

Análisis de la historia de accidentes aéreos:

Autor: Esteban Mohr

1) Descripción de la temática de los Datos:

Se recopilan datos de la historia de accidentes de aviones y fatalidades en todo el mundo desde 1908 hasta 2023. El conjunto de datos encapsula información invaluable para investigadores, entusiastas de la aviación y expertos en seguridad interesados en comprender la dinámica, tendencias y patrones de incidentes de aviación a lo largo de más de un siglo.

2) Hipótesis:

Se desea comprender la evolución del comportamiento de los accidentes aéreos a lo largo de la historia de la aviación. En concreto en base a los datos se desea conocer lo siguiente:

- Cuáles han sido los accidentes aéreos que más vidas humanas se ha llevado durante la historia de la aviación.
- Descubrir cuales modelos de aviones han sufrido una mayor cantidad de accidentes.
- Conocer cuáles son los operadores aéreos que más accidentes han sufrido, la frecuencia de sus accidentes y la cantidad de fallecimientos acumulados en sus accidentes.
- Descubrir si existen áreas geográficas (lugares, estados o países) en las cuales han acontecido más cantidad de accidentes y/o muertes en relación con otras.
- Analizar si existen rutas u horarios en los cuales se han presentado accidentes de forma más frecuente.

Posteriormente y en base a los análisis realizados se concluirá en qué condiciones existe un riesgo mayor de producirse un accidente y se generarán propuestas de cómo se podrían prevenir futuros accidentes.

3) Dataset:

El dataset utilizado se encuentra en el archivo dataset_accidentes_aereos_Esteban_Mohr.csv, el cual contiene 4998 filas y 17 columnas. Las columnas del dataset son las siguientes:

- **Date:** Fecha del accidente
- **Time:** Hora del accidente
- **Location:** Lugar del accidente
- **Operator:** Operador del vuelo
- **Flight #:** Número de vuelo
- **Route:** Ruta de vuelo
- **AC Type:** Tipo/Modelo de aeronave
- **Registration:** Número de matrícula del avión

- **cn/ln:** Número de serie/fabricación del avión
- **Aboard:** Cantidad de personas a bordo
- **Aboard Passangers:** Cantidad de pasajeros a bordo
- **Aboard Crew:** Cantidad de personas de tripulación a bordo
- **Fatalities:** Cantidad total de fallecidos
- **Fatalities Passangers:** Número de pasajeros fallecidos
- **Fatalities Crew:** Número de fallecidos en tripulación
- **Ground:** Número de fallecidos en tierra
- **Summary:** Breve descripción del accidente

4) **Objetivo del proyecto**

El objetivo de este proyecto es crear un Dashboard en Power BI que permita facilitar encontrar las respuestas a las preguntas planteadas en la sección 2 de este documento.

5) **Alcance del Proyecto**

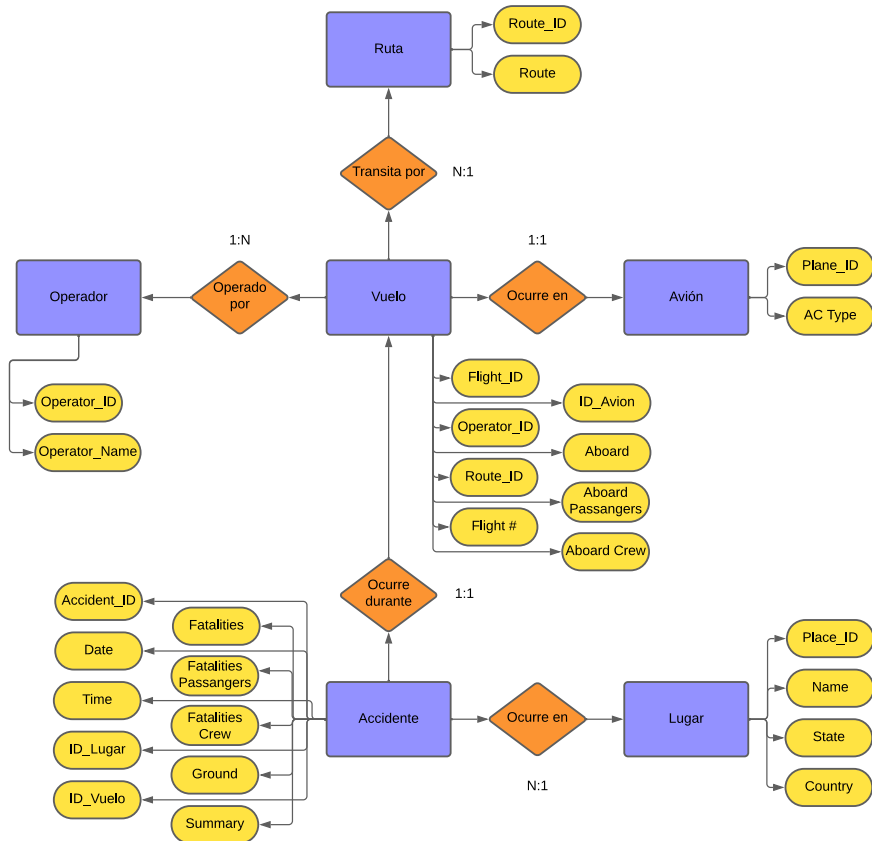
El alcance de este proyecto implica la creación de la base de datos y la creación del dashboard en Power BI. El dashboard se considerará como terminado cuando contenga todas las visualizaciones necesarias para que el usuario final pueda responder a las preguntas planteadas en la sección 2 de este documento.

6) **Usuario final y nivel de aplicación**

Los usuarios finales de esta aplicación son de tipo estratégicos, tales como gerentes de aerolíneas que desean evaluar que condiciones pueden ser riesgosa para la operación en el largo plazo o miembros de organizaciones reguladoras de aviación en busca de la creación de normas aeronáuticas que aseguren la seguridad en los vuelos.

7) Diagrama entidad-relación

En la siguiente imagen se observa el diagrama de entidad-relación del modelo de datos creado a partir del Dataset:



8) Listado de tablas

- **Accidente:** Contiene la fecha, hora,
 - PK: Accident_ID
 - FK: ID_Lugar
 - FK: ID_Vuelo
- **Lugar:** Contiene todos los lugares en los cuales ha habido un accidente. Los registros de lugares están separados por nombre, estado (solo para Estados Unidos y Canadá) y país.
 - PK: Location_ID
- **Vuelo:** Contiene los datos del vuelo en el cual ocurrió el accidente, tal como el operador, ruta, número de vuelo, modelo de avión y personas a bordo.
 - PK: Flight_ID
 - FK: Operator_ID
 - FK: Route_ID
 - FK: Plane_ID
- **Operador:** Contiene todos los operadores que han tenido algún accidente aéreo. Entiéndase Operador como la empresa u organización responsable del vuelo.
 - PK: Operator_ID

- **Ruta:** Contiene todas las rutas en las cuales ha sucedido algún accidente aéreo.
 - PK: Route_ID
- **Avión:** Contiene todos los modelos de aviones en los cuales han ocurrido accidentes.
 - PK: Plane_ID

9) Listado de columnas por tablas

A continuación, se presentan las tablas creadas en la base de datos (archivo “Base de Datos.xls”) con sus respectivas tablas, sus tipos de datos y si corresponde se indica si la columna es llave primaria (PK) o foránea:

| Accidente | | |
|------------------------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Accident_ID | Int | PK |
| Date | Date | |
| Time | Time | |
| ID_Lugar | Int | FK |
| ID_Vuelo | Int | FK |
| Fatalities | Int | |
| Fatalities Passangers | Int | |
| Fatalities Crew | Int | |
| Ground | Int | |
| Summary | Varchar | |

| Lugar | | |
|--------------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Location_ID | Int | PK |
| Name | Varchar | |
| State | Varchar | |
| Country | Varchar | |

10)

| Vuelo | | |
|--------------------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Flight_ID | Int | PK |
| Operator_ID | Int | FK |
| Route_ID | Int | FK |
| Flight # | Int | |
| Plane_ID | Int | FK |
| Aboard | Int | |
| Aboard Passangers | Int | |
| Aboard Crew | Int | |

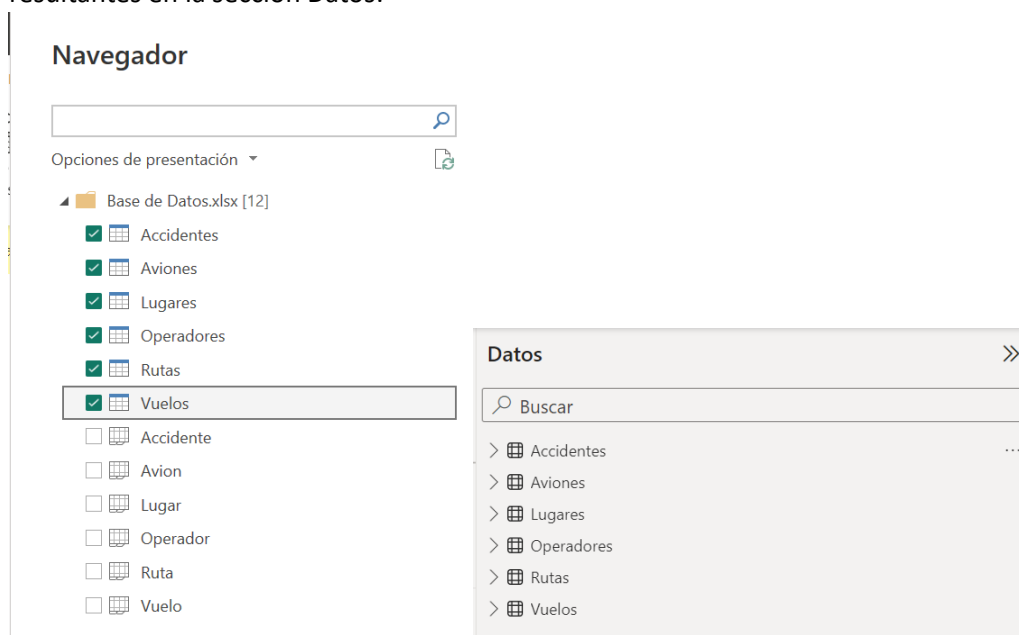
| Operador | | |
|----------------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Operator_ID | Int | PK |
| Operator_Name | Varchar | |

| Ruta | | |
|-----------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Route_ID | Int | PK |
| Route | Varchar | |

| Avion | | |
|-----------------|--------------|---------------|
| Columna | Tipo de dato | Tipo de Clave |
| Plane_ID | Int | PK |
| AC_Type | Varchar | |

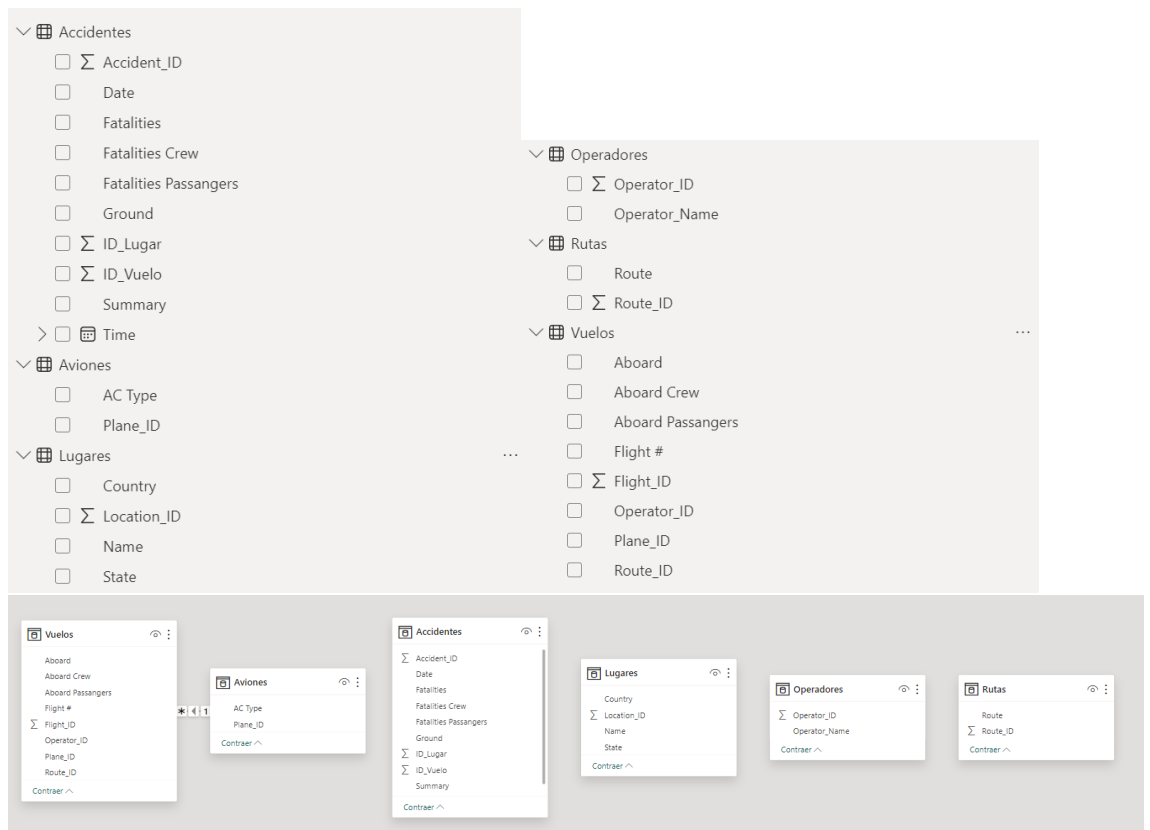
11) Carga de Datos en Power BI y ajustes iniciales

Antes de cargar los datos en Power BI se vuelve a ingresar a la base de datos (archivo “Base de Datos.xls”) para formatear los datos de cada una de las páginas como tablas para que estas puedan ser detectadas por Power BI. Habiendo realizado esto se procede a crear el archivo que contiene el informe de Power BI llamado “Análisis Accidentes Aereos.pbix” y se importan los datos de la base de datos. En las siguientes imágenes se observa cómo se seleccionan las tablas del archivo Excel y las tablas resultantes en la sección Datos:

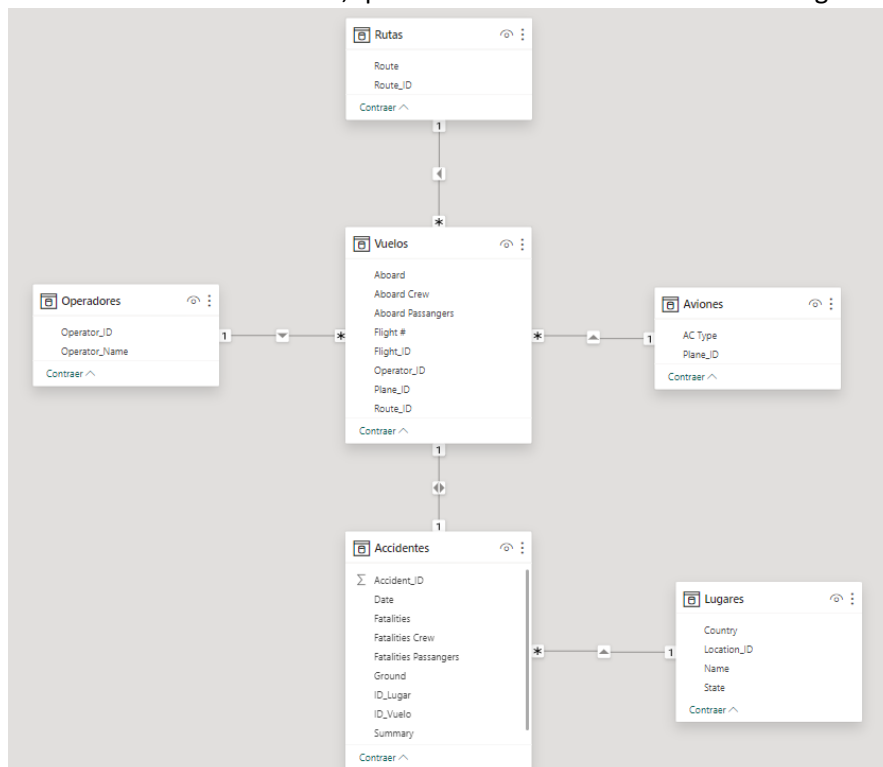


Inicialmente las relaciones no fueron correctamente debido a que existen relaciones a través de claves foráneas donde los nombres de las columnas no son exactamente las mismas, por ejemplo, en la tabla Accidentes se tiene la columna ID_Lugar y en la tabla Lugares se tiene la columna Location_ID.

En las siguientes imágenes se observa el estado inicial de las diferentes columnas y del modelo relacional.



Debido a la inconsistencia de los nombres, como arreglo temporal se ajustan manualmente las relaciones, quedando el modelo relacional de la siguiente forma:



Como todas las relaciones creadas son del tipo uno es a muchos (1:N) no es necesario crear ninguna tabla puente.

12) Transformación de Datos

A continuación, se describen las diferentes transformaciones realizadas a los datos de las diferentes tablas de la base de datos:

a. Tabla Accidentes

- i. Se cambia el tipo de la columna *Accident_ID* a número entero
- ii. Se cambia el tipo de la columna *Date* a tipo Fecha con configuración regional para formato de fecha de estados unidos. De esta manera se interpreta bajo el formato Mes-Día-Año (MM-DD-YYYY).
- iii. Se cambian el tipo de las columnas *ID_Lugar*, *ID_Vuelo*, *Fatalities*, *Fatalities Passengers*, *Fatalities Crew* y *Ground* a número entero.
- iv. Se cambia el valor de *Ground* para el accidente con ID 4330 a 0. Esto se debe a que este accidente junto al accidente con ID 1329 corresponden a los dos accidentes aéreos del atentado a las torres gemelas en 2001. La cantidad de fallecidos en tierra para ambos registros de accidente están duplicadas por lo cual para limpiar los datos es necesario realizar esta transformación. Para realizar esta transformación se crea una nueva columna temporal con la expresión:

```
= Table.AddColumn("#Tipo cambiado1", "temp", each if [Accident_ID] = 4330 then 0 else [Ground])
```

Con esta expresión la columna *temp* tiene los mismos elementos de *Ground*, pero el registro con *Accident_ID* igual a 4330 se le cambia el valor a 0. Luego se elimina la columna *Ground* original y se le cambia el nombre de la columna *temp* a *Ground*. Además, a la nueva columna *Ground* se le cambia el tipo a número entero.

- v. Se elimina la columna *Time*, ya que estos datos no serán usados en el tablero.
- vi. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

b. Tabla Aviones

- i. Se cambia el tipo de la columna *Plane_ID* a número entero y el tipo de la columna *AC Type* a texto.
- ii. Se aplica la funcionalidad de limpiar texto en la columna *AC Type*.
- iii. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

c. Tabla Lugares

- i. Se cambia el tipo de la columna *Location_ID* a número entero y el tipo de las columnas *Name*, *State* y *Country* a texto.
- ii. Se aplica la funcionalidad de limpiar texto en las columnas *Name*, *State* y *Country*.
- iii. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

d. Tabla Operadores

- i. Se cambia el tipo de la columna *Operator_ID* a número entero y el tipo de la columna *Operator_Name* texto.
- ii. Se aplica la funcionalidad de limpiar texto en la columna *Operator_Name*.

- iii. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

e. Tabla Rutas

- i. Se cambia el tipo de la columna *Route_ID* a número entero y el tipo de la columna *Route* a texto.
- ii. Se aplica la funcionalidad de limpiar texto en la columna *Route*.
- iii. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

f. Tabla Vuelos

- i. Se cambia el tipo de la columna *Flight #* a texto y el tipo del resto de las columnas a número entero.
- ii. Se elimina la columna *Flight #* ya que estos datos no serán utilizados en el análisis.
- iii. Se cambia el nombre de todas las columnas a nombres en español y más intuitivos.

13) Creación de tabla calendario

Para crear la tabla calendario se crea una tabla calculada utilizando la siguiente expresión, con la cual se crean las fechas de todos los días desde el 17 de septiembre de 1908 hasta el 29 de octubre de 2023:

```
Calendario = CALENDARAUTO()
```

Luego se crean columnas calculadas para extraer el día, el mes, el año y el día de semana de cada una de las fechas utilizando las siguientes expresiones:

```
Day = DAY(Calendario[Date])
```

```
Month = MONTH(Calendario[Date])
```

```
Year = YEAR(Calendario[Date])
```

```
Week_Day = WEEKDAY(Calendario[Date])
```

Además, para este tablero se requiere tener no solo las fechas en formato numérico, sino en también en formato de texto. Las siguientes expresiones crean columnas calculadas para extraer el mes, y el día de semanas en formato de texto.

```
Month_Text = FORMAT(DATE(2000, Calendario[Month], 1), "MMMM")
```

```
Week_Day_Text = FORMAT(Calendario[Week_Day], "dddd")
```

14) Columnas Calculadas

Se crea una columna calculada adicional a las creadas en la tabla calendario, esta es la columna *Total Fallecidos* la cual representa la cantidad total de fallecidos de un accidente de avión sumando los fallecidos en el avión y los fallecidos en tierra. Esta columna se calcula utilizando la siguiente expresión.

`Total Fallecidos = Accidentes[Fallecidos a Bordo] + Accidentes[Fallecidos en Tierra]`

15) Medidas Calculadas

En esta sección se describen las diferentes medidas calculadas para ser utilizadas en el tablero. Cada una de estas medidas se encuentran en la columna calculada creada bajo el nombre Medidas

a. Max_Fallecidos_Accidente:

Esta medida representa el accidente que más fallecidos ha generado en la historia, la cual se calcula utilizando la siguiente expresión:

`Max_Fallecidos_Accidente = MAX(Accidentes[Total Fallecidos])`

b. Promedio_Fallecidos_Por_Accidente:

Esta medida representa el promedio de fallecidos por accidente, la cual se calcula utilizando la siguiente expresión:

`Promedio_Fallecidos_Por_Accidente = AVERAGE(Accidentes[Total Fallecidos])`

c. Total_Fallecidos:

Esta medida representa el número total de fallecidos provocados por accidentes aéreos en toda la historia entre 1908 y 2023. Esta medida se calcula utilizando la siguiente expresión:

`Total_Fallecidos = SUM(Accidentes[Total Fallecidos])`

d. Total_Modelos:

Esta medida representa el total de diferentes modelos de avión que han sido parte de la historia de los accidentes aéreos. Esta medida se calcula utilizando la siguiente expresión:

`Total_Modelos = DISTINCTCOUNTNOBLANK(Aviones[Modelo de Avión])`

e. Total_Operadores:

Esta medida representa el total de diferentes operadores que han sufrido accidentes aéros. Esta medida se calcula utilizando la siguiente expresión:

`Total_Operadores = DISTINCTCOUNTNOBLANK(Operadores[Nombre Operador])`

f. Total_Paises:

Esta medida representa la cantidad total de países (bajo esta categoría se encuentran también lugares como mares y océanos) en los cuales han ocurrido accidentes aéreos. Esta medida se calcula utilizando la siguiente expresión:

`Total_Paises = DISTINCTCOUNT(Lugares[País])`

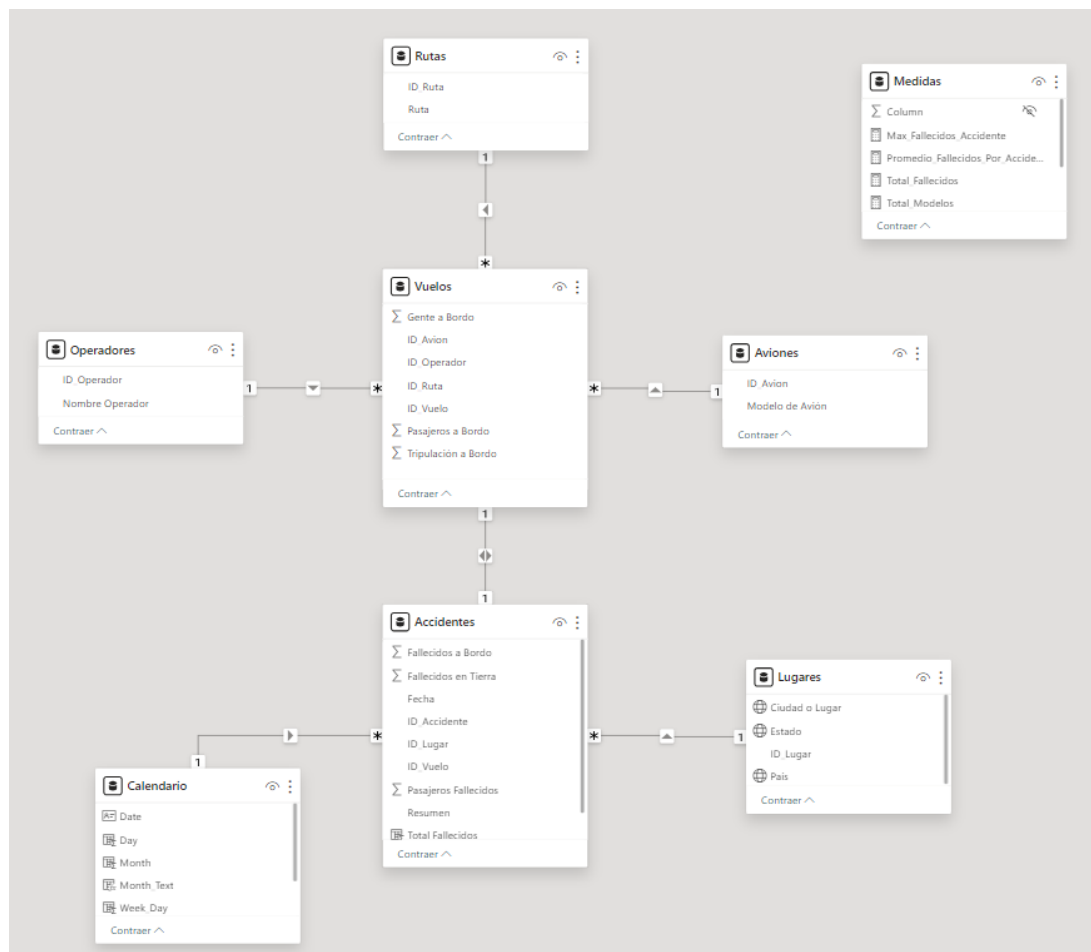
g. Total_Rutas:

Esta medida representa la cantidad total de rutas en las cuales han ocurrido accidentes aéreos.

Total_Rutas = `DISTINCTCOUNTNOBLANK(Rutas[Ruta])`

16) Modelo de datos en Power BI

Una vez realizadas todas las transformaciones y creadas las tablas calculadas, columnas calculadas y medidas, el modelo de datos relaciones queda de la siguiente forma:

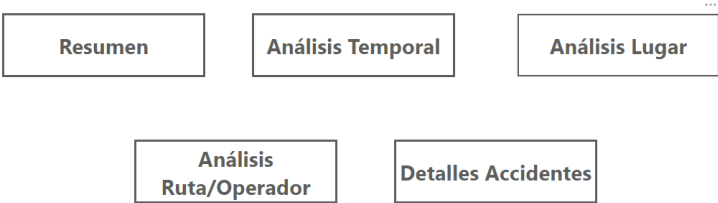


17) Visualizaciones Creadas

En esta sección se presentan las diferentes solapas creadas en Power BI. Cabe destacar que en las diferentes visualizaciones se tienen los nombres y descripciones de ejes por defecto. Cambiar estos nombres y el estilo de estas visualizaciones se encuentran entre las tareas pendientes a realizar para la versión final del tablero.

- a. Solapa Portada:
Esta solapa contiene 5 botones para acceder a cada una de las solapas.

Análisis de la historia de accidentes aéreos

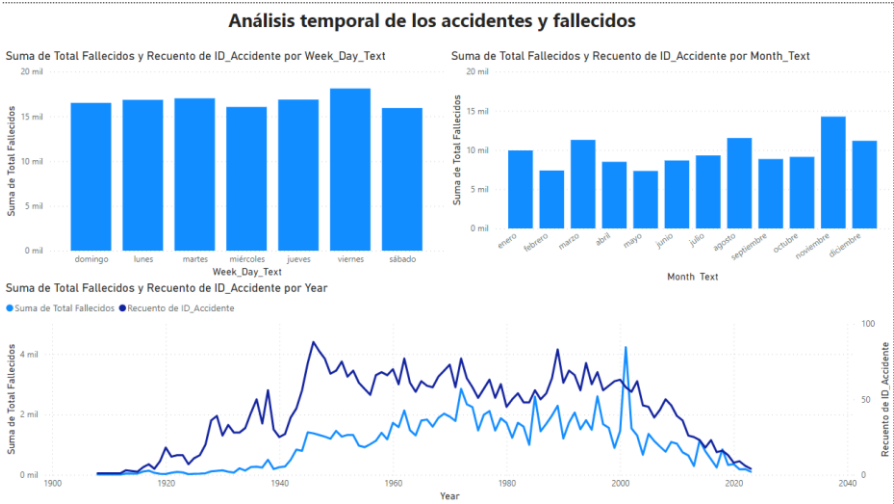


- b. Solapa Resumen:
Contiene indicadores generales del set de datos en visualizaciones de formato tarjeta. Además posee un gráfico de torta de la proporción de pasajeros fallecidos, tripulación fallecida y fallecidos en tierra.

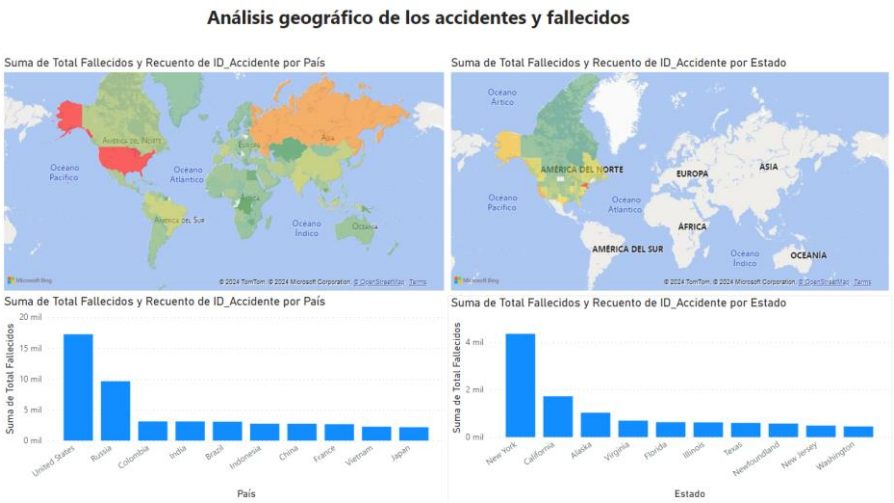


- c. Solapa Análisis Temporal:

Esta solapa corresponde a la relación entre el tiempo con el número de accidentes y fallecidos. Contiene un gráfico de líneas que muestra la cantidad de accidentes y fallecidos en cada año y dos gráficos de columnas que indican la cantidad total de fallecidos según día de semana y mes del año respectivamente. Si se coloca el cursor sobre cualquier elemento de ambos gráficos de barras se puede observar igualmente la cantidad de accidentes según la categoría.

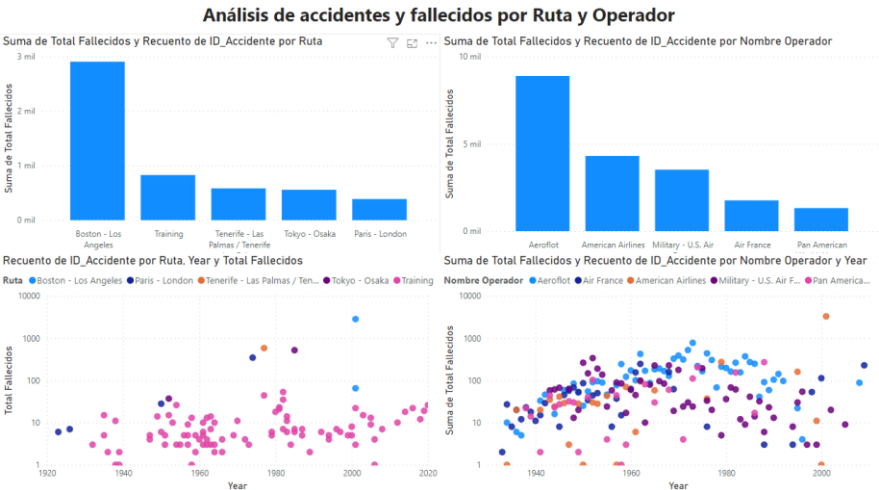


- d. **Solapa Análisis Lugar:**
- Esta solapa contiene un análisis geográfico de los accidentes, el cual contiene dos mapas, el primero contiene la distribución de accidentes y fallecidos por país y el segundo mapa ilustra la distribución por estados de Estados Unidos y Canadá. Además, esta solapa contiene dos gráficos de columnas de los top 10 países con mayor cantidad total de fallecidos y los top 10 estados con mayor cantidad de fallecidos. Si se coloca el cursor sobre cualquier elemento de ambos gráficos de barras o cualquier zona de los mapas se puede observar igualmente la cantidad de accidentes según lugar.



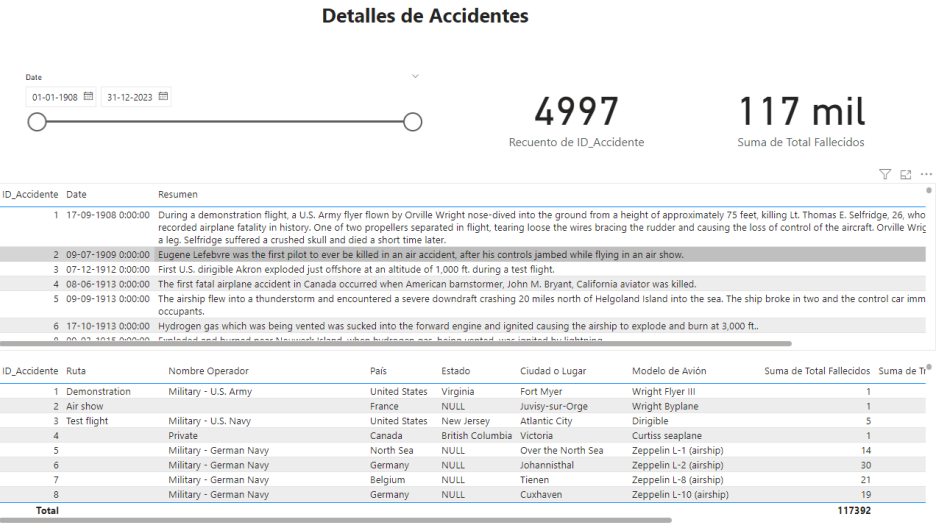
- e. **Solapa Análisis Ruta/Operador:**

Esta solapa consiste en un análisis de la cantidad de fallecidos y accidentes según ruta u operador de vuelo. La solapa contiene dos gráficos de columnas, el primero es el top 5 de rutas con mayor cantidad de fallecidos por accidentes y la segunda es el top 5 de operadores con mayor cantidad de fallecidos por accidentes. Además, tanto para el análisis de ruta y operador se tiene un gráfico de dispersión de la cantidad de fallecidos por año por ruta/operador. En cualquiera de las visualizaciones se puede observar la cantidad de accidentes colocando el cursor sobre los elementos visuales.



f. Solapa Detalles Accidentes:

Esta solapa permite al usuario buscar información en detalle sobre algún accidente en concreto, para facilitar la búsqueda en la parte superior izquierda de la solapa el usuario puede filtrar los datos de las tablas utilizando un rango de fechas. Además los indicadores de la esquina superior derecha indican la cantidad de accidentes y total de fallecidos en ese rango de tiempo.



18) Tareas pendientes:

Para dar por finalizado el tablero quedan pendientes las siguientes tareas:

- Añadir alguna visualización adicional a la solapa de resumen.
- Mejorar estética y colorimetría del tablero.
- Cambiar títulos y descripciones de ejes a las visualizaciones.
- Evaluar si es posible limpiar más los datos con alguna transformación al set de datos.
- Crear solapa de accidentes por modelo de avión o integrar visualización en alguna solapa existente.
- Añadir botones a solapas para mejorar navegación del tablero.