# Proyecto Final – Ingeniería de datos

# Universidad del Rosario, Escuela de Ingeniería Ciencia y Tecnología 2022-2

Análisis y elaboración de la base de datos a partir de la información de accidentes aéreos

Hecho por: Juan Sebastian Bernal Rojas, Juan Andrés Castro Carrillo, Juan Jose Ruiz Triana y Jose Miguel Torres Lara

**Resumen e introducción del proyecto**

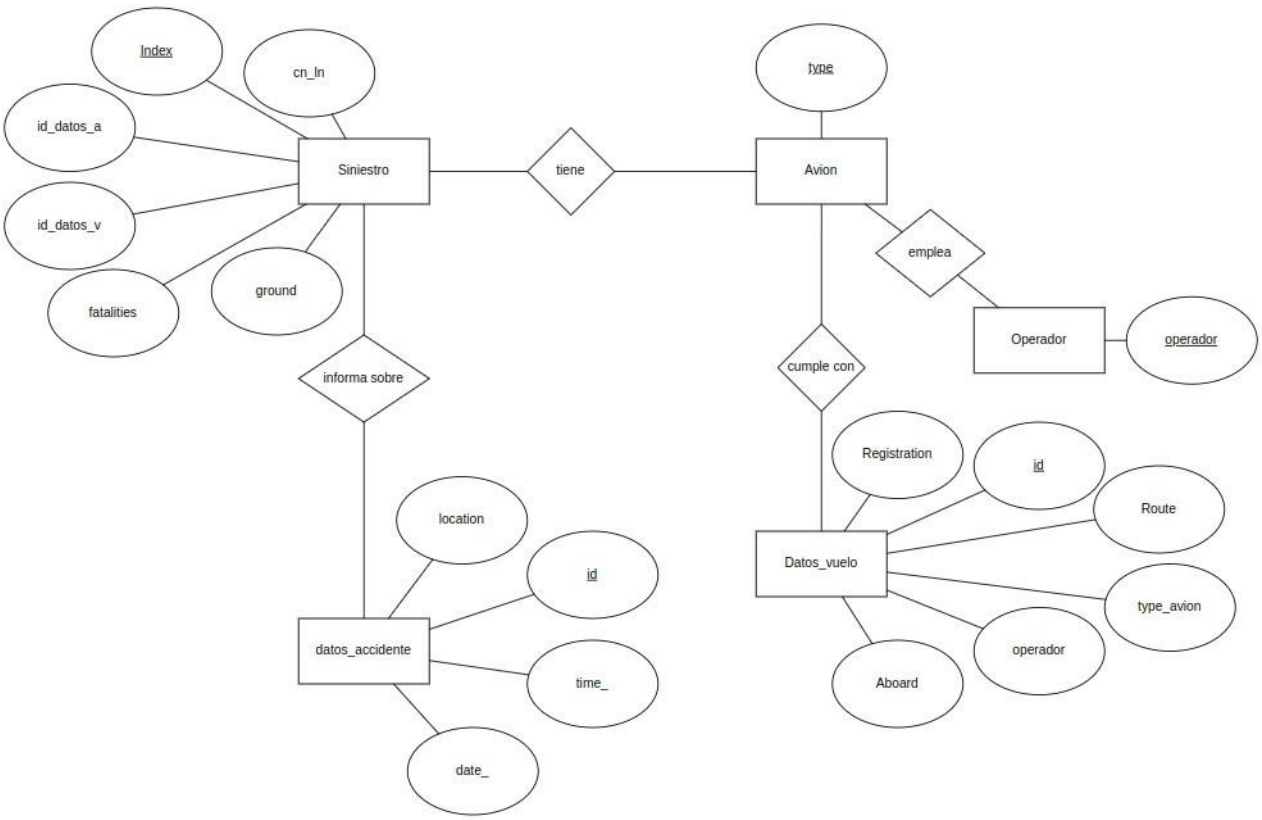
El avión es de los medios de transporte más seguros en la actualidad, sin embargo, la historia de la aerodinámica ha tenido inconvenientes que han sido resueltos de forma empírica desde los inicios hasta el día de hoy, para ello se han recaudado datos históricos de los vuelos y los accidentes que se han tenido desde 1908, para poder evaluar los cambios que ha habido en relación al paso del tiempo y los avances entre distintos tipos de aeronaves.

Es por esto que a partir de la base de datos publicada en el portal Kaggle se definieron diferentes reglas de negocio como punto de inicio del proyecto.

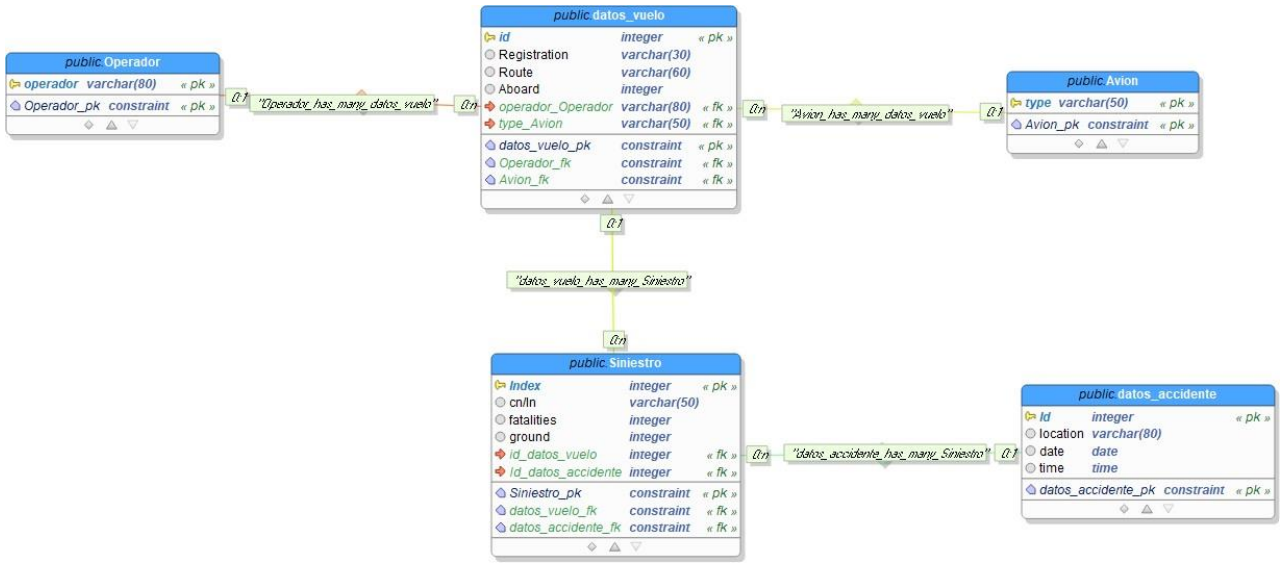
El cliente busca identificar y analizar los datos involucrados en cada accidente aéreo que tienen registrado a nivel mundial. Para organizar cada accidente se requiere un índice, el cual debe ser diferente para cada caso, además se debe describir el modelo del avión, el lugar, la fecha y la hora de los hechos, la cantidad de personas fallecidas durante el accidente, la ruta que cubría cada avión, la aerolínea a la que pertenece cada avión (llamada operador), el número de serial de cada avión (único para cada uno), el cliente solicitó que los accidentes estén organizados de manera cronológica y contengan un breve resumen de los hechos.

**Fuente de datos: [Airplane Crashes and Fatalities]**

**Diagrama del modelo Entidad-Relación (**nueva base de datos)



## Diagrama Relacional (PgModeler)




## Desarrollo de la base de datos en pgAdmin4:

Para concluir la normalización de la base de datos final se ejecutan los comandos en pgAdmin4 para crear las 5 tablas con sus respectivos atributos, para la tabla *datos_vuelos* se crea una función tipo sequence con el objetivo de diversificar cada id (primary key), este procedimiento se ejecuta también para la tabla *datos_accidente*. A medida que se van creando las tablas, se generan tantos las primary keys como las foreign keys relacionadas a cada tabla.

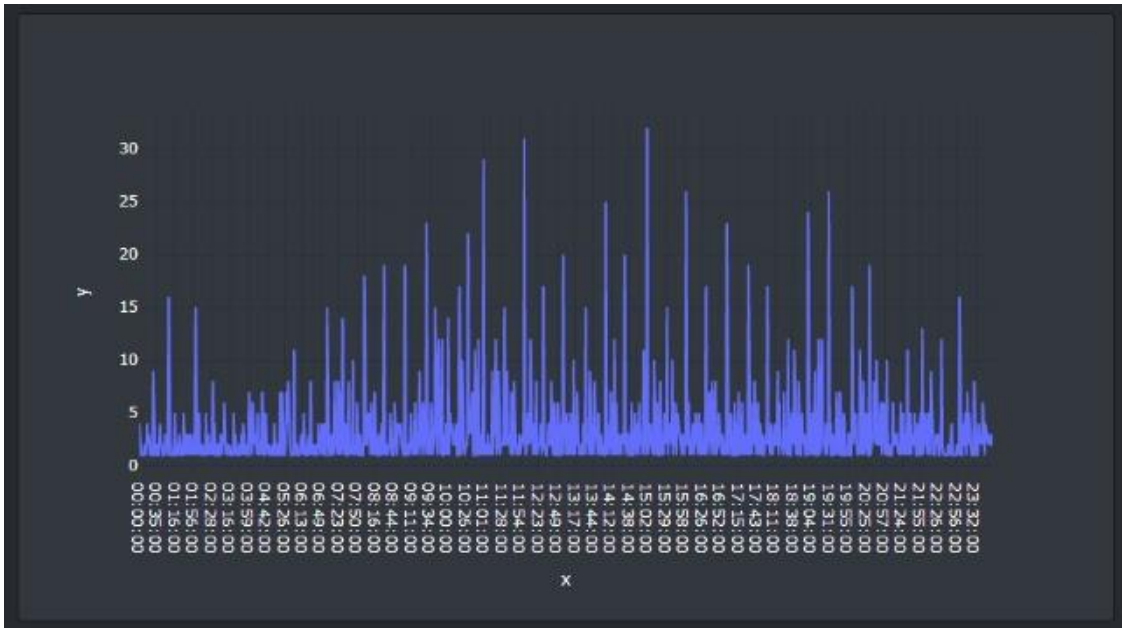
Se usa la función copy para migrar los datos de los archivos (.csv) con la información de cada tabla. Cada archivo (.csv) corresponde a una tabla, y contiene las tuplas sin repeticiones y con los formatos necesarios para cumplir los requisitos de cada atributo.

## Análisis de Datos

### Cantidad de accidentes por año

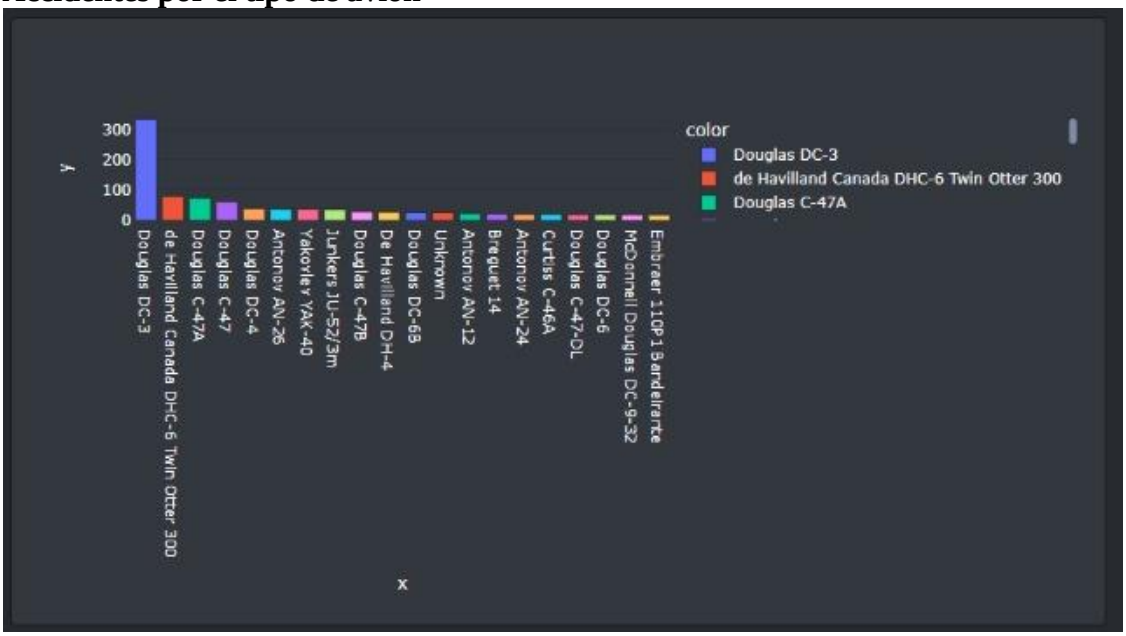
En este primer diagrama para representar la cantidad de accidentes por año que están registrados en la base de datos como grupo vimos que el que mejor nos ayudaba a plasmar los resultados era por medio del diagrama de barras ya que este fue el único que nos daba la imagen mas sencilla de interpretar y poder ver como entre 1960 y los 2000 fueron los años donde los accidentes aéreos eran mas frecuentes y como después de los 2000 se han tomado medidas como el uso de mejores tecnologías y procesos para que esta frecuencia vaya disminuyendo cada vez más.

**Franjas horarias de accidentes registrados**




Para poder representar en que horario se daban la mayor cantidad de accidentes también usamos el diagrama de barras por su sencilla interpretación y además es la que nos daba una mayor cantidad de información para poder decir que durante las noches es cuando hay menor taza de accidentes registrados.

**Accidentes por el tipo de avión**

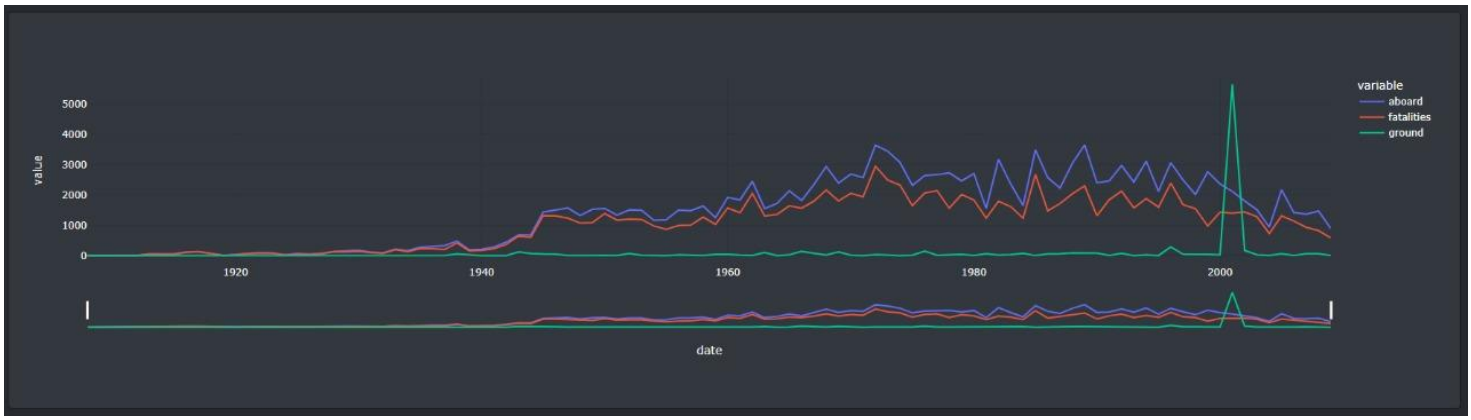


Para esta grafica necesitabamos ver que tipo de avion es el que mas accidentes ha tenido y para ello intentamos con una grafica de pastel pero solo nos mostraba un porcentaje y necesitabamos era la cantidad por lo cual en ete diagrama tambien empleamos un diagrama de barras para asi ver que aviones eran los mas accidentados, y gracias a este diagrama podemos ver que con una gran diferencia el tipo de avion mas accidentado es el Douglas DC-3

**Los 20 operadores con mas accidentes**

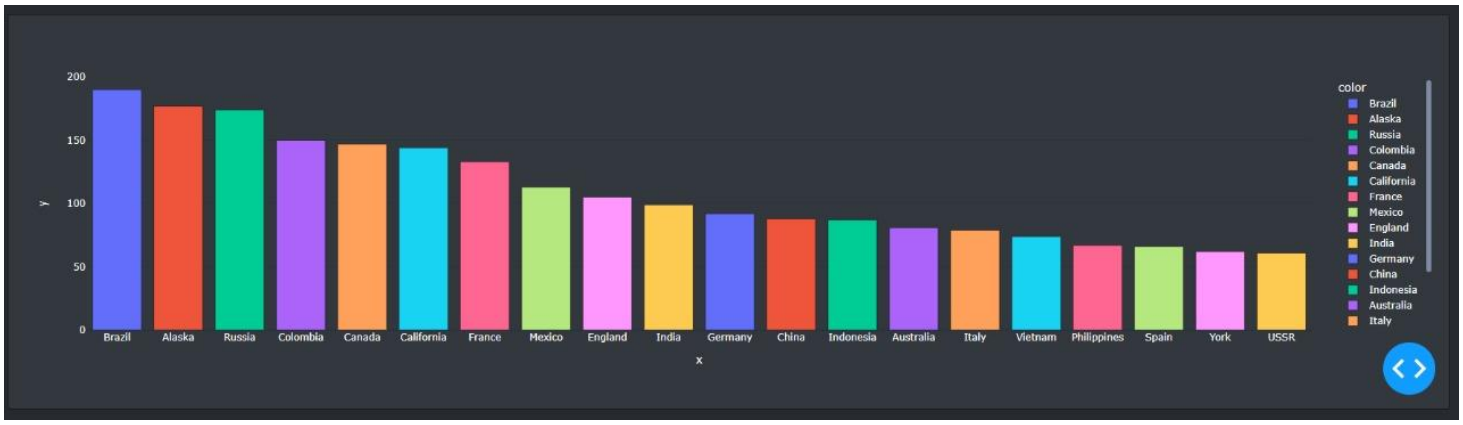

A fin de saber que operadores son los que mas han tenido accidentes para esta ocasión necesitábamos un porcentaje mas no la cantidad de accidentes, es por ello que necesitábamos que a partir del gran total de accidentes sacar que operadores sufren la mayor cantidad de accidentes y es por ello que escogimos el diagrama circular y así ver que a partir de todos los operadores poder ver cuales 20 aparecen mas en la base de datos y también pudimos obtener que los operador con mayor cantidad de accidentes fueron AB Aerotransport y ACES Colombia con el 9.68% de los accidentes cada uno.

**Total de muertes comparado con los pasajeros**




En esta grafica podemos ver la relación que hay entre el numero de personas que iban a bordo del avión durante el vuelo junto con las personas que lamentablemente fallecen debido al accidente, para ello necesitábamos ver como se iban comparando estas relaciones a medida que avanza el tiempo, el diagrama que nos dio la mejor representación de estas relaciones es el diagrama de líneas ya nos muestra como las personas que viajaban en avión fue incrementando después de 1940 y junto con ellas las personas que fallecían en accidentes aéreos, también encontramos que en los inicios de los 2000 hubo una gran cantidad de accidentes.

**Accidentes por país**




Para poder representar que países tienen la mayor cantidad de accidentes necesitábamos no ver como se relacionaban o compararlas mediante el gran total de accidentes es por lo cual nos decidimos de usar ujn diagrama de barras y asi poder ver estos datos, no empleamos la opción del mapa mundial que ofrece dash debido a que el diagrama mostrado no nos ayudaba mucho a sacar conclusiones, gracias a este diagrama podemos ver que en Brasil es el país donde mas accidentes han ocurrido pero no podemos ver una tendencia hacia que región ocurren la mayor cantidad de accidentes

**Conclusiones**

A partir del uso del repositorio de GitHub, creado con la finalidad de que cada participante del grupo tuviera la posibilidad de trabajar en el proyecto de forma dinámica, usándolo como un punto de acceso común al proyecto por lo que la plataforma GitHub usa Git un sistema de control de versiones que permite gestionar de una manera practica el desarrollo de proyectos. En adición a esto, una de las principales dificultades que se presentaron mediante todo el proceso de normalización, fue la información desorganizada que se encontraba en la base de datos original, que obstaculizaba la migración de datos a las tablas SQL, por ejemplo, se tuvo que corregir el formatos que el motor Postgres no aceptaba adicionalmente tuvimos que revisar alternativas de inserción para una tabla que presento dificultades en la agregación avanzada, se tuvo que aprender a manejar Excel de una manera eficiente muchas veces se recurrió a sala Knuth ya que la inexperticia no permitía una solución en la limpieza de los datos, la creación de diferentes diagramas relacionales debido a que inicialmente no cumplían las expectativas, sin embargo, los integrantes del equipo lograron solucionar la problemática consiguiendo seguir con el objetivo. Para la parte del análisis de datos finales, el uso de Python y librerías con pandas y Dash fueron indispensables, ya que, a partir de estas, se puede entender computacionalmente el cambio que ha tenido el transporte aéreo desde 1908 en cuanto a seguridad, un ejemplo de esto son las muertes causadas dentro del avión en cada siniestro.

Link de Github para verificar los datos: Base de datos sobre los accidentes aereos