Sprint 4 - Proyecto

DATASETS

- instacart_orders.csv
- products.csv
- order_products.csv
- aisles.csv
- departments.csv

¡Felicidades! Has terminado el sprint sobre análisis exploratorio de datos. Es momento de aplicar el conocimiento y las habilidades que adquiriste en un estudio de caso de análisis.

Cuando hayas terminado el proyecto, envía tu trabajo al equipo de revisión en la plataforma para que lo evalúen. Te darán su feedback en 48 horas. Realiza los cambios según los comentarios que recibiste, y luego envía la nueva versión de vuelta al equipo de revisión.

Es posible que recibas más comentarios sobre la nueva versión. Esto es totalmente normal. Es común que pases por varios ciclos de feedback y revisiones.

Tu proyecto se considerará completado una vez que el equipo de revisión lo apruebe.

Descripción del proyecto

Para este proyecto, trabajarás con datos de Instacart.

Instacart es una plataforma de entregas de comestibles donde la clientela puede registrar un pedido y hacer que se lo entreguen, similar a Uber Eats y Door Dash. Este conjunto de datos particular fue lanzado públicamente

(materiales en inglés) por Instacart en 2017 para una competición Kaggle *(materiales en inglés)*. Los datos reales pueden descargarse directamente de la página de la competición Kaggle.

El conjunto de datos que te hemos proporcionado tiene modificaciones del original. Redujimos el tamaño del conjunto para que tus cálculos se hicieran más rápido e introdujimos valores ausentes y duplicados. Tuvimos cuidado de conservar las distribuciones de los datos originales cuando hicimos los cambios.

Tu misión es limpiar los datos y preparar un informe que brinde información sobre los hábitos de compra de los clientes de Instacart. Después de responder a cada pregunta, escribe una breve explicación de tus resultados en una celda markdown de tu Jupyter notebook.

Este proyecto requerirá que hagas gráficos que comuniquen tus resultados. Asegúrate de que cualquier gráfico que vayas a crear tenga un título, ejes etiquetados y una leyenda si es necesario; e incluye `plt.show()` al final de cada celda con un gráfico.

Diccionario de datos

Hay cinco tablas en el conjunto de datos, y tendrás que usarlas todas para hacer el preprocesamiento de datos y el análisis exploratorio de datos. A continuación se muestra un diccionario de datos que enumera las columnas de cada tabla y describe los datos que contienen.

- `instacart_orders.csv`: cada fila corresponde a un pedido en la aplicación Instacart.
  - `'order_id'`: número de ID que identifica de manera única cada pedido.
  - `'user_id'`: número de ID que identifica de manera única la cuenta de cada cliente.

- `'order_number'`: el número de veces que este cliente ha hecho un pedido.
- `'order_dow'`: día de la semana en que se hizo un pedido (0 si es domingo).
- `'order_hour_of_day'`: hora del día en que se hizo el pedido.
- `'days_since_prior_order'`: número de días transcurridos desde que este cliente hizo su pedido anterior.

`products.csv`: cada fila corresponde a un producto único que pueden comprar los clientes.

- `'product_id'`: número ID que identifica de manera única cada producto.
- `'product_name'`: nombre del producto.
- `'aisle_id'`: número ID que identifica de manera única cada categoría de pasillo de víveres.
- `'department_id'`: número ID que identifica de manera única cada departamento de víveres.

`order_products.csv`: cada fila corresponde a un artículo pedido en un pedido.

- `'order_id'`: número de ID que identifica de manera única cada pedido.
- `'product_id'`: número ID que identifica de manera única cada producto.
- `'add_to_cart_order'`: el orden secuencial en el que se añadió cada artículo en el carrito.
- `'reordered'`: 0 si el cliente nunca ha pedido este producto antes, 1 si lo ha pedido.

`aisles.csv`

- `'aisle_id'`: número ID que identifica de manera única cada categoría de pasillo de víveres.
- `'aisle'`: nombre del pasillo.

`departments.csv`

- `'department_id'`: número ID que identifica de manera única cada departamento de víveres.
- `'department'`: nombre del departamento.

Instrucciones para completar el proyecto

Paso 1: Abre los archivos de datos (`/datasets/instacart_orders.csv`, `/datasets/products.csv`, `/datasets/aisles.csv`, `/datasets/departments.csv` y `/datasets/order_products.csv`) y echa un vistazo al contenido general de cada tabla.

Observa que los archivos tienen un formato no estándar, así que vas a necesitar establecer ciertos argumentos en `pd.read_csv()` para leer los datos correctamente. Mira los archivos CSV para tener una idea de cuáles deberían ser esos argumentos.

Ten en cuenta que `order_products.csv` contiene *muchas* filas de datos. Cuando un DataFrame tiene demasiadas filas, `info()` no imprimirá los recuentos no nulos por defecto. Si quieres imprimir los recuentos no nulos, incluye `show_counts=True` cuando llames a `info()`.

Paso 2: Preprocesa los datos de la siguiente manera:

> Verifica y corrige los tipos de datos (por ejemplo, asegúrate de que las columnas de ID sean números enteros).
> Identifica y completa los valores ausentes.
> Identifica y elimina los valores duplicados.

Asegúrate de explicar qué tipos de valores ausentes y duplicados encontraste, cómo los completaste o eliminaste y por qué usaste esos métodos. ¿Por qué crees que estos valores ausentes y duplicados pueden haber estado presentes en el conjunto de datos?

Paso 3: Una vez que los datos estén procesados y listos, haz el siguiente análisis:

[A] (deben completarse todos para aprobar)

> Verifica que los valores en las columnas `'order_hour_of_day'` y `'order_dow'` de la tabla `orders` sean razonables (o sea, `'order_hour_of_day'` va de 0 a 23 y `'order_dow'` va de 0 a 6).

Crea un gráfico que muestre el número de personas que hacen pedidos dependiendo de la hora del día.
Crea un gráfico que muestre qué día de la semana la gente hace sus compras.
Crea un gráfico que muestre el tiempo que la gente espera hasta hacer su próximo pedido, y comenta los valores mínimos y máximos.

[B] (deben completarse todos para aprobar)

¿Hay alguna diferencia en las distribuciones de `'order_hour_of_day'` en miércoles y sábados? Traza los histogramas de ambos días en el mismo gráfico y describe las diferencias que observes.
Traza la distribución del número de pedidos que hacen los clientes y las clientas (por ejemplo, cuántos clientes hicieron un solo pedido, cuántos hicieron solo dos, cuántos solo tres, etc.)
¿Cuáles son los 20 principales productos que se piden con más frecuencia (muestra su identificación y nombre)?

[C] (deben completarse todos para aprobar)

¿Cuántos artículos compra la gente por lo general en un pedido? ¿Cómo es la distribución?
¿Cuáles son los 20 principales artículos que se vuelven a pedir con más frecuencia (muestra sus nombres e identificaciones de producto)?
Para cada producto, ¿qué proporción de sus pedidos se vuelven a pedir (crea una tabla con columnas para el ID del producto, el nombre del producto y la proporción en que se ha vuelto a comprar)?
¿Cuál es la proporción de productos pedidos que se vuelven a pedir para cada cliente?
¿Cuáles son los 20 principales artículos que la gente pone en sus carritos primero (muestra las identificaciones de los productos, los nombres de los productos y el número de veces que fueron el primer artículo añadido al carrito)?