INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

UF-6 ACTIVIDAD EVALUABLE: MAPAS DE CALOR Y BOXPLOTS

Equipo 3:

José María Colombres Elguea | A01734153

Santiago Gael Gallardo Parente | A01734347

Miguel Sebastián Carreto Bahena | A01735592

Análisis de datos y herramientas de inteligencia artificial II

(Grupo 101)

Profesores: Fabiola Diaz Nieto, Candy Yuridiana Alemán Muñoz & Francisco Javier Navarro Barrón

Fecha de entrega: 03 de junio de 2023

UF-6 ACTIVIDAD EVALUABLE: MAPAS DE CALOR Y BOXPLOTS

Durante el proyecto desarrollado para el socio formador, se ha trabajado con 4 archivos principales. Estos archivos ya han sido filtrados para eliminar valores nulos, cambiar tipos de datos e incluso reemplazar outliers, sobre todo para las operaciones de modelado. También se realizaron operaciones para crear nuevas columnas categóricas y nuevos data frames específicos. El presente reporte muestra un resumen sobre las variables que no fueron contempladas respondiendo a las preguntas base, además un análisis de los archivos finales (limpios) utilizando diagramas de cajas y bigotes, histogramas y mapas de calor.

Archivos originales

- ***Facturacion_Devoluciones_Credito_Clientes.xlsx:*** Cuenta con 4 hojas diferentes: Facturación: Muestra todos los registros de facturas del periodo 2019-2023 ***(df_facturacion).*** Devoluciones: Muestra los registros de devoluciones del periodo 2019-2023. ***(df_devoluciones).*** Notas de Crédito: Muestra los registros de NC del periodo 2019-2023 ***(df_notascredito).*** Clientes: Muestra el registro de los clientes con clave, nombre y RFC (***df_clientes)***
- ***Antigüedad de saldos 17.03.2023.xlsx:*** El archivo de excel solo cuenta con una hoja por lo que se generó un único Data Frame llamado ***df_ant_saldos***. En este se muestran diferentes características de compras realizadas durante 2021, 2022 y 2023. Algunas de estas son la fecha de la compra, el vencimiento de la deuda y el monto adeudado. Así como el nombre del cliente, número del cliente y número de la factura.
- ***Gastos y costos 20-23.xlsx:*** Con un total de 4 hojas, muestra el desglose de gastos y costos para el periodo 2020-2023, un año por hoja.
- ***Detalle precios y productos fabricados. xlsx (.csv):*** Al solo tener una hoja en el documento se generó un solo data frame llamado ***df_precios*** en el cual tenemos todas las facturas de las ventas, así como un desglose de ellas, obteniendo las columnas de Cantidad de artículos vendidos, la descripción de estos, el precio unitario, costo unitario y el margen de ganancia.

Preguntas base

¿Hay alguna variable que no aporte información?

- **Facturación:** En la hoja de Facturación del documento original se identificaron 2 variables que mostraban información duplicada que no era relevante, que fueron 'FECHA_DOC' y 'FECHA_ENT'. En las hojas de Devoluciones y Notas de Crédito se identificó a la columna 'SERIE' como una variable intrascendente.
- **Gastos y Costos:** En el documento de Gastos y Costos se logra mantener una consistencia de los datos en cuanto al formato del documento. La columna 'Otros' en las hojas del año 2022 y 2023 se encontraba vacía por lo que fue la única variable que no aportaba información relevante.
- **Antigüedad de saldos:** En un principio no se encontró ninguna variable irrelevante, sin embargo, al agregar una columna nueva de nombre 'DÍAS_ADEUDO' se eliminaron dos columnas de fecha que duplicaban información.
- **Detalle de precios:** La variable 'Nombre 1' no fue de utilidad para el análisis ya que ya se tienen variables como nombre del cliente, nombre del vendedor y clave del cliente, además del nombre del producto.

Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

- **Facturación:** Durante las sesiones con el socio formador, el director Francisco Nogueras indicó que las columnas de valor para el análisis en la hoja de Facturación son 'FECHA_ELAB' y 'FECHA_VEN'. Por ello se decidió eliminar las columnas de 'FECHA_DOC' y 'FECHA_ENT'. Para la hoja Devoluciones se eliminaron las columnas de 'FECHA_DOC' y 'FECHA_ENT', además de la columna de 'SERIE' que tenía un mismo valor. También se eliminó la columna 'DES_FIN' que tenía todos los valores en 0. En la hoja de Notas de crédito se decidió eliminar las columnas de 'FECHA_DOC' y 'FECHA_ENT', además de la columna de 'SERIE' que tenía un mismo valor y en Clientes se conservaron las 3 columnas**.**
- **Gastos y Costos:** En la hoja del año 2020 se incluye una última columna que presenta el resultado de restar 'TOTALSAT' - 'TOTALMX'. Sin embargo a razón de que en las otras pestañas se omitió; dicha columna se eliminó definitivamente. Por otra parte, se identificó que para la hoja de 2022 y 2023 se agregó una columna que está en su totalidad vacía, por lo que dentro de dichas hojas se eliminó la columna "Otros" pues no aportaba nada al proyecto.

- **Antigüedad de saldos:** Las variables eliminadas fueron 'FECHA_FACTURA' y 'FECHA_VENCIMIENTO'. Estas son irrelevantes debido a que se agregó la variable "DIAS_ADEUDO". Esta columna resume las dos anteriores y muestra más claramente las compras que tuvieron financiamiento. Simplificando el entendimiento de las columnas agregando la duración del adeudo en días.
- ***Detalle de precios:*** En el archivo de Detalle de precios y productos se decidió eliminar la columna de 'Nombre 1' ya que esta no era de utilidad para el análisis. También se eliminó la descripción del artículo ya que es una columna que no aportaba datos significativos y se puede identificar el artículo con la columna de clave artículo.

¿Existen variables que tengan datos extraños?

- **Facturación:** En ninguna hoja se encontraron valores atípicos, y aquellas notas de facturación elevadas se revisaron de forma manual. En la hoja de Facturación se pudo observar que muy pocas compras están por encima de los 50000, lo que puede significar que la empresa prefiere facturar por volumen.
- **Gastos y Costos:** No se encontraron valores atípicos**.** Al tratarse de un documento que muestra los montos de las salidas de efectivo netas, no se recomienda la edición/modificación de estos debido a que no se trata de variables indicativas de tendencias sino montos totalmente representativos, que al modificarse podrían afectar los cálculos a futuro.
- **Antigüedad de saldos:** Se pudo observar que existen dos datos atípicos utilizando el promedio de las desviaciones estándar de la variable 'MONTO_ADEUDADO'. En este caso se analizaron específicamente y se decidió tomarla en consideración al ser un cliente importante para Calor y Control.
- **Detalle de precios:** No se identificaron variables con valores atípicos

Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?

- **Facturación:** Las variables cualitativas se encuentran en montos similares y a pesar de que hay una distancia amplia entre los valores más altos y bajos, a partir de la clasificación de cuartiles (por ejemplo) se logró un mejor análisis.
- **Gastos y Costos:** Los rangos no afectaron el análisis del documento.
- **Antigüedad de saldos:** En general los rangos son similares sobre todo si se comparan con las hojas de facturación, debido a que la empresa seguramente mantiene un tope límite para el crédito.

- **Detalle de precios:** Este archivo muestra el desglose de los precios por operación, por lo que siempre se mantendrá una constante en la diferencia entre el costo para la empresa y el precio final al cliente, que es el porcentaje de ganancia por producto u operación.

¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?

- **Facturación:** Estos datos están interconectados. Las variables Clave de cliente y el RFC están presentes en las 4 hojas del archivo. Al renombrar las columnas se estandarizó como clave 'CLIENTE'. Esto fue de gran utilidad para analizar las cifras y vincularlas con el nombre del cliente al que pertenece la clave. La Fecha de documento (FECHA_DOC) está presente en las hojas de cálculo de Facturación, Devoluciones, Notas de Crédito. Esta clave indica la fecha en que se realizó la factura o pedido y es la hoja de facturación la que tiene los datos "madre". La Clave de documento (CVE_DOC) está presente en las hojas de cálculo de Facturación y Devoluciones.
- **Antigüedad de saldos:** La Clave de cliente está presente en la hoja del archivo de excel, mientras que también se encuentra en el archivo de Facturación. La variable se puede observar cómo "CVE_CLPV", permitiendo un posible merge entre esos archivos. La viabilidad de esta combinación debe de ser analizada para justificar su utilidad.
- **Detalle de precios y productos fabricados:** La clave del documento; CVE_DOC se encuentra tanto en el archivo de detalle de precios y en el de facturación. La columna de nombre se encuentra en el documento antigüedad de saldos y en el de detalle de precios y la columna de FECHA_DOC se encuentra en el documento de facturación y en el de detalle de precios.

Análisis a través de gráficos

Se adjunta el link del repositorio donde se encuentra el archivo de código

- **Link de Github:** https://github.com/josemacoel/UF6_Equipo3.git