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:

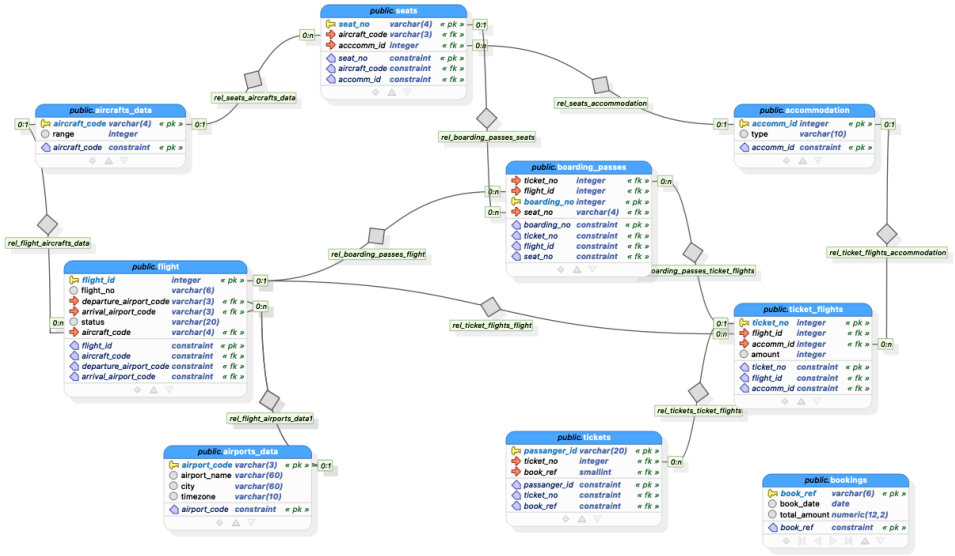
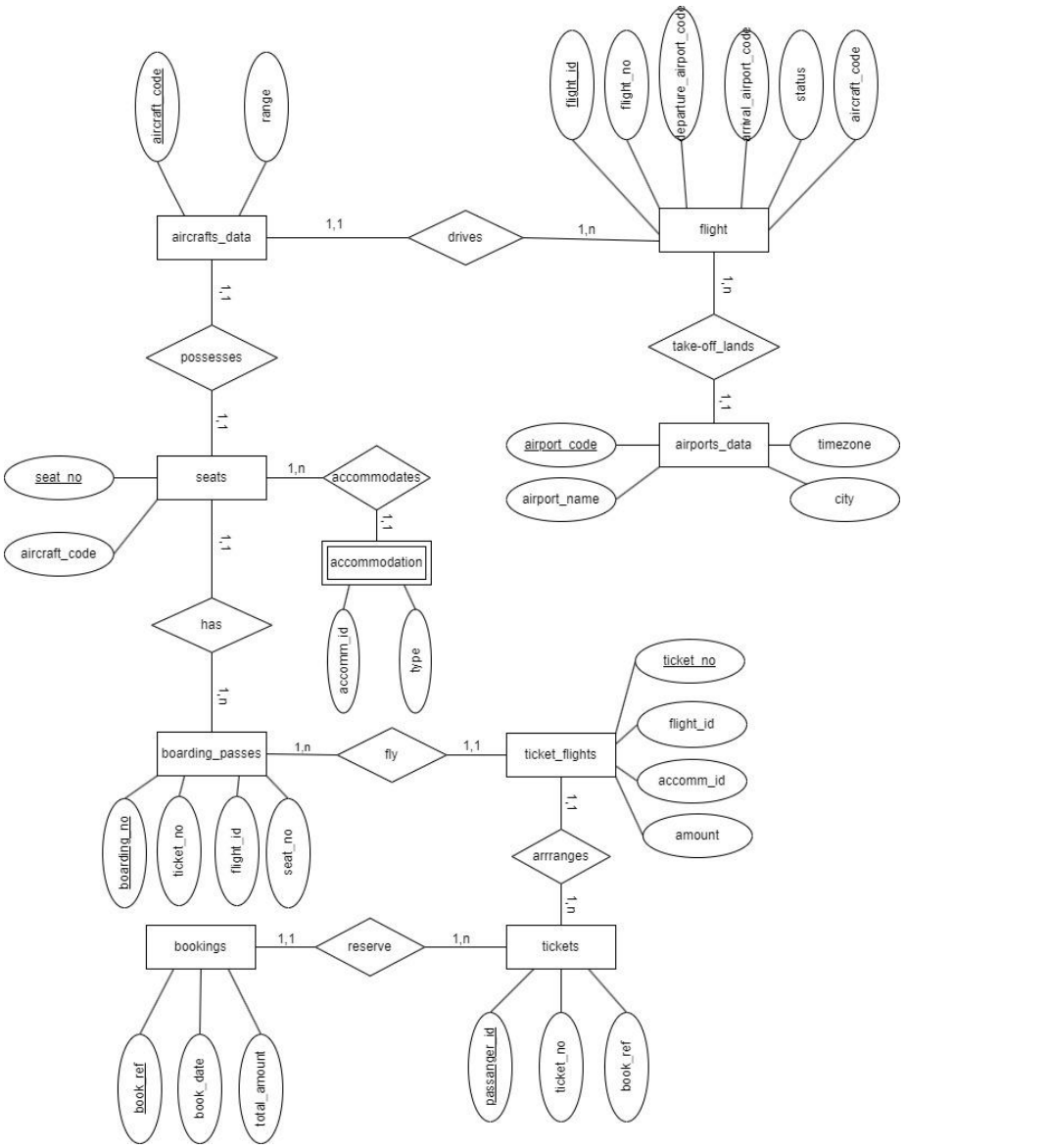


Diagrama Relacional:



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.

Identificar y describir al menos cuatro posibles escenarios de análisis que podrían realizarse con los datos cargados en la base de datos.

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.
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.
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.
4. Determinar que avión hace los vuelos más largos viendo los cambios de zonas horarias o también se podría ver comparando la cantidad de kilómetros que hay entre ciudad y ciudad. Con un ranking del range que tiene cada modelo de avión mirar si esto afecta en todos sus casos.
5. Determinar que aviones son los más usados para realizar distintos vuelos, y ver cuál es el que tiene más vuelos por cada estado, es decir, si está cancelado, agendado o si ya aterrizó.

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