Práctica Profesionalizante I

Entrega Nº 2: Exploración y Análisis de datos (EDA)

Equipo: "Data Voyagers"

BADIN, María Paula

LEDEZMA, Mariano

PERALTA, María Laura

URZAGASTE, María Gisela

Informe de Análisis de Datos

Introducción

Análisis Exploratorio de Datos

¿Qué es el análisis de datos exploratorio?

El análisis de datos exploratorio (**EDA**) lo utilizan los científicos de datos para analizar e investigar conjuntos de datos y resumir sus principales características, empleando a menudo métodos de visualización de datos.
Mediante EDA podemos:

- Determinar la mejor manera de manipular los orígenes de datos para obtener las respuestas necesarias en cuanto a: descubrir patrones, detectar anomalías, probar una hipótesis o comprobar supuestos

Usos
Se utiliza principalmente para:

1. Ver qué datos pueden revelarse más allá de la tarea de modelado formal o las pruebas de hipótesis
2. Permite conocer mejor las variables del conjunto de datos y las relaciones entre ellas.
3. También permite determinar si las técnicas estadísticas que está considerando para el análisis de datos son apropiadas.

¿Por qué es importante el análisis exploratorio de datos en la ciencia de datos?

El principal objetivo del EDA es consultar los datos antes de hacer cualquier suposición.

- Permite identificar errores obvios.
- Comprender mejor los patrones en los datos
- Detectar valores atípicos o sucesos anómalos
- Encontrar relaciones interesantes entre las variables.

Herramientas de análisis de datos exploratorio

Entre las herramientas de ciencia de datos más comunes utilizadas para crear un EDA incluyen:

1. **Python**: un lenguaje de programación interpretado y orientado a objetos con semántica dinámica. Sus estructuras de datos incorporadas de alto nivel, combinadas con la escritura dinámica y el enlace dinámico, hacen que sea muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de scripts o de unión para conectar los componentes existentes entre sí. Python y EDA se pueden utilizar conjuntamente para identificar los valores que faltan en un conjunto de datos, para que pueda decidir cómo manejarlos en machine learning.

2. **Librerías** de procesamiento automático como: **DataPrep**, **lDataProfiling** que nos permite tener un acercamiento a nuestro problema brindando información de manera rápida y así poder enriquecer nuestro desarrollo en lenguaje python

Procesos de preparación de datos

El proceso de preparación de datos puede variar según la necesidad, pero normalmente consta de los siguientes pasos: **Adquisición de datos, Exploración de datos, Limpieza de datos, Transformación de datos.**

Adquisición de datos

El primer paso en cualquier proceso de preparación de datos es adquirir los datos que un analista utilizará para llevar adelante su análisis. Estos datos en general se presentan en un formato accesible, como un documento de Excel o CSV.

El presente análisis se basa en un dataset inicial en formato .csv. El mismo consta de 180.519 filas y 53 columnas.

DataCoSupplyChainDataset.csv

Exploración de datos

Examinar y definir los datos nos ayuda a comprender cómo el análisis comenzará a tomar forma. Empleamos en primera instancia un análisis visual y a posterior un análisis de medidas estadísticas: la media, la media y la desviación estándar.

Se generaron **estadísticas descriptivas** para las columnas numéricas. Esto incluyó la mediana, la media y la desviación estándar para evaluar la distribución y la variabilidad de los datos.

- **Días de Envío (real) y Días de Envío (programado):** Los valores medios son 3.5 y 2.93, respectivamente, lo que sugiere que los envíos suelen realizarse en un plazo de aproximadamente 3 días, aunque algunos pedidos se envían el mismo día (mínimo de 0). La diferencia entre los días programados y reales podría reflejar retrasos.
- **Beneficio por Pedido:** La media es 21.97, pero la desviación estándar es alta (104.43), indicando una gran variabilidad en los beneficios por pedido. Además, hay valores negativos, lo que sugiere pérdidas en algunos pedidos.
- **Ventas por Cliente:** La media es 183.10 con un rango amplio, desde 7.49 hasta casi 1940. Esto indica que los montos de ventas varían considerablemente.
- **Riesgo de Entrega Tardía:** El valor medio de 0.55 indica que alrededor del 55% de los pedidos tienen un riesgo de entrega tardía, y es una variable binaria (0 o 1).

- ❖ **Descuentos en el Artículo del Pedido:** El descuento medio es de 20.66, con un máximo de hasta 500. Esto sugiere que algunos productos reciben descuentos considerables.
- ❖ **Cantidad de Artículos por Pedido:** El valor medio de 2.13 sugiere que la mayoría de los pedidos incluyen 1 o 2 artículos, con un máximo de 5 artículos.
- ❖ **Precios y Ganancias:** Los precios de los productos varían considerablemente, con un rango que va desde 9.99 hasta 1999.99. La ganancia por pedido también tiene una gran dispersión, lo que sugiere que algunos productos son mucho más rentables que otros.

Limpieza de datos

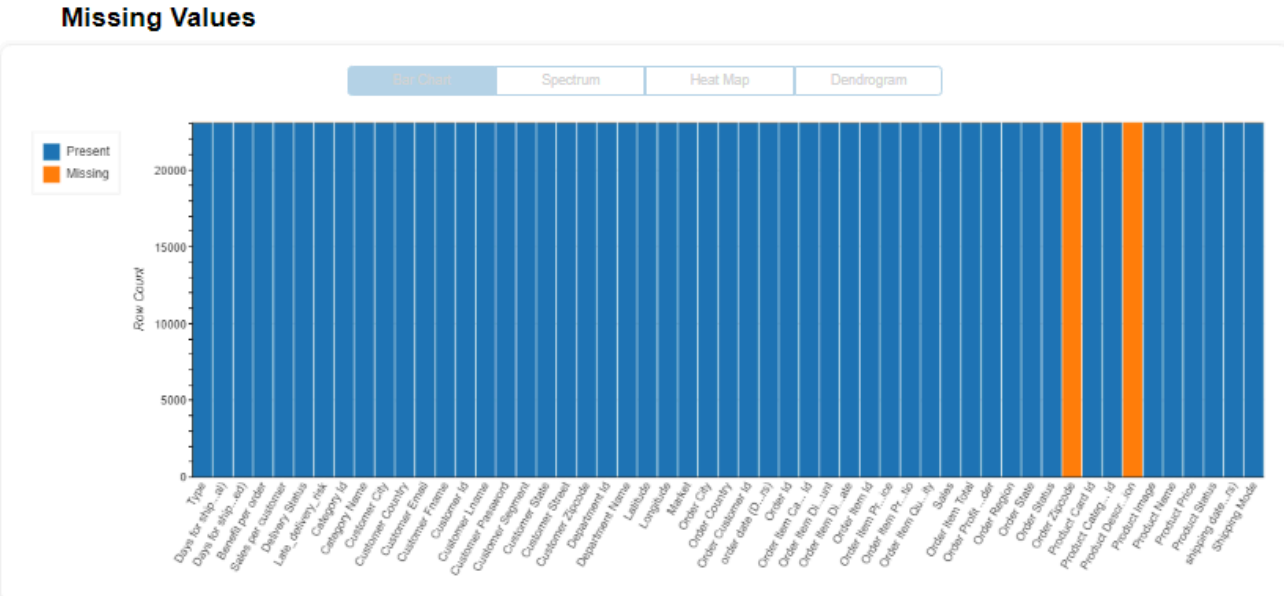
Durante la fase de exploración, es posible notar que los datos pueden estar mal estructurados y se requiere ordenarlos para mejorar su calidad. En nuestro análisis realizamos lo siguiente:

- Corrección de errores de entrada
- Eliminación de duplicados o valores atípicos
- Eliminación de datos faltantes
- Eliminacion de columnas no significativas
- Ocultación de información confidencial o sensible como nombres o direcciones

El primer paso en la exploración del dataset fue filtrar el conjunto de datos original para incluir solo registros correspondientes a tres países: Argentina, Brasil y México. Esto permitió concentrar el análisis en un subconjunto relevante del dataset. Logrando reducirlo a 23.090 filas y 53 columnas.

Este nuevo dataset incluye una variedad de tipos de datos, como float64, int64 y object. Las columnas abarcan información detallada sobre pedidos, clientes, productos y detalles de envío.

Nos encontramos con columnas que tienen valores nulos: 'Order Zipcode' y 'Product Description'.



Varias columnas del tipo “object”, que podrían ser una cadena de texto u otro tipo no numérico. Será necesario su posterior conversión a formatos numéricos, como en el caso de ‘Sales’.

Durante la exploración del dataset, se encontró que no existen filas duplicadas, esto significa que cada registro es único y no hay duplicaciones en los datos.

Adicionalmente, se verifica que el DataFrame contiene valores vacíos. Se decidió eliminar las columnas que contenían estos valores, ya que no eran relevantes para el análisis.

En la verificación realizada, no se encontraron valores negativos en las columnas clave del dataset. Sin embargo, se identificó que las columnas Sales, Order Item Total y Product Price presentan una cantidad considerable de valores no numéricos o nulos.

Pasamos al análisis y selección de las columnas que consideramos relevantes. Se realizaron dos pasos de limpieza:

1. **Columnas vacías:** se eliminaron las columnas que contienen sólo valores nulos o irrelevantes para el análisis. Las columnas eliminadas fueron: ‘Product Description’, ‘Order Zipcode’, ‘Product Status’.
2. **Columnas no necesarias:** se eliminaron columnas que no aportan información esencial para el análisis, para simplificar el dataset y enfocarnos en las variables clave. Las columnas eliminadas fueron: 'Sales per customer', 'Category Id', 'Customer City', 'Customer Email', 'Customer Fname', 'Customer Lname', 'Customer Password', 'Customer State', 'Customer Street', 'Customer Zipcode', 'Department Id', 'Department Name', 'Latitude', 'Longitude', 'Market', 'Order City', 'Order Customer Id', 'Order Item Cardprod Id', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Profit Ratio', 'Product Image'.

El **dataset final**, tras aplicar filtros y eliminaciones de columnas no necesarias, consta de 23,090 filas y 28 columnas. Este tamaño de dataset se considera óptimo para continuar con el análisis, ya que ha sido reducido a un formato manejable que mantiene la riqueza de la información relevante para el estudio. Con este conjunto de datos, se procederá a realizar un análisis más detallado y específico, asegurando la calidad y precisión de los resultados.

Luego de este proceso:

1- Se verifica que no existen filas duplicadas y espacios vacíos en el nuevo dataset.

2- Se procedió a **renombrar las columnas** del dataset para facilitar el análisis posterior. Las columnas fueron traducidas al español y se eliminaron espacios en los nombres para mantener consistencia y claridad.

3- Se realizó un **conteo de valores únicos por columna**, revelando información clave como el número de categorías distintas y la variedad en los valores de cada columna.

- **Variables con pocas categorías:** columnas como Pago, DiasEnvio(Programado), EstadoEntrega, RiesgoEntregaTardia, SegmentoCliente, ModoEnvio, PaisPedido tienen un número reducido de valores únicos, esto sugiere que son ideales para análisis categóricos o gráficos de barras.

- **<u>Variables con muchas categorías</u>:** columnas como IDCliente, IDPedido, IDArticuloPedido, NombreProducto, BeneficioPorPedido tienen una gran cantidad de valores únicos, esto implica que son variables específicas para cada registro y clave para el análisis a un nivel más detallado.
- **<u>Columnas con diversidad moderada</u>:** CategoriaProducto, PrecioProducto, TotalArticulosPedido muestran una diversidad moderada de valores, lo que puede ser relevante para identificar patrones de precios o ventas.

4- Se identificaron columnas con datos en formato **'object'** que deben ser convertidos a formatos numéricos para el análisis posterior. Por ejemplo, PrecioArticuloPedido y GananciaPorPedido se encontraban en formato object y se debieron convertir a float.

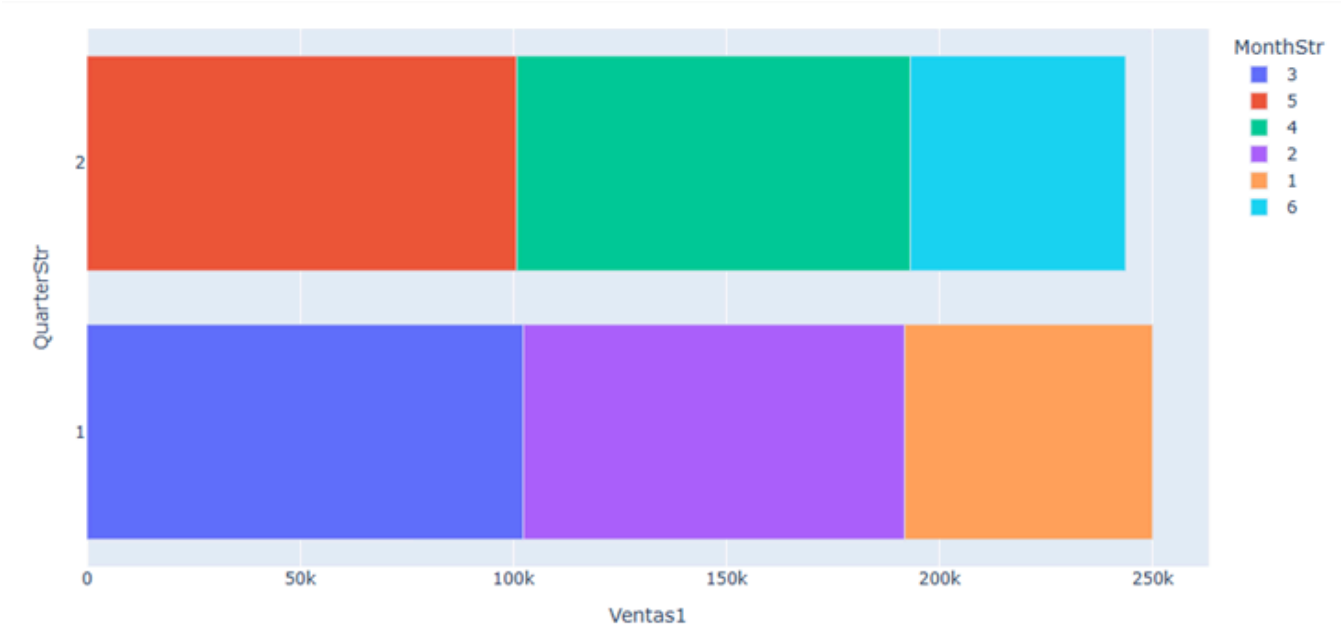
Transformación de datos

Algunos datos pueden estar listos para el análisis, mientras que otros conjuntos de datos pueden requerir un trabajo extra para transformarlos a fin de garantizar que se encuentren en un formato o una estructura acorde a la necesidad de nuestro problema para lograr resultados significativos.
En nuestro análisis realizamos transformaciones en:

1. Las columnas de datos que contienen valores de fecha en formato String:
    - Por una lado transformar estas columnas a un formato DateTime
    - Por otro, desglosar la fecha en sus componentes (año, mes y dia).
2. Las columnas que contienen precios

Estos cambios permitirán realizar distintos análisis con respecto a periodos de tiempos específicos.

Ej: análisis de ventas (anual, cuatrimestral, mensual)



Análisis Univariado

El análisis univariado se enfoca en la evaluación de una sola variable a la vez para resumir y entender sus características fundamentales. En este informe, se aplica para examinar la distribución y estadísticas descriptivas de variables como los días de envío, precios del producto, beneficio por pedido, entre otras. Este análisis ayuda a identificar tendencias, detectar valores atípicos y obtener una visión general de la variable bajo estudio.

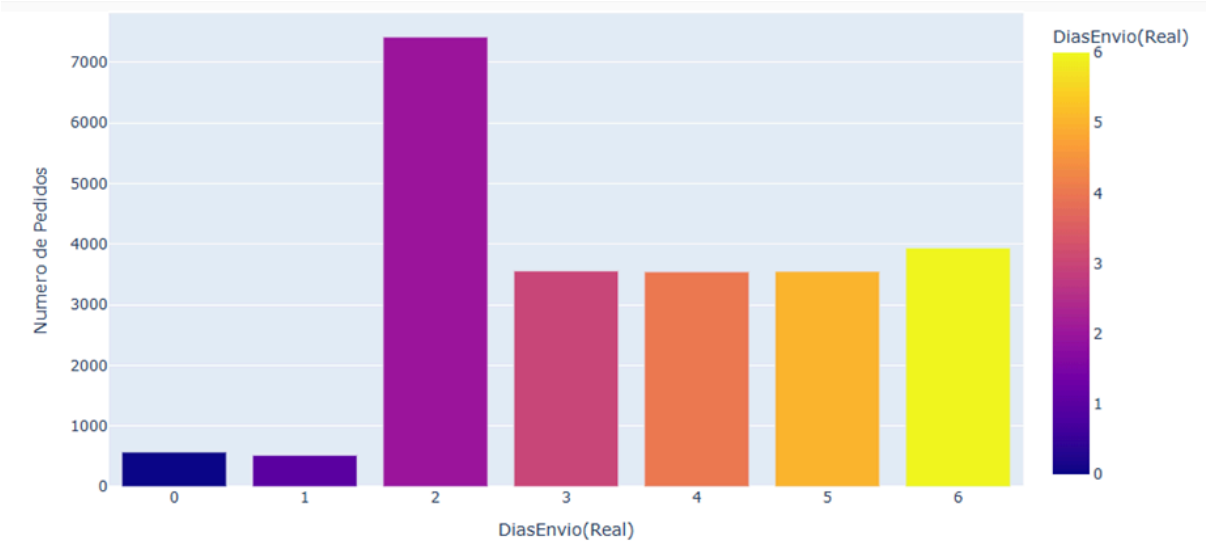
Durante la exploración inicial de los datos, se revisaron diversas **variables categóricas** clave, como categoría de producto, métodos de pago, modos de envío, segmento de clientes y estado de entrega. También **variables numéricas** como Días de envío, real y programado, beneficio por pedido, riesgo de entrega tardía y cantidad de artículos por pedido.

En el análisis de precios de productos y beneficios por pedido, se observa que estaban en formato de texto. Para solucionar esto, se implementó una función de limpieza que elimina caracteres no numéricos, convirtiéndolos en formato numérico. Esta limpieza permite calcular estadísticas clave como la mediana, media y desviación estándar. Los resultados mostraron que la media y mediana de los precios varían considerablemente entre categorías, con algunas mostrando valores muy altos debido a productos específicos de mayor valor. Similar comportamiento reflejan las estadísticas de beneficio por pedido.

Se generaron histogramas y boxplots para visualizar la distribución y los valores atípicos en las siguientes variables: días de envío, real y programado, ventas, precio del producto, beneficio por pedido y cantidad de artículos por pedido.

- ***Histogramas:*** ilustran la frecuencia de las observaciones en distintos intervalos de cada variable. Permiten observar la distribución general de los datos y la presencia de cualquier sesgo o agrupamiento.

  Ejemplo: Dias de envio reales para un pedido

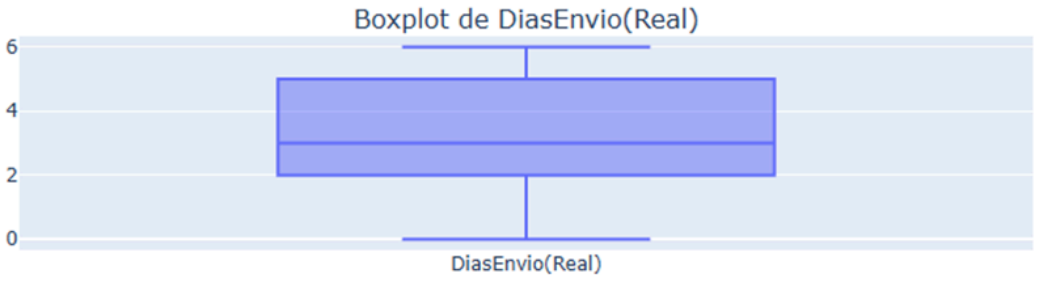


- ***Boxplots:*** muestran la mediana, cuartiles y valores atípicos para cada variable. Son útiles para identificar la variabilidad y detectar valores extremos que podrían influir en el análisis general.

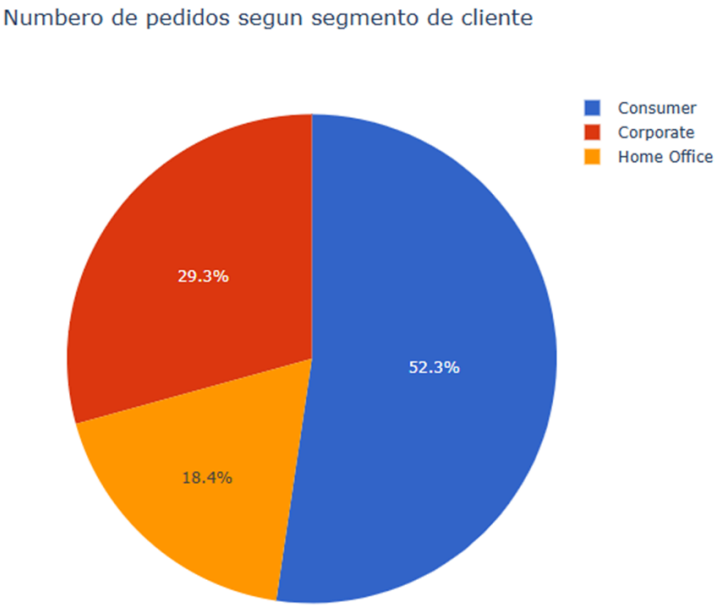
Algunos de los resultados obtenidos:

- **Días de envío(real):** la mayoría de los envíos se completan en un rango de 0 a 6 días. Los boxplots indican algunos valores atípicos, pero estos no son numerosos.
- **Días de envío (programado):** Los datos se concentran en torno a 4 días, con una alta consistencia y pocos valores atípicos.
- **Ventas:** Los datos muestran una concentración en valores específicos.
- **Precio del producto:** la mayoría de los precios se agrupan en un valor específico, con pocos valores atípicos, lo que sugiere una alta consistencia en los precios.
- **Beneficio por pedido:** el análisis revela una amplia variabilidad en los beneficios.
- **Cantidad de artículos por pedido:** los pedidos suelen contener entre 1 y 5 artículos, con una distribución relativamente uniforme. No presenta valores atípicos significativos.

Ejemplo: Boxplot de Visualización para Dias de Envio



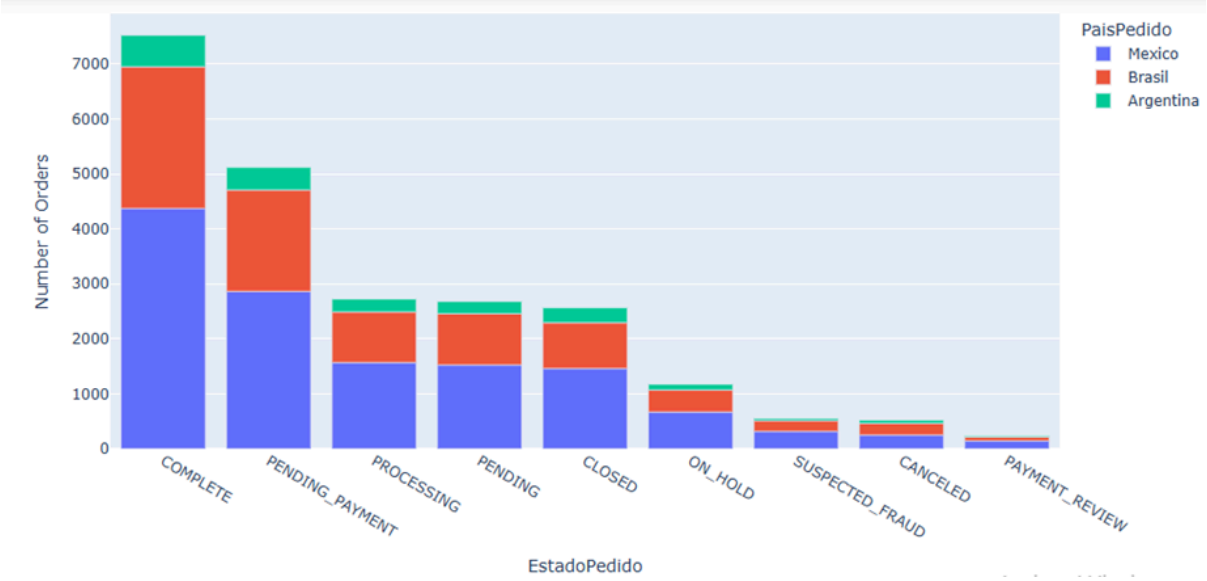
Ejemplo: Segmento de clientes según regiones



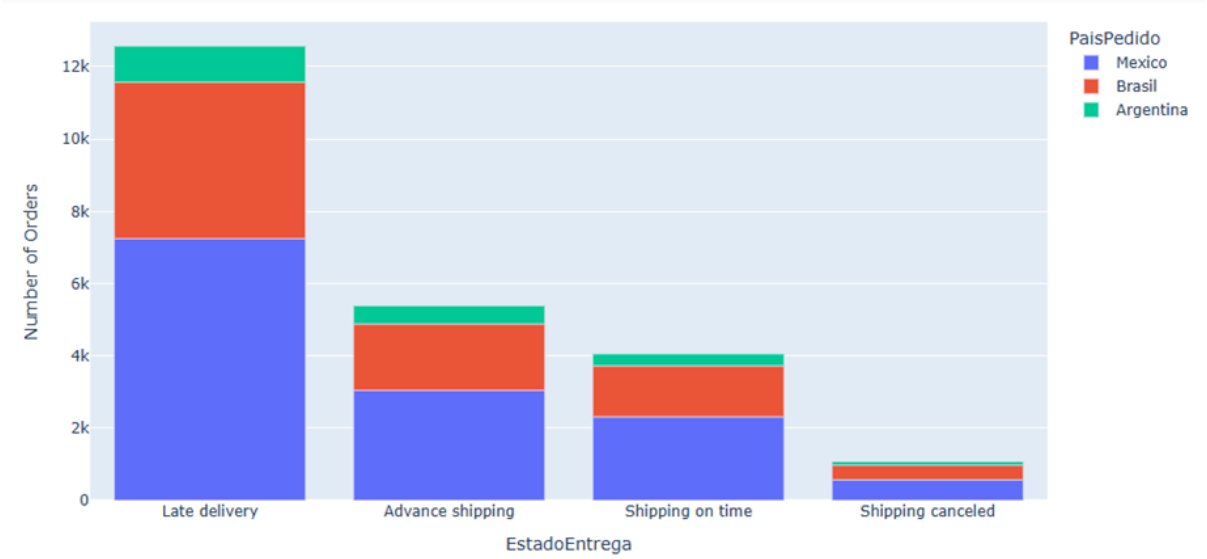
Análisis Bivariado

El análisis bivariado es una técnica estadística que examina la relación entre dos variables para entender cómo una puede influir o relacionarse con otra. En el contexto de este informe, el análisis bivariado se utiliza para explorar la interacción entre variables como el estado de las órdenes y la región de pedido, o el número de pedidos y el tiempo de entrega real. Este tipo de análisis permite identificar patrones, correlaciones y posibles áreas de mejora en el proceso de gestión y logística.

Ejemplo: Estado de los pedidos segun el pais



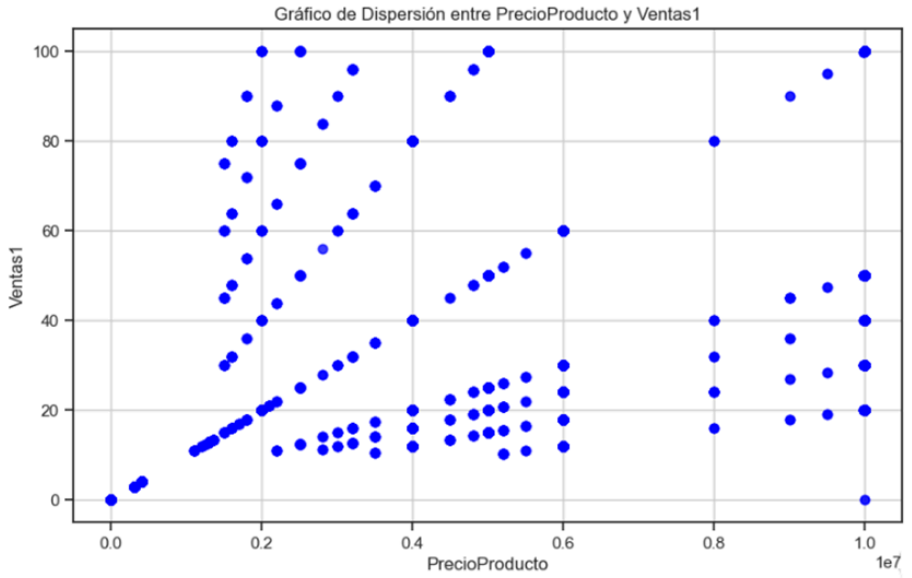
Ejemplo: Estado de las entregas según país



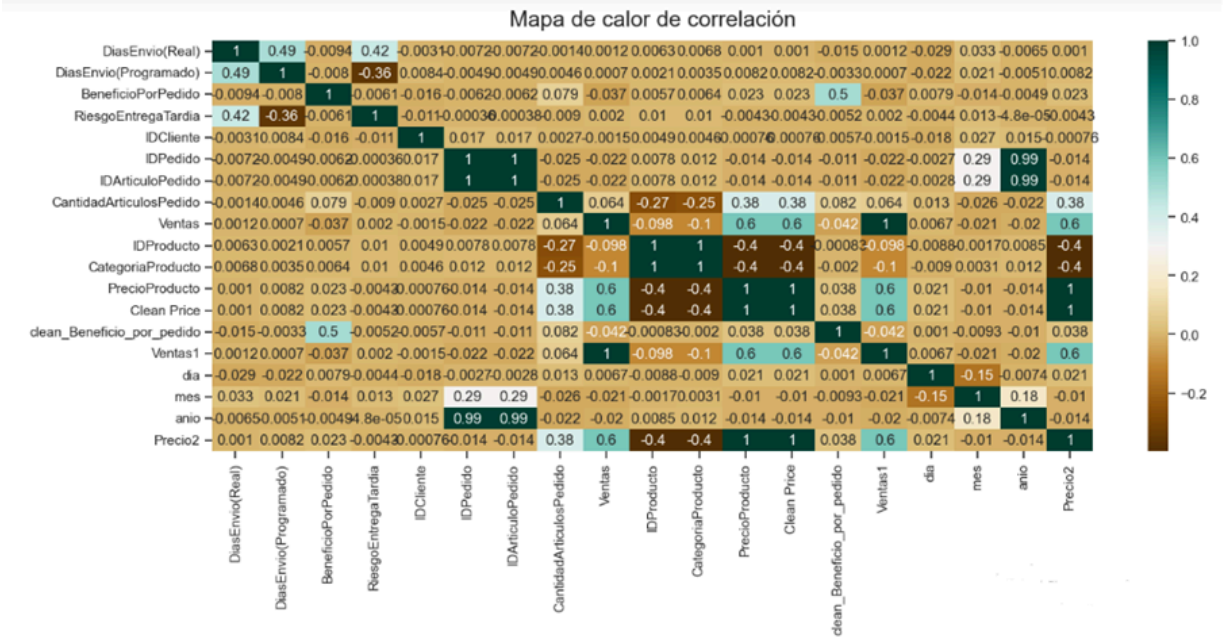
Visualizaciones multivariantes

Se utilizan para relacionar y comprender las interacciones entre los diferentes campos en los datos. En nuestro análisis incluimos:

1. **Gráfico de dispersión:** se utiliza para trazar puntos de datos en un eje horizontal y uno vertical para mostrar cuánto afecta una variable a otra.



2. **Mapa de calor:** es una representación gráfica de datos donde los valores se representan por color.



Consideraciones finales:

1. Se puede utilizar el análisis exploratorio para garantizar que los resultados que generan sean válidos y aplicables a las conclusiones y objetivos de negocio deseados. Además, permite confirmar a las partes interesadas que están haciendo las preguntas correctas. El EDA ayuda a responder las preguntas sobre desviaciones estándar, variables categóricas e intervalos de confianza.
2. Una vez que se ha completado el EDA y se ha extraído la información útil, sus características pueden utilizarse para un análisis o modelado de datos más complejo, incluido machine learning.

Link a Repositorio

https://github.com/ISPC-PP1-2024/proyecto/blob/main/codigo/2%20entrega_primera-parte/FINAL%20PARA%20ENTREGAR-LISTO/Resolucion_2Entrega_suplyChain-final%20-%20conGraficos.html

Link Trello

https://trello.com/b/ZxQZufSW/pp1cdia-2024