Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

Visualización Interactiva de la información

Laboratorio: Refinamiento de datos con Python y Visualización con Power BI

Nombre de los autores:

Andrés Felipe Vargas González

José Francisco Peñaloza González

José Gustavo Cabiativa Moreno

Universidad Internacional de la Rioja (UNIR)

Diciembre de 2024

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

Tabla de contenido

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

Documentación del código diseñado para el proceso de transformación de los datos

Para realizar el proceso de limpieza del **dataset**, se elaboró un notebook en **Jupyter Lab** de Python que contiene el script que hace el proceso. Con la finalidad de documentar el proceso de limpieza, a continuación, se describe el procedimiento realizado:

Proceso de limpieza

1. Se importan varias librerías que son útiles para el análisis y la visualización de datos, asi:
    a. **Pandas** (para trabajo de estructuras de datos en especial los dataframes),
    b. **Os** para acceso de archivos
    c. **Re** (manejo de cadenas de texto),
    d. **Matplotlib** (gráficos) y
    e. **Datetime** (manejo de fechas).
2. Posteriormente, se define la ruta de la carpeta de los datos, se carga el archivo CSV y se crea el **dataframe** con pandas.
3. Se define la función `**determinar_patron**` que toma una cadena de texto como entrada y genera un patrón de expresión regular basado en los caracteres de la cadena, y clasifica los caracteres en dígitos, letras, espacios y signos de puntuación. Adicionalmente, va eliminando duplicados consecutivos en el patrón resultante.
4. Se procede con definir una serie de funciones que realizarán la corrección de los datos de correo electrónico, producto, ciudad, categoría, fecha venta, cliente, teléfono, dirección, método de pago, estado, comentario. Al final de la aplicación de cada corrección, se cuenta la frecuencia de los datos en el **Dataframe**, se ordena los resultados en orden ascendente y se crea un gráfico de barras para visualizar la frecuencia de los productos. El proceso se describe a continuación:

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

a) La corrección del correo electrónico se hace a través de tres funciones:

b) La función `**limpiar_correo**`, que toma una dirección de correo electrónico como entrada y elimina caracteres no permitidos tanto en el usuario como en el dominio, si la dirección de correo no cumple con estos criterios, se devuelve la dirección original.

c) La función `**corregir_arroba**` que corrige errores en la dirección de correo electrónico relacionados con el uso del símbolo **@** y elimina caracteres no deseados

d) La función llamada `**validar_email**` que verifica si una dirección de correo electrónico es válida según un patrón de expresión regular. Posteriormente, se aplican las funciones al **dataframe** y se añade una nueva columna que indica si cada dirección es válida o no.

e) Para la corrección del producto, se definen las funciones `**validate_string**` y `**clean_string**`, la primera verifica si una cadena de texto contiene solo caracteres alfanuméricos y espacio y la segunda limpia una cadena de texto eliminando caracteres no deseados y ajustando los espacios, posteriormente se limpia y valida los valores de la columna Producto, y luego agrupa y cuenta los productos válidos e inválidos.

f) Para la corrección de la ciudad, se define la función `**add_space_before_mayus**`, que toma una cadena de texto como entrada y añade un espacio antes de cada letra mayúscula que sigue a una letra minúscula.

g) Luego, se limpia y se valida los valores de la columna **Categoria**, y luego se agrupa y se cuenta las categorías válidas e inválidas.

h) Para la corrección de la fecha venta, se definen dos funciones:

- `**is date**`, que verifica si una cadena de texto es una fecha válida en uno de varios formatos especificados
- `**clean_string_date**` que limpia una cadena de texto eliminando caracteres no deseados al principio y al final, ajustando los espacios.

i) Posteriormente, se limpia y se valida los valores de la columna Cliente, y luego se agrupa y se cuenta los clientes válidos e inválidos, efectuándose la corrección de los clientes.

j) Para corregir los datos del teléfono, se aplica la función `**clean_string_date**` a cada valor en la columna Teléfono del **DataFrame** y luego se genera patrones de expresión regular para los números de teléfono en la columna **Teléfono**, se almacena estos patrones en una nueva columna **TelefonoRegex**, y luego obtiene una lista de los patrones únicos. Después, se itera sobre cada patrón único de expresión regular en la columna **TelefonoRegex**, se obtiene un ejemplo de número de teléfono para cada patrón. Luego se aplica una función llamada `**clean_phone_number**`, que limpia un número de teléfono eliminando ciertos caracteres no deseados y ajustando el formato. Una tercera función denominada `**standarize_phone**`, la cual toma un número de teléfono como entrada, elimina el prefijo internacional **+1**, elimina caracteres no numéricos y lo formatea en el formato **XXX-XXX-XXXX** si tiene 10 dígitos. Finalmente, la función `**llamada delete_001**` que limpia y formatea un número de teléfono eliminando el prefijo 001 y ajustando el formato según la longitud del número.

k) Para la corrección de los datos de dirección, se implementan dos funciones:

- `**add_space_between_num_and_str**` que inserta un espacio entre letras y números en una cadena de texto.
- `**replace_apoap**`, que reemplaza la cadena **APOAP** por **APO AP** en una cadena de texto.

i) La corrección de método de pago se realiza a través de la aplicación de las funciones `**clean_string**`, `**add_space_before_mayus**` y `**validate_string**` a la columna **Metodo_Pago**, y luego agrupa y cuenta los métodos de pago que sean válidos e inválidos.

j) Se define una función `**place_encamino**` que reemplaza la cadena Encamino por En camino en una cadena de texto, con el objeto de hacer correcciones

en la columna `**estado**`, y luego se aplica esta función junto con las definidas previamente `**clean_string**`, `**add_space_before_mayus**` y `**validate_string**` sobre la columna `**estado**` para contar los estados válidos e inválidos.

k) Para la corrección de los comentarios, se reemplaza los valores nulos en la columna `**Comentario**` con cadenas vacías y luego se limpia los valores de la columna eliminando los caracteres no deseados y ajustándose los espacios.

5. Finalmente, se elimina las columnas de validación del **dataframe** y se pasa el **dataframe** a un archivo de formato CSV correspondiente al **dataset** debidamente limpiado.

Documentación del proceso de construcción de la visualización de datos

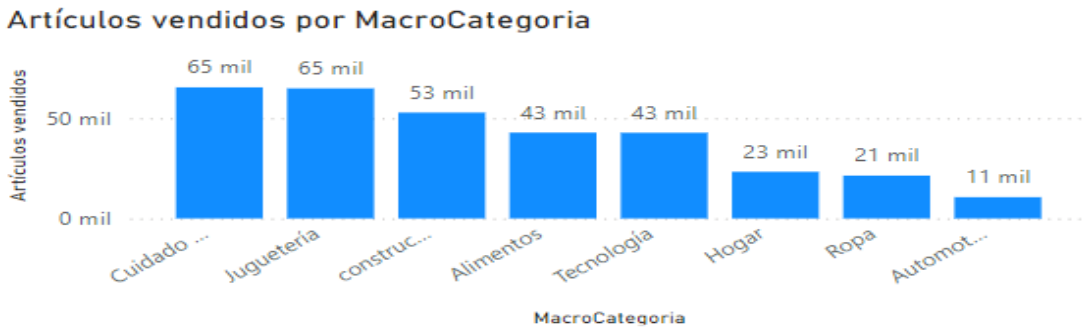
En primer lugar y previo al diseño de la visualización de datos en Power BI, se formula una serie de preguntas que se pretender responder con la visualización. A continuación, se relacionan las preguntas agrupadas por categorías y el gráfico respectivo que las resuelve:

Panel de preguntas y visualización de los datos

Ventas por Categorías y Productos

- ¿Cuál es la distribución de las ventas por macrocategoría?

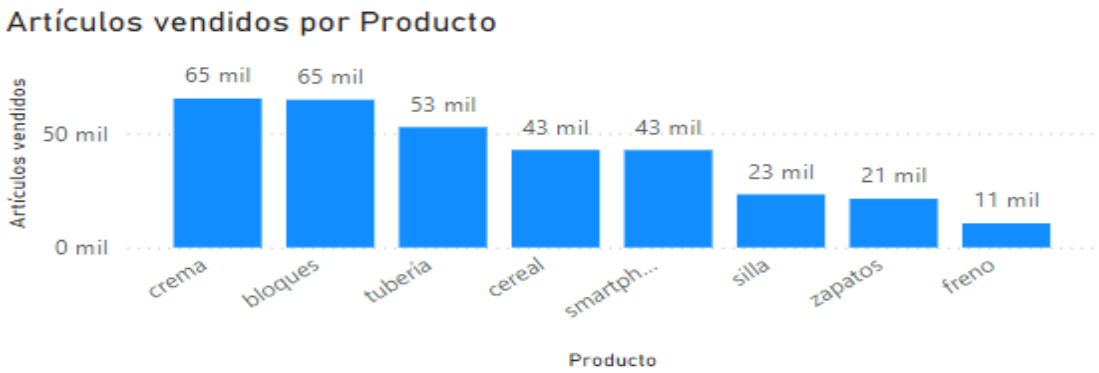
Se utilizó un gráfico de barras porque permite comparar fácilmente los valores absolutos de cada macrocategoría en términos de artículos vendidos. Este tipo de gráfico es ideal para mostrar diferencias claras entre categorías.



- ¿Qué productos tienen mayor volumen de ventas?

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

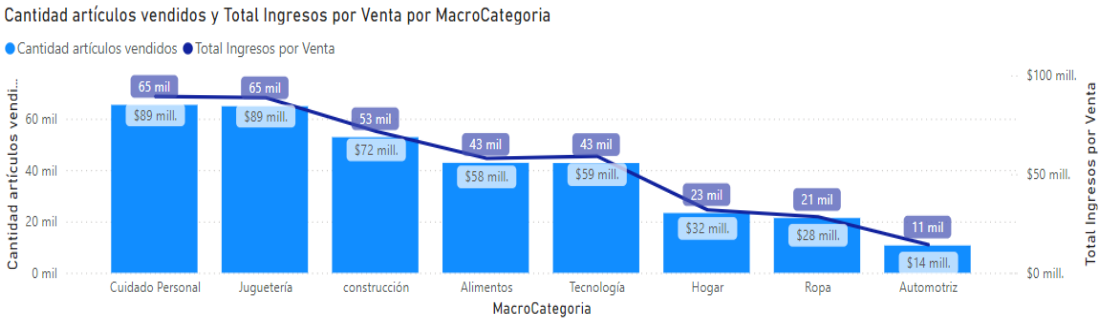
El gráfico de barras facilita observar qué productos específicos lideran en ventas. Es perfecto para detectar los artículos más demandados y analizar su contribución en comparación con otros.



- ¿Qué macrocategoría genera mayores ingresos en ventas?

Se eligió un gráfico combinado de barras y líneas porque permite observar dos métricas relacionadas en un solo lugar:

1. **Barras**: Representan el número de artículos vendidos, mostrando el volumen de ventas.
2. **Línea**: Representa el ingreso total, destacando la relación entre la cantidad vendida y los ingresos generados. Este enfoque permite evaluar tanto el rendimiento por volumen como por valor.



Análisis Geográfico

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	

- ¿Qué ciudades generan mayores ingresos en ventas?

Se utilizó un gráfico de mapa de calor, que permite determinar fácilmente el volumen de ventas por sector geográfico. Este tipo de gráfico es ideal para observar que ciudades generan mayor o menor volumen de ventas y ubicarlas geográficamente, para determinar si influye su ubicación, clima o factor logístico y tomar acciones basadas en datos informados

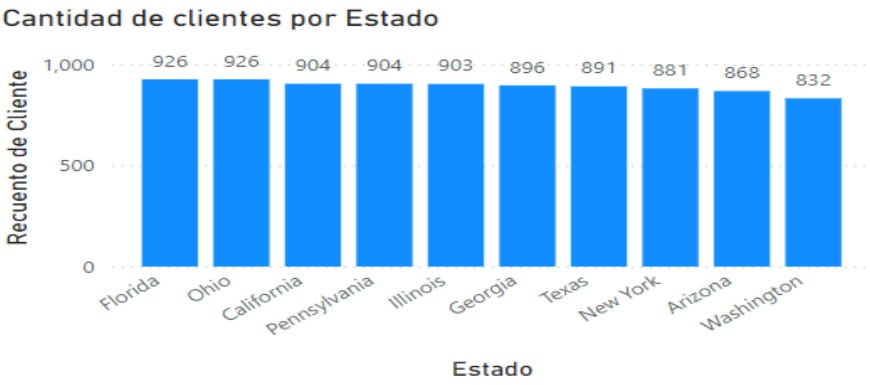
- ¿Cuántas ventas se realizaron por estado?

Se utilizó un gráfico de barras porque permite comparar fácilmente los valores de ventas por cada estado del país. Este tipo de gráfico es ideal para mostrar diferencias claras de ingresos entre estados

- ¿Qué estados tienen el mayor volumen de clientes VIP?

Este tipo de gráfico facilita la identificación rápida de los estados que tienen el mayor número de clientes VIP, lo que es esencial para la toma de decisiones estratégicas. Al comparar los

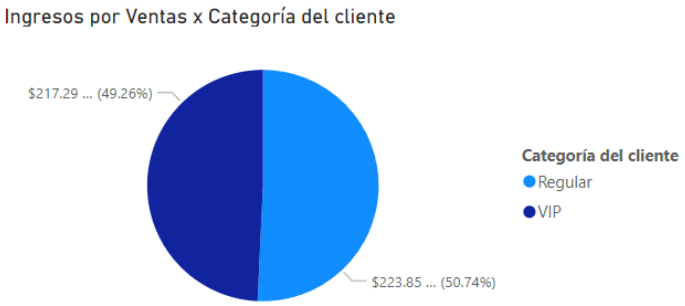
volúmenes de clientes VIP, se puede inferir cuáles estados generan mayores ingresos, lo que es crucial para focalizar esfuerzos de marketing y servicio al cliente.



Comportamiento del Cliente

- ¿Cómo se distribuyen los clientes por categoría (Regular, VIP)?

El Gráfico de Pastel muestra de manera clara y proporcional cómo se distribuyen los clientes entre las categorías Regular y VIP, facilita la comparación visual de las proporciones de cada categoría y proporciona una visión rápida y clara de la distribución total de los clientes.



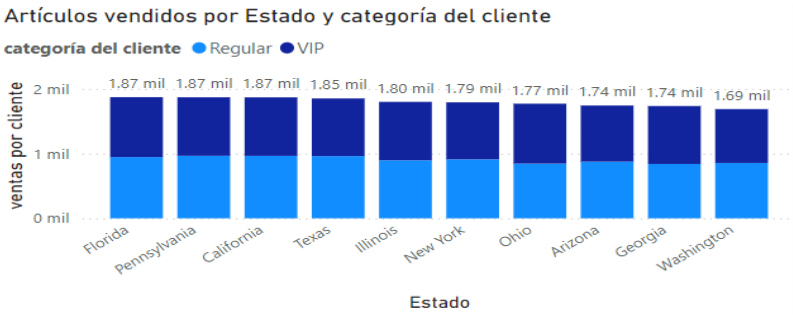
- ¿Qué ciudades tienen el mayor volumen de clientes VIP?

El mapa de calor es muy útil en este objetivo ya que permite una comparación directa de los volúmenes de clientes VIP entre diferentes ciudades, facilita la identificación de las

ciudades con mayores volúmenes de clientes VIP, por lo que podemos usarlo para establecer estrategias de focalización de marketing y recursos.

- ¿Cuál es la proporción de estados asociados con clientes VIP?

Gráfico de Barras apiladas, es útil ya que muestra claramente la proporción de clientes VIP en diferentes estados y facilita la comparación visual de estas proporciones.



Métodos de Pago y Descuentos

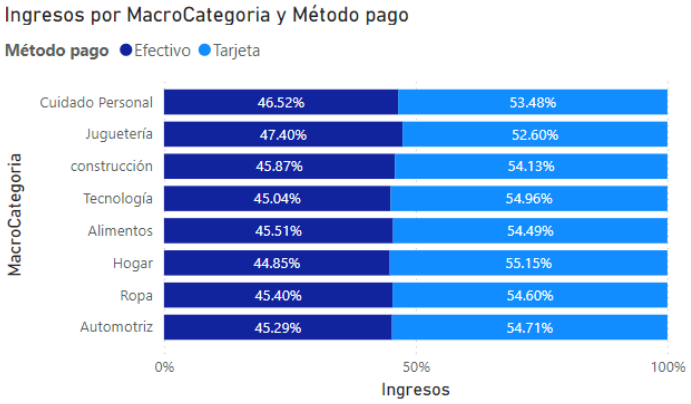
- ¿Cuáles son los métodos de pago más utilizados?

El diagrama de pastel muestra cada método de pago como un segmento del total, lo que permite ver de manera clara y proporcional qué métodos predominan en el uso. Es fácil

identificar las secciones grandes y pequeñas a simple vista y permite comparar directamente la participación relativa de cada método de pago en el conjunto total.

- ¿Qué porcentaje de ventas tuvo cada método de pago por macrocategoría?

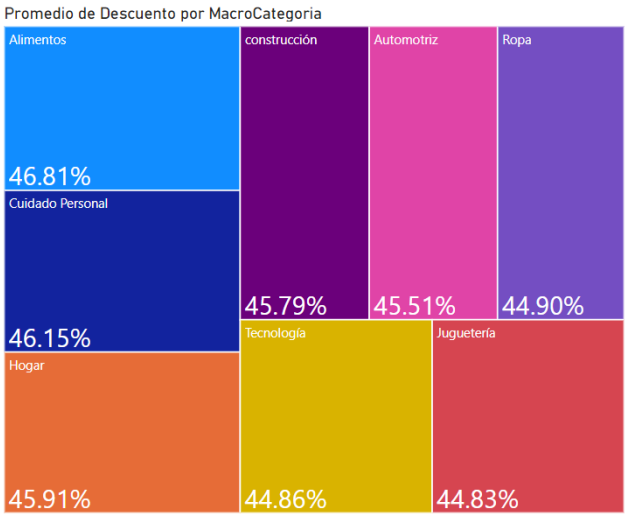
El gráfico de barras horizontales apiladas permite comparar tanto el total de ventas por macrocategoría como la proporción de cada método de pago dentro de cada macrocategoría y se puede ver al mismo tiempo en un solo gráfico cómo se distribuyen los métodos de pago dentro de cada macrocategoría



- ¿Qué descuentos son más comunes en las ventas? (Gráfico TreeMap)

Un gráfico TreeMap es ideal para visualizar los descuentos más comunes en las ventas, mostrando cada tipo de descuento como un rectángulo proporcional a su frecuencia. Esto permite una rápida identificación de los descuentos más y menos comunes y facilita la comparación entre múltiples categorías en una sola visualización.

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	



Tendencias Temporales

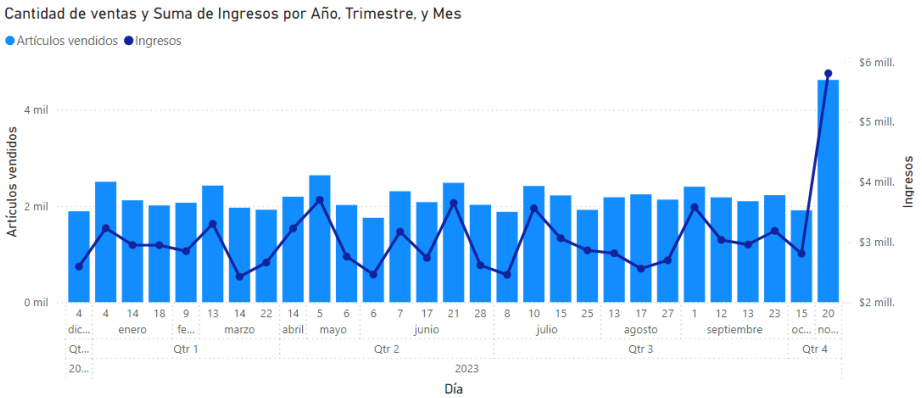
- ¿Cuáles son los ingresos mensuales por macrocategoría?
- ¿Cuál es la tendencia mensual en la cantidad de productos vendidos?

Se eligió un gráfico combinado de barras y líneas porque permite observar dos métricas relacionadas en un solo lugar:

Barras: Representan el número de artículos vendidos, mostrando el volumen de ventas.

Línea: Representa el ingreso total, destacando la relación entre la cantidad vendida y los ingresos generados. Ayuda a identificar picos, caídas y patrones estacionales en los ingresos mensuales.

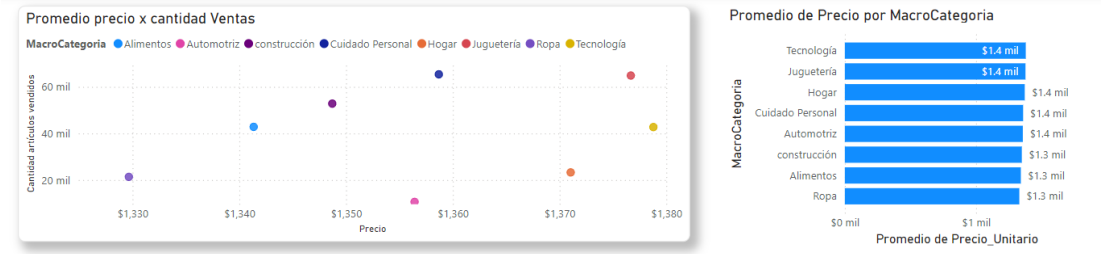
Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	



Relación entre Variables

- ¿Cuál es la relación entre cantidad de productos vendidos y su precio promedio?

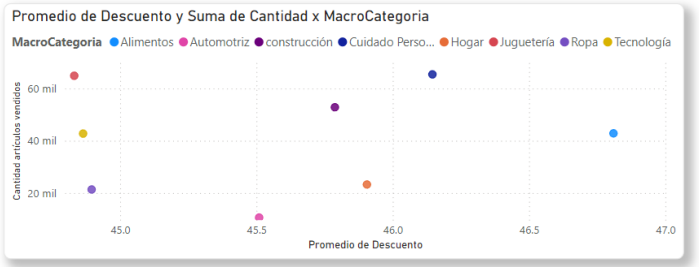
En el diagrama de dispersión, cada punto representa un producto. El eje horizontal (X) representa el precio promedio, y el eje vertical (Y) la cantidad de productos vendidos. Esto ayuda a visualizar si hay una correlación entre el precio y la cantidad vendida. En el gráfico de barras, cada barra representa un rango de precios y la altura de la barra muestra la cantidad total de productos vendidos en ese rango. Esto facilita la comparación de cuántos productos se venden a diferentes niveles de precios.

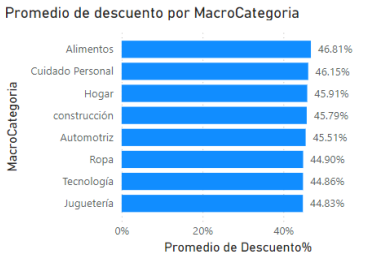


- ¿Cuál es la relación entre el descuento aplicado y el volumen de compra?

En el diagrama de dispersión, cada punto representa una transacción. El eje horizontal (X) representa el porcentaje de descuento aplicado, y el eje vertical (Y) el volumen de compra. Esto ayuda a visualizar si hay una correlación entre el descuento y el volumen de compra. En el gráfico de barras apiladas horizontales, cada barra representa un rango de volumen de compra, y cada sección dentro de la barra representa un porcentaje de descuento aplicado.

Esto facilita una comparación detallada y proporciona una visión clara de cómo el descuento afecta el volumen de compra.





Contenido de presentable

El presentable es un archivo comprimido .zip que incluye los siguientes elementos:

- **Carpeta Datos ventas:** Esta carpeta contiene los archivos csv del ejercicio:
  - **Datos Ventas:** Es el dataset limpio.
  - **Datos Ventas Clean:** Es el dataset limpio
- **Archivo Notebook.ipynb:** Es el cuadernillo de Jupyter que a traves de Python se consigue limpiar la data.
- **Visualización.pbix:** Es el archivo de Power BI donde se cargans las visualizaciónes de este ejercicio.
- Diccionario de datos: Es el archivo donde se cambian los valores genericos del ejercicio por nombres de productos más realistas para mejorar la visualización del ejercicio.
- Estados: Es la asignación de Estados de USA para que el dataset pueda permitir hacer graficas de Geolocalización en Power BI.

También se podrá disponer de este contenido en nuestro repositorio haciendo click aquí.

Así mismo también el Power BI queda publicado en el workspace de la universidad al cual se podrá consultar a través de este link, usando las credenciales de Office 365 de la Unir.

Asignatura	Datos del alumno	Fecha
Visualización Interactiva de la Información	Apellidos: Vargas, Cabiativa y Peñaloza	2024-12-09
	Nombre: Andrés, Gustavo y José	