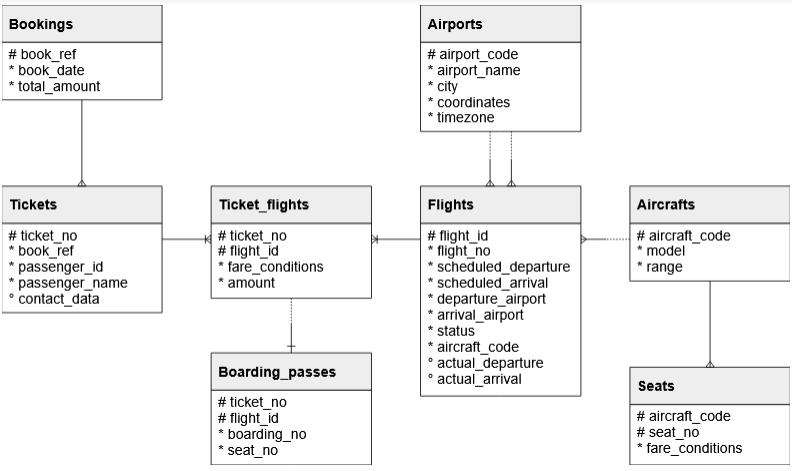
# Introduction

The aviation industry is one of the largest industries worldwide, with *Statista* estimating it’s worth at 762.8 billion U.S. dollars in 2023 (<https://www.statista.com/markets/419/topic/490/aviation/#overview>), with consistent growth year on year (barring the COVID-19 period). It goes without saying that analysing the trends in industry data is of high importance to companies operating in the sector.

In this project, I will look at data sourced from the *postgres demo database* (<https://postgrespro.com/community/demodb>). I will be looking at trends and correlations in the following factors: **Ticket price, Ticket fare type, Flight duration,** and **Ratio of sold vs unsold tickets**. I will be aiming to showcase these in a dashboard that would allow industry professionals to get an overview of our dataset, whilst also allowing the user to isolate flight data of their choice.

# Data Source and Environment

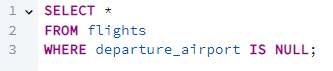
Data is sourced from the *PostgreSQL Demonstration Database* sourced from: <https://postgrespro.com/community/demodb>. The database is made up of 8 tables as described in the schema diagram below:

Based on our objectives, we will be utilising the *Flights*, *Ticket\_flights* and *Seats* tables primarily. We will also be sourcing airport location data, using data sourced from [**http://www.ip2location.com**](http://www.ip2location.com).

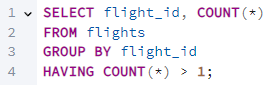
We will use *PostgreSQL* to explore our data primarily, utilise *Microsoft Excel* for the airport geographical data, and visualise our data in *Power BI*.

# Data Preparation

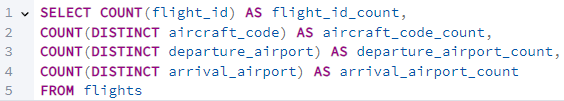
After importing the database, we used *SQL* code to check for *NULL* values in our columns of interest. Thankfully, there were no *NULL* values raised.



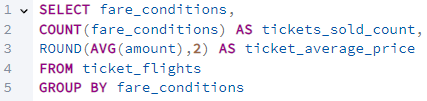
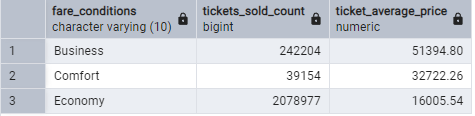
We also checked for duplicate rows by grouping by *flight\_id*, as this column should have a unique value for every row in our dataset. Again, no duplicates were raised.



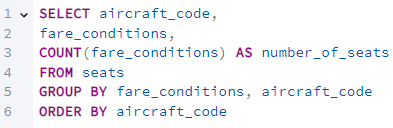
After this, we did some exploratory analysis on our clean database. Firstly, we looked into the number of *flight\_ids*, unique *aircraft\_codes*, and *unique airports*. We can see there are a total of *65,664 flights* in our database, belonging to *8 aircraft codes*, with *104 airports* being used as departure or arrival locations.

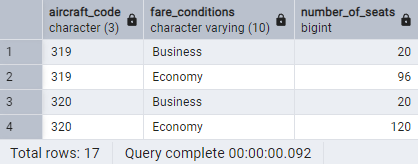


Next, we looked at the *ticket\_flights* table. We looked at what *fare\_conditions* there were, how many tickets were sold for that fare, and the average price of tickets for that fare. This showed we had 3 fares, *Business*, *Comfort*, and *Economy*. Based on the number of tickets sold, it appears that *Comfort* fares are not a feature on all flights, as the number of tickets sold for that fare is considerably less than the more expensive *Business* fare. We can see that the average ticket price increases as you go from *Economy* to *Comfort*, and *Comfort* to *Business*.



Lastly, we did some exploratory analysis of the *seats* table. This table shows the available seats for each *aircraft\_code*. We looked at the number of seats for each *fare\_condition* for each individual *aircraft\_code*. This confirmed the previous hypothesis that the *Comfort* fare is not on all aircrafts, with it only being available on flights with *aircraft\_code 773.* There are also two *aircraft\_codes*, *CN1* and *CR2*, that only have *Economy* seats available.





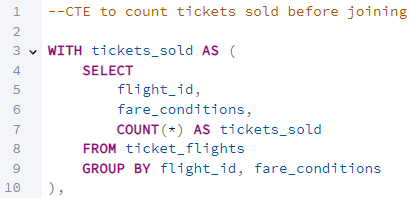
# Data Joining

After exploring our data, the next step was to join the data into a single comprehensive table with which we could import into Power BI to create our visualisations. The ideal join would give us a table with the following information:

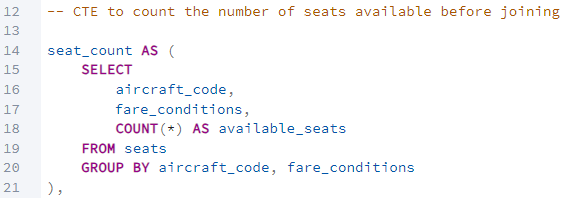
1. Include *flight\_id*, *aircraft\_code,* and *flight\_duration* to allow for filtering visualisations
2. Divide flights into different rows for each individual *fare\_condition*
3. Display the number of seats available in each *fare\_condition* and the number of seats filled
4. Track the *average cost* of tickets for that *fare\_condition* and *flight\_id*
5. Record the *departure* and *arrival airports*

In order to achieve all these goals in a single *SQL query*, we decided to use *Common Table Expressions* (CTEs) to help isolate sections of our query that need to be grouped from other selections that cannot be used with a *GROUP BY* statement.

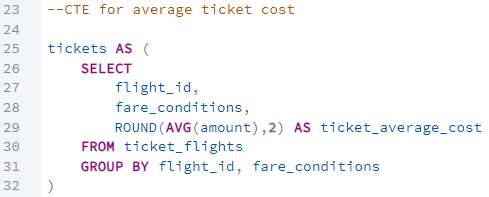
First of our CTEselections is counting the number of *tickets sold* grouped by *fare* and *flight\_id*.



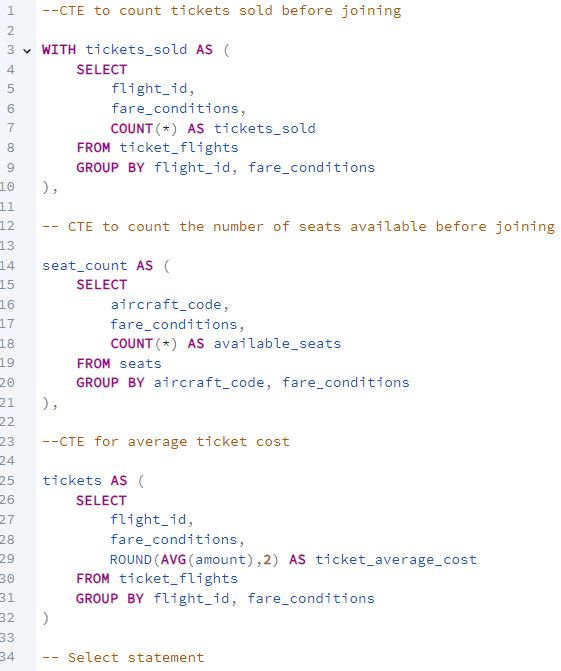
Next, we select the number of *available seats* grouped by *fare* and *aircraft\_code*.



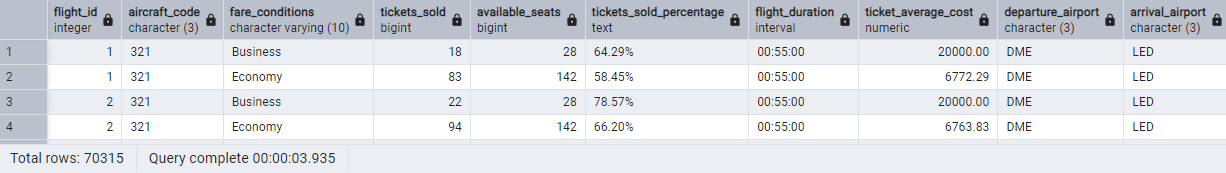
Our last CTE is to find the *average ticket cost* grouped by *fare* and *flight\_id*. We rounded this value to 2 decimal places given that it is a currency.



Following this, we simply use a *SELECT* statement to pick our desired columns from our *flights* table, joining the CTEs on the *GROUP BY* conditions. We also created a column to show the *percentage of tickets sold* to assist with our visualisations. Our final query and table are shown below:

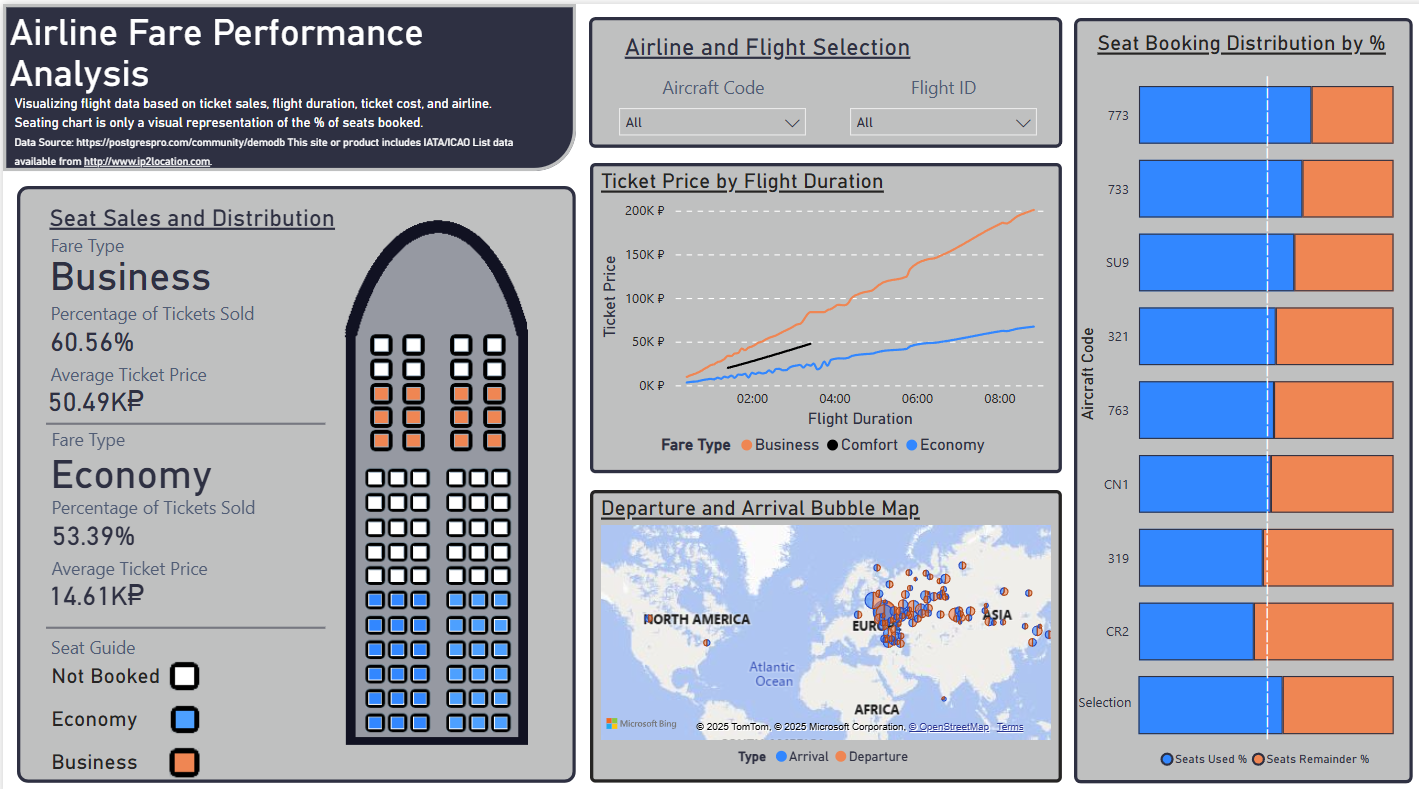






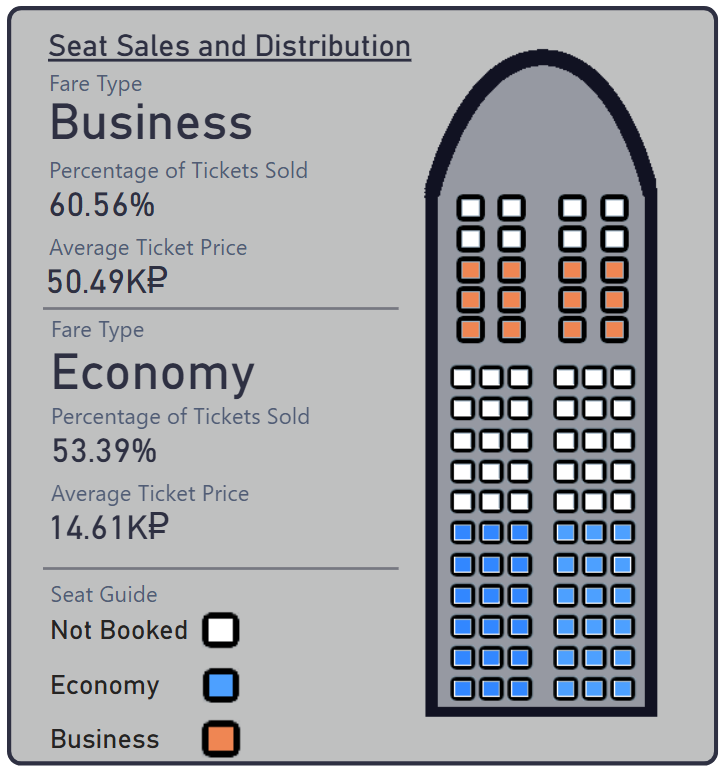
# Dashboard Design

We created our dashboard in *Power BI*, using a cool grey colour scheme with orange and blue being used for key data. Each visual was designed with a specific purpose in mind, which we will break down below.



## Seating Plan

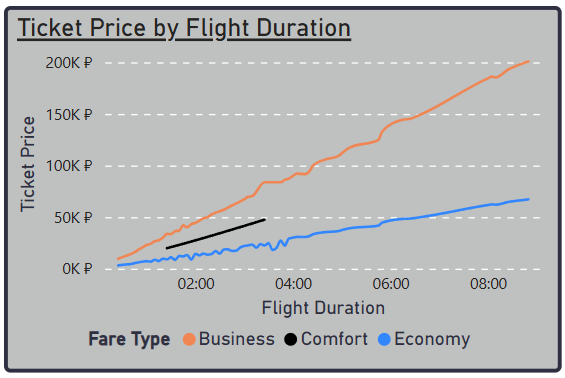
This visual contains information on the *percentage of tickets sold* for each fare, the *average price* that each ticket was sold for, and a visual representation of an airline seating plan showing the *percentage of tickets sold visually*.

The airline seating plan visual is designed to catch the viewer’s attention, whilst the statistics on ticket sales allow for a more analytical view of the data provided.

## Ticket Price by Flight Duration Line Chart

This visual provides a view of the relationship between *Ticket Price* and *Flight Duration*, divided by the different *fares* selected.

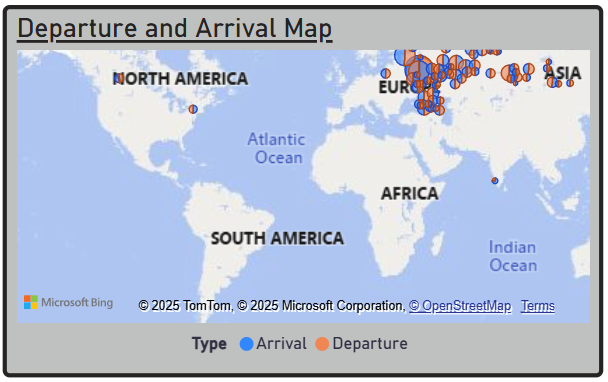
Whilst the visual is quite small in terms of space on the dashboard, the line chart provides an easy-to-read display of information that is particularly useful when filtering the data being viewed. Of note is the short *Comfort fare* line on the chart – This fare is only available on a few different airlines so it does not have a large representation of data. However, it was decided that it would be most appropriate to include this despite the smaller range of data.



## Departure and Arrival Map

This visual shows a representation of *flight departure* and *flight arrival* locations in a bubble map format. The larger the bubble, the more flights arrive or depart from that location. You can hover over the two segments of the bubble to see the exact number of flights for that location.

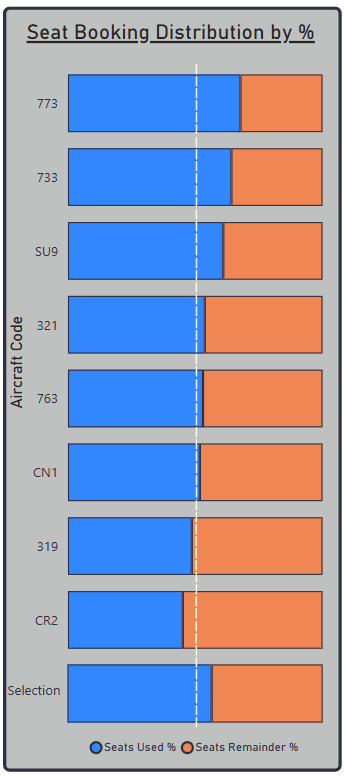
Another smaller visual, the map doesn’t take much space on the dashboard as it is not always of interest to the user. However, it can allow the user to get an overview of where exactly their selection is based.



## Seat Booking Distribution Stacked Bar Chart

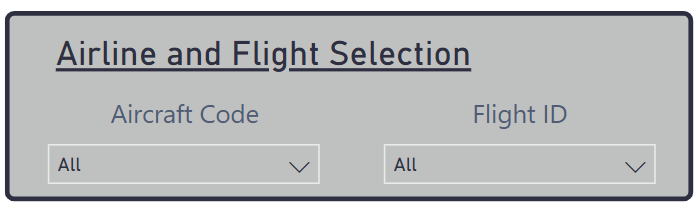
The last visual, our stacked bar chart, compares the *percentage of seats booked* against the *percentage of seats not booked*, divided by airline. The bottom bar shows these statistics for the filtered selection. The white dashed line shows the 50% mark.

The purpose of this visual is to give an overview of the dataset, whilst still allowing the user to filter their own selection to focus onto data of interest.



# Filtering and Interactivity

The main source of data filtering in our dashboard is with the *Aircraft Code* slicer and *Flight ID* slicer. Users can select any combination of *Aircraft Codes*, and then can further filter based on *Flight IDs* from the selected airlines. These slicers affect all visuals.



The user can also filter data by selecting location bubbles on the map visual.

# Limitations and Further Research

There are a number of limitations of our project developed here that could be improved if someone wanted to conduct a more developed project:

1. Geographical Scale

Our data is mostly sourced from Eastern-Europe and Asia. If data was sourced from a greater number of airlines around the world, more reliable conclusions could be drawn on a global scale.

1. Airline Quantity

Even within the smaller geographical scale mentioned above, the quantity of airlines in our dataset is small. Increasing the number of airlines we source data from could make for more comprehensive conclusions.

1. Airline Classifications

Further classification of our airlines could allow for more deep examination of our dataset and more robust conclusion making. For example, splitting our airlines into *Luxury*, *Regular* and *Budget* airlines would give us an additional layer of depth to our data.

**Must include; This site or product includes IATA/ICAO List data available from** [**http://www.ip2location.com**](http://www.ip2location.com)**.**