Maestría en Management & Analytics

Programación para el Análisis de Datos – trabajo final

Consigna

El trabajo final consiste en aplicar los conocimientos adquiridos durante la materia para realizar un análisis exploratorio de un dataset.

En principio, pueden elegir entre dos dataset: un dataset sobre ventas y ganancias de un supermercado y un dataset sobre churn bancario. Alternativamente, pueden trabajar con un dataset de vuestra elección, en el caso que prefieran hacerlo. En ese caso, me tienen que informar sobre cuál dataset quieren usar, detallando sus características, para que les pueda dar el ok.

En el caso del dataset del supermercado, el análisis debe orientarse a identificar cuáles productos, regiones, categorías y segmentos son más o menos rentables, de modo tal de ofrecer recomendaciones al supermercado sobre dónde concentrar sus recursos.

En el caso de churn bancario, el análisis debe orientarse a identificar cuáles características hacen más probable que un cliente deje el banco.

Más allá de estas indicaciones, el trabajo es libre, en el sentido de que pueden encarar el análisis como mejor lo consideren conveniente. Tienen que aplicar las técnicas de manipulación, transformación, limpieza y visualización de datos vistos en clase para realizar un informe con insights sobre el dataset.

Recomendaciones:

- Realicen una descripción de las diferentes variables
- Analicen las relaciones entre las variables.
- Realizar diferentes agrupaciones de datos y realizar análisis sobre los datos agrupados.
- Crear columnas nuevas con variables que les permitan obtener información relevante sobre los datos.
- Apliquen las técnicas de visualización aplicadas en clase.
- Documenten y comenten los resultados obtenidos. Es importante que incluyan una interpretación de los hallazgos que realicen.

En el caso que trabajen con un dataset elegido por ustedes, deberán definir ustedes mismos los objetivos del análisis e informarlos de manera explícita en el trabajo. Asimismo, deberán seguir las recomendaciones enunciadas arriba.

Entregable

El entregable del TP será una notebook de Jupyter con el código en Python del análisis realizado y los comentarios y las conclusiones en las celdas en markdown. En el caso que hayan desarrollado código de soporte (por ejemplo un módulo con funciones definidas por ustedes), deberán adjuntarlo también. Los grupos que trabajen con datasets diferentes a los sugeridos, deberán entregar también el dataset.

Justifiquen los diferentes análisis que vayan realizando, explicando por qué los realizan y cuáles conclusiones les permiten obtener. En el caso que realicen transformaciones a los datos o desarrollen variables nuevas, también justifiquen el motivo. Al finalizar, realicen un análisis de cierre con las conclusiones principales de vuestro trabajo.

Asegúrense de que el código corra correctamente y no haya bugs, ya que se penalizará la entrega de código con errores. La notebook debe poderse ejecutar celda por celda sin que se generen errores, de modo tal de poder reproducir los resultados de sus análisis.

Fecha de entrega

La fecha límite de entrega será el lunes 1ro de mayo de 2023 a las 18:30 horas.

Información sobre los datasets

1) Supermercado

El dataset del caso del supermercado está almacenado en el file Superstore.csv.

Tip: para leerlo usar el encoding 'windows-1252'.

Las columnas del dataset son:

- Row ID: ID único para cada fila.
- Order ID: ID de pedido único para cada cliente.
- Order Date: fecha de pedido del producto.
- Ship Date: fecha de envío del producto.
- Ship Mode: modo de envío especificado por el cliente.
- Customer ID: ID único para identificar a cada cliente.
- Customer Name: nombre del cliente.
- Segment: segmento al que pertenece el cliente.
- Country: país de residencia del cliente.
- City: ciudad de residencia del cliente.
- State: Estado de residencia del cliente.
- Postal Code: código postal del cliente.
- Region: región a la que pertenece el cliente.
- Product ID: ID único del producto.
- Category: categoría del producto ordenado.
- Sub-Category: sub-categoría del producto ordenado.
- Product Name: nombre del product.
- Sales: valor en USD de las ventas del producto.
- Quantity: cantidad del producto vendidas
- Discount: descuento de la operación
- Profit: ganancia o pérdida de la venta

2) Churn bancario

El dataset del caso del supermercado está almacenado en el file Churn_Modelling.csv.

Las columnas del dataset son:

- RowNumber: número de fila
- CustomerId: ID del cliente
- Surname: apellido del cliente
- CreditScore: score de crédito
- Geography: país de residencia
- Gender: género
- Age: edad
- Tenure: años como cliente
- Balance: saldo en la cuenta
- NumOfProducts: cantidad de productos que el cliente tiene con el banco
- HasCrCard: si tiene una tarjeta de crédito (1= sí, 0 = no)
- IsActiveMember: si es un miembro activo (1= sí, 0 = no)
- EstimatedSalary: salario estimado
- Exited: si dejó el banco (1= sí, 0 = no)