INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

ACTIVIDAD 2.1 (VALORES NULOS) Y ACTIVIDAD 2.2 (VALORES NULOS)

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 I

(Grupo 101)

Profesores: Candy Yuridiana Alemán Muñoz, Alfredo García Suárez & Francisco Javier Navarro Barrón

Fecha de entrega: 22 de abril de 2023

ACTIVIDAD 2.1 (VALORES NULOS) Y ACTIVIDAD 2.2 (VALORES NULOS)

El presente reporte muestra un análisis detallado del procesamiento de datos nulos para los archivos de: Datos de Facturación.xlsx (actualizado), Detalle precios y productos fabricados 2022.xslx y Gastos y costos 20-23. Se aplicaron diversos métodos de reemplazamiento para cada columna de los datasets y data frames, con el objetivo de generar 3 archivos listos para su posterior análisis.

Archivo 1:

Facturación, devoluciones, notas de crédito & clientes (Facturacion_D_NC_C.xlsx)

Este archivo de Excel contiene 4 hojas de cálculo, por lo que se dividió en 4 Dataframes diferentes con el objetivo de procesar los datos nulos.

Facturación (df_facturacion)

La primera operación que se realizó para la hoja de FACTURACIÓN fue eliminar las columnas de FECHA_DOC y FECHA_ENT. En la sesión con el socio formador, Francisco (director de Calor & Control) indicó que las columnas de valor para el análisis en la hoja de facturación son: FECHA_ELAB y FECHA_VEN, por lo que se tomó la decisión de borrarlas. De esta forma, la identificación de valores nulos fue la siguiente:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10961 entries, 0 to 10960
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   CVE_DOC        10961 non-null  object
 1   CLIENTE        10961 non-null  object
 2   STATUS         10961 non-null  object
 3   VENDEDOR       10913 non-null  float64
 4   FECHA_ELAB     10961 non-null  datetime64[ns]
 5   FECHA_VEN      10961 non-null  datetime64[ns]
 6   FECHA_CANCELA  358 non-null    datetime64[ns]
 7   CAN_TOT        10961 non-null  float64
 8   DES_TOT        10961 non-null  float64
 9   TOTAL          10961 non-null  float64
 10  RFC            10961 non-null  object
dtypes: datetime64[ns](3), float64(4), object(4)
memory usage: 942.1+ KB
```

Como se observa en la imagen anterior, la función info muestra que las columnas con valores nulos con la de VENDEDOR y FECHA_CANCELA. Para la columna de VENDEDOR se decidió asignar el valor de 0 a las filas nulas, pues las claves van del 1 al 12 y la clave 0 es viable para identificar operaciones con vendedor no identificado sin cambiar el tipo de dato a la columna.

```
3   VENDEDOR       10961 non-null  float64
```

Por otro lado, la columna FECHA_CANCELA indica la fecha en que se realizó la cancelación de una factura, pero tiene muchos valores nulos. Ya que contiene valores importantes se decidió crear un nuevo data frame filtrado solo con las que tienen valores no nulos y se borró del df original

```
df_facturacionfiltro = df_facturacion[df_facturacion['FECHA_CANCELA'].notnull()]
```

```
df_facturacion = df_facturacion.drop("FECHA_CANCELA",axis=1)
df_facturacion
```

Devoluciones (df_devoluciones)

Como primer paso dentro de esta hoja, se decidió eliminar las columnas de FECHA_DOC y FECHA_ENT, siguiendo la lógica indicada por el socio formador para el archivo de facturación. Posteriormente se eliminaron las columnas de SERIE, que tenía un mismo valor para todas las columnas y la de DES_FIN que tenía todos los valores en 0. A partir de ello, se identificaron las columnas con valores nulos:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194 entries, 0 to 193
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   TIP_OP         194 non-null    object
 1   CVE_DEV        194 non-null    object
 2   CLIENTE        194 non-null    object
 3   STATUS         194 non-null    object
 4   VENDEDOR       191 non-null    float64
 5   CVE_PEDI       188 non-null    object
 6   FECHA_VEN      194 non-null    datetime64[ns]
 7   FECHA_CANCELA  7 non-null      datetime64[ns]
 8   CAN_TOT        194 non-null    float64
 9   FECHA_ELAB     194 non-null    datetime64[ns]
 10  RFC            194 non-null    object
 11  FOLIO          194 non-null    int64
 12  CVE_DOC        183 non-null    object
dtypes: datetime64[ns](3), float64(2), int64(1), object(7)
memory usage: 19.8+ KB
```

Siguiendo el paso en el df_facturacion, se decidió asignar el valor de 0 a las filas nulas de la columna VENDEDOR, pues las claves van del 1 al 12 y la clave 0 es viable para identificar operaciones con vendedor no identificado sin cambiar el tipo de dato a la columna.

```
df_devoluciones['VENDEDOR'] = df_devoluciones['VENDEDOR'].fillna(0)
```

Para la columna CVE_PEDI, se decidió realizar un proceso más complejo. Como la columna CVE_PEDI y CVE_DOC son similares, se decidió que antes de colocar una clave específica los registros nulos de la columna CVE_PEDI el código busque en la columna de CVE_DOC el formato de clave del pedido que es F12345 (F seguido de de 5 números). Este proceso también se realizó a la inversa ya que ambas tenían valores nulos pero no necesariamente en la misma fila. Una vez realizado este proceso, la información de las columnas quedo asi:

```
5   CVE_PEDI        194 non-null    object
6   FECHA_VEN       194 non-null    datetime64[ns]
7   FECHA_CANCELA   7 non-null      datetime64[ns]
8   CAN_TOT         194 non-null    float64
9   FECHA_ELAB      194 non-null    datetime64[ns]
10  RFC             194 non-null    object
11  FOLIO           194 non-null    int64
12  CVE_DOC         185 non-null    object
```

Para llenar los valores nulos de la columna CVE_DOC se decidió utilizar una clave genérica que identifique a aquellos elementos que no cuentan con una clave, que fue SI0000 (Sin identificar seguido de cuatro 0, para respetar el formato alfanumérico de 6 dígitos). Posteriormente se eliminó la columna FECHA_CANCELA, no sin antes hacer un segundo dataframe con el filtro de aquellas filas que si tuvieran valores en esa columna. Tanto el filtro de facturación como el de devoluciones se guardaron en el Excel limpio de nulos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194 entries, 0 to 193
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   TIP_OP      194 non-null    object
 1   CVE_DEV     194 non-null    object
 2   CLIENTE     194 non-null    object
 3   STATUS      194 non-null    object
 4   VENDEDOR    194 non-null    float64
 5   CVE_PEDI    194 non-null    object
 6   FECHA_VEN   194 non-null    datetime64[ns]
 7   CAN_TOT     194 non-null    float64
 8   FECHA_ELAB  194 non-null    datetime64[ns]
 9   RFC         194 non-null    object
 10  FOLIO       194 non-null    int64
 11  CVE_DOC     194 non-null    object
dtypes: datetime64[ns](2), float64(2), int64(1), object(7)
memory usage: 18.3+ KB
```

Notas de crédito (df_notascredito)

Este documento cuenta con 497 valores nulos en la columna FECHA_CANCELA ya que estas compras no fueron canceladas y por lo tanto están vacías. Esta columna es de gran relevancia por lo que se guardaron las instancias con un valor en esa columna en un nuevo data frame (filtro_cancelaciones) para despues borrar la columna del df original. Cabe mencionar que el filtro realizado se agregó al excel en una hoja nueva llamada NOTAS DE CREDITO-DEVOLUCIONES. La columna VENDEDOR contaba con 10 valores nulos ya que estaban vacíos y cada número correspondía a un vendedor. Para no perder valores relevantes se reemplazaron los espacios vacíos por el número 0. Por último, la columna CVE_PEDI contaba con 11 valores nulos, similarmente se utilizó el método de reemplazo para cambiar esos valores en blanco por una clave genérica SI0000 (sin identificar) para mantener los registros en el documento.

Clientes (df_clientes)

El data frame contaba con 15 valores nulos en la columna RFC y uno en la columna NOMBRE. El valor vacío en NOMBRE corresponde a las ventas realizadas en el mostrador (por lo que no es un cliente en específico) por lo que se agregó el nombre MOSTRADOR. En cuanto a los valores nulos en RFC son datos que están incompletos para algunos clientes por lo que se agregó un RFC genérico (000000000000) con la función de reemplazo.

Archivo 2:

Detalle precios y productos fabricados 2022.xlsx

Este archivo contiene tan solo una hoja por lo que se convirtió en un data frame. En este caso el archivo solo contaba con dos valores nulos en la columna NOMBRE_VENDEDOR, ya que dos transacciones no contaban con un nombre. Se consideró eliminar las dos instancias ya que al ser un número reducido no afectan mucho al análisis general. Pero al contar con información completa se optó por cambiar esos valores nulos por el nombre VENDEDOR ANONIMO para de esta manera no perder ningún dato.

```
CVE_DOC                       0
FECHA_DOC                     0
NOMBRE_VENDEDOR               2
NOMBRE_CLIENTE                0
CANT                          0
CVE_ART                       0
DESCR                         0
PRECIO_UNITARIO               0
COSTO_UNITARIO                0
COSTO_UNITARIO_CALCULADO      0
SUBTOTAL_PARTIDA              0
COSTO_TOTAL_CALCULADO         0
MARGEN_UNITARIO_CALCULADO     0
MARGEN_TOTAL_CALCULADO        0
```

Archivo 3:

Gastos y costos 2020-2023.xlsx

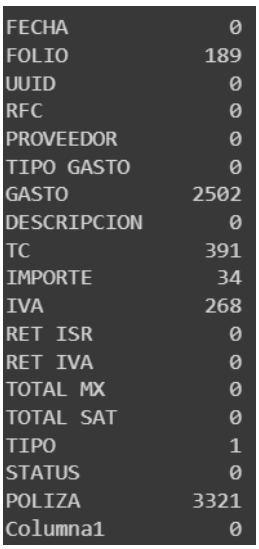
El archivo de excel muestra un desglose de todas las operaciones de facturación y los conceptos que se registran dentro de cada una de ellas. La mayoría de las columnas guardan valores numéricos como cantidades monetarias, tipo de cambio, pero también existen unas cuantas que almacenan conceptos como el registro de municipio, concepto de costo, etc.

Al analizar los datos correspondientes se encontraron los siguientes datos sobre datos nulos en cada hoja del archivo:

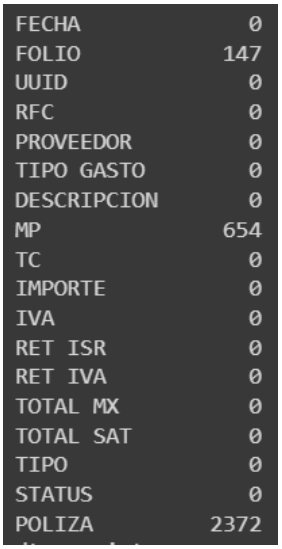
- ☐ 2020: 7 columnas que contaban con la presencia de datos nulos.
- ☐ 2021: 3 columnas que contaban con la presencia de datos nulos.
- ☐ 2022: 5 columnas que contaban con la presencia de datos nulos.
- ☐ 2023: 6 columnas que contaban con la presencia de datos nulos.

El proceso que se siguió para reemplazarlos por datos que no interfieran con la extracción de la información se realizó de la siguiente manera: para las columnas que almacenan datos numéricos las sustituciones se realizaron por valores “0” cero, siempre y cuando no interfirieran con operaciones o cálculos propios de la columna. Para aquellas columnas que contienen datos en formato str o conceptos, para aquellos valores nulos se les sustituyó por un concepto que define la omisión de la información a través de la palabra “Omitido.”

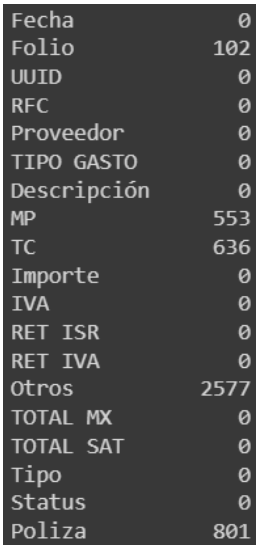
Cantidad de valores nulos por columna, antes de la limpieza:



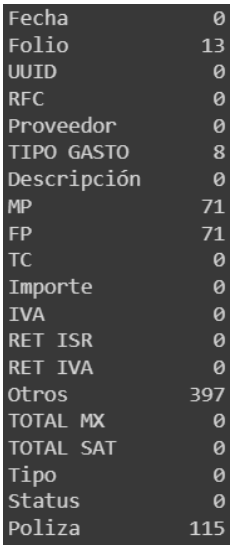
FECHA	0
FOLIO	189
UUID	0
RFC	0
PROVEEDOR	0
TIPO GASTO	0
GASTO	2502
DESCRIPCION	0
TC	391
IMPORTE	34
IVA	268
RET ISR	0
RET IVA	0
TOTAL MX	0
TOTAL SAT	0
TIPO	1
STATUS	0
POLIZA	3321
Columna1	0



FECHA	0
FOLIO	147
UUID	0
RFC	0
PROVEEDOR	0
TIPO GASTO	0
DESCRIPCION	0
MP	654
TC	0
IMPORTE	0
IVA	0
RET ISR	0
RET IVA	0
TOTAL MX	0
TOTAL SAT	0
TIPO	0
STATUS	0
POLIZA	2372



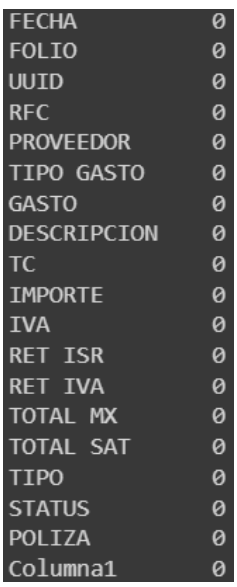
Fecha	0
Folio	102
UUID	0
RFC	0
Proveedor	0
TIPO GASTO	0
Descripción	0
MP	553
TC	636
Importe	0
IVA	0
RET ISR	0
RET IVA	0
Otros	2577
TOTAL MX	0
TOTAL SAT	0
Tipo	0
Status	0
Poliza	801



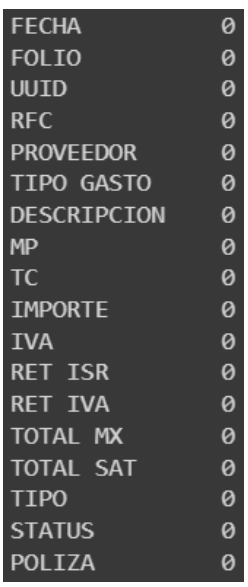
Fecha	0
Folio	13
UUID	0
RFC	0
Proveedor	0
TIPO GASTO	8
Descripción	0
MP	71
FP	71
TC	0
Importe	0
IVA	0
RET ISR	0
RET IVA	0
Otros	397
TOTAL MX	0
TOTAL SAT	0
Tipo	0
Status	0
Poliza	115

(2020) (2021) (2022) (2023)

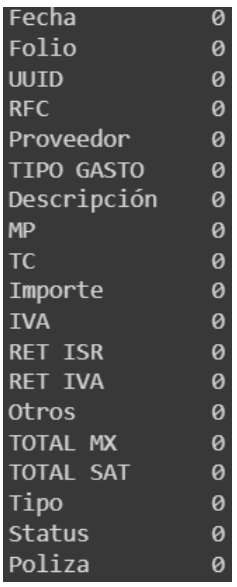
Cantidad de valores nulos por columna, después de la limpieza:



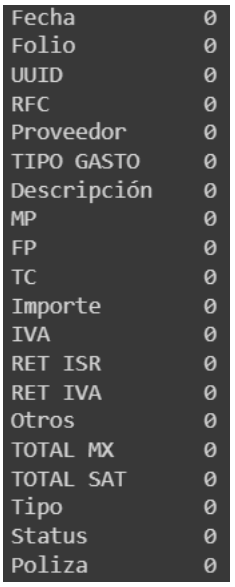
FECHA	0
FOLIO	0
UUID	0
RFC	0
PROVEEDOR	0
TIPO GASTO	0
GASTO	0
DESCRIPCION	0
TC	0
IMPORTE	0
IVA	0
RET ISR	0
RET IVA	0
TOTAL MX	0
TOTAL SAT	0
TIPO	0
STATUS	0
POLIZA	0
Columna1	0



FECHA	0
FOLIO	0
UUID	0
RFC	0
PROVEEDOR	0
TIPO GASTO	0
DESCRIPCION	0
MP	0
TC	0
IMPORTE	0
IVA	0
RET ISR	0
RET IVA	0
TOTAL MX	0
TOTAL SAT	0
TIPO	0
STATUS	0
POLIZA	0



Fecha	0
Folio	0
UUID	0
RFC	0
Proveedor	0
TIPO GASTO	0
Descripción	0
MP	0
TC	0
Importe	0
IVA	0
RET ISR	0
RET IVA	0
Otros	0
TOTAL MX	0
TOTAL SAT	0
Tipo	0
Status	0
Poliza	0



Fecha	0
Folio	0
UUID	0
RFC	0
Proveedor	0
TIPO GASTO	0
Descripción	0
MP	0
FP	0
TC	0
Importe	0
IVA	0
RET ISR	0
RET IVA	0
Otros	0
TOTAL MX	0
TOTAL SAT	0
Tipo	0
Status	0
Poliza	0

(2020) (2021) (2022) (2023)