INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

CAMPUS QUERÉTARO

Analítica de Datos y Herramientas de Inteligencia Artificial II

UF 3

Limpieza de Bases de Datos

Natalia María Ovando Flores A01368118

Andrea Cosset Hernández Lugo A01707744

Lucia Castañeda Medeguin A01706644

Nicole Aryam Rodríguez A00831569

21 de Abril del 2023

Datos de Facturación

Después de leer el archivo nos dimos cuenta que las columnas con datos nulos fueron las de clave vendedor (con 48 datos), fecha de entrada (con 2 datos) y la fecha de cancelación (con 10,537 datos). La columna de fecha de cancelación tiene más datos nulos, indicando que estas facturas no están canceladas.

Para la clave vendedor, la sustituimos por 0 porque no existe el registro de quién realizó esas venta y con el propósito de que se tengan datos reales de quién realizó cada venta se puso un 0. Asignándoles a “otro vendedor”.

Para la columna de fecha de entrada, nos dimos cuenta que está en orden cronológico por lo que, decidimos sustituir con "forward fill", ya que, aunque no sea el dato exacto es un buen aproximado y de igual manera, no afectará a nuestro análisis en el futuro.

Finalmente, en la columna de fecha de facturación, hay valores nulos debido a que son facturas que no fueron canceladas, es decir si fueron efectivas, por lo que sustituimos por un "0", para que si haya un registro, pero tomando en cuenta que esa es la razón por la que son datos nulos.

Detalle de precios y productos

Este archivo únicamente tiene datos nulos en el nombre del vendedor. Decidimos rellenar los valores nulos con la moda, es decir el vendedor más repetido (LETICIA RAMÍREZ HERNÁNDEZ). Por la lógica de cómo es la vendedora con más ventas, la probabilidad de que ella haya hecho esas ventas son mucho mayores. Además, no consideramos que sea un monto realmente considerable para que impacte con mucha diferencia en otro vendedor.

Gastos y Costos 20-23

Primeramente se decidió ir limpiando hoja por hoja, es decir año por año.

Para el año 2020 obtuvimos datos nulos en las columnas:

- Folio: el cuál se sustituyó agregando un folio 0 para todos los datos nulos. Dando a entender que son datos con un folio faltante y posteriormente se deberá analizar uno por uno para determinar un por qué, pero se clasifican desde ahora.
- Gasto: Igualmente se rellenó con un 0, ya que se intuye que al no tener un monto, esta factura no representó un gasto para la empresa.
- Tipo de Cambio: Se indica que se desea rellenar con un tipo de cambio promedio del tipo de cambio usado. Es decir, alrededor de 18 pesos.
- Importe: Rellenamos con una mediana para mantener la estacionalidad de los datos de importe.
- IVA: Se rellenó con un promedio para evitar la dispersión de los datos.
- Tipo: Se rellenó con la letra I, ya que nos dimos cuenta que las facturas que tienen esta letra, tienen contenido en la descripción, al igual que las facturas sin datos en esta columna.
- Póliza: Al tener dato string, simplemente agregamos la leyenda “SIN POLIZA”.

Para la columna 2021 obtuvimos datos nulos en las siguientes columnas:

- Folio: Se utilizó la misma lógica que para 2020.
- MP: Tiene el dato string de PUE y PDD, significando pago en una sola exhibición. Inferimos todos los faltantes, fueron hechos en una sola exhibición, para facilidad del análisis.
- Póliza: Se utilizó la misma lógica que para 2020.

Para la columna 2022 obtuvimos datos nulos en las siguientes columnas:

- Folio: Se utilizó la misma lógica que para 2020.
- MP: Se utilizó la misma lógica que para 2021.
- TC: Se utilizó la misma lógica que para 2020.
- Otros: Solo tenemos celdas vacías. Lo rellenamos con un 0 para tener un dato que indique ausencia de algo.
- Póliza: Se utilizó la misma lógica que para 2020.

Finalmente, para la columna 2023 obtuvimos datos nulos en las siguientes columnas:

- Folio: Se utilizó la misma lógica que para 2020.
- MP: Se utilizó la misma lógica que para 2021.
- FP: Los datos de esta columna son numéricos, así que decidimos rellenarlos con la mediana de los datos para mantener la estacionalidad de los datos.
- Otros: Se utilizó la misma lógica que para 2022.
- Póliza: Se utilizó la misma lógica que para 2020.