MSc Data Science

Coursework Assessment 2:
CI7320 Database & Data Management

Title: - Flight Punctuality Exercise

Submitted by: -
Name: - Jay Santosh Dugad

TABLE OF CONTENTS

1.) BENEFITS OF BUILDING A DATA WAREHOUSE FOR THE DATA SET PROVIDED

Data about the punctual arrival and departure of aircraft coming and departing from various airports must be gathered, stored, and arranged to build a data warehouse for the punctuality statistics of UK airports. Information like the planned and actual arrival and departure timings, delays, cancellations, and the reasons behind them, along with other pertinent metrics, might all be included in this data.
By consolidating this data into a specific location, analysts and researchers will have access to thorough and trustworthy information for a range of uses.

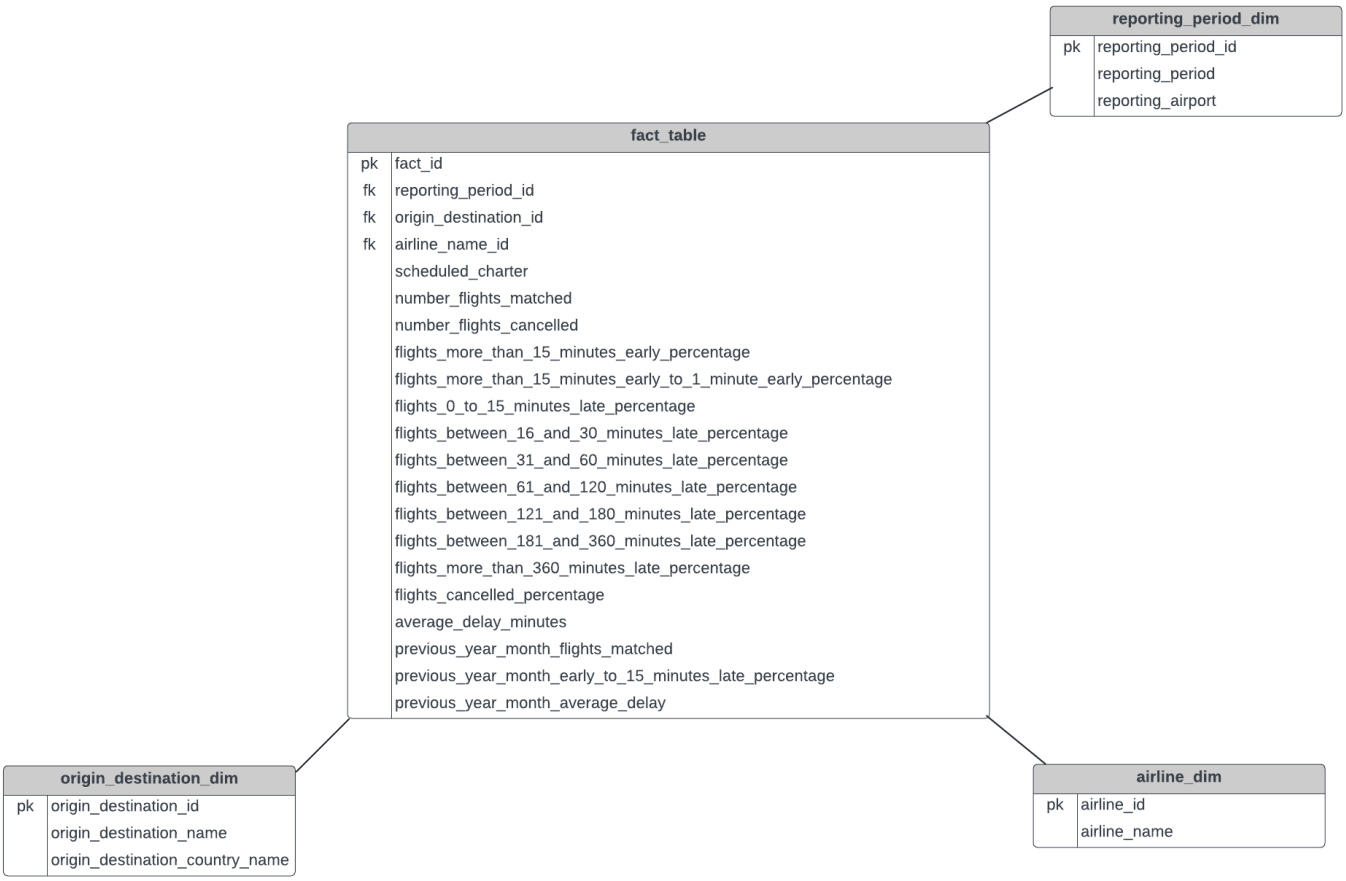
1. **Performance Analysis: -** By examining the historical trends in punctuality at various airports, researchers can spot trends and anomalies. By identifying areas for improvement, this research can assist airlines and airports in improving their operations. A data warehouse offers a single, easily accessible location for integrating data from multiple sources, such as combining punctuality information from several airports. A data warehouse's historical punctuality data allows for the investigation of patterns and trends over time. By identifying both short- and long-term trends in timeliness, this analysis can offer insightful information about how various airports operate.

2. **Enhancing the Passenger Experience: -** By understanding the statistics on punctuality, airports and airlines may better handle the expectations of their passengers. They can improve the general satisfaction and experience of passengers by giving accurate information about anticipated wait times and probable delays. By utilizing past data and performance patterns, airports and airlines can make well-informed decisions to better their operations, such as modifying flight schedules and allocating resources more efficiently to increase punctuality. Additionally, airlines and airports can take proactive steps to solve these problems by identifying areas for improvement, such as particular airports or times of the day with high rates of delays. For example, dedicating more resources during peak hours can help reduce delays.

3. **Data Integration: -** To guarantee quality and consistency, data from several sources is combined in the data warehouse. Through the integration of timeliness statistics with other data sets like meteorological conditions, aircraft schedules, or passenger demographics, scholars can acquire a more comprehensive understanding of the elements that impact airport punctuality. Researchers can examine how atmospheric conditions and timeliness statistics interact to determine how flight punctuality is impacted by weather patterns. For example, they are able to determine the relationships between unfavourable weather—like storms or dense fog—and longer wait times. By being aware of these trends, airlines and airports can proactively modify schedules or assign more staff to reduce the impact of weather-related delays.

4. **Environmental Impact Assessment: -** Fuel consumption and emissions linked to delays and inefficiencies, as well as other environmental effects of aviation operations, can be evaluated using punctuality data. Initiatives to promote sustainability and develop methods to reduce environmental impact can both benefit from this information. Airports and airlines can save expenses and increase operational efficiency by increasing timeliness. Reducing delays, for instance, can result in increased customer satisfaction, fuel savings, and greater resource usage. Initiatives for sustainability and the creation of strategies to lessen the environmental effect of aviation operations can greatly benefit from the insights gained from punctuality data.

2.)DATA WAREHOUSE USING STAR SCHEMA & JUSTIFY YOUR DESIGN



Link for Star schema diagram

In the above diagram PK is referred as Primary Key and FK is referred as Foreign Key.

JUSTIFICATIONS: -

Fact Table Selection: - The selection of the fact table includes quantitative data like the number of flights, the percentages of cancellations and delays, and the average delay times. These measurements are suited for the fact table since they may be combined and analysed. Because these metrics are simple to aggregate and interpret, they are appropriate for the fact table. A number of analytical processes, including aggregation, ratio calculation, and comparison, can be carried out by stakeholders by inserting them in the fact table in order to gain insightful information. Making better informed decisions is made possible by this configuration, which offers a thorough picture of the punctuality statistics.

- **Dimension Tables:-**

Reporting Period Dimension: - This dimension permits comparisons and trend analysis across several reporting periods as well as analysis over a range of time periods. This table also consists reporting_airport row as it reduces number of table which decreases complexity.

Origin-Destination Dimension: - The schema is made simpler by combining origin_destination and origin_destination_country into a single dimension. It reduces complexity by enabling analysis based on both the origin and the destination variables in a single table.

Airline Dimension: - By offering data about airlines, this dimension makes it possible to do analysis using metrics unique to each airline.

- **Primary & Foreign Keys: -**
  Each dimension table's primary key, usually called _id, is essential to preserving data integrity. It gives every record in the dimension table a distinct identity. In order to prevent duplication and guarantee the accuracy and consistency of the data, this uniqueness is essential. Furthermore, the fact table has foreign keys that reference each dimension table. By establishing link among the fact table and dimension tables, these keys enable drill-down and data aggregation along several dimensions. To create associations between the fact table and the dimension tables, foreign keys from the dimension tables are included in the fact table. This makes it possible to efficiently aggregate and analyse data across several dimensions. Analysis of data along many dimensions, including aggregation and drill-down, can be carried out by connecting the fact table to the dimension tables.

3.) CREATE TABLE STATEMENT FOR TABLE IN STAR SCHEMA

Airline Table (Dimension Table)

```
CREATE TABLE  airline_dim(
    airline_id INT PRIMARY KEY,
    airline_name VARCHAR(100)
);
```

Reporting period Table (Dimension Table)

```
CREATE TABLE reporting_period_dim(
    reporting_period_id INT PRIMARY KEY,
    reporting_period VARCHAR(100),
    reporting_airport_name VARCHAR(100)
);
```

Origin destination Table (Dimension Table)

```
CREATE TABLE origin_destination_dim(
    origin_destination_id INT PRIMARY KEY,
    origin_destination_name VARCHAR(100),
    origin_destination_country VARCHAR(100)
);
```

Fact Table

```
CREATE TABLE fact_table (
  fact_id INT PRIMARY KEY,
  reporting_period_id INT,
  origin_destination_id INT,
  airline_id INT,
  scheduled_charter VARCHAR(10),
  number_flights_matched INT,
  number_flights_cancelled INT,
  flights_more_than_15_minutes_early_percentage FLOAT(10),
  flights_more_than_15_minutes_early_to_1_minute_early_percentage FLOAT(10),
  flights_0_to_15_minutes_late_percentage FLOAT(10),
  flights_between_16_and_30_minutes_late_percentage FLOAT(10),
  flights_between_31_and_60_minutes_late_percentage FLOAT(10),
  flights_between_61_and_120_minutes_late_percentage FLOAT(10),
  flights_between_121_and_180_minutes_late_percentage FLOAT(10),
  flights_between_181_and_360_minutes_late_percentage FLOAT(10),
  flights_more_than_360_minutes_late_percentage FLOAT(10),
  flights_cancelled_percentage FLOAT(10),
  average_delay_minutes FLOAT(10),
  previous_year_month_flights_matched INT,
  previous_year_month_early_to_15_minutes_late_percentage FLOAT(10),
  previous_year_month_average_delay FLOAT(10),
  FOREIGN KEY (reporting_period_id) REFERENCES reporting_period_dim(reporting_period_id),
  FOREIGN KEY (origin_destination_id) REFERENCES origin_destination_dim(origin_destination_id),
  FOREIGN KEY (airline_id) REFERENCES airline_dim(airline_id)
);
```

4.) STEPS OF CREATING AND POPULATING DATABASE

Data cleaning: - To clean up data, every dataset must be carefully examined for mistakes or missing numbers. Data cleaning ensures the quality of the data by removing or resolving these issues. When analysing timeliness statistics for specific airport datasets in the UK, data cleansing is an essential step in guaranteeing the precision and dependability of the information. A thorough assessment and correction of mistakes and missing data will enable stakeholders to use the data with confidence for reporting, analysis, and decision-making. This is a crucial stage in creating a solid data warehouse. It guarantees the accuracy, consistency, and completeness of the data, providing a strong basis for further data analysis. Stakeholders may trust the insights gleaned from the data warehouse by cleaning the data, which promotes better decision-making and on-time arrival.

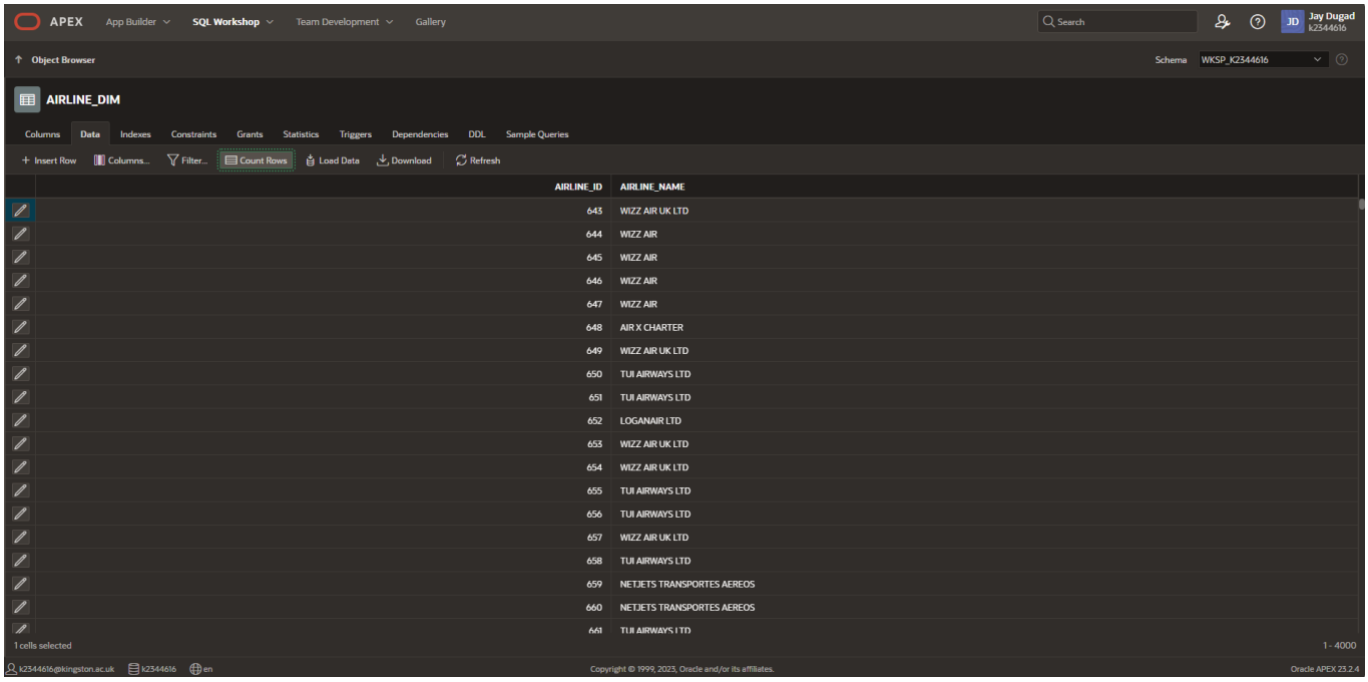
Integrating Data: - Combining or merging datasets from various sources into a single dataset is referred to as "integrating data." In order to give a uniform perspective of the data and enable thorough analysis and reporting, it entails combining the datasets. Deeper insights and improved decision-making are made possible by thorough analysis and reporting, which is made possible by a single dataset. For data integration I used a built-in feature of Excel that is the Data tab's ability to combine datasets. This approach is appropriate for smaller datasets and simple integration requirements. I used Power Query which is an effective data connection tool in Excel's Data tab, that lets users combine data from various sources, clean, transform, and reorganize it to meet complex integration requirements.

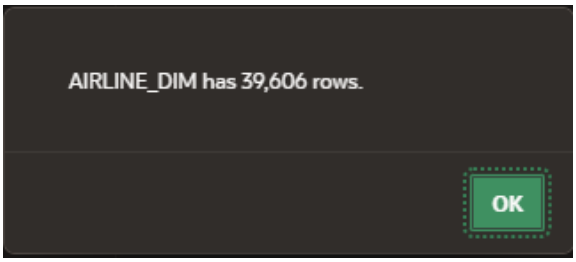
Export to CSV: - Exporting the data to a CSV file is the next step once it was being cleaned and prepared in Excel. Because CSV files are widely accepted and simple to construct, they are a common format for sharing data. This file format is straightforward and comprises of plain text with commas used to divide the values within each row of data. Each line represents a row of data. Data can be simply loaded into Oracle Apex for additional processing and analysis if it is exported to a CSV file.

Upload to Oracle APEX: - Using the SQL Workshop, I imported the CSV file into database table in Oracle APEX. Once the CSV file has been chosen, I mapped the columns to the matching columns in the database table. Because it guarantees that the data is successfully imported into the database, this step is essential. Then CSV file's contents were posted into the database table after the import was finished.

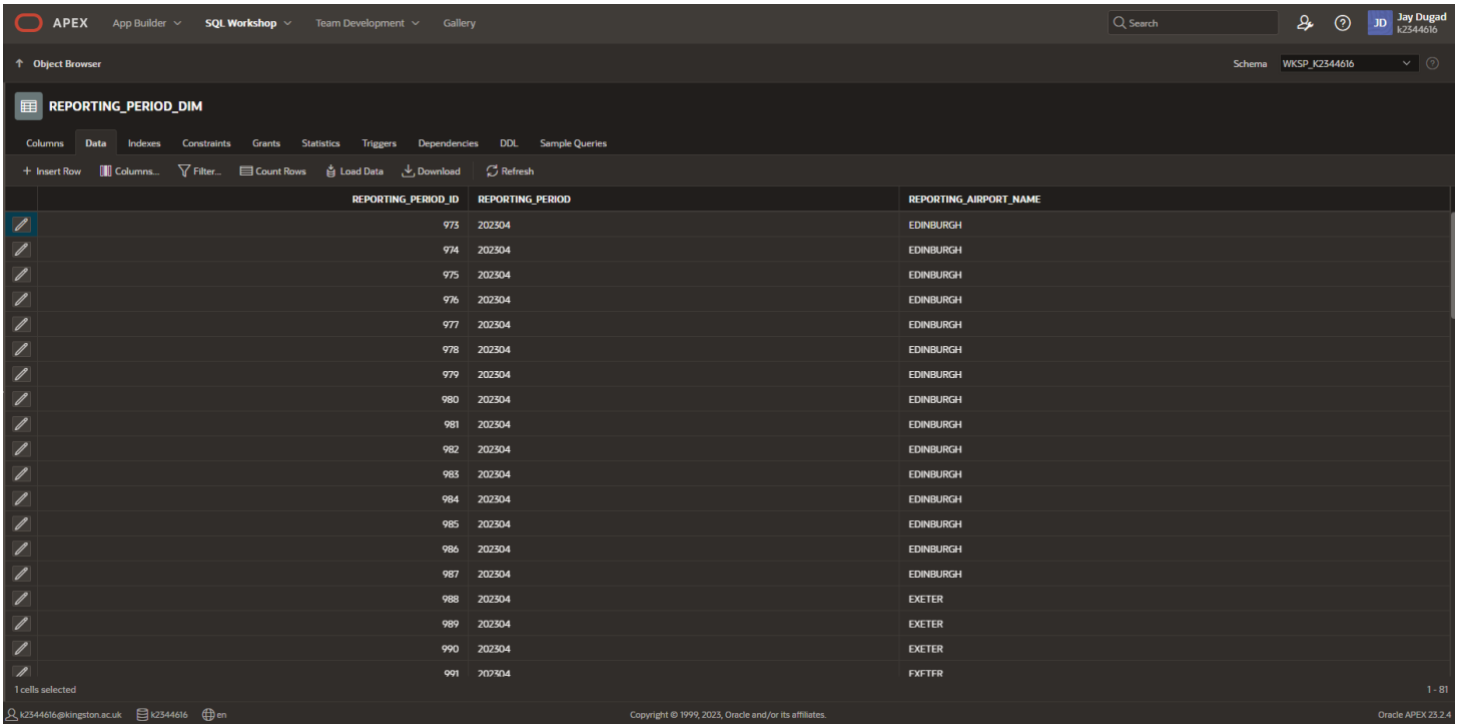
Data Validation: - An essential stage in the process of importing and uploading data into a data warehouse is data validation. It is crucial to confirm that the data satisfies the standards and has been appropriately imported after it has been posted. Validating data is carefully reviewing and confirming the information to make sure it is true, comprehensive, and consistent. To preserve data reliability and integrity, this procedure is necessary.

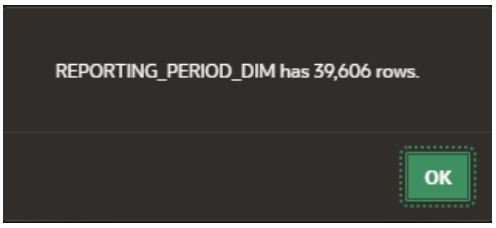
- **<u>Airline Table: -</u>** Airline table consists of airline_id as primary key. This table also includes column of airline_name. This table consists of 39,606 rows of data.



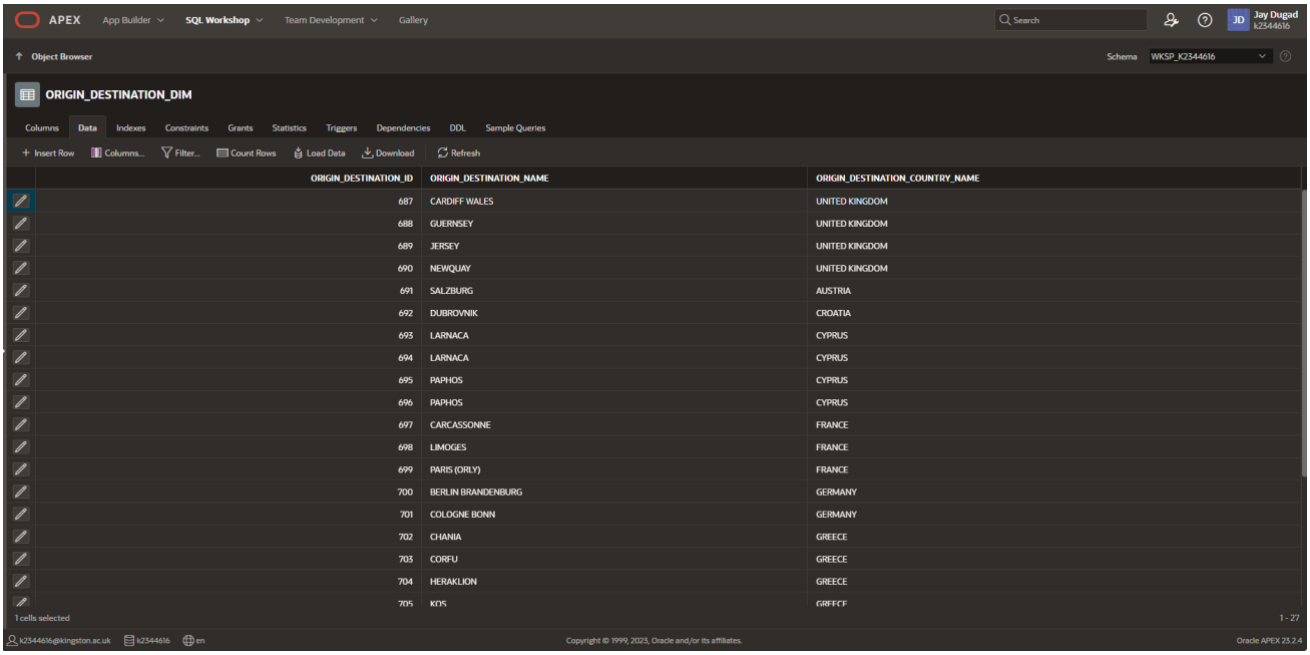


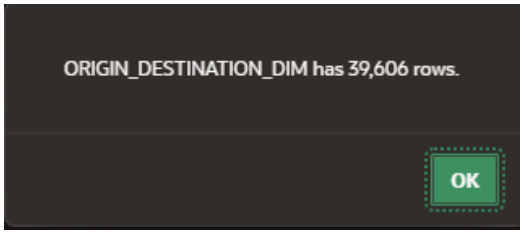
- **<u>Reporting period Table: -</u>** Reporting period table consists of reporting_period_id as primary key. This table also includes reporting_period and reporting_airport_name columns. This table consists of 39,606 rows of data.



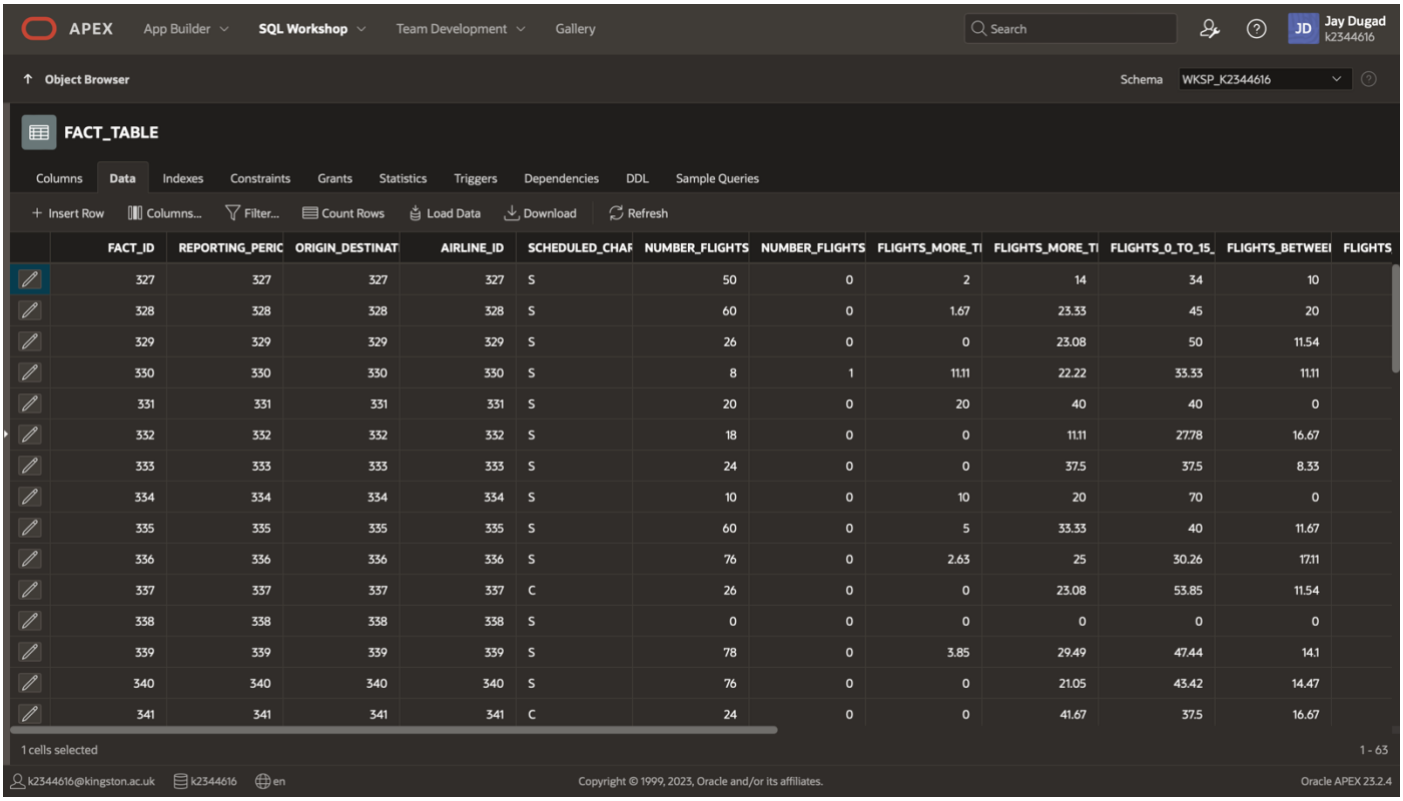


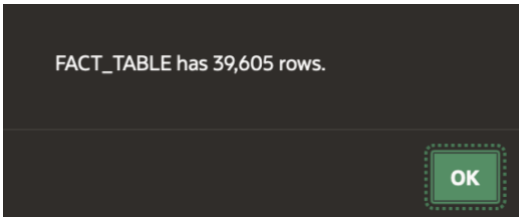
- **<u>Origin destination Table: -</u>** Origin destination table consists of origin_destination_id as primary key. This table also includes origin_destination_name and origin_destination_country_name columns. This table consists of 39,606 rows of data.





- **Fact Table: -** Fact table consists of fact_id as a primary key. This table has 'reporting_period_id', 'origin_destination_id', 'airline_id' as foreign keys. This table also includes many columns. Fact_table has all the quantitative values such as percentage, average, count etc. As fact table makes it easy for the person/company/stake holder to make decision based on these quantitative data. This table consists of 39,605 rows of data.





5.) ADVANTAGES AND EXAMPLES OF OLAP CUBES

The airline business can benefit greatly from OLAP (Online Analytical Processing) cubes since they can effectively evaluate large amounts of data from many angles. The following are some benefits of OLAP cubes and two instances of how they are used in the aviation sector:

Advantages of OLAP: -

- **Quick and Flexible Analysis: -** OLAP cubes let users examine large amounts of data from several angles in a timely manner. This adaptability is especially useful in the aviation industry's dynamic and data-rich environment, where stakeholders must evaluate performance, see patterns, and act quickly. For instance, in order to spot trends and improve scheduling, airlines might examine flight punctuality data according to several criteria, including departure time, destination, aircraft type, and day of the week.

- **Enhanced Decision Support: -** OLAP cubes provide well-informed decision-making across a range of functions within the aviation sector by offering a consolidated and user-friendly view of data. Decision-makers can react to operational difficulties, market dynamics, and customer requests with effectiveness by using real-time or almost real-time insights. To evaluate overall business performance and develop strategic initiatives, airline executives, for instance, can use OLAP cubes to track key performance indicators (KPIs) including load factor and on-time performance.

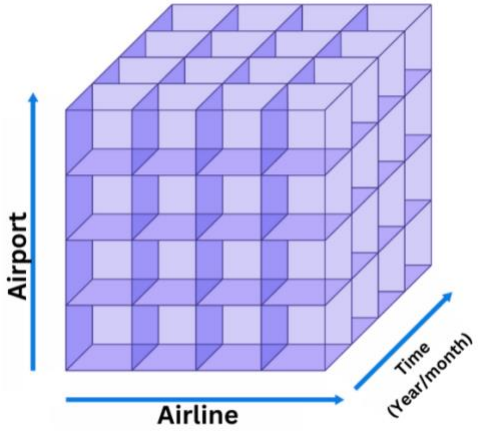
- **Data Drill-down and Exploration: -** OLAP cubes enable interactive drill-down and exploration features that enable users to go further into the data and find new insights. Through transactional data drilling, airlines may go beyond high-level summaries and enable root cause investigation and performance optimization. Airlines can investigate individual transactions or events that underlie performance indicators by going beyond aggregate statistics with the help of transactional data drilling. Analysts can delve further into the data to determine the causes of delays, such as maintenance problems with the aircraft, heavy air traffic, or unfavourable weather conditions, if an airline notices a drop in on-time performance on a particular route.

- **Cost Reduction and Revenue Maximization: -** By using OLAP cubes, airlines can find operational inefficiencies and make changes to cut expenses. For instance, airlines can lower fuel expenditures by optimizing flight routes through the analysis of fuel usage data. By using dynamic pricing methods based on demand, airlines can maximize revenue through the use of OLAP cubes. Airlines are able to maximize income by adjusting ticket rates in real-time by studying demand patterns and previous data. Airlines can examine data on luggage costs and customer preferences using OLAP cubes. Airlines can boost ancillary revenue by providing a variety of baggage fee choices, including as pre-paid baggage fees or tiered pricing based on luggage weight.

EXAMPLES OF OLAP CUBES

Cubes can include three axes, or perhaps less or more, as the diagram below illustrates. The analyst in this example can add extra dimensions, such as month-over-month change, or even more airports, if they need to examine more data.

OLAP Cube

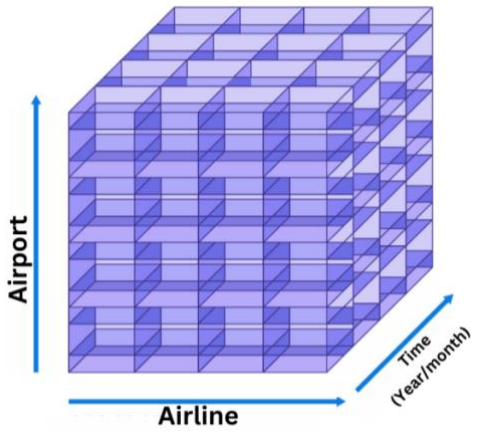


The analyst can do the many procedures present in any OLAP software when data has been entered into a cube.

EXAMPLE 1:

1. **Overview of On-Time Arrivals: -** The airline looks at a summary of flight punctuality first. They have an OLAP cube that collects information about flight delays in a number of dimensions. The airline can rapidly determine the total number of delays and their primary causes using this perspective. They can see that there have been 320 flights delayed overall during this time at any particular airport or of any particular airlines.
   Using a variety of variables, we can drill down from here. As an example, it could depend on the airline, a certain airport, or whether the aircraft is schedule charter. This kind of data allows the airline to determine that, out of all the routes operated by EasyJet UK LTD, the route from Cardiff to Bristol has the greatest average delay minutes -1454 overall. The airlines may make adjustments to their timetable or strive to be more punctual in light of this information.
   This method uses **Drill down approach** which provides more precision and information from the data. The analyst can dig further into their data to find various insights related to avg. delay time, numbered of cancelled flight, etc.
   The provided diagram shows an example of a three-dimensional OLAP cube and shows how it is drilled down, which facilitates data analysis. Expandable with extra axes to accommodate more dimensions for more in-depth analysis, a three-axis cube is perfect for more straightforward analysis and provides an insightful picture of the data.

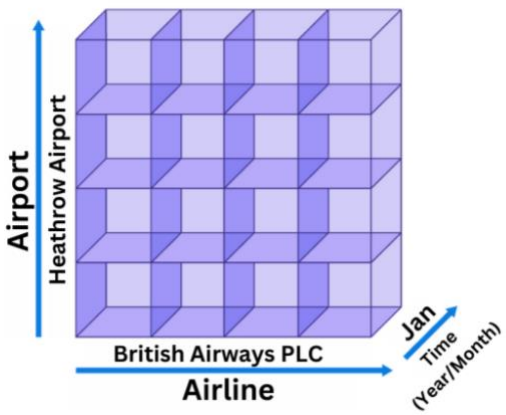
Drill down



EXAMPLE 2:

2. **Improved planning & forecasting:-** Another example of OLAP in airline sector is that it can help a airline/stake holders to analyse data separately for a certain subset like any type of airline or airport. For example, when compared to other airports in the same nation, we could observe that certain airports have larger arrival and departure delays. Airlines may be able to adjust the number of flights at that airport as a result.
   The airline ID 982 from Antalya to Edinburgh has the lowest average delayed minutes as compared to the previous year, as can be shown by slicing the OLAP cube. This study may provide insight on how they can improve other route's delay operated by the same airlines. The airline can then make use of these insights to put tactics into place that will help minimize delays, like making schedule adjustments, improving crew readiness, or pre-arranging alternate routes.
   This analysis can be done through the process of slicing. **Slicing approach** is a way to examine multidimensional data by taking a single value from a single dimension. By assigning a single value or a set of values to one or more dimensions, it enables users to concentrate on a particular subset of data. One of the basic processes used to examine OLAP cubes is slicing, along with dicing and drilling.
   We can see that one slice is removed from a cube in the diagram below. This makes it easier to focus on a single or particular subset of the data. Here, the data has been sliced so that it only includes January month and specific airport Heathrow

Slice



6.) VISUALISATIONS USING TABLEAU

VISUALISATION 1: -

- **Aim of visualisation: -** The aim of this visualisation is to analyse and illustrate the differences between the better delay performance from the prior year and the current year for a certain reporting airport. The purpose of creating this visualisation is to identify trends, patterns, and insights that demonstrate the progress being made in delay performance.

- **Steps taken to create the visualisation: -**

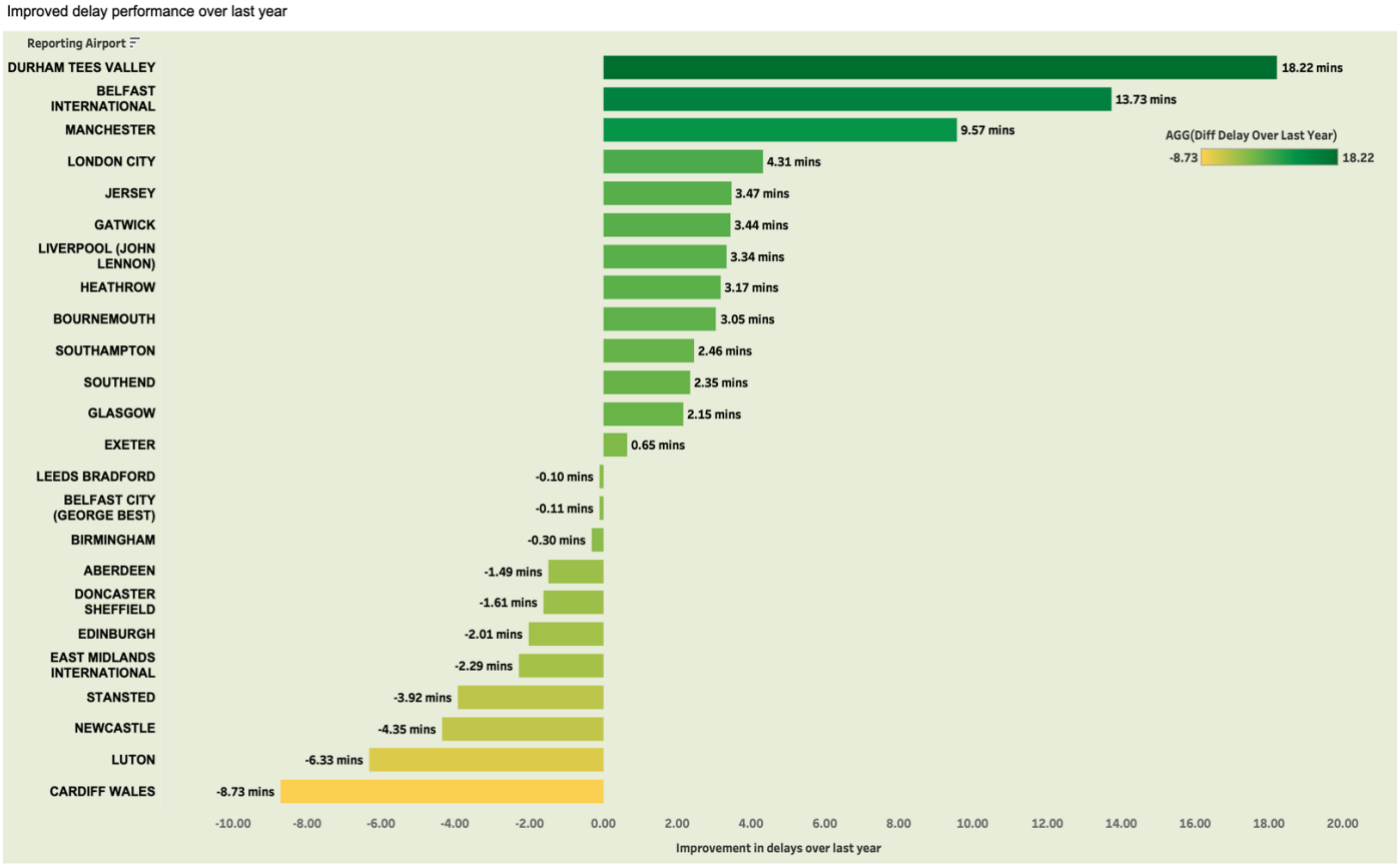
  Step 1: Connect to Data: - I started by connecting Tableau to the combined database.

  Step 2: Prepare the Data: - I created a calculated field to calculate the difference in delay times between the current and prior years.
  Name: - [Diff Delay Over Last Year]
  Formula: - (AVG([Delay Mins CurrentYear])-AVG([Delay Mins Prev Year])) * -1

  Step 3: Create the Visualisation: - Now I created a bar chart to visualise the differences in delay performance between the prior year and the current year for each airport.
  Dragged [Reporting Airport] to Rows.
  Dragged [Diff Delay Over Last Year] to Columns.
  After this I added [Diff Delay Over Last Year] and [Reporting Airport] to filters.

  Step 4: Add Additional Information: - Then I used gradients of two colours to make it more visually attractive and also make it easy to understand. After that I displayed the legends in the top right corner for a greater understanding. I also used format option to change font, colour and size of the headers. I also changed background colour of the visualisation to look more decent.

- **Key findings from the visualisation: -** For certain airports, the average delay time has dropped significantly from the prior year to the current one. The graph below shows how various airports, including Durham Tees Valley, Belfast International, Manchester, London City, and others, have reduced their delay times from approximately 18 minutes to 4 minutes.
  However, several airports like Cardiff Wales, Luton, New Castle, etc. have seen increases in delay times compared to the previous year, ranging from 8.7 to 4 minutes. A few airports, such as Leeds Bradford and Exeter, haven't changed all that much either.
  Thus, this knowledge can assist airports and airlines in better scheduling and flight management as well as increased timeliness.



VISUALISATION 2: -

- **Aim of the Visualisation: -** This visualisation aims to throw light on airline performance as measured by the quantity of flights and delays. We can compare the punctuality of various airlines by examining the visualization that displays the percentage of flights that arrive on time or early. This will eventually result in less financial and resource losses for airlines and help them improve their punctuality timings.

- **Steps taken to create the visualisation: -**

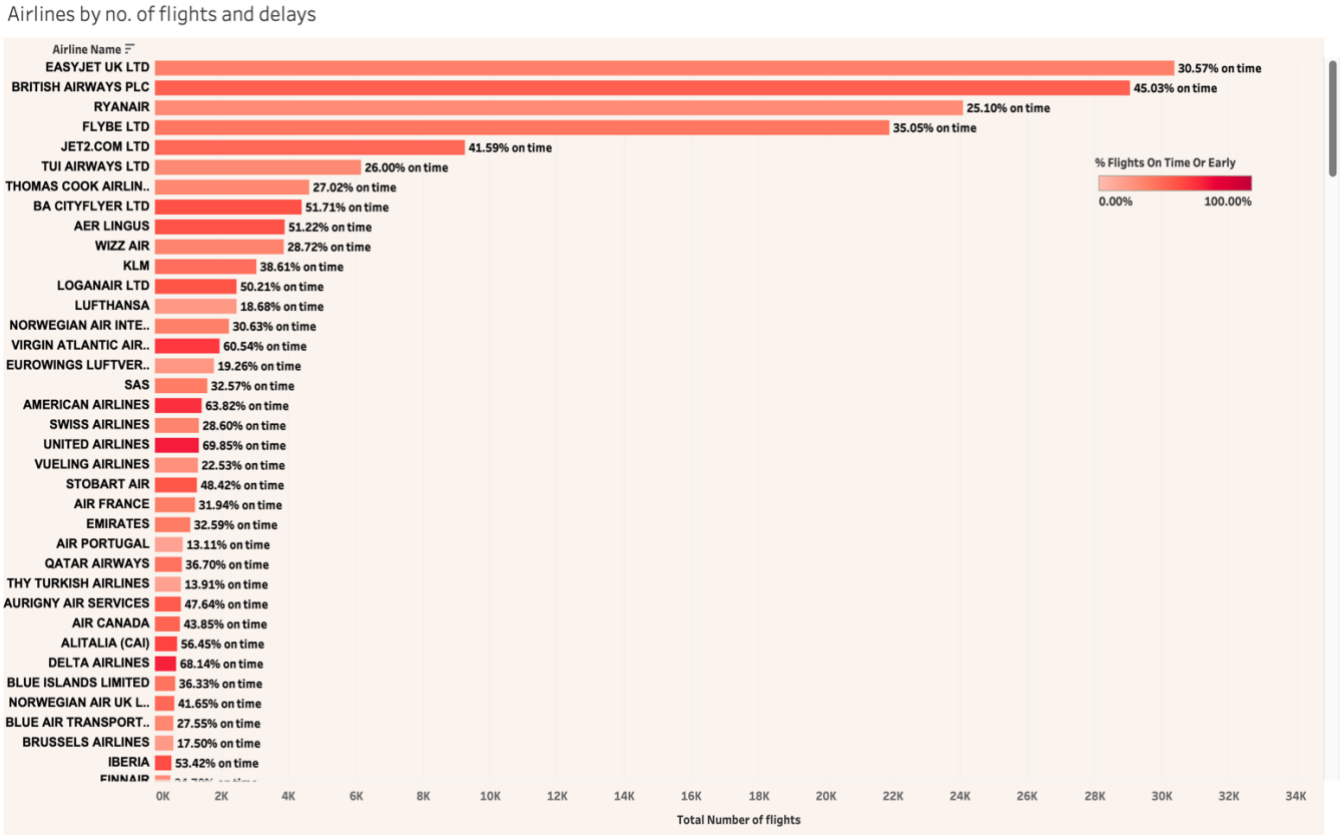
  Step 1: Connect to Data: - I started by connecting Tableau to the combined database.

  Step 2: Prepare the Data: - I created a calculated field to calculate the total number of flights.
  Name: - [Total Num of Flights]
  Formula: - [Number Flights Matched] + [Flights Cancelled] + [actual_flights_unmatched]

  Step 3: Create the Visualisation: - Now I created a bar chart to visualise the comparison among airlines showing number of flights and percentage of flights which were on time or early.
  Dragged [Airline Name] to Rows.
  Dragged SUM([Total Num of Flights]) to Columns.

  Step 4: Add Additional Information: - Then I used different shades of same colour to make it more visually attractive and also make it easy to understand. After that I displayed the legends in the top right corner for a greater understanding. I also used format option to change font, colour and size of the headers. I also changed background colour of the visualisation to look more decent.

- **Key findings from the visualisation: -** Some airlines, such as EasyJet UK LTD, British Airways PLC, Ryanair, and many more, have a higher percentage of early or on-time , ranging from 20% to 35%. We can also see from the visualisation that EasyJet UK LTD has the most flights 30,372 than any other airline. However, there are a lot of airline that have zero flights. The quantity of flights and the length of delays fluctuate greatly between carriers. A more accurate picture of airline punctuality can be obtained by analysing the data by proportion of early or on-time flights. This can aid airlines in controlling the quantity of flights they operate. An airline may decide to increase or decrease the number of flights based on these analyses.



VISUALISATION 3: -

- **Aim of the Visualisation: -** The aim of the visualization is to analyse and compare the average flight delay times for airports across the UK. This will help identify which airports have the highest and lowest average delays, providing valuable insights for travellers and aviation authorities.

- **Steps taken to create the visualisation: -**

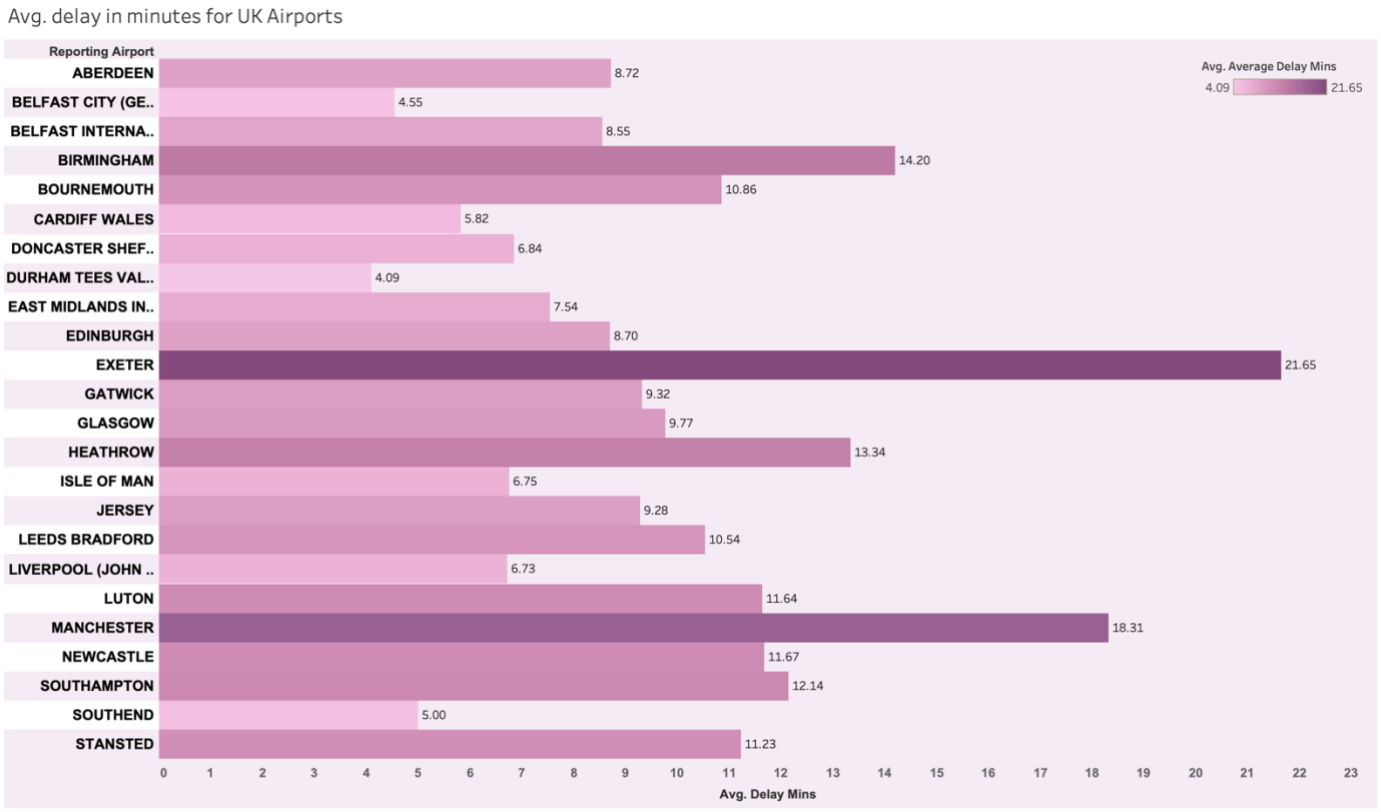
  Step 1: Connect to Data: - I started by connecting Tableau to the combined database.

  Step 2: Prepare the Data: - I didn't created any calculated field here as it was not necessary for this visualisation.

  Step 3: Create the Visualisation: - Now I created a bar chart to visualise the number of flights in UK showing Average delay minutes.
  Dragged AVG([Average Delay Mins]) to Columns
  Dragged [Reporting Airport] to Rows.

Step 4: Add Additional Information: - Then I used different shades of same colour to make it more visually attractive and also make it easy to understand. After that I displayed the legends in the top right corner for a greater understanding. The legend showed minimum as well as maximum level of average Delay minutes. I also used format option to change font, colour and size of the headers. I also changed background colour of the visualisation to look more decent.

- **Key findings from the visualisation: -** With the use of this visual aid, it is evident to all of us that Durham Tees Valley has the lowest average delay minutes, at around 4 minutes, while Exeter Airport has the most average delay minutes, at about 21.5 minutes followed by Manchester having around average 18 minutes delay. The airline and the airport administration can decide how to shorten delays based on these data. However, it would also assist travellers in choosing an aircraft that will arrive at the airport with the least amount of delay.



VISUALISATION 4: -

- **Aim of the Visualisation: -** This visualisations aim is to examine the flow of flights from origin to destination based on the quantity of flights between various nations. We can determine the busiest flight routes and the movement between various nations by visualising this data.

- **Steps taken to create the visualisation: -**

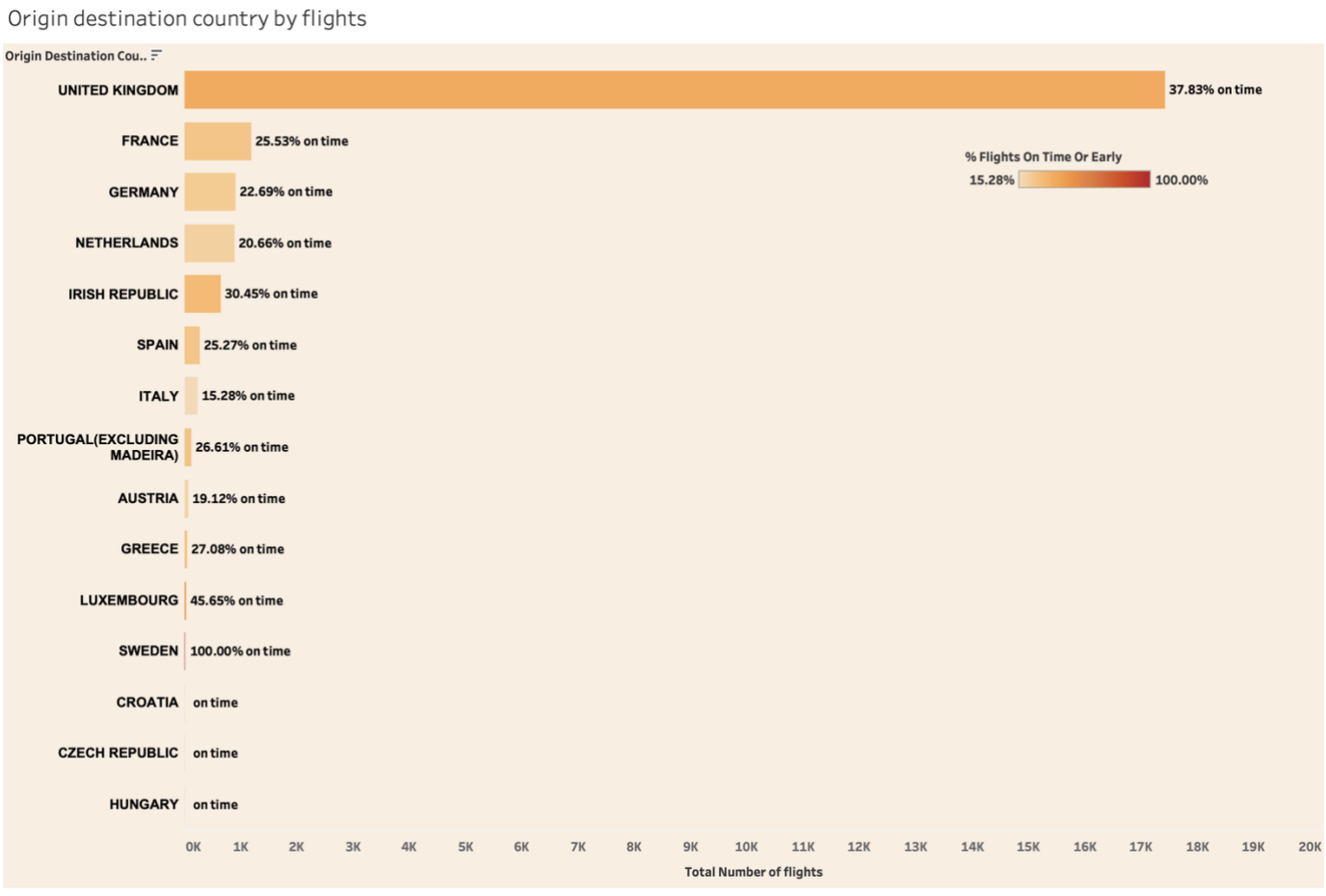
  Step 1: Connect to Data: - I started by connecting Tableau to the combined database.

  Step 2: Prepare the Data: - I used SUM([Total Num of Flights]) calculation field which I have created in the 2nd visualisation.

Step 3: Create the Visualisation: - Now I created a bar chart to visualise the number of flights in countries showing that how many percentage of them were early or on time.
Dragged SUM([Total Num of Flights]) to Columns
Dragged [Origin Destination Country] to Rows.

Step 4: Add Additional Information: - Then I used different shades of same colour to make it more visually attractive and also make it easy to understand. After that I displayed the legends in the top right corner for a greater understanding. The legend showed minimum as well as maximum level of percentage of flights which were on time or early. I also used format option to change font, colour and size of the headers. I also changed background colour of the visualisation to look more decent.

- **Key findings from the visualisation: -** Each line represents the flow between an origin and a destination country, with the thickness of the line indicating the number of flights. This visualization attempts to show the nations whose flights arrive early or on schedule in addition to the quantity of flights for each country's particular origin destination. Airlines can use this data to see which countries have the longest delays and how many planes are going to those countries. This can assist airlines in altering the quantity of flights.



Steps added to create this drop down of all the airport within a country: - I created a hierarchy named Origin Destination. This hierarchy consisted of two columns that were Origin Destination Country and Origin Destination Airport. This made it possible to create a drop down view in the same visualisation.

The graph below shows the names of the airports located within each nation within the same graph mentioned above.
This graph also allows us to view all of the airports in a country by clicking on a specific country's name. It also shows the number of flights that arrive at each airport, along with the proportion of flights that arrive early or on time.