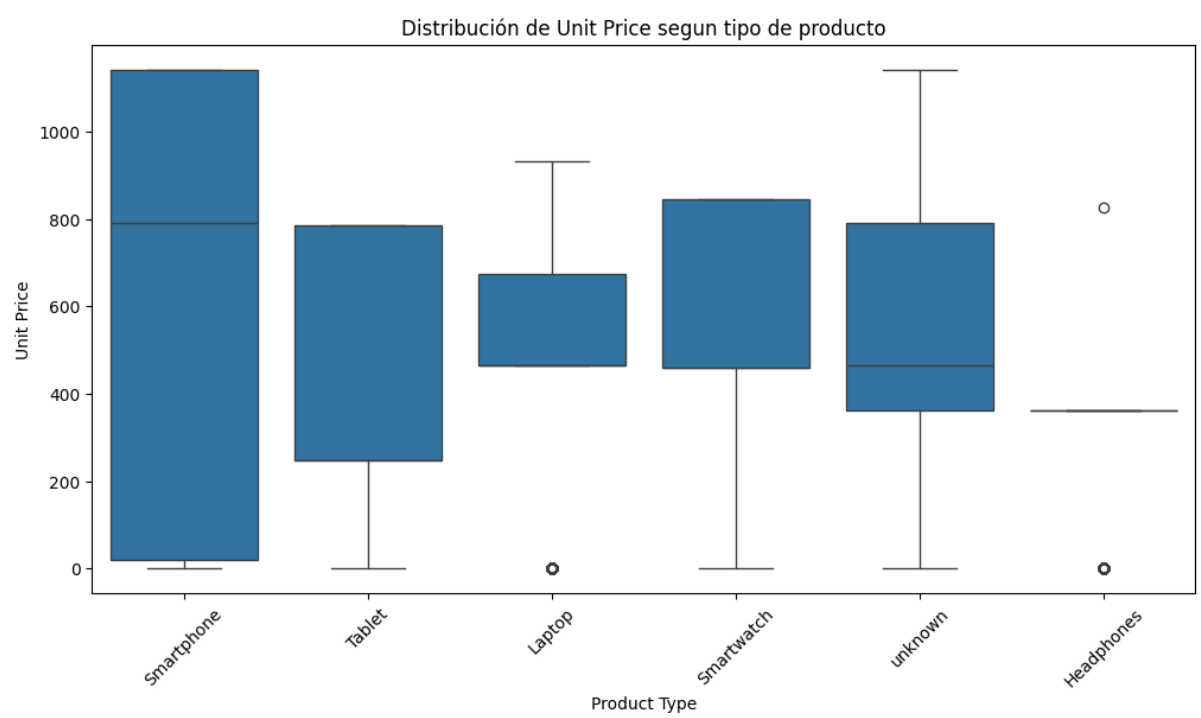
Se observa que la mayoría de los productos están en un mismo punto debido a la edad media de los usuarios (46), el que hay que observar de manera más delicada debería ser el unknown sin embargo ya se ha dicho en repetidas ocasiones por qué puede pasar.

Precio total según el tipo de producto



- **Smartphone:**
Se puede denotar que es el rango de precio más ancho, al mismo tiempo es el que tiene el precio total mas alto
- **Tablet y Laptop:** Los precios son similares y muestran una dispersión más baja que los smartphones.
- **Smartwatch:** Tiene un rango de precios ligeramente mayor que las tabletas y laptops, pero sigue siendo más moderado en comparación con los smartphones.
- **Headphones :** Tienen los precios más bajos, con una distribución estrecha y sin presencia de valores extremos.
- **Desconocido:**

Esta categoría muestra una distribución amplia y comparable a la de los smartwatch. También tiene un par de outliers, sugiriendo que los productos no clasificados incluyen algunos con precios significativamente más altos



En esta grafica se puede observar que otra vez el smartphone es el que tiene mas rango a comparación de los demás. El principal dato que sorprende es el de audífonos ya que sus datos son outliers

Observaciones y Hallazgos Importantes

Variable que se quiere encontrar: Tipo de producto

Variables relacionadas: Se observan varias variables que afectan al tipo de producto, tales como:

- **Unit Price:** Esta variable suele tener una alta importancia en la predicción del tipo de producto. Diferentes tipos de productos tienen rangos de precios distintos.
- **Total Price:** Similar al precio por unidad, el precio total también es un indicador significativo del tipo de producto.
- **Quantity:** La cantidad de productos comprados puede variar según el tipo de producto, afectando así su clasificación.
- **Add-on Total:** El total de complementos adquiridos puede diferir entre tipos de productos, proporcionando información adicional para la predicción.

- **Age:** La edad del cliente puede tener una influencia menor pero aún relevante en el tipo de producto adquirido.

Relaciones inesperadas:

Cancelaciones y Métodos de Pago:

- Parece haber una alta tasa de cancelaciones cuando se utiliza PayPal y Transferencia Bancaria como métodos de pago. Esto podría indicar problemas específicos con estos métodos o una mayor probabilidad de cancelación entre los usuarios que los eligen.
- **Productos y Frecuencia de Cancelación:**
Los auriculares (Headphones) y los smartwatches tienen una tasa de cancelación notablemente alta en comparación con otros productos como smartphones y laptops. Esto podría sugerir que estos productos tienen más problemas de satisfacción del cliente o que los clientes cambian de opinión con más frecuencia después de realizar el pedido.
 - El bajo consumo de add ons para la gente que es “miembro leal”

Patrones interesantes:

Género y Preferencias de Compra:

Las mujeres parecen comprar más tablets y smartwatches, mientras que los hombres compran más laptops y smartphones. Esta diferencia en las preferencias de compra podría ser útil para campañas de marketing dirigidas.

También parece que las transacciones con un precio total más alto parecen tener una mayor tasa de cancelación. Esto podría deberse a la indecisión del cliente o a problemas con la aprobación de pagos para montos más altos.

Implicaciones para el modelo de machine learning:

Estos hallazgos pueden servirnos para ver si nuestro modelo encuentra alguna relación importante de las que se discutieron o quizás nuevas.

Modelo de machine learning:

Precisión: 62.74%

Recall: 93.56%

RMSE: 0.61

Principal grupo de venta para cada tipo de producto electrónico:

Product Type

Headphones Male

Laptop Male

Smartphone Male

Smartwatch Male

Tablet Male

unknown Male

Acada podemos ver los resultados del modelo de machine learning propuesto, vemos que hay un problema con la precisión ya que no es un muy buen rendimiento. Sin embargo podemos suponer que de la mayoría de productos lo compran más el publico masculino, tal vez a excepción de alguno que otro por el margen de error,

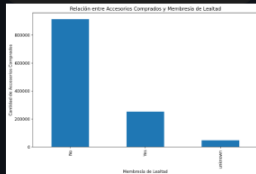
Dashboard:

El objetivo del dashboard es ver de manera mas simple nuestro trabajo dando bastantes datos importantes y graficas que se utilizaron en este mismo

Dashboard de Análisis de ventas de dispositivos electronicos

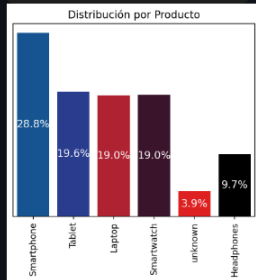
Total de Clientes

20762



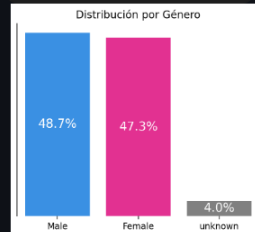
Edad Promedio

46.0



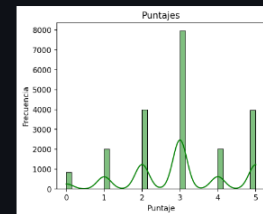
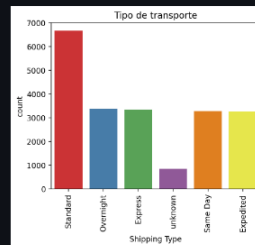
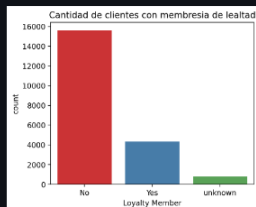
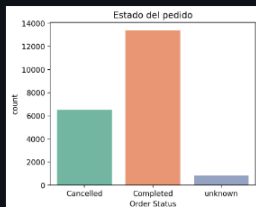
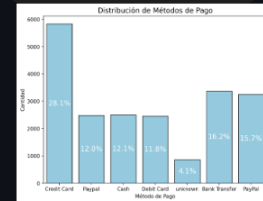
Cantidad de productos comprados promedio

5.16



Accesorios comprados promedio

58.42



Cada cosa que es representada es importante para poder aumentar las ventas de los dispositivos electrónicos, ya que se intenta buscar ciertas poblaciones objetivo en cada una y poder ver que es lo que prefieren los clientes.

Conclusiones y Futuras líneas de trabajo:

En este proyecto se pudo analizar ciertos comportamientos que podrian cambiar las ventas de manera positiva para la empresa. También se buscó otros factores negativos como motivos de por qué se cancelan tantos pedidos, o si el programa de miembros leales es útil.

Mejoras:

Se puede mejorar el modelo de machine learning para tener una mejor prediccion de los datos.

Hacer más tablas de comparación con tablas categóricas con numéricas por que principalmente se hicieron de tipo de producto.

Posibles Direcciones para Investigaciones Futuras:

Puede buscarse de manera mas detallada razones y motivos de porque los pedidos se cancelan o porque los miembros leales no compran add ons.

Referencias:

Base de datos:

<https://raw.githubusercontent.com/IHawkl/project/refs/heads/main/dflimpio--1>

Fuente internet

<https://www.kaggle.com>

Bibliotecas usadas:

- Pandas: Librería utilizada para la manipulación y análisis de datos.
Página oficial: <https://pandas.pydata.org/docs/>
- Matplotlib: Librería utilizada para la creación de visualizaciones básicas en Python.
Página oficial: <https://matplotlib.org/stable/contents.html>
- Seaborn: Herramienta utilizada para mejorar visualizaciones de datos en Python.
Página oficial: <https://seaborn.pydata.org/>
- Scikit-learn: Herramienta utilizada para el entrenamiento y evaluación del modelo de machine learning.
Página oficial: <https://scikit-learn.org/stable/>

Anexo:

Código fuente:

<https://github.com/IHawkl/project/blob/main/ProyectoFinalC11.ipynb>

Graficos no usados en reporte principal:

