INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

REPORTE DATOS NULOS ACTIVIDAD 2.2

Unidad de formación

Analítica de datos y herramientas de inteligencia artificial II (Gpo 501)

Profesores

Alfredo García Suárez

Candy Yuridiana Alemán Muñoz

Francisco Javier Navarro Barrón

Fabiola Díaz Nieto

Equipo 8

Ruth Maya López A01707467

Leonardo Isaías Guevara Pioquinto A01275327

Valeria Medina Martinez A01275070

Fecha de entrega

Abril 22, 2023

DETALLE PRECIOS Y PRODUCTOS FABRICADOS 2022

En la base de datos “Detalle precios y productos fabricados 2022” iniciamos creando un DataFrame con el nombre “productos” para tener un mejor manejo de los datos. Posteriormente para tener una mejor idea de los datos, aplicamos un .info(), con el cual observamos el tipo de dato que contiene este DataFrame.

Después de realizar la exploración de los datos, sumamos los nulos de cada columna para poder ver si existían algunos datos los cuales debemos de sustituir o eliminar.

Realizando la función observamos que existían dos nulos para la columna “NOMBRE VENDEDOR”. Al ser un dato único (no sustituible) decidimos colocar “--” para no sustituir con datos que no tuvieran sentido y no modificar los resultados que podríamos obtener posteriormente.

Al aplicar esta función verificamos cuántos nulos había todavía y ya todas las columnas mostraban cero.

De igual manera observamos que existían columnas que no eran relevantes dentro de esa base de datos y que eliminarlas nos harían tener una base de datos más limpia y un mejor manejo de datos, por lo que eliminamos las columnas: “DESCR”, “COSTO_UNITARIO”,“SUBTOTAL_PARTIDA”,
“MARGEN_UNITARIO_CALCULADO” y “NOMBRE_VENDEDOR”.
Finalmente guardamos los cambios en un csv nuevo para poder usarlo posteriormente.

DATOS DE FACTURACIÓN

Analizando a detalle la base de datos denominada “Datos de facturación” del socio formador *Calor y Control* logramos identificar valores nulos gracias al desarrollo de un código de programación en el que con ayuda de diferentes librerías logramos la limpieza de este primer archivo. Dentro de estas librerías se encuentran pandas, la cual es una biblioteca que nos permite el análisis y manipulación de bases de datos. Es fundamental esta librería para la parte de exploración de datos en el proceso de “Análitica de Datos” ya que nos permite generar en python estructuras visiblemente más atractivas de los datos tales como ‘DataFrames’ o ‘Series’. Posterior a la importación de las librerías, analizamos valores nulos del DataFrame, tanto en columna como en conjunto global y pudimos identificar que existían “10,587” valores nulos. Gracias a la herramienta de info observamos que estas instancias eran

de tipo object y datetime. Tres columnas específicamente contenían estos valores y pertenecían a “CVE_VEND” (Clave del vendedor), “FECHA_ENT” y “FECHA_CANCELA”. Cabe destacar que son valores únicos (cualitativos) que no pueden ser sustituidos por métodos como el de la media o la mediana, por lo tanto se optó por el método de sustitución de valores nulos por un string en concreto, ya que este método de imputación reemplaza todos los valores nulos en las columnas seleccionadas con un string (“--”) especificado. En este caso la columna “CVE_VEND” contiene códigos de identificación de los vendedores y los valores nulos pueden representar situaciones únicas o que desconocemos de la base de datos original, por ende, infringir en la manipulación a nuestro propio criterio podría influir negativamente en la información de la empresa. De igual forma con las fechas específicas, en el caso “FECHA_CANCELA” describe si una factura fue cancelada, al no tener un valor datetime indica que no fue cancelada por lo tanto puede ser reemplazada por doble guión y las misma lógica se empleó para “FECHA_ENT”.

A manera de conclusión, el método de sustitución por un string específico es una técnica eficaz y común en la limpieza y procesamiento de datos, especialmente cuando los Nan son únicos o no pueden ser reemplazados por otras técnicas. En el caso específico de esta base de datos la aplicación de este método nos permitió mantener la integridad de los datos en el conjunto.

GASTOS Y COSTOS 20-23

2020

Para este año, se observa que existen varios valores nulos pertenecientes a este DataFrame, de forma general observamos que los valores de las variables “FOLIO” que mostró 189 valores nulos, “IMPORTE” que mostró 34 valores nulos e“IVA” que mostró 268 valores nulos, fueron sustituidas por “0” de tal forma que gracias a que estos valores son numéricos y se podrían considerar únicos, no es posible aplicar operaciones mediante métricas como la mediana o la media para asignarles nuevos valores. Por esta razón se decidió mantener las cantidades con este valor que no representa alguna cantidad y poder mantener el formato de valores para los registros y operaciones futuras.

Por otro lado, para los valores nulos de las columnas “GASTO” que mostró 2502 valores nulos, “TC” que mostró 391 valores nulos, “TIPO” que mostró 1 valor nulo y “POLIZA” que mostró 3321 valores nulos, se realizó el reemplazo por “- -” de tal forma que

dichos valores al ser categóricos o que integran claves, no es posible rellenar con alguna abreviación o valor en específico, porque se alteraría la secuencia de valores y no se puede confirmar el comportamiento de ciertos registros.

2021

Después de revisar los registros incluidos en el documento, se encontraron algunos datos nulos que fueron reemplazados por otros valores con el fin de mejorar la interpretación y evitar la inclusión de información innecesaria en el análisis posterior. Se identificaron 147 valores nulos en la columna "FOLIO", los cuales fueron sustituidos por el valor "0", ya que esta columna muestra un número único que no puede ser reemplazado por el promedio u otra métrica. Si se hubiera utilizado otra estrategia, los valores no coincidirían con los registros y las entradas del resto del documento.

Por otra parte, se aplicó la estrategia de reemplazar los valores nulos en las columnas "MP" y "POLIZA" con la cadena "--", ya que estos valores no son numéricos y cualquier otro reemplazo afectaría la interpretación de los resultados. No se cuenta con información acerca del origen de estos registros y tratar de interpretar información no confirmada afectaría en los resultados. Se identificaron 654 valores nulos en la columna "MP" y 2372 en la columna "POLIZA".

En conclusión, se aplicó una estrategia específica para el reemplazo de los valores nulos en función de las características de cada columna y de los posibles efectos que pudieran tener en la interpretación de los resultados. Con estas medidas, se logró mejorar la calidad y confiabilidad del análisis de los datos.

2022

Tomando la misma base de datos pero ahora utilizando los datos del 2022 , observamos que se encuentran datos nulos, mismos que reemplazamos por distintos valores para tener un mejor manejo de los datos y que la base de datos se mantenga limpio.

Primeramente se obtuvieron *102* valores nulos en la columna “*Folio*” los cuales fueron reemplazados por “*0*”, ya que como se mencionó anteriormente este valor es único e irremplazable. Misma situación que ocurre con la columna “*Otros*” la cual presentó un total de *2577* registros nulos.

Por otro lado, para las columnas "*MP*" que mostró *553* valores nulos, "*TC*" que mostró *636* valores nulos, y "*Póliza*" que mostró *801* valores nulos, se sustituyó por "- -" debido a que no son valores numéricos y por lo tanto para que tuviera sentido se reemplazó por una que no afecta a la base de datos y ayuda a tener un mejor entendimiento y manejo de la información.

2023

Tomando en cuenta los registros pertenecientes a este documento, podemos encontrar que conservan algunos datos nulos, mismos que reemplazamos por distintos valores para mejorar la interpretación y evitar contar con información no necesaria para el posterior análisis.

Primeramente se obtuvieron *13* valores nulos en la columna "*Folio*" los cuales fueron reemplazados por "*0*", debido a que este valor muestra un número único que muy difícilmente puede ser reemplazado por el promedio o alguna otra métrica, pues si se llegara a aplicar los valores no podrían coincidir con los registros y entradas del resto del documento. Misma situación que ocurre con la columna "*Otros*" la cual presentó un total de *397* registros nulos.

Por otro lado, para las columnas "*Tipo de gasto*" que mostró *8* valores nulos, "*MP*" que mostró *71* valores nulos, "*FP*" que mostró *71* valores nulos y "*Póliza*" que mostró *115* valores nulos, se aplicó el reemplazo de estos mediante "- -" debido a que estos valores no son numéricos y por lo tanto, el reemplazo de estos por alguna otra categoría o abreviación afectará en el desarrollo y la toma de decisiones posteriores, ya que se desconocen los orígenes de estos registros e interpretar información que no es confirmada, afectará en los resultados.