



Tecnológico de Monterrey

**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY
CAMPUS QUERÉTARO**

Actividad 2.2 (Valores Nulos)

Unidad de formación.

Analítica de datos y Herramientas de

Inteligencia Artificial II

Grupo 101

Integrantes.

Carolina Solis Flores A01708072

Maria Fernanda Martinez Ríos A01067198

Renata Pilar Gómez Castillo A01351806

Jose Ignacio Hernández Rodríguez A01703130

Profesor.

PhD. Alfredo García Suárez

Abril 2023

De acuerdo a Nora Casillas y Francisco Noguera, la empresa busca la realización de un **“Reporte de control operativo”** en el que pueda visualizar semana a semana el estado financiero operativo de la empresa y el flujo de efectivo operativo. Esto con el afán de conocer que tan rentable son las operaciones de la empresa. Para poder realizar el reporte de control operativo se debe llevar a cabo una serie de pasos, es por esto que durante esta actividad se realizará la sustitución o eliminación de datos nulos.

Los archivos utilizados para la realización de esta actividad fueron los siguientes:

1. FACTURACIÓN, DEVOLUCIONES, NOTAS DE CRÉDITO, CLIENTES. xlsx
2. Detalle de precios y productos fabricados 2022. xlsx
3. Gastos y costos 20-23. xlsx

Para una mejor visualización, estos archivos fueron convertidos en nueva dataframes.

1. clientes
2. devoluciones
3. facturacion
4. notas_credito
5. precios
6. gastos_20
7. gastos_21
8. gastos_22
9. gastos_23

Antes de comenzar con el reemplazo de datos dentro de los data frames, se realizaron cambios a algunos nombres de las columnas para un mejor entendimiento del contenido.

Para los data frames de “facturación” “notas de crédito” y “devoluciones” se realizaron tres modificaciones.

- FECHAELAB → FECHA_ELAB
- FECHA_ENT → FECHA_ENTREGA
- CAN_TOT → CANTD_TOT

Para los data frames de “Gastos y Costos”, se realizaron siete modificaciones:

- Fecha → FECHA
- Folio → FOLIO
- Proveedor → PROVEEDOR
- Descripción → DESCRIPCION
- Status → STATUS
- Tipo → TIPO
- Poliza → POLIZA

Mientras que para el df de antigüedad de saldos, solo se realizó un cambio el cual fue:

- No. CLIENTE → CVE_CLPV

Después de realizar estos cambios, se continuó con la preparación de los datos. Para este paso decidimos hacer una profunda limpieza y detallada selección de los datos que sean más útiles para cumplir con el objetivo del mismo, ya que por lo general, las bases de datos contienen diversos errores que puede cometer el mismo programa que se encarga de la recolección de los datos, así como de los usuarios que ingresan dicha información. Los errores encontrados más comunes suelen estar relacionados con la existencia de datos duplicados, valores nulos, valores atípicos, así como índices que impiden un correcto análisis de la información. Por lo tanto, es muy importante que previamente a la manipulación de los datos, se complete la etapa de selección y limpieza. En este caso, nos centraremos en el reemplazo o eliminación de datos nulos.

Detalle de precios y Productos fabricados

Dentro del data frame de precios, se encontró que no existe ningún dato nulo por lo que no se realizó ningún cambio ni eliminación de datos.

FACTURACIÓN, DEVOLUCIONES, NOTAS DE CRÉDITO, CLIENTES.

Hablando del data frame **facturación**, se encontraron datos nulos en las columnas de CVE_VEND, FECHA_ENTREGA, FECHA_CANCELA. Para estas columnas se tomó en consideración que los valores son de tipo datetime64 y float64. Se decidió que los reemplazos realizarán el cuarto método de sustitución de valores nulos, donde se sustituyen los datos nulos por un string en concreto.



Para las columnas que contienen columnas vacías se decidió rellenarlas con "--", se tomó esta decisión pues los valores dentro de estas columnas son fechas y el número del vendedor que realizó la venta, que en este caso no pueden ser reemplazadas con otro método de sustitución ya que no son datos a los cuales se les pueda realizar un cálculo de promedio o media, o rellenar con los datos de abajo o arriba, esto debido a que son datos específicos, como por ejemplo, dentro de FECHA_CANCELA, solo se coloca la fecha de los pedidos cancelados, por lo que rellenar las celdas vacías con alguna otra fecha podría afectar el resultado final del control operativo. Asimismo, los datos de las columnas CVE_VEND y FECHA_ENTREGA, sólo podrían ser reemplazados con datos reales (no utilizar "--") si se tuviera un acercamiento con la persona que recopiló los datos, en este caso, cuando se tuvo un acercamiento que se tuvo con Nora Casillas, nos mencionó que CVE_VEND no era tan importante, por lo que rellenar los datos no es primordial.

Continuando con el dataframe de **devoluciones**, se encontró que las columnas que cuentan con datos nulos son DOC_ANT, CVE_PEDI, FECHA_CANCELA, CVE_VEND. Al igual que el dataframe de facturación, los datos nulos fueron reemplazados por "--". Esto debido a que son datos únicos e irremplazables. Ejemplificando, dentro de DOC_ANT se encuentra un número que permite identificar el documento del pedido que fue realizado, este número es diferente para cada pedido.

Pasando al data frame de **clientes**, se encontraron datos nulos dentro de las columnas RFC y nombre. Estas celdas vacías fueron sustituidas con el método número cuatro, donde se sustituyen los datos nulos por un string en concreto, en este caso por "--". Esto debido a que en el caso del RFC, esta es una clave única de registro que sirve para identificar a toda aquella persona que realiza una actividad económica, si se utilizará una RFC diferente podrían existir problemas legales, pues se estaría diciendo que esa clave única está realizando una compra que no ha realizado.

En cuanto a el data frame de **notas de crédito**, se tienen datos nulos dentro de las columnas de CVE_PEDI, FECHA_CANCELA y CVE_VEND. Para estos datos también se utilizó el método cuatro, donde se sustituye por valores concretos. Donde no hay valores, se tomó la decisión de reemplazarlo por "--". Esto debido a que dentro FECHA_CANCELA, solo se ingresan los pedidos que son cancelados, por lo que rellenar con fechas todos los datos



faltantes podría dar a entender que todos los pedidos han sido anulados. Como se mencionó anteriormente, la clave del vendedor no es primordial, pero reemplazarlo con algún número podría afectar, pues se infiere que un vendedor realizó la compra cuando en realidad fue realizada por otra persona. Y por último, la clave del pedido, es única para cada pedido a realizar, pues permite un seguimiento de este, por lo que reemplazarlo por un fill forward o un fill forward podría afectar a el correcto seguimiento de la venta.

Gastos y costos 20-23.

Hablando del data frame gastos se encontraron datos nulos en las siguientes columnas: Folio, Gasto, TC , Póliza, Importe, IVA y Tipo. Para estas columnas se tomó en consideración que los valores son de tipo float64 y object. Se decidió que los reemplazos realizarán el cuarto método de sustitución de valores nulos, donde se sustituyen los datos nulos por un string en concreto o un valor.


Para la columna de **Folio**, se determinó sustituir los valores nulos por "--". Esto se debe a que esta columna hace referencia al folio de operación, por lo que si la modificación se realizara por otro valor. Esto modificaría el sentido de la información.

Para la columna **Gasto** se determinó sustituir los valores nulos por "0". Esto se debe a que esta columna indica el total del gasto, por lo que al no contar con la información, determinamos que debería utilizarse un "0". Ya que, en caso de utilizar otro valor o promedio, no se estaría contando con información del todo fiable ni verídica.

Para la columna **TC**, se determinó que los valores nulos se sustituyeran por "--". Esto se debe a que esta columna hace referencia a la forma de pago de la transacción, por lo que si la modificación se realizara por otro valor. Esto modificaría el sentido de la información.

Para la columna **Póliza**, se determinó que los valores nulos se sustituyeran por "--". Esto se debe a que esta columna hace referencia a la póliza de la operación por lo que si la modificación se realizara por otro valor. Esto modificaría el sentido de la información.





Para la columna **Tipo**, se determinó que los valores nulos se sustituyeran por “I”. Esto se debe a que esta columna hace referencia al tipo de la operación, por lo que si la modificación se realizara por otro valor. Esto modificaría el sentido de la información.

En cuanto a los valores de la columna **IVA** ahí se hizo una modificación de otro tipo, la cuál consistió en sustituir los valores nulos por la mediana del total de los valores de la columna. Esto se realizó de esta forma, con la finalidad de tener un dato numérico en concreto que tenga relación directa con la información de la columna.

Hablando de la columna “Otros”, todos los datos de esta son valores nulos por lo que vamos a eliminar esa columna debido a que no contienen información con la que pudiéramos reemplazar, sacarle media/promedio o escribirla manualmente, puesto que no sabemos a que se refiere.

Cabe mencionar que, este proceso se realizó para el periodo de tiempo 2020-2023. Por lo que se realizó varias veces, a pesar de esto, se identificaron los valores nulos en las mismas columnas.