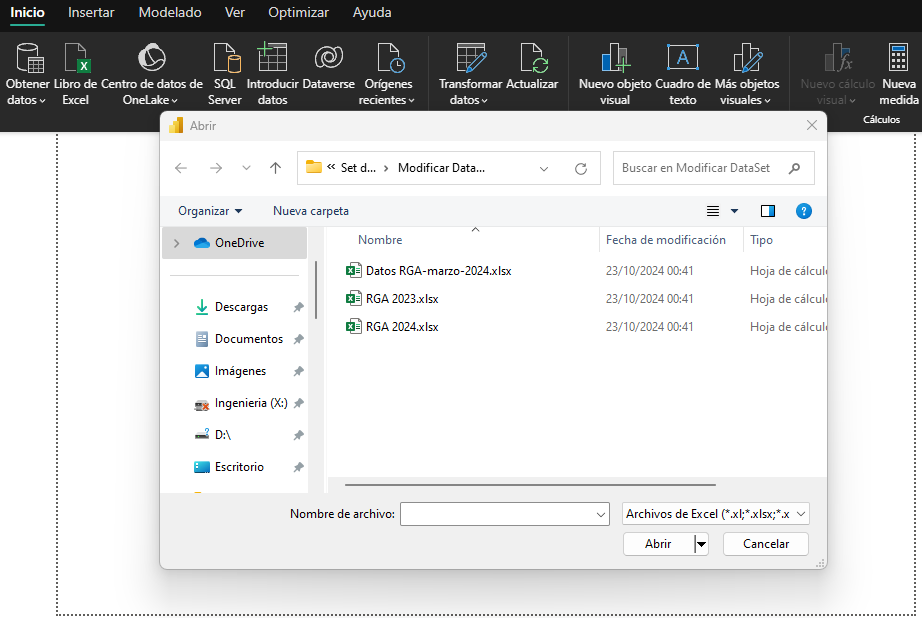
En este documento se describirá el proceso ETL para integrar y transformar tres conjuntos de datos distintos en un único dataset combinado en Power BI, facilitando el análisis y la visualización de datos relacionados con vuelos, región, y pasajeros. Este proceso permitirá obtener insights significativos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

**1. Proceso de Carga de Archivos**

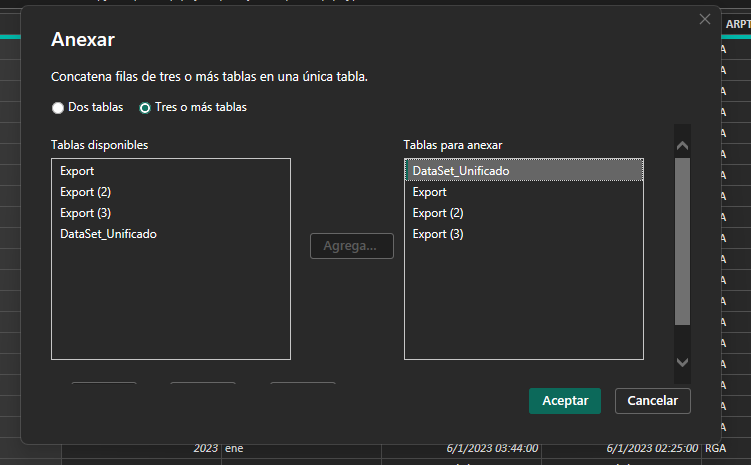
El primer paso del proceso ETL es la **extracción** de los tres conjuntos de datos y su **carga** en Power BI.

**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.

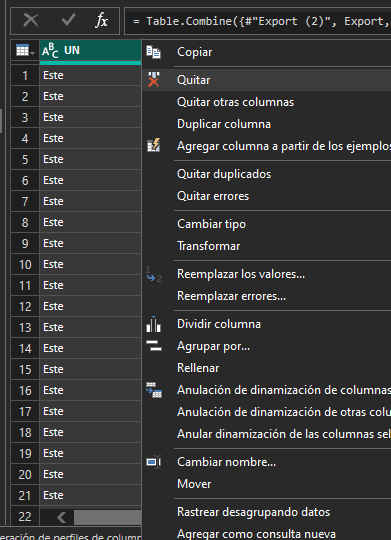
|  |  |  |  |
| --- | --- | --- | --- |
| **Columnas** | **Tipo de valor** | **Tamaño** | **Descripción** |
|  |
| ID\_OP\_FLIGHT | Entero | 7 | ID de registro de vuelos |  |
| AÑO | Entero | 4 | Años de la fecha de cuando se voló |  |
| UN | Caracter | 4 | Filtro de consulta anual |  |
| MES | Caracter | 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. **Transformación de dataSet, Agregar columnas necesarias**:

Para mejorar la claridad y la funcionalidad de nuestras visualizaciones, hemos decidido agregar algunas columnas adicionales que faciliten el análisis de los datos. Entre ellas, se incluye una columna llamada “Mes\_N”, que contiene valores numéricos correspondientes a cada mes del año (1 para enero, 2 para febrero, y así sucesivamente). Esta columna servirá como un índice de ordenación en las visualizaciones, permitiéndonos organizar los datos cronológicamente de manera precisa. Al contar con esta referencia, podremos realizar análisis de tendencias y patrones a lo largo del tiempo, facilitando la comparación mensual y mejorando la experiencia visual de los informes.

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

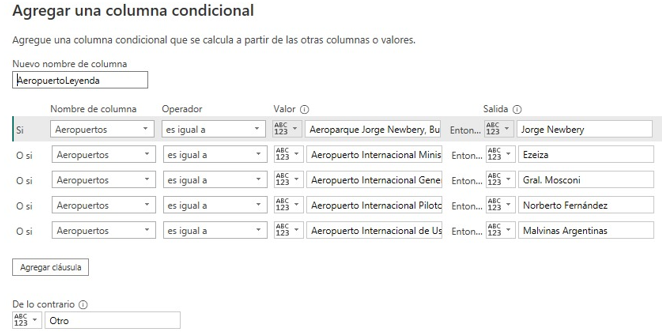
    de lo contrario  si  [MES]  =  "nov"  entonces  11

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

    de lo contrario  null

**Columna de Latitud y Longitud, Aeropuertos, Aeropuerto Leyenda, Ciudad**

En este caso, hemos añadido dos columnas adicionales que nos permitirán almacenar las coordenadas geográficas de diferentes aeropuertos. Estas columnas, una para la latitud y otra para la longitud, nos ayudarán a representar la ubicación exacta de cada aeropuerto en el mapa. Con esta información, podremos generar visualizaciones geoespaciales, facilitando el análisis visual de la distribución y proximidad de aeropuertos en diversas regiones. Esto resulta especialmente útil para identificar patrones y tendencias de ubicación que pueden enriquecer estudios logísticos, de tráfico aéreo o de accesibilidad regional.



Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

**Conclusión ETL**

El proceso ETL llevado a cabo para integrar y transformar los tres conjuntos de datos relacionados con vuelos, regiones y pasajeros ha permitido consolidar toda la información en un único dataset en Power BI, optimizando tanto su análisis como su visualización. A lo largo del proceso ETL, hemos conseguido no solo una carga eficaz de los datos, sino también su limpieza y estructuración, asegurando así una base de datos robusta y preparada para el análisis.

Además de eliminar datos irrelevantes y gestionar valores nulos y errores, lo que garantiza la confiabilidad del dataset final, también hemos incorporado columnas adicionales para enriquecer las visualizaciones y facilitar los análisis. Estas nuevas columnas incluyen información de latitud y longitud, que nos permite localizar geográficamente los aeropuertos, así como la ciudad en la que se encuentran, mejorando la capacidad de realizar análisis geoespaciales. También hemos añadido una columna de ordenación mensual ("Mes\_N"), que organiza los datos cronológicamente, permitiendo observar tendencias mensuales de manera más precisa.

Al resolver problemas como registros duplicados y optimizar la estructura de los datos, hemos alcanzado un dataset final con 7,972 registros y 38 columnas que es limpio, bien organizado y completamente funcional. Este dataset resulta ideal para la exploración de datos y la toma de decisiones estratégicas, proporcionando una base fiable para el análisis avanzado de tendencias y patrones en los datos de vuelos, regiones y pasajeros.