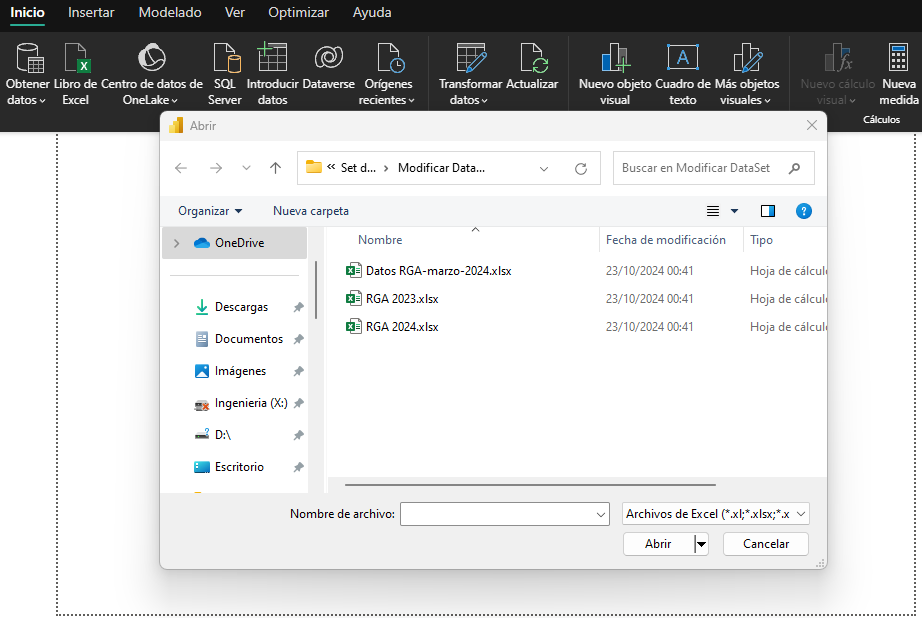
En este documento se describirá el proceso ETL para integrar y transformar tres conjuntos de datos distintos en un único dataset combinado en Power BI, facilitando el análisis y la visualización de datos relacionados con vuelos, región, y pasajeros. Este proceso permitirá obtener insights significativos a partir de la combinación de diferentes fuentes de información.

**1. Proceso de Carga de Archivos**

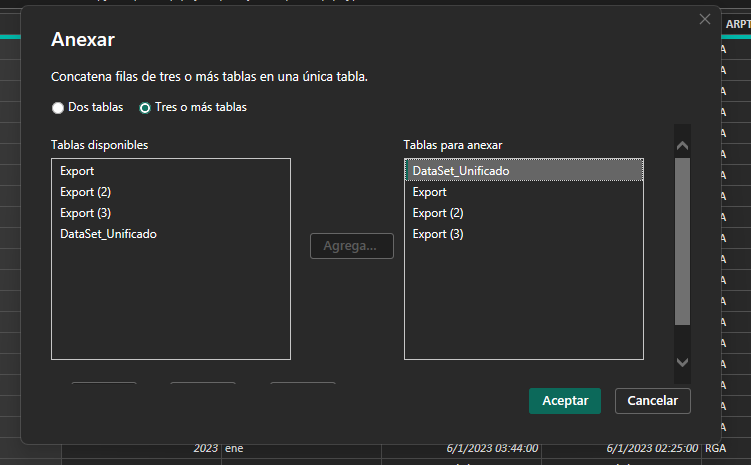
El primer paso del proceso ETL es la **extracción** de los tres conjuntos de datos y su **carga** en Power BI.

**Importar Datos:**

* En la pestaña de Inicio, selecciona la opción Obtener datos.
* Aparecerá una ventana con diversas opciones de fuentes de datos. Dependiendo del formato de los datasets, selecciona el tipo de fuente de datos adecuado, en este caso formato .xlsx.
  + Excel: Si los datasets están en formato .xlsx.



* Anexar los tres dataset para convertirlo en uno solo.



**2. Transformación de Datos**

La transformación de los datasets implica limpiar, modificar, y combinar los datos para preparar un único dataset estructurado

EL siguiente data set unificado contiene 8.466 registros y 35 columnas.

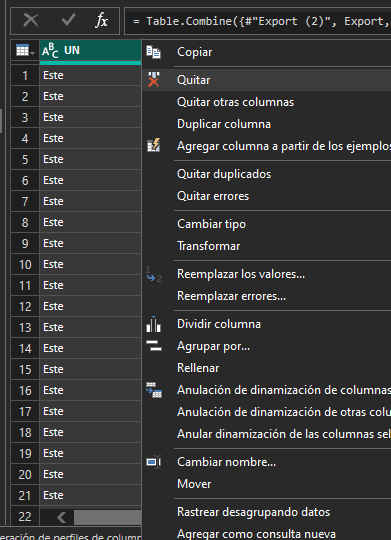
|  |  |  |  |
| --- | --- | --- | --- |
| **Columnas** | **Tipo de valor** | **Tamaño** | **Descripción** |
|  |
| ID\_OP\_FLIGHT | Entero | 7 | ID de registro de vuelos |  |
| AÑO | Entero | 4 | Años de la fecha de cuando se voló |  |
| UN | Caracter | 4 | Filtro de consulta anual |  |
| MES | Caracter | 3 | Mes de la fecha de cuando se voló |  |
| FECHA Y HORA REAL | DateTime | 19 | Fecha de cuando se voló |  |
| FECHA Y HORA PROG | DateTime | 19 | Fecha programación |  |
| ARPT | Caracter | 3 | Aeropuerto origen (código IATA de 3 a 4 caracteres) |  |
| ID LLAA | Caracter | 3 | ID aerolínea (código IATA de la aerolínea) |  |
| LLAA | Caracter | 50 | Descripción de aerolínea |  |
| NRO. VUELO | Entero | 4 | Número de vuelo |  |
| STAND | Entero | 1 | Consultas al Cliente |  |
| MOV TIPO | Caracter | 10 | Tipo de vuelo (arribo/partida) |  |
| ARPT1 | Caracter | 4 | Aeropuerto destino (código IATA de 3 a 4 caracteres) |  |
| PAIS | Caracter | 9 | Nacional |  |
| VUELO REGION | Caracter | 20 | Regionalidad del vuelo |  |
| VUELO TIPO | Caracter | 30 | Tipo de vuelo (comercial, aviación general, taxi aéreo, etc.) |  |
| VUELO CLASE | Caracter | 20 | Clase de vuelo |  |
| GATE | Caracter | 5 | Consultas al Cliente |  |
| VUELOS | Entero | 1 | Solo un valor “1” |  |
| PAX TOTALES | Entero | 4 | Pasajeros totales |  |
| PAX | Entero | 4 | Pasajeros que pagan |  |
| PAX TOB | Entero | 4 | Pasajero en tránsito (no se baja del avión en la conexión) |  |
| PAX TOC | Entero | 4 | Pasajero en transferencia (se baja del avión en la conexión) |  |
| PAX TSC | Entero | 4 | Pasajero en transferencia (se baja del avión en la conexión) |  |
| PAX INFOAS | Entero | 4 | Pasajeros infantes |  |
| PAX EXENTOS | Entero | 4 | Los Diplomáticos (es decir, los que no pagan tasa) |  |
| MATRÍCULA | Caracter | 10 | Matrícula del avión |  |
| ID ACFT | Caracter | 10 | ID equipo/modelo de aeronave |  |
| ACFT DESC | Caracter | 50 | Descripción equipo/modelo de aeronave |  |
| BODY | Caracter | 2 | Consultas al Cliente |  |
| ASIENTOS TOTALES | Entero | 4 | Asientos totales |  |
| ASIENTOS BSN | Entero | 4 | Asientos business |  |
| ASIENTOS ECY | Entero | 4 | Asientos economy |  |
| ASIENTOS 1ST | Entero | 4 | Asientos First class |  |
| FACTOR OCUPACIÓN | Decimal | 5,2 | % de asientos ocupados respecto a la cantidad de asientos |  |

1. **Eliminar datos, filas y columnas Irrelevantes**:

En el dataset, revisamos las columnas y elimina aquellas que no sean necesarias para el análisis

\* ID\_OP\_FLIGHT: Se quita ya que en los demás archivos dataset no tenemos el ID de registros.

\* UN: Se quita, la información de esta columna solo es “ESTE”, y se utilizó para la generación de dataset mediante el filtro anual



1. **Manejo de Datos Nulos**:

Durante este proceso, se identificaron tres filas que contenían la mayoría de sus columnas con valores "NULL". Estas filas solo mostraban información relacionada con la sumatoria de las cantidades en las columnas **ASIENTOS TOTALES**, **ASIENTOS 1ST**, **ASIENTOS ECY**, y **ASIENTOS BSN**.

Continuando con el análisis, detectamos valores "NULL" en las columnas **STAND**, **GATE** y **FACTOR OCUPACIÓN**. Dado que la cantidad de celdas afectadas era considerablemente alta, y con el fin de no perder registros valiosos para el análisis y poder investigar las posibles causas del error, se decidió reemplazar los valores "NULL" por "0". Este valor no está registrado en las columnas mencionadas, lo que facilitará su uso posterior como filtro para consultas y aclaraciones con el cliente.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamenteCaptura de pantalla de un celular

Descripción generada automáticamente

1. **Manejo de Datos Error**

Se identificaron celdas con el valor "Error" en las columnas **NRO. Vuelo** y **STAND**, las cuales seguían un patrón: estos errores aparecían en vuelos privados y todas las demás columnas que deberían contener información estaban vacías. Además, se detectó que la columna **ARTP1** presentaba un formato anómalo con un símbolo de negativo ("-AR"), diferente al del resto de los registros.

Dado que estos registros no aportan información relevante y su cantidad es limitada (33 registros), se decidió eliminarlos del dataset.

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

Luego de inspeccionar todas las columnas que no tenga errores, se procede a utilizar la herramienta “Quitar errores” para asegura que no quedan valores erróneos.

Imagen de la pantalla de un video juego

Descripción generada automáticamente con confianza media

1. **Formateo de Tipos de Datos**:

Para las 33 columnas, debemos configurar los valores de acuerdo con lo establecido en el archivo previamente definido: **"09. Diccionario de datos Aeropuerto 2000.xlsx"**, utilizando los valores correspondientes a cada columna.













1. **Registros duplicados**

Como parte del proceso de transformación de datos, es necesario eliminar los registros duplicados. Para ello, utilizaremos una técnica que consiste en concatenar las columnas **FECHA Y HORA REAL** y **FECHA Y HORA PROG**. Esta concatenación se almacenará en una nueva columna llamada **Personalizado**, que nos permitirá identificar y eliminar los duplicados. El total de registros afectados es de 560.

Captura de pantalla de un celular

Descripción generada automáticamente

1. **Creación de columnas extras**:

En este caso, tomamos la decisión de agregar una columna extra llama “Mes\_N” que nos servirá para utilizar de ordenamiento de mes cuando se realicen las visualizaciones.

Formula implementada

  si  [MES]  =  "ene"  entonces  1

   de lo contrario  si  [MES]  =  "feb"  entonces  2

    de lo contrario  si  [MES]  =  "mar"  entonces  3

    de lo contrario  si  [MES]  =  "abr"  entonces  4

    de lo contrario  si  [MES]  =  "may"  entonces  5

    de lo contrario  si  [MES]  =  "jun"  entonces  6

    de lo contrario  si  [MES]  =  "jul"  entonces  7

    de lo contrario  si  [MES]  =  "ago"  entonces  8

    de lo contrario  si  [MES]  =  "sep"  entonces  9

   de lo contrario  si  [MES]  =  "oct"  entonces  10

    de lo contrario  si  [MES]  =  "nov"  entonces  11

   de lo contrario  si  [MES]  =  "dic"  entonces  12

    de lo contrario  null

**Conclusión ETL**

El proceso ETL realizado para integrar y transformar los tres conjuntos de datos relacionados con vuelos, región, pasajeros, etc. ha permitido consolidar la información en un único dataset en Power BI, optimizando su análisis y visualización. A través de una serie de pasos sistemáticos, hemos logrado no solo la carga efectiva de los datos, sino también su limpieza y estructuración.

La eliminación de datos irrelevantes y la gestión de valores nulos y errores ha garantizado que el dataset resultante, con 7.942 registros y 35 columnas, sea fiable y útil para análisis posteriores. Al abordar los registros duplicados y al añadir una columna adicional para facilitar la ordenación por mes, se ha mejorado la calidad y la usabilidad de los datos.