Proyecto Final Ing. De Datos

Gerónimo Rojas, Karen Meléndez y Daniel Alarcón

Repositorio:

https://github.com/kmelendez04/Proyecto_datos

Contextualización y planteamiento del problema

Para este proyecto vamos a utilizar una base de datos llamada Airline Dataset de la página kaggle.com en la cual se puede acceder fácilmente a base de datos sobre diferentes temas. Esta tabla proporciona una gran cantidad de datos relacionados con operaciones de aerolíneas y datos de pasajeros. La tabla ofrece información sobre aviones, aeropuertos, pases de abordar, reservas, vuelos, tiquetes y asientos.

Airline Database, es un recurso valioso para las aerolíneas y vendedores de servicios turísticos. Ya que se podría analizar los destinos con más rentables, los asientos más vendidos, los modelos de los aviones más usados, etc.

Cabe destacar que esta base de datos no solo es valiosa por la amplitud de datos que ofrece, sino también por su relevancia y actualidad. Proporciona información en tiempo real sobre aviones, rutas, pasajeros y otros aspectos fundamentales. Además, su documentación se distingue por su accesibilidad y claridad. Está cuidadosamente organizada de tal manera que facilita al lector la comprensión de las complejas relaciones entre entidades, atributos y tablas. De esta forma, se convierte en una herramienta esencial para el análisis y la toma de decisiones informadas en el ámbito de la aviación y el turismo.

Reglas de negocio

1. El avión (aircrafts_data) tendrá un código único (aircraft_code)
2. El avión (aircrafts_data) tendrá una distancia máxima que puede recorrer (range).
3. Los aeropuertos (airports_data) tendrán un código único (airport_code)
4. Los aeropuertos (airports_data) tendrán un nombre (airport_name), ciudad donde están ubicados (city), y zona horaria del continente donde están ubicados (timezone).
5. Las reservas (bookings) tendrán una referencia única (book_ref).
6. Las reservas (bookings) tendrán fecha de reserva (book_date) y precio total (total_amount)
7. El vuelo (flight) tendrá un id único (flight_id)
8. El vuelo tendrá el número de vuelo (flight_no), aeropuerto de salida (departure_airport_code), aeropuerto de llegada (arrival_airport_code), estado (status) y código del avión (aircraft_code)
9. La acomodación tendrá un id único (accomm_id)
10. La acomodación tendrá el tipo de acomodación (type)
11. La acomodación estará restringida para que solo pueda tomar uno de los siguientes valores: Business, comfort o economy
12. La silla (seats) tendrá un número único (seat_no)

13. La silla (seats) tendrán el código del avión (aircraft_code) y el id de la acomodación (accomm_id)
14. El tiquete de vuelo (tiquet_flight) tendrá un número único (ticket_no)
15. El tiquete de vuelo (tiquet_flight) tendrá la referencia de la reserva (book_ref), el id de la acomodación (accomm_id), el id del vuelo (flight_id) y el costo (amount).
16. El tiquete de pasajero (tickets) tendrá id perteneciente al pasajero único (passanger_id)
17. El tiquete de pasajero (tickets) tendrá el número de tiquete (ticket_no) y la referencia de reserva (book_ref)
18. Los pases de abordar (boarding_passes) tendrán un numero único (boarding_no)
19. Los pases de abordar tendrán el número de tiquete (ticket_no), el id del vuelo (flight_id) y numero de asiento (seat_no)
20. Los aeropuertos podrán tener uno o más aviones.
21. Uno o más asientos pueden tener la misma acomodación
22. Uno o más tiquetes pueden tener la misma acomodación

Identificación y descripción de las entidades, relaciones y atributos

Con base a los datos, se pudo establecer las siguientes entidades con sus respectivos atributos:

- **aircraft_data**: aircraft_code , range
- **airports_data**: airport_code, airport_name, city, timezone
- **bookings**: book_ref, book_date, total_amount
- **flights**: flight_id, flight_no, departure_airport_code, arrival_airport_code, status, aircraft_code
- **accommodation**: accom_id, type
- **seats**: aircraft_code, seat_no, accom_id
- **ticket_flights**: ticket_no, flight_id, accom_id, amount
- **tickets**: ticket_no, book_ref, passenger_id
- **boarding_passes**: ticket_no, flight_id, boarding_no, seat_no

Diagrama Entidad Relación:

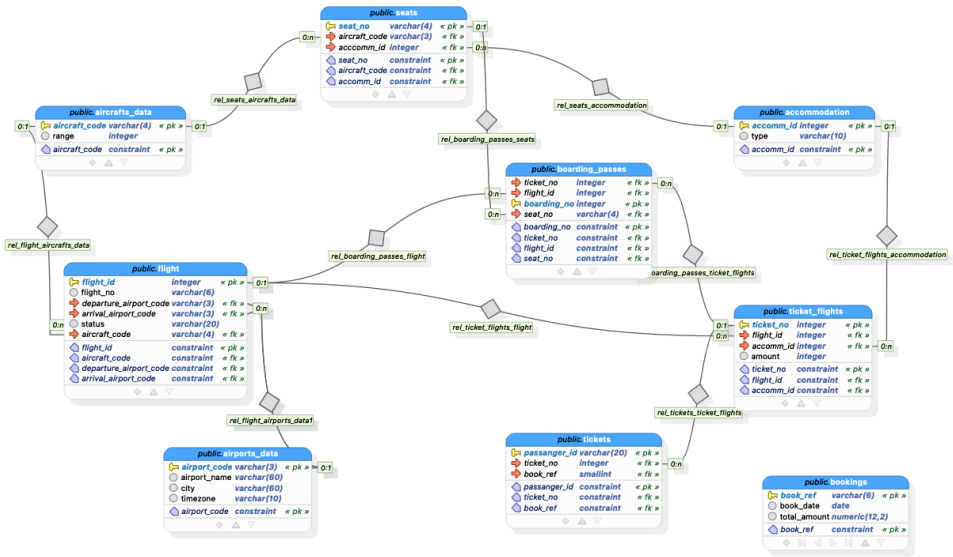
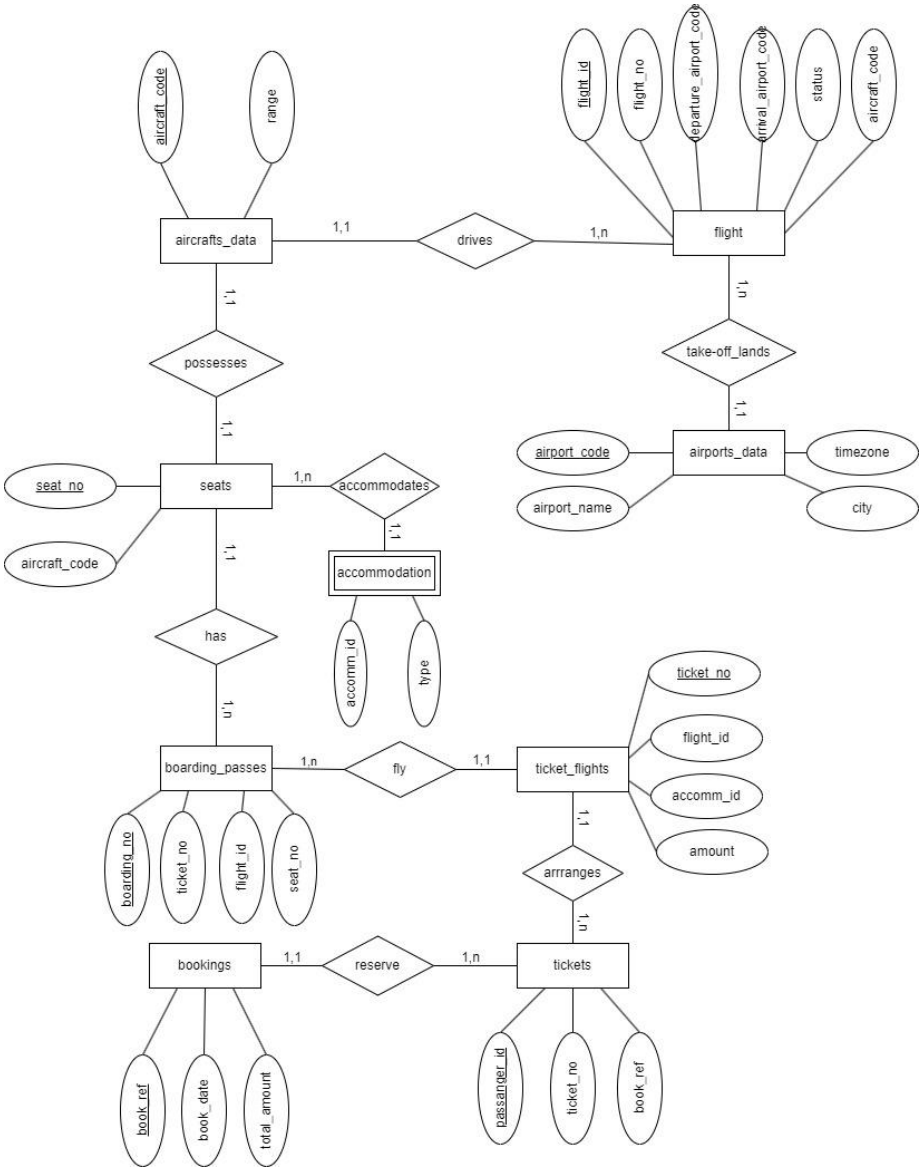


Diagrama Relacional:



Proceso de carga de información:

Se elige una base de datos llamada *airline_daataset,* la cual tiene un total de 8 tablas, pero al momento de normalizarla en un Excel quedó finalmente con un total de 9 tablas. Para hacer la carga masiva de los datos se convierten los archivos de las tablas a formato CSV, para luego utilizar la sentencia Copy en sql.

Antes de hacer el proceso de carga masiva se tuvieron que cambiar algunos datos ya que en la base de datos original ciertos datos a los que se les hacía referencia como foreign key no estaban presentes en la tabla original donde estaban como primary keys, por lo tanto producia problemas a la hora de hacer la carga.

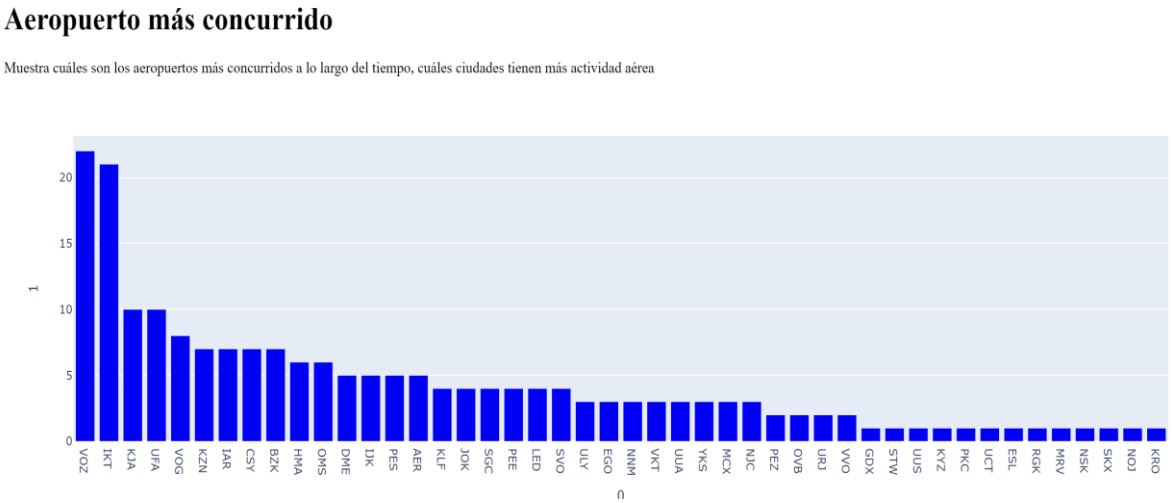
Además, se creó la tabla de accommodation con una columna de ID único y los tipos de acomodación, para relacionarla con las tablas seats y ticket_flights, pues en ambas tablas la columna accom_id se puede repetir varias veces. La razón por la cual no se hizo este proceso

con los códigos de los aeropuertos es que al ser referenciada en la tabla flight, cada vuelo es diferente. Si se hiciera este proceso se podría perder información sobre el vuelo y por consiguiente se perdería información en las tablas relacionadas con esta.

Discusión escenarios de análisis

Sean los análisis específicos por personas denotadas de la siguiente manera: G (Gerónimo Rojas), D (Daniel Alarcón), K (Karen Meléndez).

1. Determinar mediante los datos en que aeropuerto despegan y aterrizan mayor cantidad de vuelos, teniendo en cuenta todos los vuelos que están en la base de datos. Obteniendo así las zonas más concurridas y que tienen más movimiento aéreo a lo largo del tiempo.



Ventajas y desventajas de las gráficas implementadas:

- Usar un gráfico de barras en la comparación de aeropuertos y el tránsito en ellos es una forma óptima de observar cuál de estos lugares es el más concurrido. Aquí podemos observar fácilmente en el eje x cada código de los aeropuertos y en el eje y la cantidad de tránsito que hay en estos. Sin embargo, al tener tantos aeropuertos puede que sea difícil diferenciar los códigos en la parte inferior con este gráfico.

Análisis del Grupo:

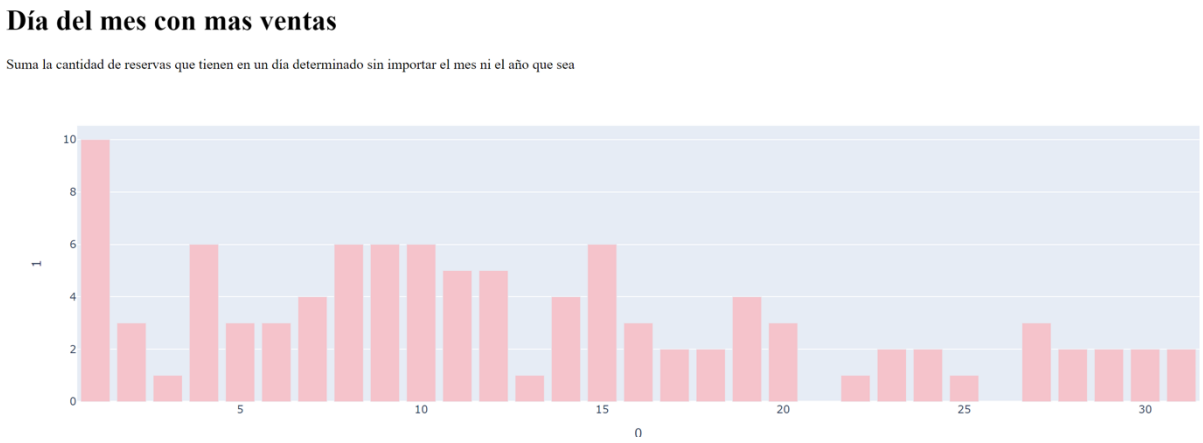
G: En la gráfica se puede evidenciar los aeropuertos más concurridos son VOZ e IKT, los cuáles son aeropuertos ubicados en Rusia. Lo que nos hace pensar que estas zonas en Rusia son bastante transitadas, puede ser como sitio turístico o como aeropuerto para hacer escalas, esto comparándolos con el resto de los aeropuertos que se pueden ver en la base de datos. Los terminales aéreos mencionado anteriormente cuentan con una gran ventaja con respecto a los demás, esto se puede ver cuando se nota que estas dos duplican a sus sucesores que son los aeropuertos con códigos KJA y UFA que cuentan con 10, ne cambio VOZ cuenta con 22 e IKT con 21.

K: Como se puede observar en la gráfica, el aeropuerto con código VOZ (Voronezh International Airport) es el más transitado, con un conteo de 22 veces mencionado entre despegues y aterrizajes. Sin embargo, el aeropuerto con código IKT (Irkutsk Airport) está cerca con un conteo de 21. A comparación de los otros aeropuertos, es notable destacar que

estos dos, doblan y hasta triplican el tránsito en lugares como Yemelyanovo Airport (KJA) y Domodedovo International Airport (DME).

D: Para el tercer y cuarto puesto hay un empate entre el aeropuerto de Krasnoyarsk (KJA) y el aeropuerto de Ufá (UFA) siendo ambos aeropuertos 10 veces contados. Estos dos aeropuertos rusos sumando en conjunto de las veces que fueron contados, no llegan a siquiera a pasar el segundo aeropuerto que es el de Irkusysk (IKT), también ruso. Esto claramente representa una preferencia significativa de viaje en los pasajeros que vuelan en el país. Y podemos apreciar en la gráfica, que esta tendencia es ligeramente decreciente en los siguientes aeropuertos.

2. Determinar en un ranking que día del mes se hacen más reservas. Esto ubicándolos del mayor al menor sumando la cantidad de reservas que tienen en un día determinado sin importar el mes ni el año que sea. Así se podría determinar en qué días del mes se debería hacer más publicidad para llegar a la audiencia en esas fechas y que la publicidad sea más efectiva y se compren más tiquetes.



Ventajas y desventajas de las gráficas implementadas:

- Usar un gráfico de barras es óptimo para ver los días del mes cuando más se hacen reservas, ya que se puede visualizar fácilmente en el eje x los días del mes y en el eje y el número de reservas. Aunque es claro cuál eje representa las variables, puede ser confuso si el máximo de reservas fuera 31.

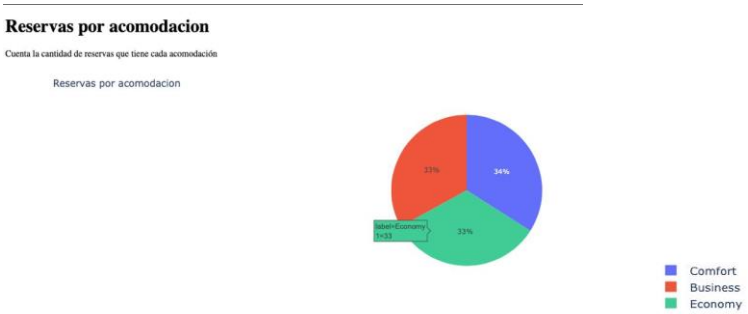
Análisis del Grupo:

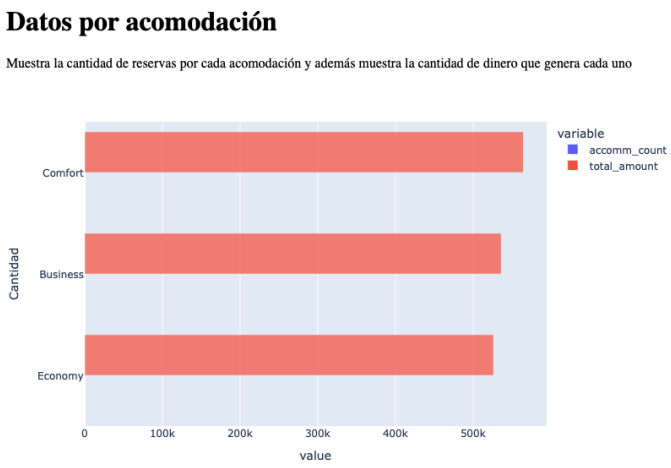
G: En la gráfica se puede ver que el día que más se hacen reservas es 1 de cada mes con 10 reservas, aunque la gráfica sea un poco irregular y no se ve una tendencia muy clara, podemos notar que los primeros 15 días del mes se hacen más reservas, mientras que en la segunda quincena no hay días con más de 4 reservas mientras que en la primera hay 8 días con más de esta cantidad de reservas.

K: Como podemos ver en la gráfica, el día que más se hacen reservas es el primero de cada mes. El cual alcanza 10 reservas a comparación del resto de días que llegan máximo a 6 reservas. Esto puede ser debido a los días de pago o descuentos, pero para saber en realidad por qué sucede esto tendríamos que realizar otras investigaciones basándose en los días de pago de las personas o los descuentos realizados por las aerolíneas.

D: El comportamiento en reservas es irregular, pues el comportamiento no sigue ningún patrón. Pero, al menos es posible afirmar que durante los primeros 15 días del mes es cuando se hacen la mayor cantidad de reservas y existe al menos una reserva diaria. Y puede ser una coincidencia que el promedio en cantidad de reservas para tiquetes aéreos en días impares es mayor que los días pares o, por otro lado, puede ser porque las personas tienen una mayor capacidad de endeudamiento los primeros días del mes.

3. Analizar cuales acomodaciones son las que más se compran, y de esta manera ver la diferencia de ingresos que genera cada una de las acomodaciones. Si business al ser más costosa genera más ingresos a comparación de las demás o si por la cantidad tan reducida de sillas que hay se generan menos ingresos que en comfort o economy.





Ventajas y desventajas de las gráficas implementadas:

- El pie chart es una forma efectiva de ver las reservas por acomodación, se puede ver fácilmente el número de estas. Además, como los números son tan parecidos en otro tipo de gráfico, sería difícil diferenciar la cantidad de reservas. Por lo contario el grafico de barras es efectivo para ver la cantidad de ingresos totales que genera cada acomodación no son números enteros como en el pie chart, por lo tanto, es más fácil ver la cantidad de dinero.

Análisis del Grupo:

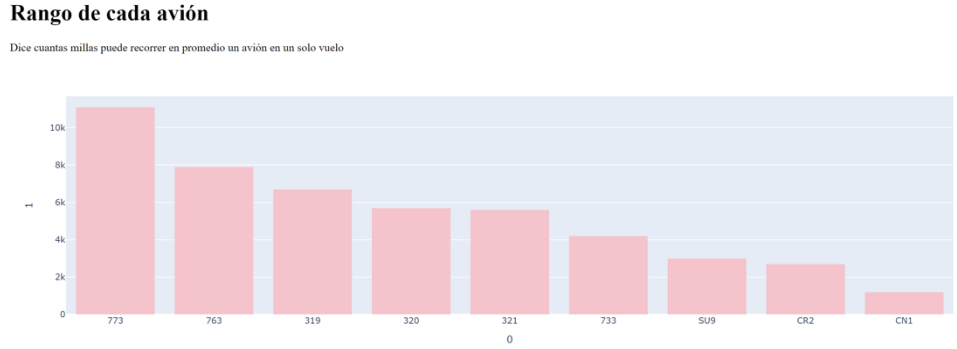
G: De las dos gráficas se puede analizar que tanto en el número de reservas y los ingresos son muy similares. Comfort es la que tiene más comprar y también la que más ingresos genera,

por otra parte, aunque business y economy se compren en la misma cantidad se puede evidenciar que business al ser una clase más costosa genera más ingresos que el economy.

K: En la gráfica podemos ver que todas las acomodaciones tienen una cantidad de compra similar, sin embargo, la acomodación número 3 (economy) es la más comprada con un total de 34 compras y las otras dos acomodaciones tienen 33 compras. Sin embargo, en la segunda gráfica se puede ver claramente que la acomodación número 3 (comfort) genera más ingresos ($564800.00).

D: La gráfica nos demuestra que los pasajeros no tienen gran preferencia al momento de escoger asientos, de hecho, la mayoría opta por comprar la que es más "equilibrada" porque la categoría Comfort no es tan estrecha como Economy, pero tampoco tan extravagante como lo puede llegar a ser "Business". Es interesante notar que la cantidad de personas que compraron en Business es la misma que quienes adquirieron sus tiquetes en Economy, pues esto puede significar que la diferencia en precios tampoco es mucho. Ya que, los precios por lo general de la clase Economy son más del doble o triple que un economy, pero en este caso solo hay $9500 de diferencia en ganancias para la compañía.

4. Determinar que avión hace los vuelos más largos viendo los cambios de zonas horarias. Con un ranking del range que tiene cada modelo de avión mirar si esto afecta en todos sus casos.



Ventajas y desventajas de las gráficas implementadas:

- Al usar diagrama de barras se puede ver la diferencia entre los aviones que hacen más cambios de horario, pues en la base de datos hay dos zonas horarias, Asia y Europa. También se puede ver la diferencia entre los rangos de los aviones.

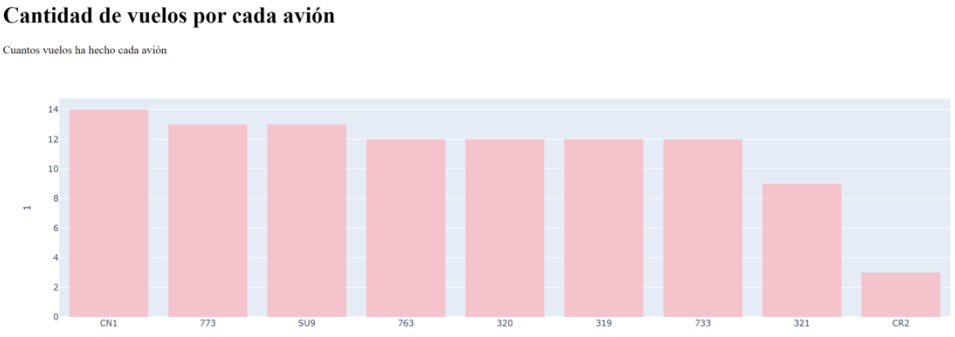
Análisis del Grupo:

G: Analizando ambas gráficas se puede ver que el avión con más cambios de zonas horarios que es el 773 también es el avión que más rango, esto también coincide en el segundo lugar con el 763 pero es muy raro el caso del CN1 que es el tercer avión con más cambios de zona horaria, pero es el avión con menos rango por lo que al revisar por qué cambia tanto de zona horaria se notó que hace vuelos cortos cambiando de la zona europea de Rusia a la de Asia por lo que es en distintas zonas horarias.

K: En las gráficas podemos observar que el avión que hace más cambios horarios, en otras palabras, que hace los viajes más largos es el que tiene el código 773. Al compararlo con la gráfica del rango podemos ver qué este mismo avión es el que puede recorrer más distancia. El segundo avión que más cambia de zona horaria también es el segundo con mayor rango (763). Por el contrario, uno de los aviones con menos rango, el CN1, es el tercero con mayor cambio de horario. Por esta razón, no se puede concluir que hay una relación directa entre los aviones con mayor recorrido o cambio de zona horaria y el rango de los mismos. Se tendría que analizar la ubicación de los aeropuertos, los aviones más comunes en estos lugares y otros factores para poder llegar a una conclusión sobre la relación entre los dos factores analizados.

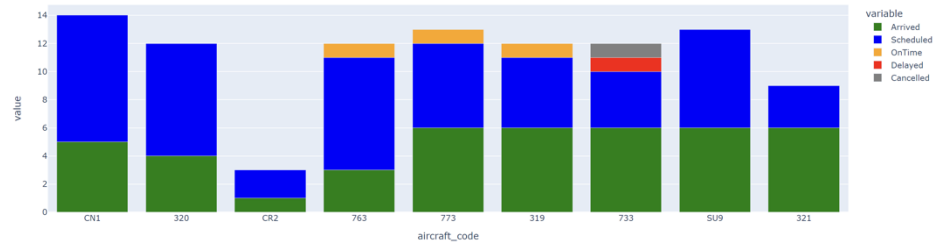
D: Se puede evidenciar que las avionetas, como la CN1, CN2 y los aviones pequeños como el SU9 (que abarcan menos de 150 pasajeros), son las que tienen menor cantidad de millas en recorrido, pero no necesariamente hacen menor cantidad de cambio de horario que los aviones grandes como el 733. De hecho, están muy a la par considerando su tamaño. Adicionalmente, los aviones viejos como el 733 y el 321 hacen un recorrido en millas aceptable pero el cambio horario no es tan frecuente como las avionetas. También, podemos notar el caso especial de los súper jets SU9, que tienen una cantidad elevada de millas recorridas y al ser más veloces y atraviesan una mayor cantidad de distancia en menos tiempo, puede que sea la razón para atravesar con mayor frecuencia cambios horarios.

5. Determinar que aviones son los más usados, y ver cuál es el que tiene más vuelos por cada estado, es decir, si está cancelado, agendado o si ya aterrizó.



Cantidad de vuelos por cada avión teniendo en cuenta cada estado del vuelo

Muestra cuántos vuelos tiene cada aeronave en cada estado (despegó, aterrizó, a tiempo, demorado, cancelado)



Ventajas y desventajas de las gráficas implementadas:

- Los diagramas de barras son útiles para analizar la cantidad de vuelos entre los diferentes tipos de aviones y para mostrar fácilmente la diferencia entre los estados del vuelo. Sin embargo, en la segunda en la cual se comparan los estados, los colores pueden ser un poco confusa para algunas personas, así que otra opción puede ser hacer un pie chart para cada avión y estado.

Análisis del Grupo:

G: En las gráficas se pueden ver la cantidad de vuelos que ha tenido cada tipo de avión y además los estados de vuelo que tiene cada uno de los aviones. Se puede ver un claro dominio de los vuelos que aterrizaron y los que están agendados, mientras que los que estan a tiempo, retrasados o cancelados están presentes en una menor proporción. En el caso de del 733 se puede ver que es el único tipo de avión que tiene vuelos retrasados y cancelados.

K: En las gráficas podemos ver que la mayoría de los aviones tienen la misma cantidad de vuelos que llegaron. Sin embargo, sí se puede ver una diferencia entre los vuelos programados. Pues se nota que los aviones que han hecho más vuelos tienen más estados como a tiempo, cancelado o demorado.

D: Aquí podemos notar que el tamaño sí hace la diferencia, pues las avionetas y el SU9 al tener una menor capacidad de carga, no tienen problemas de cancelación o retraso haya el momento. Es curioso notar que el CN2 tiene considerablemente una menor cantidad de vuelos en ruta que el predecesor CN1, en general, los aviones tienen una mayor cantidad de vuelos realizados de los que han sido agendados a excepción del SU9, probablemente por ser de uso militar, tiene que ser un uso equilibrado y ya pensado con mucha antelación. Adicionalmente, podemos observar que los aviones viejos como el 319 y 320 tienen casi la misma cantidad de vuelos a realizar que los aviones grandes como el 733. En general, las avionetas de uso personal o militar tienen un mayor uso que los aviones de uso civil.

Conclusiones

1. Los aeropuertos más transitados son aquellos con código VOZ (Voronezh International Airport) e IKT (Irkutsk Airport) mencionados 22 y 21 veces respectivamente en la base de datos. Estos dos aeropuertos doblan en número de menciones con respecto a los otros de la gráfica.
2. Los datos sugieren que los primeros 15 días del mes es cuando más se realizan reservas de vuelos. Llegando a hasta 10 reservas el primer día del mes.

3. También se puede decir que las acomodaciones se compran en cantiades similares, pues la 1 y la 2, *bussines* y *comfort* respectivamente se reservan un total de 33 veces, mientras que la 3, *economy*, se reserva un total de 34 veces. Esta acomodación es la que genera más ingresos totales.
4. Por último, hay una relación entre los aviones con más rango y la cantidad de veces que cambian de zonas horarias. Sin embargo, cuando el rango de los aviones no se diferencia tanto, la distancia recorrida tampoco lo hace.

Apreciaciones Finales

G: La base de datos elegida fue muy apropiada para el desarrollo del proyecto, fue muy breve normalizar las tablas que tenía la tabla debido a que estaban muy bien resumidas y muy fáciles de entender. Además, la base permitió hacer consultas en SQL que dejaban aplicar muchos de los conceptos que se fueron aprendiendo en el curso de Ingeniería de Datos, se permitió el uso de un nuevo concepto como lo es el COALESCE y ayudó a poner en práctica los conocimientos que se obtuvieron.

K: La base de datos nos facilitó la normalizacion de las tablas ya que la mayoria de datos ya estaban organizados, la información que nos proporcionó esta base fue fácil de entender y analizar. Sin embargo, a la hora de pasar los datos con carga masiva de datos nos dimos cuenta de que había algunos errores en los datos como tal, así que tuvimos que hacer cambios para poder ejecutar la función de *copy* en SQL.

Considero que la parte más difícil fue usar DASH para crear las gráficas, ya que se debía instalar diferentes extensiones desde la terminal. Sin embargo, todos estos procesos nos permitieron profundizar lo aprendido en clase de una forma efectiva.

D: Este tipo de base de datos es ideal para el aprendizaje, ya que los conceptos se pueden dimensionar fácilmente y para la comprensión de los datos no es necesario tener conocimientos técnicos. Los análisis realizados pueden ayudar a cualquier consumidor a sabes escoger de manera pertinente cuándo puede ser mejor reservar un vuelo e incluso qué avión puede ser de más confianza**.**

Gracias a la materia de Ingeniería de Datos fue posible amoldar esta información que sin manipulación puede ser insignificante o tediosa de analizar. Gracias a las funciones aprendidas en PostgreSQL fue posible ordenar, clasificar y visualizar datos de registro público de manera eficaz.

https://www.kaggle.com/datasets/mohammadkaiftahir/airline-dataset