KINGSTON UNIVERSITY LONDON

DATA WAREHOUSE (CI7320)

COURSEWORK – II

NAME	SUJAY GRAMA SURESH KUMAR
KU NUMBER	K2201621
KU EMAIL ID	K2201621@KINGSTON.AC.UK
ASSOCIATE PROFESSOR NAME	Dr. BERYL JONES
SENIOR LECTURER NAME	Dr. PUSHPA KUMARAPELI

TABLE OF CONTENTS

1. Discuss the benefits of building a data warehouse for the data set provided.

- In the present era, as data plays an increasingly important role in decision-making, businesses require the capability to store and analyze vast quantities of data. The airline industry in UK is no exception, as it generates a vast amount of data from multiple sources, including airport punctuality statistics.
- In this context, building a data warehouse can offer significant benefits, such as enhancing data accuracy, simplifying data retrieval, providing deeper insights into historical data, and enabling more informed decision-making. This section will discuss the advantages of building a data warehouse for the United Kingdom airport industry for the dataset provided.
- The Real-time data warehouses can be utilized in the airline industry to consolidate data sources into a single system, resulting in improved control over information. By analyzing large volumes of data in real-time, decision-making abilities can be enhanced.
- To develop a data warehouse for the given data sets, a recommended method is to establish multiple dimension tables and a single fact table. These dimension tables can include specifics such as Airport Reporting Name, Name of the Airline, The Origin Destination, Scheduled or Charter, and Period of reporting. On the other hand, the fact table, titled "Punctuality Statistics for Jan to Dec" , can hold the punctuality data of airports for each month.
- By implementing such strategies, data from different sources can be accessed and made available across the entire system. This leads to improved management of accounting information and revenue collection, and also helps in creating more efficient payroll systems with a reduced number of forged transactions.
- Through the implementation of a data warehouse and the utilization of the aforementioned tables, we can simplify the process of accessing airport punctuality statistics and enhance data accuracy. This, in turn, can result in improved decision-making abilities and quicker report generation.
- The fact table offers a single picture of punctuality statistics across airports and airlines, while dimension tables assist in organising and standardising data from multiple sources. As a result, decision-makers will be better able to recognise trends and patterns in punctuality statistics over time and take proactive actions to enhance performance.
- A data warehouse can also offer significant benefits in terms of insights, analysis, and decision-making capabilities within the United Kingdom airport industry and also by supplying customers with prompt and precise data guarantees that they are able to make knowledgeable choices, which results in an enhancement of their transactional decisions.

- The construction of a data warehouse for the United Kingdom airport sector can offer numerous advantages, and also this data from the dataset can be error-prone due to being sourced from various locations, but through data standardization, uniformity can be ensured, and a centralized view of the data can be provided.
- Consequently, improving data quality can be crucial to enable informed decision-making across all tables. Data warehouses offer a convenient way to retrieve all punctuality statistics, as they are stored in a unified location, eliminating the necessity to look for data from multiple sources.
- The convenience of having data consolidated in a single location within a data warehouse enables analysts and decision-makers to perform data analysis without the need to gather data from multiple sources.
- The inclusion of historical data in data warehouses offers a more comprehensive understanding of airport performance over time, allowing analysts to identify trends, patterns, and irregularities in punctuality statistics. Decision-makers can leverage this information to identify and prioritize areas for improvement and to make data-driven decisions.
- In conclusion, data warehouses enable fast and efficient reporting through optimized storage and structured organization of data, and can handle growing amounts of data without affecting performance, allowing organizations to analyze data more effectively and make informed decisions promptly. The establishment of a data warehouse in the UK airport sector is a crucial step towards enhancing overall performance and customer satisfaction.

2. Design a data warehouse using a star schema. You must justify your design decisions.

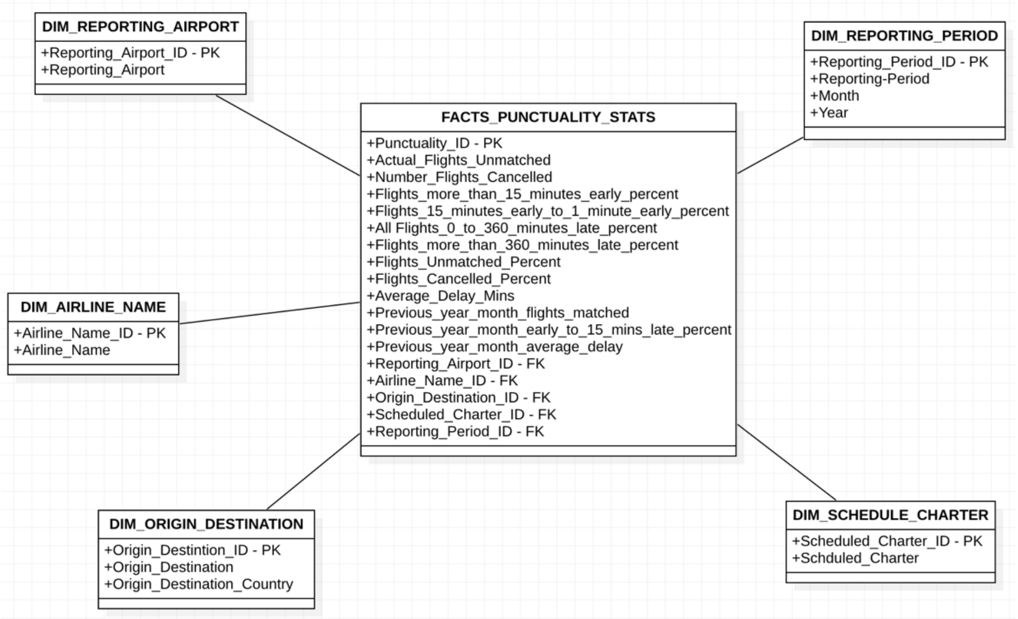

DIM_REPORTING_AIRPORT
+Reporting_Airport_ID - PK
+Reporting_Airport
DIM_REPORTING_PERIOD
+Reporting_Period_ID - PK
+Reporting-Period
+Month
+Year
FACTS_PUNCTUALITY_STATS
+Punctuality_ID - PK
+Actual_Flights_Unmatched
+Number_Flights_Cancelled
+Flights_more_than_15_minutes_early_percent
+Flights_15_minutes_early_to_1_minute_early_percent
+All Flights_0_to_360_minutes_late_percent
+Flights_more_than_360_minutes_late_percent
+Flights_Unmatched_Percent
+Flights_Cancelled_Percent
+Average_Delay_Mins
+Previous_year_month_flights_matched
+Previous_year_month_early_to_15_mins_late_percent
+Previous_year_month_average_delay
+Reporting_Airport_ID - FK
+Airline_Name_ID - FK
+Origin_Destination_ID - FK
+Scheduled_Charter_ID - FK
+Reporting_Period_ID - FK
DIM_AIRLINE_NAME
+Airline_Name_ID - PK
+Airline_Name
DIM_ORIGIN_DESTINATION
+Origin_Destintion_ID - PK
+Origin_Destination
+Origin_Destination_Country
DIM_SCHEDULE_CHARTER
+Scheduled_Charter_ID - PK
+Schduled_Charter

Fig 1a: Star Schema with Primary and Foreign Keys

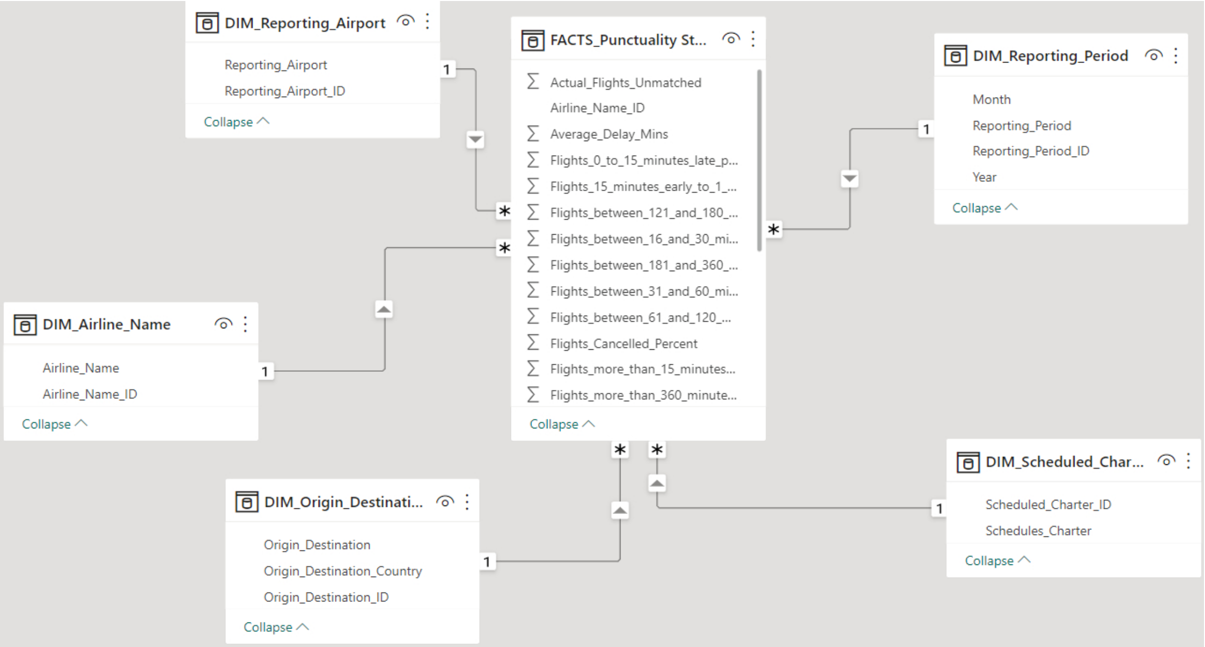


Fig 1b: Final Star Schema

Justification on the design decisions made:

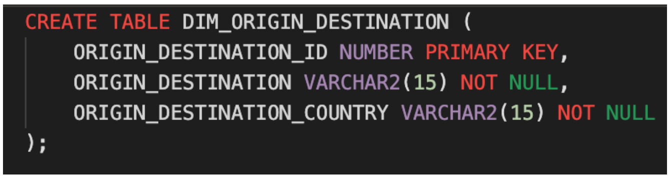
- **Dimension table Reporting Period**: This dimension table has primary key named Reporting_Period_ID and will contain information about the reporting period, such as the month and year. This dimension is used in data warehousing and analytics, as it allows users to slice and dice data by time and compare performance over different periods. It will be linked to the fact table on the Reporting Period field.
- **Dimension table Reporting Airport**: This dimension table has primary key named Reporting_Airport_ID and will contain information about the reporting airport, such as name of the reporting airport . This can be useful for identifying trends and patterns in performance across different airports, as well as for comparing the performance of different airlines operating at the same airport. It will be linked to the fact table on the Reporting Airport field.
- **Dimension table Origin Destination**: This dimension table has primary key named Origin_Destination_ID and will contain information about the origin and destination countries for each flight, as well as the specific origin and destination airports. It will be linked to the fact table on the Origin Destination fields.
- **Dimension table Airline Name**: This dimension table has primary key named Airline_Name_ID and will contain information about the airline name. By analyzing the performance of different airlines, users can identify which airlines have the best and worst punctuality rates through the airline name. It will be linked to the fact table on the Airline_Name field.
- **Dimension table Scheduled Charter**: This dimension table has primary key named Scheduled_Charter_Id and will contain information about whether the flight was scheduled or chartered. It will be linked to the fact table on the Scheduled_Charter field.
- **Fact table Punctuality Statistics**: The fact table also has primary key named Punctuality_Id and also in this schema will be named as "Punctuality Statistics" and will contain information about the number of flights matched, actual flights unmatched, number of flights cancelled, and the percentage of flights that fall into various punctuality categories, such as flights more than 15 minutes early or between 31 and 60 minutes late. It will also include the percentage of flights that were unmatched or cancelled, and the average delay time. This fact table will be connected to the dimension tables using foreign keys.

3. Write the CREATE table statements for the tables in your star schema (include all primary and foreign keys).

◊ These are the snapshots of create table SQL queries for the Fact and Dimension tables created in Oracle Apex. Details for accessing Oracle Apex are provided after the references.

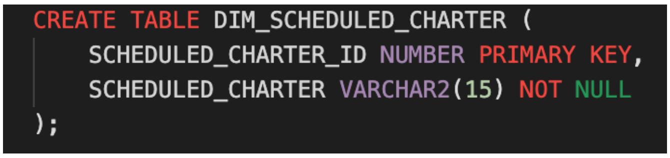
i. DIMENSION TABLES SQL QUERIES:

- Origin Destination:



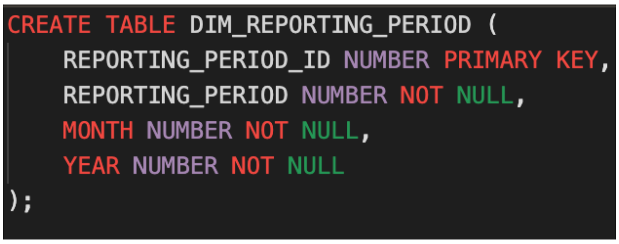
```
CREATE TABLE DIM_ORIGIN_DESTINATION (
    ORIGIN_DESTINATION_ID NUMBER PRIMARY KEY,
    ORIGIN_DESTINATION VARCHAR2(15) NOT NULL,
    ORIGIN_DESTINATION_COUNTRY VARCHAR2(15) NOT NULL
);
```

- Scheduled Charter:



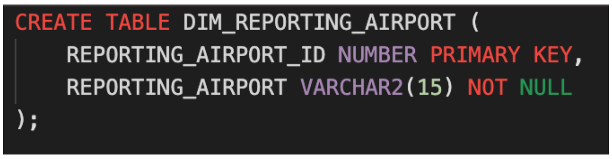
```
CREATE TABLE DIM_SCHEDULED_CHARTER (
    SCHEDULED_CHARTER_ID NUMBER PRIMARY KEY,
    SCHEDULED_CHARTER VARCHAR2(15) NOT NULL
);
```

- Reporting Period:



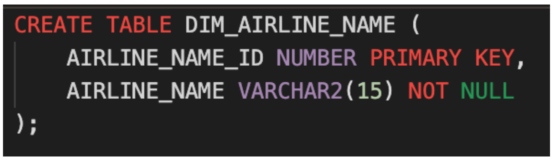
```
CREATE TABLE DIM_REPORTING_PERIOD (
    REPORTING_PERIOD_ID NUMBER PRIMARY KEY,
    REPORTING_PERIOD NUMBER NOT NULL,
    MONTH NUMBER NOT NULL,
    YEAR NUMBER NOT NULL
);
```

- Reporting Airport:



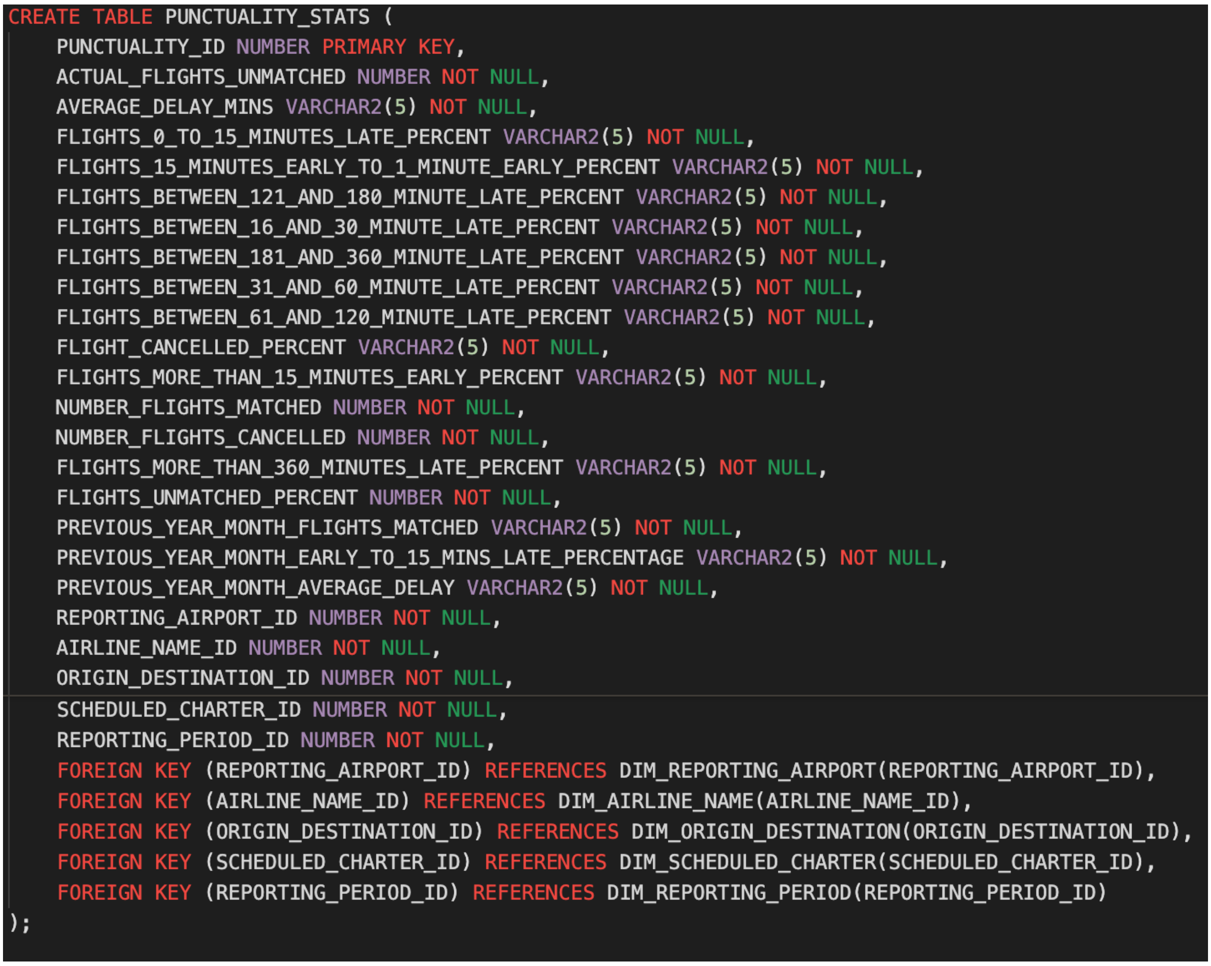
```
CREATE TABLE DIM_REPORTING_AIRPORT (
    REPORTING_AIRPORT_ID NUMBER PRIMARY KEY,
    REPORTING_AIRPORT VARCHAR2(15) NOT NULL
);
```

- Airline Name:



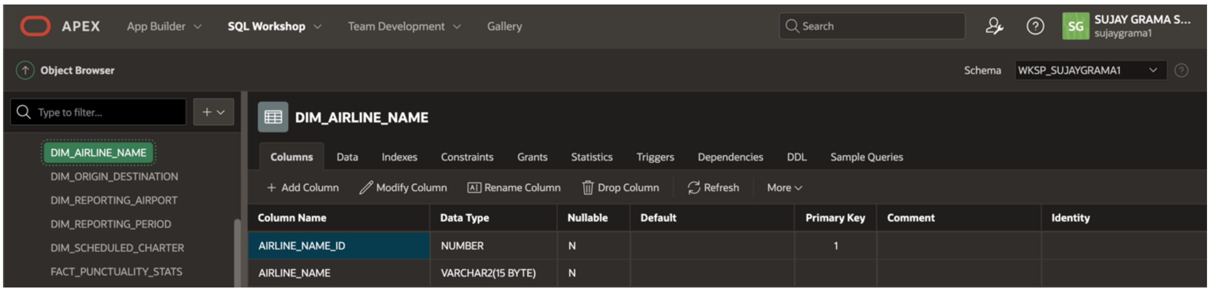
```
CREATE TABLE DIM_AIRLINE_NAME (
    AIRLINE_NAME_ID NUMBER PRIMARY KEY,
    AIRLINE_NAME VARCHAR2(15) NOT NULL
);
```

ii. FACT TABLE SQL QUERIES:



```sql
CREATE TABLE PUNCTUALITY_STATS (
    PUNCTUALITY_ID NUMBER PRIMARY KEY,
    ACTUAL_FLIGHTS_UNMATCHED NUMBER NOT NULL,
    AVERAGE_DELAY_MINS VARCHAR2(5) NOT NULL,
    FLIGHTS_0_TO_15_MINUTES_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_15_MINUTES_EARLY_TO_1_MINUTE_EARLY_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_BETWEEN_121_AND_180_MINUTE_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_BETWEEN_16_AND_30_MINUTE_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_BETWEEN_181_AND_360_MINUTE_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_BETWEEN_31_AND_60_MINUTE_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_BETWEEN_61_AND_120_MINUTE_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHT_CANCELLED_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_MORE_THAN_15_MINUTES_EARLY_PERCENT VARCHAR2(5) NOT NULL,
    NUMBER_FLIGHTS_MATCHED NUMBER NOT NULL,
    NUMBER_FLIGHTS_CANCELLED NUMBER NOT NULL,
    FLIGHTS_MORE_THAN_360_MINUTES_LATE_PERCENT VARCHAR2(5) NOT NULL,
    FLIGHTS_UNMATCHED_PERCENT NUMBER NOT NULL,
    PREVIOUS_YEAR_MONTH_FLIGHTS_MATCHED VARCHAR2(5) NOT NULL,
    PREVIOUS_YEAR_MONTH_EARLY_TO_15_MINS_LATE_PERCENTAGE VARCHAR2(5) NOT NULL,
    PREVIOUS_YEAR_MONTH_AVERAGE_DELAY VARCHAR2(5) NOT NULL,
    REPORTING_AIRPORT_ID NUMBER NOT NULL,
    AIRLINE_NAME_ID NUMBER NOT NULL,
    ORIGIN_DESTINATION_ID NUMBER NOT NULL,
    SCHEDULED_CHARTER_ID NUMBER NOT NULL,
    REPORTING_PERIOD_ID NUMBER NOT NULL,
    FOREIGN KEY (REPORTING_AIRPORT_ID) REFERENCES DIM_REPORTING_AIRPORT(REPORTING_AIRPORT_ID),
    FOREIGN KEY (AIRLINE_NAME_ID) REFERENCES DIM_AIRLINE_NAME(AIRLINE_NAME_ID),
    FOREIGN KEY (ORIGIN_DESTINATION_ID) REFERENCES DIM_ORIGIN_DESTINATION(ORIGIN_DESTINATION_ID),
    FOREIGN KEY (SCHEDULED_CHARTER_ID) REFERENCES DIM_SCHEDULED_CHARTER(SCHEDULED_CHARTER_ID),
    FOREIGN KEY (REPORTING_PERIOD_ID) REFERENCES DIM_REPORTING_PERIOD(REPORTING_PERIOD_ID)
);
```

iii. DIMENSION AND FACT TABLES VIEW IN ORACLE APEX:



Schema: WKSP_SUJAYGRAMA1

DIM_AIRLINE_NAME
DIM_ORIGIN_DESTINATION
DIM_REPORTING_AIRPORT
DIM_REPORTING_PERIOD
DIM_SCHEDULED_CHARTER
FACT_PUNCTUALITY_STATS

DIM_AIRLINE_NAME

Column Name	Data Type	Nullable	Default	Primary Key	Comment	Identity
AIRLINE_NAME_ID	NUMBER	N		1		
AIRLINE_NAME	VARCHAR2(15 BYTE)	N				

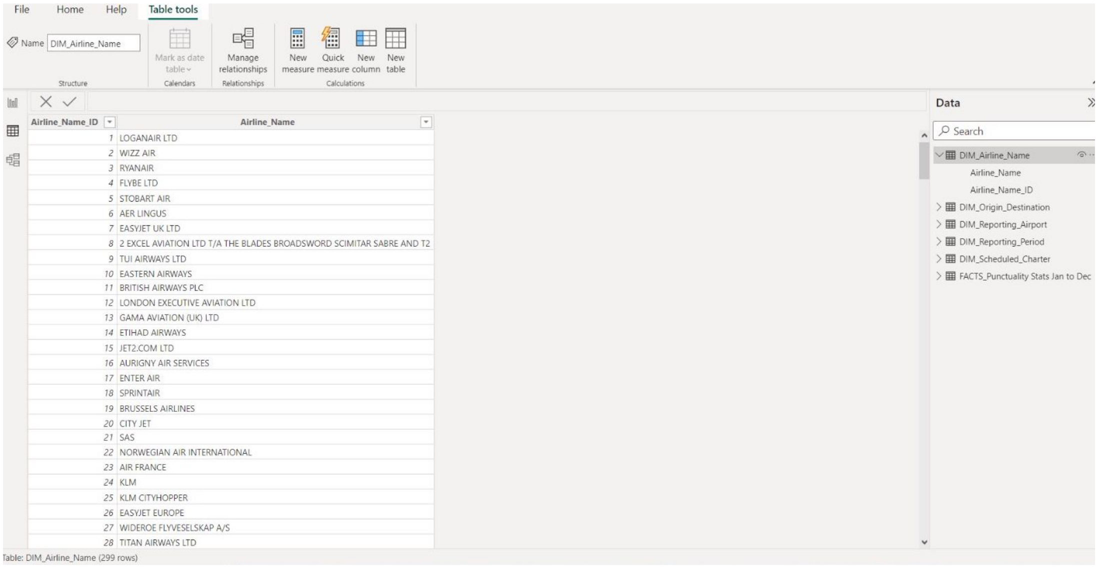
Fig 2 : Created Tables View

4. Discuss the steps you took in creating and populating the database. This should include the steps you took in preparing the data and the transformation tasks performed.

- Creating a database involves several steps including preparing the data and performing transformations to populate the tables. In this case, the given dataset contains information about airline flights and their performance, and it was needed to transform this data into a star schema for efficient querying.
- In a data modelling tool called Power BI, we can use the "Query Editor" to perform ETL (Extract, Transform, Load) tasks. This involves extracting data from various sources, transforming it as necessary, and then loading it into the target database. We can perform tasks such as merging tables, filtering data, splitting columns, and pivoting data to prepare the data for analysis.
- Once the ETL tasks are completed and the data is transformed and normalized, we can populate the database using the "Import" option in Power BI. This will create the database based on the dimension and fact tables created earlier.
- The first step is to prepare the data by cleaning it and removing any duplicate or irrelevant information, filling in missing values, and transforming data types as it ensures that the data is accurate, reliable and to obtain the consistency across the data.. Then, a need to create a logical model for the database, which in this case will be a star schema consisting of a fact table and several dimension tables.
- To create the dimension tables, the Power BI tool was used to extract, transform, and load (ETL) the data. The ETL process will involve several steps, including:
  i) **Extraction:** Extracting the data from the given dataset into Power BI by importing it from the CSV, SQL Server, Excel, etc.
  ii) **Transformation:** Once the data is prepared, we can create the dimension tables in Power BI using the "Enter Data" option by transforming the data by creating separate tables for each of the dimension tables (Reporting Period, Reporting Airport, Origin Destination Country, and Origin Destination, Airline Name, and Scheduled Charter) and removing any duplicate values. Assign unique primary keys for each table and each value attribute.
  iii) **Loading:** We will load the transformed data into the dimension tables in Power BI by assigning the primary keys.
- Then the fact table is created by using the same ETL process to transform the data into a numerical value table with foreign keys referencing the appropriate dimension tables. This helps to normalize the data and reduce redundancy. Then the unique primary key is also created for the identification of each row in the fact table.
- The Oracle Apex was used to write queries and create the database for the star schema model. The fact table will be at the center of the schema with the dimension tables branching out from it, creating a star shape. This schema allows for efficient querying and analysis of the airline flight data.
- Finally, the data was visualized using various charts and graphs in Tableau. This helped to analyze the data and gain insights into the trends and patterns in the airline data.

- In summary, the database creation and population involved preparing and transforming the data using Power BI's ETL process to create a star schema consisting of a fact table and several dimension tables, which allows for efficient querying and analysis of the airline flight data. By following the steps of importing the data, creating dimension and fact tables, performing ETL tasks, populating the database, and visualizing the data, I was able to analyze the given dataset. This helped me to gain insights into the performance of airlines during the reporting period and identify trends and patterns in the data.

◊ Snapshots of Preparation of the data and transformation tasks performed to obtain Dimension and Fact tables in the data modelling.

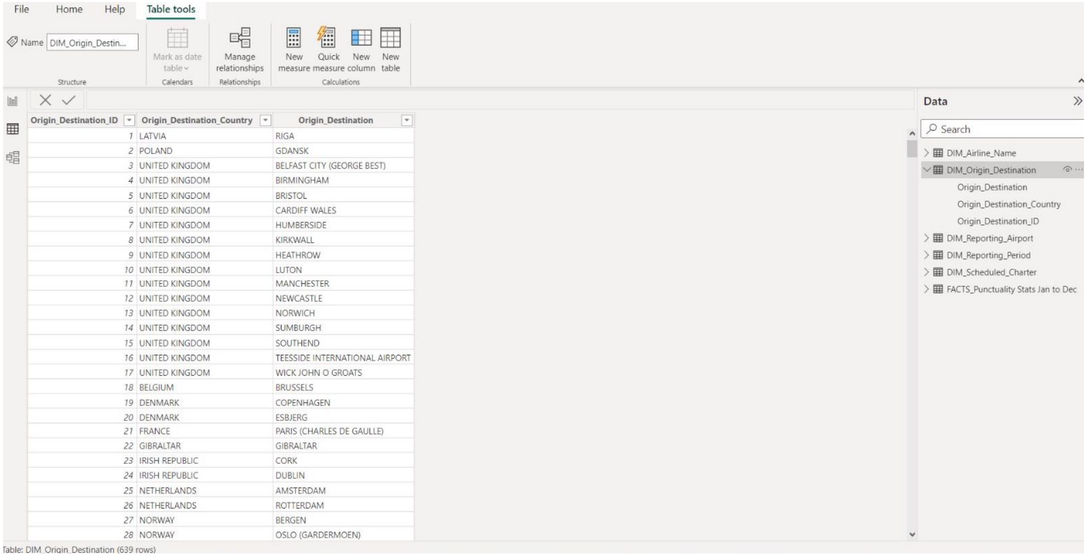


Name DIM_Airline_Name

Airline_Name_ID	Airline_Name
1	LOGANAIR LTD
2	WIZZ AIR
3	RYANAIR
4	FLYBE LTD
5	STOBART AIR
6	AER LINGUS
7	EASYJET UK LTD
8	2 EXCEL AVIATION LTD T/A THE BLADES BROADSWORD SCIMITAR SABRE AND T2
9	TUI AIRWAYS LTD
10	EASTERN AIRWAYS
11	BRITISH AIRWAYS PLC
12	LONDON EXECUTIVE AVIATION LTD
13	GAMA AVIATION (UK) LTD
14	ETIHAD AIRWAYS
15	JET2.COM LTD
16	AURIGNY AIR SERVICES
17	ENTER AIR
18	SPRINTAIR
19	BRUSSELS AIRLINES
20	CITY JET
21	SAS
22	NORWEGIAN AIR INTERNATIONAL
23	AIR FRANCE
24	KLM
25	KLM CITYHOPPER
26	EASYJET EUROPE
27	WIDEROE FLYVESELSKAP A/S
28	TITAN AIRWAYS LTD

Table: DIM_Airline_Name (299 rows)

Fig 3: Dimension Airline Name

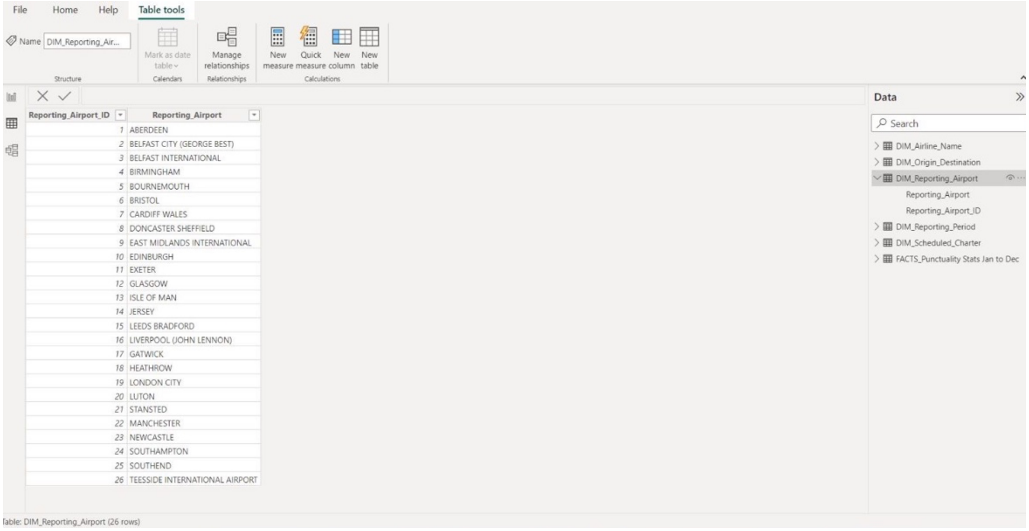


Name DIM_Origin_Destin...

Origin_Destination_ID	Origin_Destination_Country	Origin_Destination
1	LATVIA	RIGA
2	POLAND	GDANSK
3	UNITED KINGDOM	BELFAST CITY (GEORGE BEST)
4	UNITED KINGDOM	BIRMINGHAM
5	UNITED KINGDOM	BRISTOL
6	UNITED KINGDOM	CARDIFF WALES
7	UNITED KINGDOM	HUMBERSIDE
8	UNITED KINGDOM	KIRKWALL
9	UNITED KINGDOM	HEATHROW
10	UNITED KINGDOM	LUTON
11	UNITED KINGDOM	MANCHESTER
12	UNITED KINGDOM	NEWCASTLE
13	UNITED KINGDOM	NORWICH
14	UNITED KINGDOM	SUMBURGH
15	UNITED KINGDOM	SOUTHEND
16	UNITED KINGDOM	TEESSIDE INTERNATIONAL AIRPORT
17	UNITED KINGDOM	WICK JOHN O GROATS
18	BELGIUM	BRUSSELS
19	DENMARK	COPENHAGEN
20	DENMARK	ESBJERG
21	FRANCE	PARIS (CHARLES DE GAULLE)
22	GIBRALTAR	GIBRALTAR
23	IRISH REPUBLIC	CORK
24	IRISH REPUBLIC	DUBLIN
25	NETHERLANDS	AMSTERDAM
26	NETHERLANDS	ROTTERDAM
27	NORWAY	BERGEN
28	NORWAY	OSLO (GARDERMOEN)

Table: DIM_Origin_Destination (639 rows)

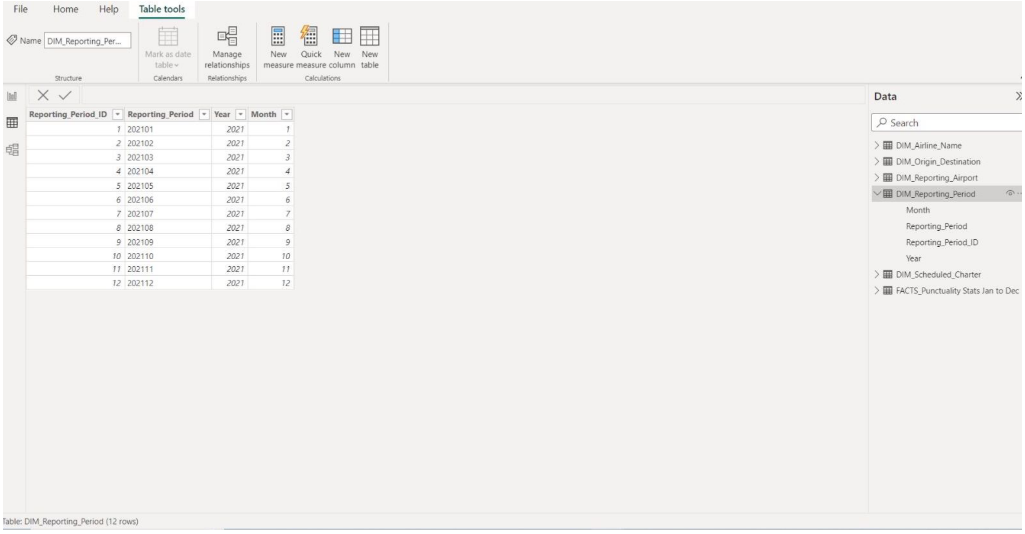
Fig 4: Dimension Origin Destination



Reporting_Airport_ID	Reporting_Airport
1	ABERDEEN
2	BELFAST CITY (GEORGE BEST)
3	BELFAST INTERNATIONAL
4	BIRMINGHAM
5	BOURNEMOUTH
6	BRISTOL
7	CARDIFF WALES
8	DONCASTER SHEFFIELD
9	EAST MIDLANDS INTERNATIONAL
10	EDINBURGH
11	EXETER
12	GLASGOW
13	ISLE OF MAN
14	JERSEY
15	LEEDS BRADFORD
16	LIVERPOOL (JOHN LENNON)
17	GATWICK
18	HEATHROW
19	LONDON CITY
20	LUTON
21	STANSTED
22	MANCHESTER
23	NEWCASTLE
24	SOUTHAMPTON
25	SOUTHEND
26	TEESSIDE INTERNATIONAL AIRPORT

Table: DIM_Reporting_Airport (26 rows)

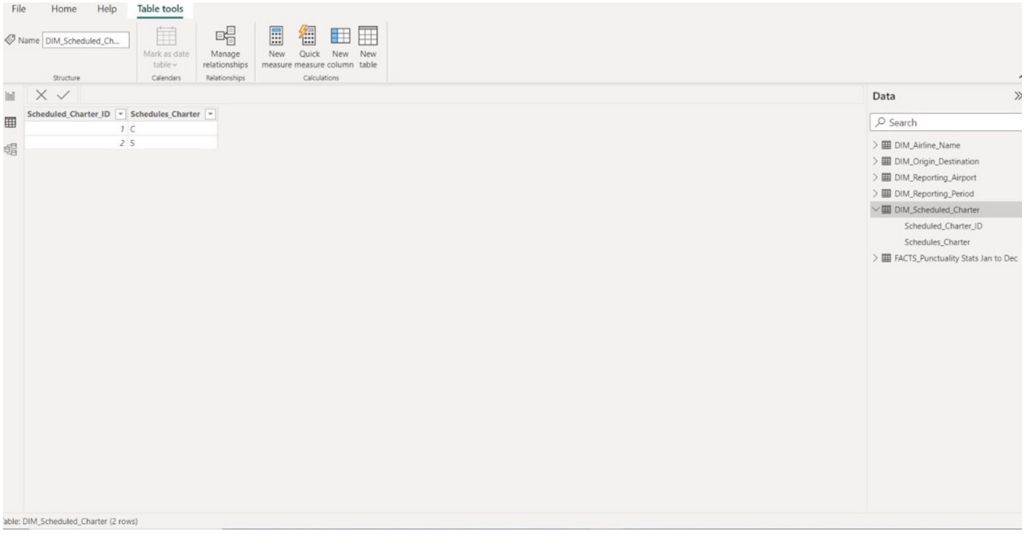
Fig 5: Dimension Reporting Airport



Reporting_Period_ID	Reporting_Period	Year	Month
1	202101	2021	1
2	202102	2021	2
3	202103	2021	3
4	202104	2021	4
5	202105	2021	5
6	202106	2021	6
7	202107	2021	7
8	202108	2021	8
9	202109	2021	9
10	202110	2021	10
11	202111	2021	11
12	202112	2021	12

Table: DIM_Reporting_Period (12 rows)

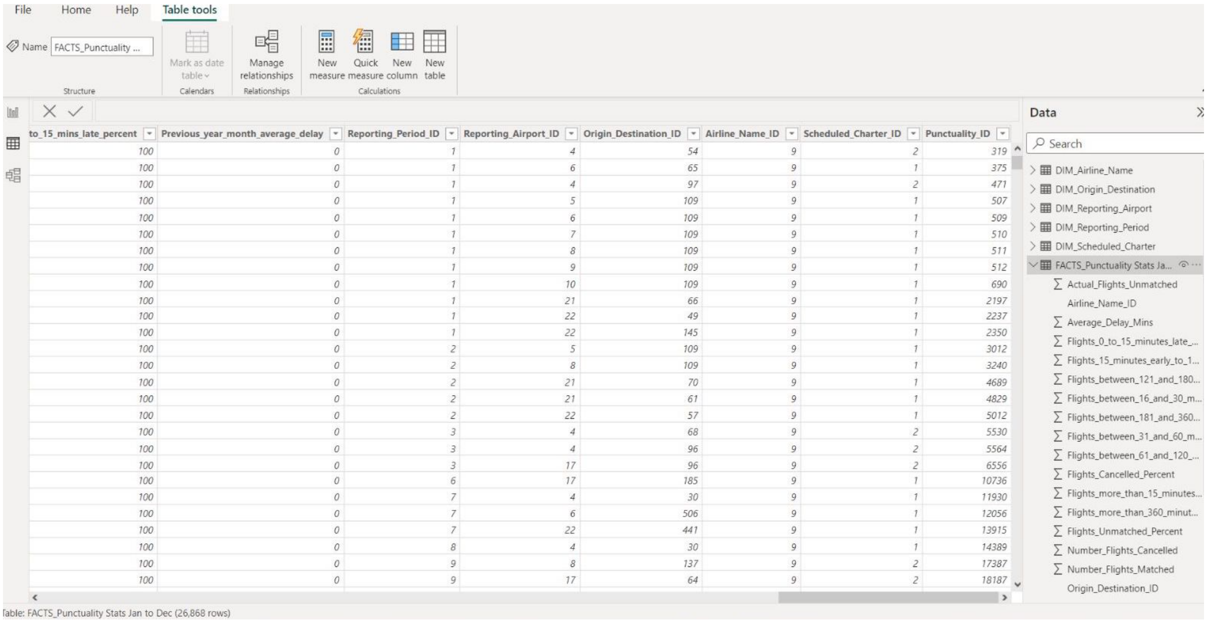
Fig 6: Dimension Reporting Period



Scheduled_Charter_ID	Schedules_Charter
1	C
2	S

Table: DIM_Scheduled_Charter (2 rows)

Fig 7: Dimension Scheduled Charter



to_15_mins_late_percent	Previous_year_month_average_delay	Reporting_Period_ID	Reporting_Airport_ID	Origin_Destination_ID	Airline_Name_ID	Scheduled_Charter_ID	Punctuality_ID
100	0	1	4	54	9	2	319
100	0	1	6	65	9	1	375
100	0	1	4	97	9	2	471
100	0	1	5	109	9	1	507
100	0	1	6	109	9	1	509
100	0	1	7	109	9	1	510
100	0	1	8	109	9	1	511
100	0	1	9	109	9	1	512
100	0	1	10	109	9	1	690
100	0	1	21	66	9	1	2197
100	0	1	22	49	9	1	2237
100	0	1	22	145	9	1	2350
100	0	2	5	109	9	1	3012
100	0	2	8	109	9	1	3240
100	0	2	21	70	9	1	4689
100	0	2	21	61	9	1	4829
100	0	2	22	57	9	1	5012
100	0	3	4	68	9	2	5530
100	0	3	4	96	9	2	5564
100	0	3	17	96	9	2	6556
100	0	6	17	185	9	1	10736
100	0	7	4	30	9	1	11930
100	0	7	6	506	9	1	12056
100	0	7	22	441	9	1	13915
100	0	8	4	30	9	1	14389
100	0	9	8	137	9	2	17387
100	0	9	17	64	9	2	18187

Table: FACTS_Punctuality Stats Jan to Dec (26,868 rows)

Fig 8: Fact Table Punctuality Statistics

5. Discuss how the airline industry in general can benefit from OLAP cubes giving examples of cubes in your discussion.

- OLAP, uses a multidimensional approach to organize and analyze data, is an efficient tool for the airline industry. With OLAP, data is organized into dimensions, such as product, market, and time, reflecting how business users typically think of the business. The main advantage of OLAP is its ability to execute queries at a high speed.
- All data is stored in tables connected to the star schema in the center, which organizes a cube with multiple dimensions, making it easy and fast to navigate through vast amounts of information. Airlines can use OLAP cubes to analyze revenue data from various angles, such as by route, by class, by time, and by customer segment. Analyzing this data can help airlines optimize their pricing strategies, manage seat inventory more effectively, and maximize revenue.
- In addition to revenue data, airlines can use OLAP cubes to analyze safety and security data from various sources, such as flight data recorders, security cameras, and passenger feedback. By analyzing this data, airlines can identify potential safety and security risks and take proactive measures to mitigate them.
- Moreover, airlines can use OLAP cubes to analyze customer data such as booking patterns, travel preferences, and loyalty program usage. Understanding customer behaviour can help airlines tailor their marketing and promotional activities to meet customer needs and improve customer loyalty.
- For instance, a UK-based airline can use an OLAP cube to analyze its revenue data for the past year. They can create a multidimensional structure that allows them to view the data in different ways, such as by month, by route, by aircraft type, and by customer segment.

- Several airlines have already benefited from using OLAP cubes. Delta Air Lines, for example, uses an OLAP cube to analyze customer data, including customer demographics, purchasing behaviour, and loyalty program participation. By using the OLAP cube, Delta has been able to increase customer loyalty and revenue.
- Similarly, British Airways uses an OLAP cube to analyze revenue data, including revenue by route, by class, and by time. By using this data, British Airways has been able to optimize pricing strategies, manage seat inventory more effectively, and maximize revenue.
- Southwest Airlines uses an OLAP cube to analyze operational data, including flight schedules, crew scheduling, and aircraft maintenance. By using this data, Southwest has been able to identify inefficiencies, reduce costs, and improve the overall customer experience.
- Emirates Airlines uses an OLAP cube to analyze safety data, including flight data recorders and passenger feedback. By using this data, Emirates has been able to identify potential safety risks and take proactive measures to mitigate them.
- Overall, the airline industry has benefited greatly from using OLAP cubes to analyze data from multiple perspectives. OLAP cubes have allowed airlines to optimize pricing strategies, manage seat inventory more effectively, improve operational efficiency, increase customer loyalty, and improve safety and security. The use of OLAP cubes has led to more efficient and effective decision-making processes, ultimately leading to increased profitability for airlines.
- In conclusion, the airline industry can benefit from OLAP cubes by using them to analyze data related to revenue management, operations management, customer behaviour, and safety and security. By using OLAP cubes, airlines can gain valuable insights that can lead to more efficient and effective decision making processes.

6. Discuss the benefits of using a data warehouse in combination with a business intelligence tool like Tableau.

- A data warehouse serves as a centralized storage location for data collected from multiple sources within an organization, while a business intelligence (BI) tool like Tableau is a software application that helps to analyze and visualize this data, enabling businesses to make informed decisions.
- Combining these two technologies enhances data mining and allows the extraction of useful information from raw data, identifying patterns and trends that are important for business intelligence. This results in better performance metrics and data reliability.
- In the field of business intelligence and data warehousing, it is common for analysts and business teams to perform data queries in order to validate its accuracy and validity.
- The main objective of BI is to enable businesses and organizations to pose and respond to inquiries regarding their data, ensuring that they have access to dependable, quantitative information to aid in their decision-making process. There are several benefits to using a data warehouse in conjunction with a BI tool such as Tableau.

- Firstly, data inconsistencies and inaccuracies are eliminated as all users are working with the same version of data stored in a central location.
- Secondly, data warehouses are designed to handle large amounts of data and complex queries, enabling users to analyze and visualize large amounts of data without impacting the performance of source systems.
- Thirdly, Tableau can quickly generate reports and visualizations from the stored data, allowing for faster decision-making without the need for data processing. Additionally, combining BI tools with data warehouses enhances the visualization of large amounts of data, increasing the approach of visualizing large amounts of data.
- Fourthly, data warehouses store historical data, making it possible to perform trend analysis and other advanced analytics. Tableau can leverage this historical data to provide insights into long-term trends and patterns. Finally, a data warehouse can integrate data from multiple sources, such as CRM, ERP, and accounting systems, providing a more complete picture of the business.
- Overall, the integration of a data warehouse with BI tools like Tableau creates a robust platform for data analysis and visualization, providing organizations with a competitive edge. By having a single source of truth, improved performance, faster reporting, advanced analytics, and data integration, businesses can make more informed decisions. Using a data warehouse in conjunction with BI tools can result in improved data quality, enhanced reporting and analysis, faster data retrieval, and scalability. With BI tools like Tableau connecting to data warehouses to generate visualizations such as bar charts, users can gain insights into their data and make informed decisions. The combination of a data warehouse and BI tools enables organizations to make data-driven decisions, improve business performance, and achieve better outcomes while having a reliable and accurate source of data.

7. **Create 3 visualizations using Tableau. For each visualization, you should include the following:**
   - **i. Aim of the visualization**
   - **ii. The steps you took to create the visualisation**
   - **iii. Key findings from the visualisation**

⇒ VISUALIZATION 1

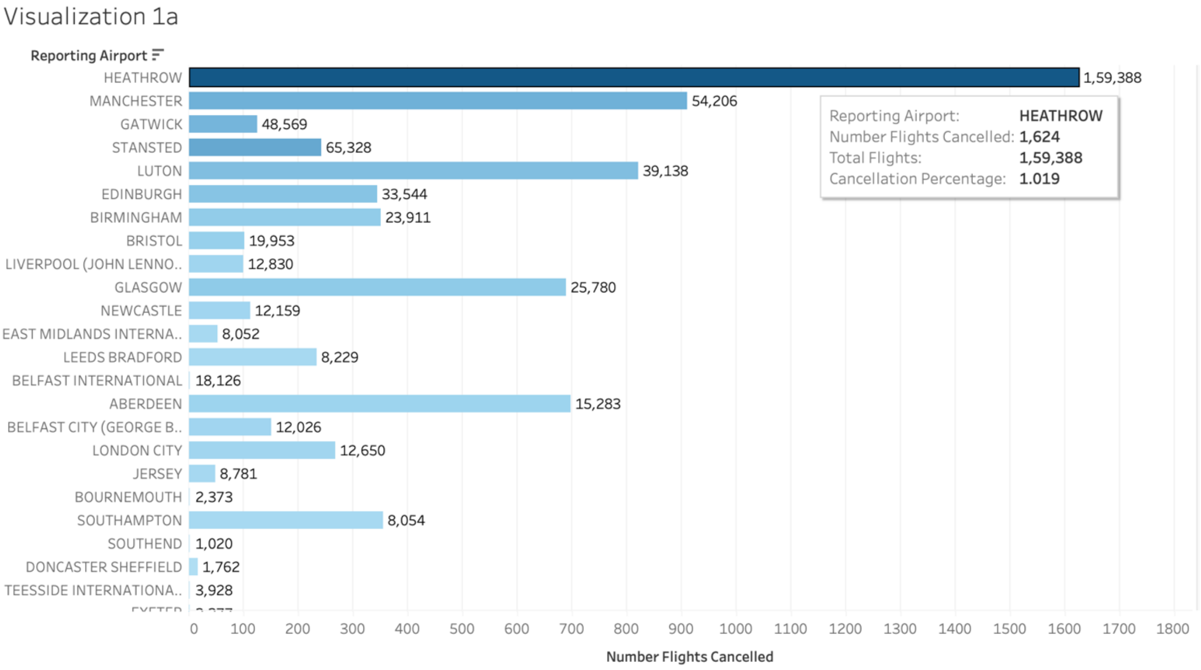


Fig 9: Visualization 1a

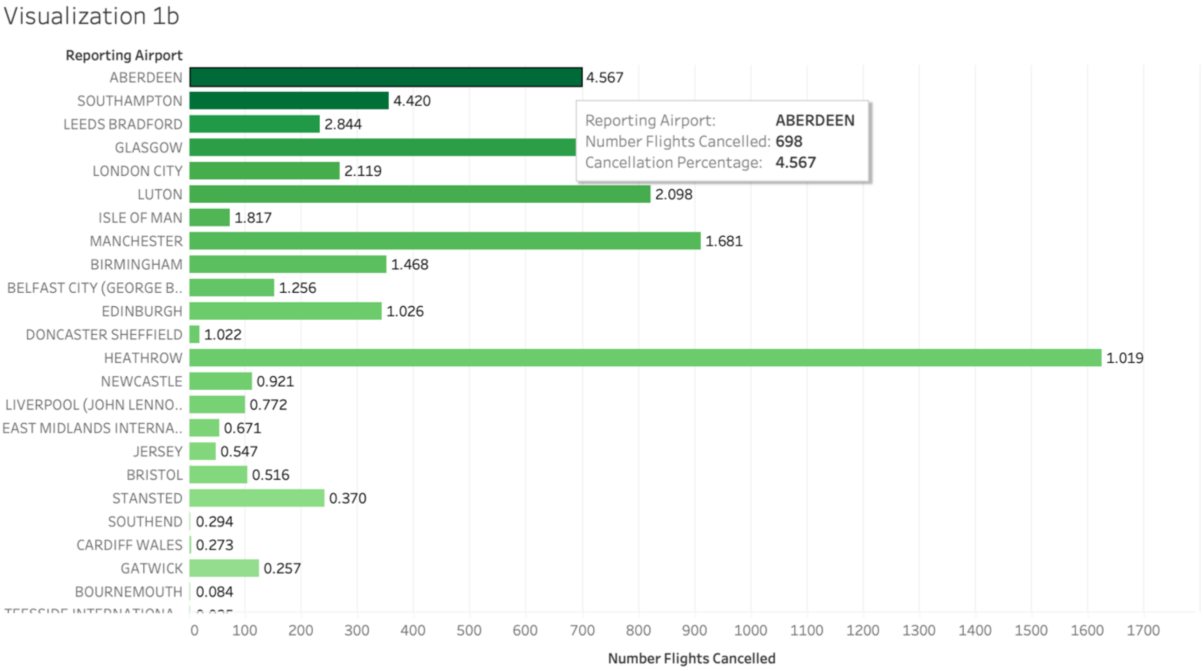


Fig 10: Visualization 1b

Aim of the visualization:

To identify the airport with the most number of flight cancellations by calculating the total number flights and the cancellation percentage at each airport for different reporting periods using appropriate formulas.

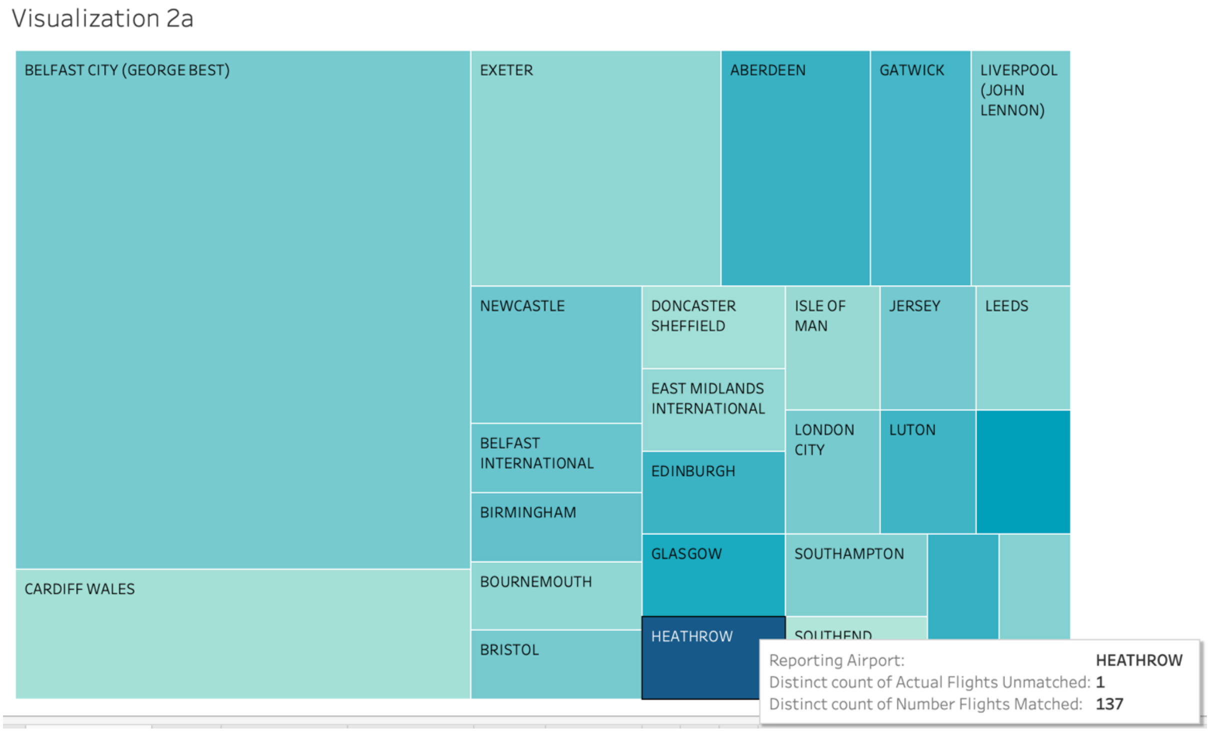
Steps to create the visualization:

- Connecting the Tableau to the dataset and create a new worksheet named Visualization 1a and 1b.
- Choosing the horizontal bar graph to get a clear picture of the data visualized.
- Drag and drop the "Reporting_Airport" and "Number_Flights_Cancelled" fields onto the Rows and Columns shelves, respectively and set measure for "Number_Flights_Cancelled" as Sum.
- Creating a calculated field for total number of flights by clicking on the "Analysis" menu in the top menu bar and selecting "Create Calculated Field". In the dialog box, name the new calculated field as **"Total Flights"** and by using the formula **"SUM([Number_Flights_Matched]) + SUM([Actual_Flights_Unmatched])".**This formula will calculate the total number of flights from Jan to Dec. Click "OK" to create the calculated field
- Creating a calculated field for cancellation percentage by clicking on the "Analysis" menu again and select "Create Calculated Field". Naming the new calculated field as **"Cancellation_Percentage"** and now by using the formula **"SUM([Number_Flights_Cancelled]) / SUM([Total Flights])"**. This formula will calculate the percentage of flights that were cancelled at each airport. Click "OK" to create the calculated field.
- Drag and Drop the calculated "Total flights" field to the Colour column in the "Marks" shelf to get the information based on the colours and also Drag and Drop the calculated "Cancellation Percentage" field to the text label in the "Marks" shelf to display the percentage of cancellation of flights at each of the airport.
- Now adding a filter for scheduled vs charter flights by dragging the "Scheduled_Charter" field to the "Filters" shelf and selecting "Scheduled" to filter out charter flights or selecting "Charter" to filter out scheduled flights.
- Then adding a filter for the reporting period by dragging the "Reporting_Period" field to the "Filters" shelf and select the desired reporting period according to the requirement.
- Now, sorting the bars in descending order by the Right-click on the "Reporting_Airport" axis and selecting "Sort". In the dialog box, selecting "Number_Flights_Cancelled" from the "Sort by" dropdown and choose "Descending" order respectively.

Key findings from the visualization:

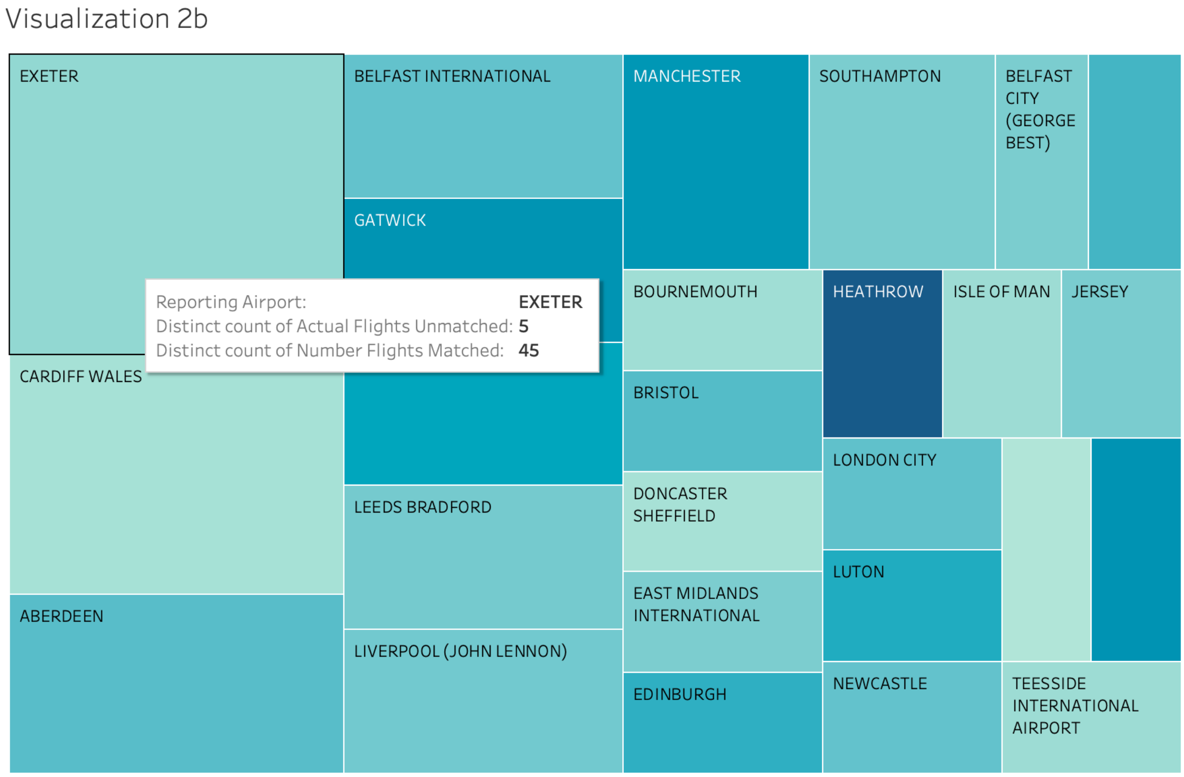
- The analysis in the Fig 9 shows that Heathrow airport has the highest number of total bookings, approximately 160,000, followed by Manchester airport with approximately 55,000 bookings, for the selected reporting period.
- The analysis in the Fig 10 shows the highest number of cancellation percentage is observed at Aberdeen airport, which is approximately 4.7%, followed by Southampton airport with approximately 4.5% cancellation rate, for the selected reporting period.
- The analysis also represent that some airports have higher than average cancellation rates, indicating a need for further investigation into their operations and policies. The airline companies operating at these airports can analyze the reasons for the cancellations and develop strategies to improve the punctuality and reliability of their flights.
- Overall, the findings of this analysis can assist airline companies and airports in identifying areas of improvement and developing strategies to enhance the customer experience.

⇒ **VISUALIZATION 2**



Reporting Airport: HEATHROW
Distinct count of Actual Flights Unmatched: 1
Distinct count of Number Flights Matched: 137

Fig 11: Visualization 2a



Reporting Airport: EXETER
Distinct count of Actual Flights Unmatched: 5
Distinct count of Number Flights Matched: 45

Fig 12: Visualization 2b

Aim of the visualization:

The aim of Visualization 2 is to compare the total number of flights matched and the actual number of flights unmatched in the reporting airports during different periods and find the airport that performs consistently well with on-time flights.

Steps to create the visualization:

- Connect the Tableau to the dataset and create a new worksheet named Visualization 2.
- Choose horizontal tree maps as they use color and size to encode data values, making it easy to quickly identify patterns and outliers in the data and compare the proportions of different categories.
- Drag and drop the "Actual Flights Unmatched" field to the size column in "Marks" shelf and set the "Measure" value to "Count Distinct" to determine the number of unique values in that column.
- Drag and drop the "Number Flights Matched" field to the color column in "Marks" shelf to indicate the values of a categorical variable identified by unique colors and set the "Measure" value to "Count Distinct.".
- Drag and drop the "Reporting Airport" field to the text column in "Marks" shelf to add and display additional information to the data.
- Add filters by dragging and dropping "Airline Name," "Origin Destination Country," and "Reporting Period" to filter the data based on user requirements.

Key findings from the visualization:

- The first analysis in Fig 11 shows that Belfast City (George Best) airport indicates by its size that the most number of flights are unmatched (around 20) when compared to other airports in the period of Jan to Jul, whereas Heathrow airport indicates the only airport with the most number of flights matched (around 137) represented by the color and also only one flight unmatched during the same period.
- The second analysis in Fig 12 shows that Exeter airport indicates the most number of flights unmatched (around 5), and another airport that is almost having the most number of flights unmatched is Cardiff Wales airport (around 4), when compared to other airports in the period of Aug to Dec. Once again, Heathrow airport indicates the only airport with the most number of flights matched (around 247) and only one flight unmatched during the same period.
- Overall, the analysis indicates that Belfast City (George Best) airport has the highest number of unmatching flights, while Heathrow airport is the most accurate airport with a higher number of flights matched.
- To address the issue of unmatched flights, airlines and airports can adopt several measures, such as implementing more efficient scheduling and operations, investing in better technology, and improving communication between airlines and airports.
- In conclusion, Visualization 2 provides insights into the performance of airports regarding flight matching accuracy. The analysis highlights the need for airports and airlines to work towards reducing the number of unmatching flights, with Heathrow airport standing out as the most accurate airport.

⇒ **VISUALIZATION 3**

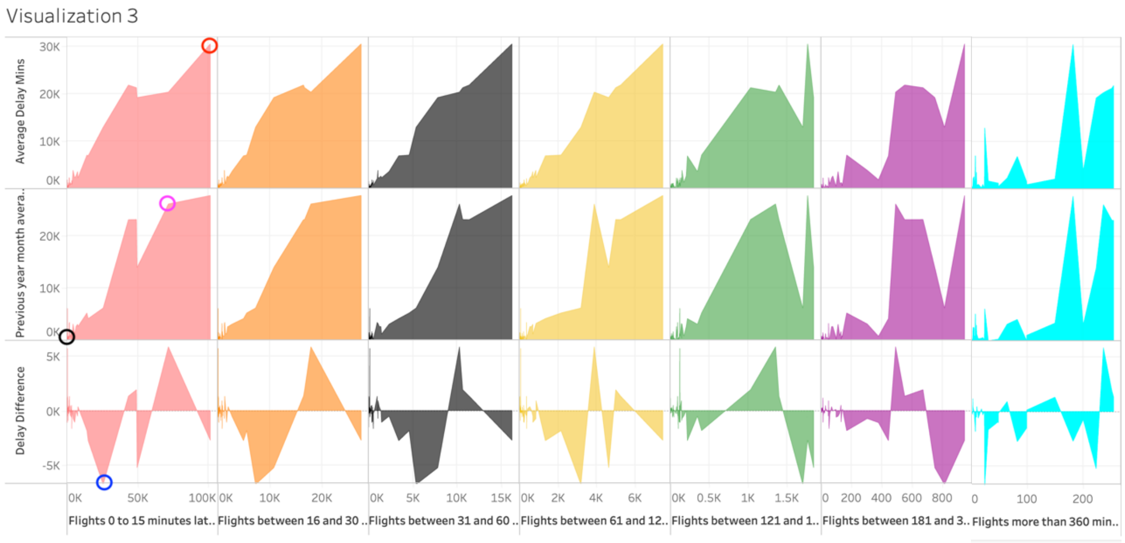
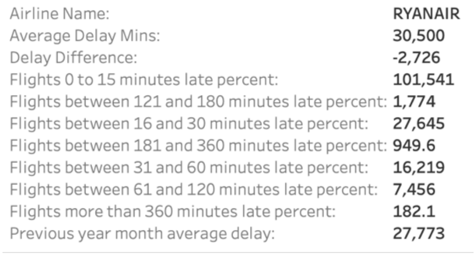


Fig 13: Visualization 3

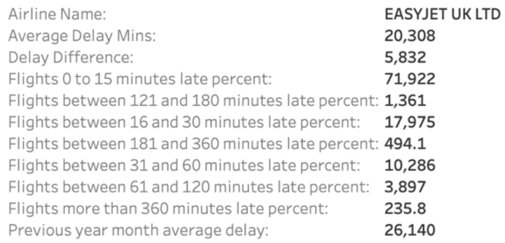
○ The red circle provides information about Ryanair Airlines.



Airline Name:	RYANAIR
Average Delay Mins:	30,500
Delay Difference:	-2,726
Flights 0 to 15 minutes late percent:	101,541
Flights between 121 and 180 minutes late percent:	1,774
Flights between 16 and 30 minutes late percent:	27,645
Flights between 181 and 360 minutes late percent:	949.6
Flights between 31 and 60 minutes late percent:	16,219
Flights between 61 and 120 minutes late percent:	7,456
Flights more than 360 minutes late percent:	182.1
Previous year month average delay:	27,773

Fig 8: Ryanair Airlines Info

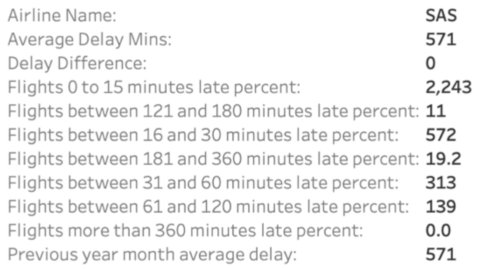
- The pink circle provides information about EasyJet UK LTD Airlines.



Airline Name:	EASYJET UK LTD
Average Delay Mins:	20,308
Delay Difference:	5,832
Flights 0 to 15 minutes late percent:	71,922
Flights between 121 and 180 minutes late percent:	1,361
Flights between 16 and 30 minutes late percent:	17,975
Flights between 181 and 360 minutes late percent:	494.1
Flights between 31 and 60 minutes late percent:	10,286
Flights between 61 and 120 minutes late percent:	3,897
Flights more than 360 minutes late percent:	235.8
Previous year month average delay:	26,140

Fig 9: Easyjet UK LTD Airlines Info

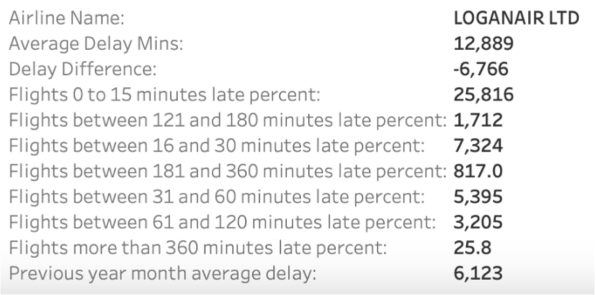
- The black circle provides information about SAS Airlines.



Airline Name:	SAS
Average Delay Mins:	571
Delay Difference:	0
Flights 0 to 15 minutes late percent:	2,243
Flights between 121 and 180 minutes late percent:	11
Flights between 16 and 30 minutes late percent:	572
Flights between 181 and 360 minutes late percent:	19.2
Flights between 31 and 60 minutes late percent:	313
Flights between 61 and 120 minutes late percent:	139
Flights more than 360 minutes late percent:	0.0
Previous year month average delay:	571

Fig 10: SAS Airlines Info

- The blue circle provides information about Loganair LTD Airlines.



Airline Name:	LOGANAIR LTD
Average Delay Mins:	12,889
Delay Difference:	-6,766
Flights 0 to 15 minutes late percent:	25,816
Flights between 121 and 180 minutes late percent:	1,712
Flights between 16 and 30 minutes late percent:	7,324
Flights between 181 and 360 minutes late percent:	817.0
Flights between 31 and 60 minutes late percent:	5,395
Flights between 61 and 120 minutes late percent:	3,205
Flights more than 360 minutes late percent:	25.8
Previous year month average delay:	6,123

Fig 11: Loganair LTD Airlines Info

Aim of the visualization:

This visualization aims to provide an overview of the airline delays and compare the average delay times in the current reporting period to the average delay times in the previous year's reporting period, and to analyze the airlines that have reduced and increased the delay times compared to the previous year using a formula.

Steps to create the visualization:

- Connecting the Tableau to the dataset and create a new worksheet named Visualization 3.
- Choosing the Area graph to get a clear picture of the data.
- Then, drag and drop the Flights_0_to_15_minutes_late_percent, Flights_16_to_30_minutes_late_percent, Flights_31_to_60_minutes_late_percent, Flights_61_to_120_minutes_late_percent, Flights_121_to_180_minutes_late_percent, Flights_181_to_360_minutes_late_percent, and Flights_more_than_360_minutes_late_percent fields to the columns shelf and set each field measure as Sum.
- Adding a new calculation field to get the delay difference between the previous and current year month delay field named **Delay Difference** using the formula **[Previous year month average delay]-[Average Delay Mins]** and click on OK.
- Then, drag and drop the Average Delay Mins, the Previous Year's Month Average Delay, and the newly calculated Delay Difference to the rows shelf and set the measure for all three fields as Sum.
- To get a more accurate visualization, drag and drop the Scheduled Charter and Reporting Period to the filter shelf data, which enabled the option to choose both Scheduled and Charter or choose one among Scheduled or Charter. We can also obtain information about the airlines' delay for every month in Reporting Period by selecting the appropriate month.
- Lastly, putting the Airline Name field to the “Detail” in the Marks shelf to obtain the airline names that had been causing the delays by hovering the cursor on the graph.

Key findings from the visualization:

- The analysis shows that Easy Jet UK LTD reduced their delay time by almost 6000 Mins compared to the previous year, whereas Loganair LTD's delay time increased by almost 6000 Mins compared to last year.
- It is also observed that SAS is the only airline where the current actual delay time and previous delay time are the same, hence the delay difference is 0Mins.
- It is also found that Ryanair has the highest delayed flight between 0-180 minutes, and Jet2.Com LTD and British Airways have the highest delayed flight between 180-360 minutes and more than 360 minutes, respectively.

- Finally, it can be observed that Ryanair has a higher percentage of delayed flight minutes compared to other airlines, while Easy Jet UK LTD has shown significant improvement compared to the previous year's delay time. On the other hand, Loganair LTD has shown a substantial increase in delay time.
- Overall, the visualization provides a comprehensive overview of airline delays, comparing the average delay times in the current reporting period to the average delay times in the previous year's reporting period. The analysis of airlines that have reduced their delay times and those that have increased their delay times compared to the previous year can be used by airline companies to improve their operations and reduce delays. It is essential to continue to monitor and analyze airline delays to identify areas of improvement and ensure a better travel experience for passengers.

8. Conclusion

- In conclusion, building a data warehouse for the United Kingdom airport industry using the provided data set can offer significant benefits, such as improving data accuracy, simplifying data retrieval, providing deeper insights into historical data, and enabling more informed decision-making ultimately leading to improved performance and customer satisfaction.
- Then also got a clear picture of consolidating data from various sources into a single system, the decision-makers can leverage the data to identify trends, patterns, and irregularities in punctuality statistics, prioritize areas for improvement, and make data-driven decisions. Overall, a data warehouse can be a crucial tool for enhancing the overall performance and customer satisfaction in the UK airport sector.
- The star schema design is an effective approach for organizing data in a data warehouse for reporting and analysis purposes. The design decision of creating dimension tables for Reporting Period, Reporting Airport, Origin Destination, Airline Name, and Scheduled Charter enables users to analyze performance across different time periods, airports, airlines, and flight types, providing valuable insights into business operations. The fact table Punctuality Statistics serves as the central table that captures key metrics and is connected to the dimension tables using foreign keys, allowing users to easily navigate and query the data. Overall, the star schema design ensures data is organized in a way that is easy to understand, use, and maintain for decision-making purposes.
- This coursework also helped in creating and populating a database which involves several steps, including preparing and transforming the data using an ETL process, creating a logical model for the database, and populating the database with the dimension and fact tables.
- Then a star schema model was used for this particular dataset, which allows for efficient querying and analysis of the airline flight data. By visualizing the data using various charts and graphs, we can gain insights into the performance of airlines during the reporting period and identify trends and patterns in the data. Overall, the process of creating and populating a database is crucial for analyzing large datasets and extracting meaningful insights.
- Overall, both the airline industry and other organizations can benefit from the combination of OLAP cubes and business intelligence tools like Power BI and Tableau, as it provides a powerful platform for storing, managing, analyzing, and visualizing data. This integration allows for improved data quality, enhanced reporting and analysis, faster data retrieval, and scalability, and ultimately leads to more informed decision-making and better business outcomes, resulting in increased profitability and a competitive edge.
- Creating and populating a database using an ETL process, was a bit complex and time-consuming task. However, by carefully following the steps and using appropriate tools and techniques, the process was completed effectively and more accurately by enabling efficient querying and analysis of the airline flight data.

- To ensure a successful data warehousing project, I learned how it should be started by carefully planning the project objectives, identifying and prioritizing relevant data sources, using appropriate data modeling techniques, paying close attention to data quality, and also by selecting suitable business intelligence tools.

REFERENCES

[1] Dropbase.io, "Data Warehouse Concepts".
Available: https://www.dropbase.io/post/what-is-a-data-warehouse-and-how-can-it-benefit-organizations [Accessed 22-Apr-2023].

[2] Javatpoint.com, "What is Star Schema".
Available: https://www.javatpoint.com/data-warehouse-what-is-star-schema .
[Accessed 24-Apr-2023].

[3] Techtarget.com, "Power BO",
Available: https://www.techtarget.com/searchcontentmanagement/definition/Microsoft-Power-BI [Accessed 27-Apr-2023].

[4] Ecapitaladvisors.com, "Why OLAP",
Available: https://ecapitaladvisors.com/blog/why-olap/ .
[Accessed 30-Apr-2023].

[5] Researchgate.net, "Airline Applications of BI Systems"
Available:
https://www.researchgate.net/publication/282966174_Airline_Applications_of_Business_Intelligence_Systems. [Accessed 3-May-2023].

[6] Tableau.com, "Data Visualization",
Available: https://www.tableau.com/en-gb/learn/articles/data-visualization.
[Accessed 5-May-2023].

[7] Papers.ssrn.com, "Data Warehousing in Airline Industry"
Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2519737
[Accessed 6-May-2023].

ORACLE APEX

Workspace Name: sujaygrama1
Email-id: K2201621@KINGSTON.AC.UK
Password: kingston@1234