CI7320 Assignment 2

Gaikwad Kunal Suresh

K2367430

Table of Contents

[Report 2](#_Toc164808163)

[Chapter 1: Benefits of building Data warehouse 2](#_Toc164808164)

[Chapter 2: Data warehouse schema Design 4](#_Toc164808165)

[Chapter 3: Create Table Statements 5](#_Toc164808166)

[Chapter 4: Database Creation and Population 6](#_Toc164808167)

[Chapter 5: Discussion on Advantages of OLAP Cubes in airline industry 11](#_Toc164808168)

[Chapter 6: Visualisations using Tableau. 16](#_Toc164808169)

[Conclusion: 25](#_Toc164808170)

# Report

## Chapter 1: Benefits of building Data warehouse

Building a data warehouse for the flight punctuality data set offers several benefits, enhancing data accessibility, analysis, and decision-making processes within the airline industry:

Data Mining and Analytics:

Data warehouses serve as a foundation for data mining and advanced analytics initiatives. By providing a centralized repository of clean, integrated data, data warehouses enable organizations to perform sophisticated analytical techniques such as predictive modelling, clustering, and segmentation to uncover valuable insights and patterns within the data.

Centralized Data Repository:

By consolidating punctuality statistics from various UK airports into a single data warehouse, airlines and aviation authorities can access comprehensive and centralized data for analysis and reporting. This reduces the need to gather information from disparate sources, streamlining data management processes.

### Return on Investment:

In the airline industry, data warehouses offer a substantial return on investment by allowing airlines to leverage insights from data warehouses to optimize route planning, pricing strategies, aircraft utilization, and customer segmentation, leading to increased profitability and market share.

### Data Quality Management:

Data warehouses play a vital role in data quality management within the airline industry by providing robust mechanisms for data cleansing, validation, and enrichment. Reliability in data quality enables airlines to make informed decisions based on trustworthy data, enhances the effectiveness of marketing campaigns, improves customer satisfaction, and minimizes risks associated with inaccurate or incomplete data.

Historical Analysis and Trend Identification:

Flight punctuality data collected over time can provide valuable insights into historical performance trends. A data warehouse enables stakeholders to analyze past performance, identify patterns, and anticipate future trends. This information can be utilized to optimize flight scheduling, resource allocation, and operational planning, leading to improved efficiency and customer satisfaction.

### Compliance and Regulatory Reporting:

Consolidating data from various sources into a centralized repository, data warehouses facilitate the creation of compliant reports related to safety, security, financial transactions, and passenger information. This capability ensures transparency, accountability, and adherence to regulatory standards in airline operations, thereby mitigating legal risks and maintaining the trust of regulatory authorities, passengers, and stakeholders.

### Role-Based Access Control:

Data warehouses support role-based access control (RBAC), allowing airlines to restrict access to sensitive data based on users' roles and responsibilities. RBAC enables granular control over who can access, view, modify, or delete data within the warehouse, reducing the risk of unauthorized access and data breaches.

Support for Decision-Making:

By providing timely access to comprehensive and relevant data, a data warehouse empowers decision-makers to make informed decisions. Whether it's optimizing flight schedules, allocating resources, or addressing operational challenges, stakeholders can leverage actionable insights derived from the data warehouse to enhance operational efficiency, minimize disruptions, and improve overall performance.

Scalability and Flexibility:

As the volume and complexity of data grow, a data warehouse offers scalability and flexibility to accommodate evolving business needs. Whether it's integrating additional data sources, expanding analytical capabilities, or supporting new use cases, the architecture of a data warehouse allows for seamless scalability without compromising performance or reliability.

### Real-time Monitoring and Alerts:

Data warehouses enable real-time monitoring of key performance indicators (KPIs) related to flight operations, such as departure and arrival times, gate utilization, turnaround times, and crew availability. Airlines can set up automated alerts within the data warehouse to notify stakeholders of deviations from expected performance metrics, allowing them to respond promptly and address issues that may affect flight punctuality.

### Performance Benchmarking:

Data warehouses enable airlines to benchmark their flight punctuality performance against industry standards and best practices. By comparing performance metrics with peer airlines and identifying areas for improvement, airlines can establish performance targets, track progress over time, and implement continuous improvement initiatives to enhance flight punctuality and operational reliability.

In conclusion, building a data warehouse for the flight punctuality data set provides numerous benefits, including centralized data management, improved data quality, historical analysis, advanced analytics, support for decision-making, and scalability. By leveraging these benefits, stakeholders in the airline industry can gain valuable insights, optimize operations, and enhance overall performance.

## Chapter 2: Data warehouse schema Design

OneDrive link to view/download Data warehouse schema: [datawarehouse-schema-link](https://kingstonuniversity-my.sharepoint.com/:i:/g/personal/k2367430_kingston_ac_uk/EVVsD26dqAhBlCHsbkF7gQkBl7-vObbDWEcfR9yINo4FRA?e=LR6tut)

### Entities and their justification:

1. FLIGHT\_PUNCTUALITY\_FACT: This entity represents factual information about flight punctuality, including various metrics such as the number of flights matched, number of flights cancelled, flight types, percentage of flights delayed by different durations, and more. It serves as a main component in analysing the performance and reliability of airline services.
2. AIRPORT\_DIM: An airport dimension provides detailed information about airports, including attributes such as airport code, airport name, and other relevant data. It serves as a reference point for understanding airport characteristics and facilitating airport-related analyses and reporting.
3. DATE\_DIM: The date dimension is essential for organizing and analysing data based on dates and time periods. It includes attributes such as date, day, month, year, quarter. It facilitates time-based analysis, trend analysis, and comparisons across different time periods in the data warehouse. Another reason to add Date Dimension is to avoid any user errors while entering date since date is very crucial for our analysis.
4. AIRLINE\_DIM: An airline dimension contains information about airlines, including attributes such as airline code, airline name. It allows for analysis of airline performance, market share, route networks, and other airline-related metrics in the context of the data warehouse.
5. ORIGIN\_DESTINATION\_DIM: This entity provides information about the destination cities of flights, enabling analysis of flight routes, passenger flows, and geographic patterns. It includes attributes such as destination city names, identifiers, and possibly geographic coordinates, facilitating route analysis and geographical insights.
6. ORIGIN\_DESTINATION\_COUNTRY\_DIM: This entity focuses specifically on countries associated with flight destination countries. It provides details about countries such as country code, name. It enables analysis of international flight patterns, market trends, and country-specific performance metrics within the data warehouse context.

## Chapter 3: Create Table Statements

#### AIRPORT\_DIM

CREATE TABLE AIRPORT\_DIM (

REPORTING\_AIRPORT\_CODE VARCHAR2(3) PRIMARY KEY,

REPORTING\_AIRPORT\_NAME VARCHAR2(250)

);

#### DATE\_DIM

CREATE TABLE DATE\_DIM (

DATE\_PK NUMBER PRIMARY KEY,

DATE\_COLUMN DATE,

QUARTER VARCHAR2(10),

DAY VARCHAR2(20)

);

#### AIRLINE\_DIM

CREATE TABLE AIRLINE\_DIM (

AIRLINE\_CODE VARCHAR2(5) PRIMARY KEY,

AIRLINE\_NAME VARCHAR2(250)

);

#### ORIGIN\_DESTINATION\_COUNTRY\_DIM

CREATE TABLE ORIGIN\_DESTINATION\_COUNTRY\_DIM (

ORIGIN\_DESTINATION\_COUNTRY\_ID NUMBER PRIMARY KEY,

ORIGIN\_DESTINATION\_COUNTRY VARCHAR2(250)

);

#### ORIGIN\_DESTINATION\_DIM

CREATE TABLE ORIGIN\_DESTINATION\_DIM (

ORIGIN\_DESTINATION\_ID NUMBER PRIMARY KEY,

ORIGIN\_DESTINATION VARCHAR2(100),

ORIGIN\_DESTINATION\_COUNTRY\_ID NUMBER,

FOREIGN KEY (ORIGIN\_DESTINATION\_COUNTRY\_ID) REFERENCES ORIGIN\_DESTINATION\_COUNTRY\_DIM(ORIGIN\_DESTINATION\_COUNTRY\_ID)

);

#### FLIGHT\_PUNCTUALITY\_FACT

CREATE TABLE FLIGHT\_PUNCTUALITY\_FACT (

FLIGHT\_PUNCTUALITY\_ID NUMBER PRIMARY KEY,

NUMBER\_FLIGHTS\_MATCHED NUMBER,

NUMBER\_FLIGHTS\_CANCELLED NUMBER,

FLIGHTS\_CANCELLED\_PERCENT NUMBER,

FLIGHT\_TYPE VARCHAR2(100),

FLIGHTS\_MORE\_THAN\_15\_MINUTES\_EARLY\_PERCENT NUMBER,

FLIGHTS\_15\_MINUTES\_EARLY\_TO\_1\_MINUTE\_EARLY\_PERCENT NUMBER,

FLIGHTS\_0\_TO\_15\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_BETWEEN\_16\_AND\_30\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_BETWEEN\_31\_AND\_60\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_BETWEEN\_61\_AND\_120\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_BETWEEN\_121\_AND\_180\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_BETWEEN\_181\_AND\_360\_MINUTES\_LATE\_PERCENT NUMBER,

FLIGHTS\_MORE\_THAN\_360\_MINUTES\_LATE\_PERCENT NUMBER,

PREVIOUS\_YEAR\_MONTH\_FLIGHTS\_MATCHED NUMBER,

PREVIOUS\_YEAR\_MONTH\_EARLY\_TO\_15\_MINS\_LATE\_PERCENT NUMBER,

PREVIOUS\_YEAR\_MONTH\_AVERAGE\_DELAY NUMBER,

FLIGHTS\_UNMATCHED\_PERCENT NUMBER,

AVERAGE\_DELAY\_MINS NUMBER,

REPORTING\_PERIOD NUMBER,

REPORTING\_AIRPORT\_CODE VARCHAR2(3),

ORIGIN\_DESTINATION\_ID NUMBER,

ARLINE\_CODE VARCHAR2(5),

CONSTRAINT FK\_FC\_REPORT\_PR FOREIGN KEY (REPORTING\_PERIOD) REFERENCES DATE\_DIM(DATE\_PK),

CONSTRAINT FK\_FC\_REPORT\_AC FOREIGN KEY (REPORTING\_AIRPORT\_CODE) REFERENCES AIRPORT\_DIM(REPORTING\_AIRPORT\_CODE),

CONSTRAINT FK\_FC\_ORG\_DEST\_ID FOREIGN KEY (ORIGIN\_DESTINATION\_ID) REFERENCES ORIGIN\_DESTINATION\_DIM(ORIGIN\_DESTINATION\_ID),

CONSTRAINT FK\_FC\_ARL\_CD FOREIGN KEY (ARLINE\_CODE) REFERENCES AIRLINE\_DIM(AIRLINE\_CODE)

);

## Chapter 4: Database Creation and Population

### Data Processing

#### Combining files:

1. **Pandas** library and **Microsoft Excel** was for data processing. Following code was used to combine all the months into single file.A screen shot of a computer program

   Description automatically generated
2. There was problem with encoding of “Jul2023\_Punctuality\_Statistics\_UK\_airports”. So, I had to use different encoding to read that file.
3. After combining all files in one csv three extra columns **‘previous\_year\_month\_flights\_matched’, ‘reporting\_period’ ‘flights\_unmatched\_percent’** were created.
4. Duplicate column with name **'reporting\_period'** was created because of encoding error. It contained only the values of July month, and values for rest of months were empty. **'ï»¿reporting\_period'** contained reporting period of all other month. So manually pasted all the values of July month in to **'ï»¿reporting\_period'** using **Microsoft Excel** and renamed it to ‘reporting\_period. Then deleted the duplicate column.
5. The column with name **‘previous\_year\_month\_flights\_matched’** is missing from January file.There is an extra column with name **‘flights\_unmatched\_percent’** is only present for February. Blank cells were filled with zero for these columns.

#### Column Transformations:

1. After combining all files together, the column transformation were performed on the combined file using **pandas library**. Following code was used to do column transformations: A screen shot of a computer code

   Description automatically generated

A screen shot of a computer code

Description automatically generated

1. The column **reporting\_period** was converted to proper date format.
2. In the column **‘scheduled\_charter’** value **S** was replaced with **“Scheduled”** and **C** was replaced with **“Charter”** for better understanding. Later this column was renamed to a better name **“flight\_type”.**
3. All the columns with percent values like **'flights\_more\_than\_15\_minutes\_early\_percent', 'flights\_15\_minutes\_early\_to\_1\_minute\_early\_percent', 'flights\_0\_to\_15\_minutes\_late\_percent',** etc were divided by 100 to convert them to proper percent value.
4. A new column **‘reporting\_airport\_code’** was added which contains IATA code for the airports.
5. A new column **‘airline\_code’** was added which contains IATA code for the airlines. Final file with all transformations is shown below: A screenshot of a computer

   Description automatically generated

The Data for all the tables was generated with the help of ChatGpt and the provided dataset files. Appropriate prompts were given to ChatGPt and data obtained in the form of csv. This data was then loaded on to the tables in Apex Oracle by following some simple steps:

1. Go to the **Object Browser** tab in Apex Oracle.
2. Then on the left hand there is a **Tables** drawer. Select the table where you want to load the data.
3. Click on the **Data** tab on the Right-Hand side of the screen.
4. Then click on the **Load Data** option. A pop-up window will appear. Then select the csv file that you have generated.
5. After selecting the csv file make sure that all the columns of the table are mapped properly with the csv file. Click on the **Load Data button** on the bottom right hand on the screen. In this way, the data is loaded on the table using a csv file.

#### AIRLINE\_DIM:

A screenshot of a computer

Description automatically generated

#### AIRPORT\_DIM

A screenshot of a computer

Description automatically generated

*ORIGIN\_DESTINATION\_DIM*

A screenshot of a computer

Description automatically generated

#### ORIGIN\_DESTINATION\_COUNTRY\_DIM

A screenshot of a computer

Description automatically generated

#### DATE\_DIM

A screenshot of a computer

Description automatically generated

#### FLIGHT\_PUNCTUALITY\_FACT

A black and white screen with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a black and white screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A black and grey striped background

Description automatically generated

A screen shot of a black and white screen

Description automatically generated

## Chapter 5: Discussion on Advantages of OLAP Cubes in airline industry

### Efficient Data Analysis:

* OLAP cubes allow airlines to analyse large volumes of data efficiently. By pre-aggregating data into multidimensional structures, complex queries can be answered quickly.
* For airlines, this means faster decision-making, better resource allocation, and improved operational efficiency.

### Flexible Reporting and Querying:

* OLAP cubes provide a flexible framework for querying and reporting. Airlines can slice and dice data along various dimensions (such as time, routes, aircraft types, and customer segments) to gain insights.
* Users can create custom reports, drill down into details, and explore data interactively.

### Support for Trend Analysis and Forecasting:

* OLAP cubes enable trend analysis by allowing airlines to track performance over time. Historical data can be easily accessed and compared.
* Forecasting models can leverage OLAP cubes to predict future demand, optimize flight schedules, and allocate resources effectively.

### Enhanced Decision Support:

* Airlines can use OLAP cubes to support strategic, tactical, and operational decisions. For example:

Strategic: Analyzing market trends, identifying growth opportunities, and planning expansion.

Tactical: Optimizing crew schedules, pricing strategies, and route planning.

Operational: Monitoring real-time flight performance, crew availability, and maintenance schedules.

### Consolidation of Data Sources:

* OLAP cubes consolidate data from various sources (such as reservations, flight operations, crew management, and financial systems).
* This integration simplifies data access, reduces redundancy, and ensures consistent reporting across the organization.

### Improved User Experience

* OLAP cubes provide a user-friendly interface for exploring data. Business analysts, executives, and operational staff can access relevant information without complex SQL queries.
* Interactive dashboards and visualizations enhance the user experience.

### Scalability and Performance:

* OLAP cubes are designed for scalability. As airlines grow and collect more data, the cube structure remains efficient.
* Performance remains stable even with large datasets, making OLAP suitable for handling the airline industry's data volume.

### Industry-Specific Metrics and Insights:

* Solutions like Oracle Airlines Data Model offer pre-built OLAP cubes tailored to the airline industry.
* These cubes provide industry-specific metrics, such as load factors, revenue per available seat mile (RASM), and passenger segmentation.

### Examples of OLAP applications:

#### OLAP Cube with Fare Structure, Operational Efficiency and Customer Behaviour

1. Fare Structure Dimension:

* This dimension provides crucial insights into the pricing strategies and revenue management practices of the airline industry.
* By analysing data along this dimension, airlines can understand the impact of different fare classes, fare types, and fare rules on overall revenue generation and passenger behaviour. For instance, airlines can evaluate the performance of premium fare classes versus economy fare classes, identify pricing strategies that maximize yield per passenger, and optimize inventory allocation to maximize revenue on high-demand flights.
* Furthermore, analysis of auxiliary revenue components within the fare structure dimension enables airlines to identify opportunities for upselling additional services and enhancing overall ancillary revenue streams.

1. Customer Behavior Dimension:

* This dimension offers valuable insights into the preferences, behaviours, and booking patterns of airline passengers. By examining data along this dimension, airlines can segment customers based on factors such as booking channel, booking frequency, and seat preferences, allowing for targeted marketing campaigns and personalized customer experiences.
* For example, airlines can tailor promotional offers to specific customer segments, optimize distribution channels based on customer preferences, and design seating configurations that align with passenger preferences.
* Additionally, analysis of customer behaviour data enables airlines to forecast demand more accurately, optimize pricing strategies based on customer segments, and improve overall customer satisfaction by catering to the individual needs and preferences of passengers throughout the booking journey.

1. Operational Efficiency Dimension:

* The Operational Efficiency Dimension provides critical insights into the performance and effectiveness of airline operations, encompassing various aspects such as airport operations, ground handling processes, and turnaround times. By delving into data along this dimension, airlines can identify inefficiencies, bottlenecks, and areas for improvement within their operational processes. For example, analysis of turnaround times at different airports can help airlines optimize aircraft utilization and minimize ground time, leading to cost savings and increased operational efficiency.
* Moreover, the Operational Efficiency Dimension plays a crucial role in enhancing the overall customer experience by ensuring smooth and efficient operations throughout the passenger journey.

A diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of

Description automatically generated

A diagram of a diagram of a business class

Description automatically generated

#### OLAP Cube with Passenger Demographic, Flight Characteristics,

1. Passenger Demographics Dimension:

* Attributes: Age Group, Gender, Nationality, Travel Purpose
* This dimension provides insights into the demographic characteristics of passengers traveling with the airline. Analysis along this dimension can help airlines understand the preferences, behaviors, and needs of different passenger segments. For example, airlines can tailor marketing campaigns, onboard services, and loyalty programs to specific demographic groups, enhancing customer satisfaction and loyalty.

1. Flight Characteristics Dimension**:**

* Attributes: Flight Distance, Flight Duration, Aircraft Type, Cabin Class
* This dimension focuses on the characteristics of individual flights operated by the airline. Analysis along this dimension allows airlines to evaluate the performance of different flight segments, optimize aircraft deployment, and tailor services based on flight duration and cabin class. For instance, airlines can adjust pricing strategies, seat configurations, and in-flight amenities to match the preferences of passengers on long-haul versus short-haul flights.

1. Revenue Streams Dimension**:**

* Attributes: Ticket Revenue, Cargo Revenue, Partner Revenue
* This dimension tracks the various revenue streams generated by the airline's operations. Analysis along this dimension enables airlines to assess the contribution of different revenue sources to overall profitability, identify opportunities for revenue growth, and optimize pricing and revenue management strategies. For example, airlines can analyze trends in ancillary revenue from services such as baggage fees, seat upgrades, and in-flight purchases to maximize ancillary revenue generation.

A diagram of a diagram of a cube

Description automatically generated with medium confidence

A diagram of a cube

Description automatically generated with medium confidence

## Chapter 6: Visualisations using Tableau.

### Punctuality of Top UK Airlines

#### Aim:

The goal of this visualization is to evaluate and compare the punctuality of top UK airlines by illustrating the percentage of flights that are either early or on-time against those that are delayed. This assessment aims to provide travellers, airline operators, and industry regulators with clear insights into which airlines are more proficient at managing their flight schedules and which may require operational improvements to enhance their punctuality.

#### How to load data into Tableau:

### After finishing processing and loading data into our database, open **Tableau Desktop**.

### On the side bar under Connect, pick the right option based on the type of file you have. For example, if your file is a CSV, choose **"Text file."** If it's an Excel file (XLSX), pick **"Microsoft Excel."**

### Once you've chosen the right option, drag your file onto the empty space on the right side.

### On the bottom half you can preview your CSV file and even perform some processing using various Tableau tools like Data Interpreter, Pivot, split, etc.

### You can also connect more datasets if you want and use them together in Tableau.

### When you're satisfied with your data, go to the bottom menu, and click on **"Sheet 1"** next to **"Data Source."**

1. You'll then see all the dimensions and measures from your dataset on the right side of the screen.

#### Steps required to create the visualisation:

1. Begin by creating a calculated field titled 'Early or On-time %'. Navigate to the top menu bar and select 'Analysis', followed by 'Create Calculated Field'. Input the following formula into the designated text box:

AVG([Flights More Than 15 Minutes Early Percent])+

AVG([Flights 15 Minutes Early To 1 Minute Early Percent])+

AVG([Flights 0 To 15 Minutes Late Percent])

1. Proceed to create another calculated field named 'Late %' utilizing the formula:

AVG([Flights Between 16 And 30 Minutes Late Percent])+

AVG([Flights Between 31 And 60 Minutes Late Percent])+

AVG([Flights Between 61 And 120 Minutes Late Percent])+

AVG([Flights Between 121 And 180 Minutes Late Percent])+

AVG([Flights Between 181 And 360 Minutes Late Percent])+

AVG([Flights More Than 360 Minutes Late Percent])

1. Drag the dimension 'Airline Name' to the columns bar, and subsequently, place both calculated fields into the row section.
2. Tableau will generate two distinct bar graphs. To stack them, right-click on the X-axis of the lower graph and select 'Dual Axis'.
3. Right click again on the same axis and select 'Synchronize Axis' option. This option ensures uniformity in the appearance of the bars.
4. As the dataset may contain an extensive list of airline names, rendering each one in the visualization could result in clutter. To mitigate this, implement a filter. Drag the 'Airline Name' dimension to the Filter pane, select the desired airline names, and then opt for 'Show Filter' in the context menu to display the filter on the sheet.
5. Hide the axis situated on the right-hand side by right-clicking on it and deselecting 'Show Header'.
6. Drop the measures that are present in the rows shelf into their respective Label section to display the exact labels.

A graph of blue and orange bars

Description automatically generated

#### Key findings from visualisation:

1. On-time Performance:
   * Jet2.com Ltd has the highest early or on-time percentage at 64.1%, indicating a strong record of punctuality.
   * Virgin Atlantic International follows closely with an early or on-time percentage of 56.7%.
   * EasyJet UK Ltd also demonstrates a notable on-time performance with an early or on-time percentage of 55.6%.
2. Late Performance:
   * EasyJet UK Ltd has the highest late percentage at 34.7%, suggesting a higher frequency of delays compared to other airlines.
   * British Airways PLC follows with a late percentage of 33.0%, indicating a significant portion of flights experiencing delays.
   * Wizz Air UK Ltd and Jet2.com Ltd have comparatively lower late percentages at 25.2% and 28% respectively.
3. Overall Performance:
   * While some airlines prioritize on-time performance, others struggle with higher rates of delays, suggesting potential areas for improvement in operational efficiency and scheduling.

### Flights Delay Pattern: Classifying Short Delay and Long Delays

#### Aim:

The aim of this visualization is to provide insights into the performance of various airports in terms of long delay and short delay percentages. By visualizing this data, stakeholders such as airport authorities, airlines, and travellers can gain a better understanding of which airports experience higher rates of delays and the magnitude of those delays, both in terms of long delays and short delays. This information can be used to inform decisions related to flight scheduling, airport infrastructure improvements, and travel planning.

#### Steps required to create the visualisation:

1. Start by creating a calculated field titled 'Short Delay%'. Navigate to the top menu bar and select 'Analysis', followed by 'Create Calculated Field'. Or Right click on the empty area where all the measures and dimensions are present and select 'Create Calculated Field'. Input the following formula into the designated text box:

AVG([Flights 0 To 15 Minutes Late Percent])+

AVG([Flights More Than 15 Minutes Early Percent])+

AVG([Flights Between 16 And 30 Minutes Late Percent])+

AVG([Flights Between 31 And 60 Minutes Late Percent])

1. Proceed to create another calculated field named 'Long Delay %' utilizing the formula:

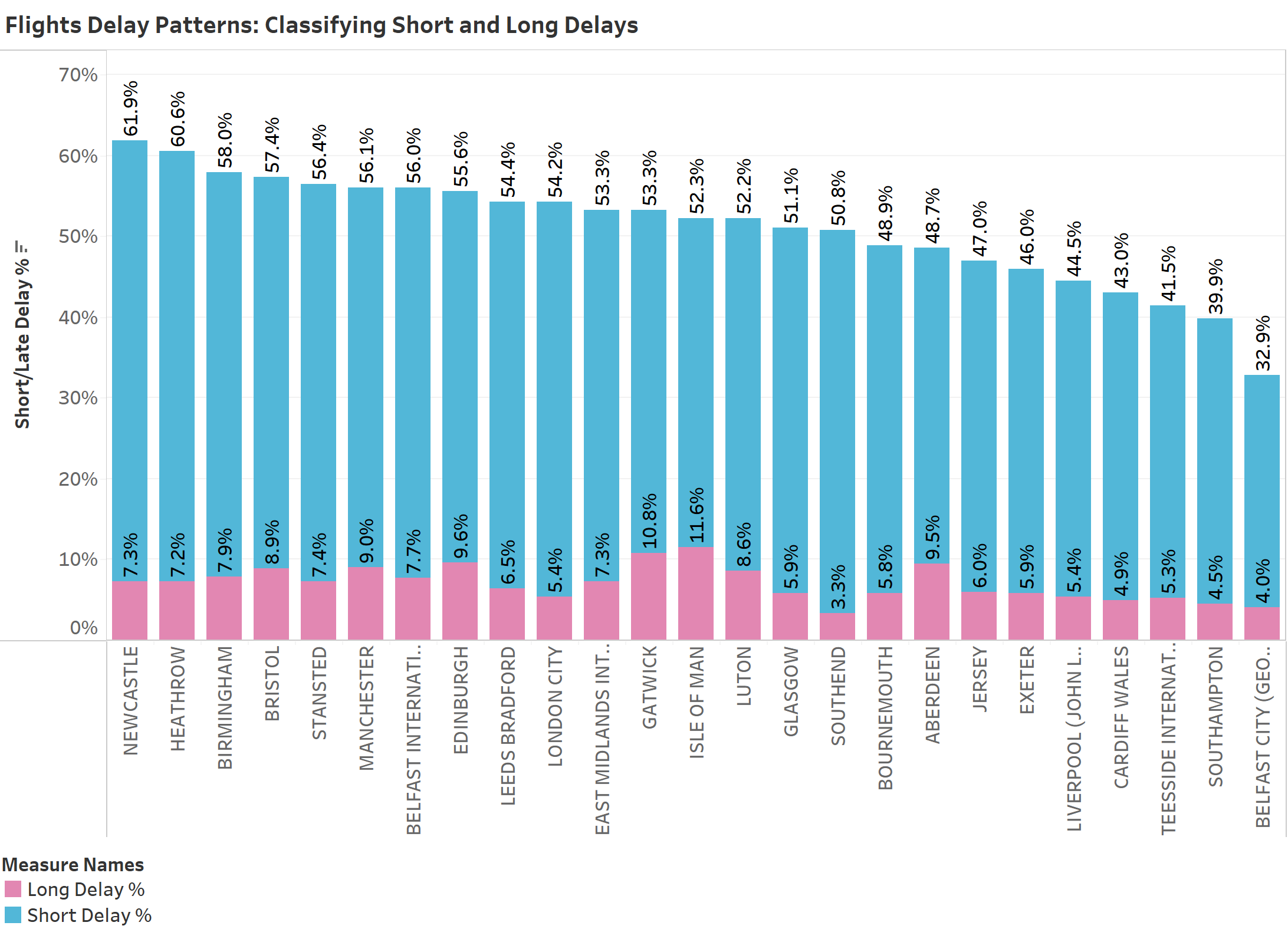
AVG([Flights Between 61 And 120 Minutes Late Percent])+

AVG([Flights Between 121 And 180 Minutes Late Percent])+

AVG([Flights Between 181 And 360 Minutes Late Percent])+

AVG([Flights More Than 360 Minutes Late Percent])

1. Place the dimension 'Reporting Airport' into the columns section.
2. Add 'Short Delay %' and 'Long Delay %' to the rows section. This will likely generate two separate graphs. To merge them into a single graph, right-click on the x-axis of the lower graph and choose Dual axis.
3. Right-click on the same axis again and choose the 'Synchronize Axis' option. This ensures that the bars look consistent across the graph.
4. Drop the measures that are present in the rows shelf into their respective Label section to display the exact labels.
5. Hide the axis situated on the right-hand side by right-clicking on it and deselecting 'Show Header'.



#### Key findings from visualisation:

* Variability in Delay Rates:

While some airports such as London Heathrow and London Gatwick exhibit relatively high short delay percentages (60.64% and 53.30% respectively), others like Southampton and Southend have notably lower short delay percentages (39.87% and 50.78% respectively). This indicates varying levels of efficiency in flight operations and infrastructure management.

* Performance Trends Across Regions:

By examining the data, trends in delay rates across different regions or airport sizes may emerge. Larger international airports like Heathrow and Gatwick may experience higher delay rates compared to smaller regional airports like Exeter and Isle of Man. Understanding these trends can assist stakeholders in identifying areas for improvement and allocating resources effectively.

* Opportunities for Improvement:

The visualization can highlight airports with consistently high delay rates, indicating areas where interventions may be needed to improve operational efficiency and passenger experience. Additionally, airports with lower delay rates may serve as benchmarks for best practices that can be adopted by others to mitigate delays and improve overall performance.

### Trend Analysis of Average Delay Minutes Over Time

#### Aim:

The aim of this visualization is to analyze the trend of average delay minutes over time across different reporting periods (months) spanning from 2022 to 2023. By visualizing this data, stakeholders such as airline operators, airport authorities, and travelers can gain insights into the seasonal variations and overall trajectory of average delay minutes. This information can inform decision-making related to flight scheduling, resource allocation, and passenger communication strategies.

#### Steps required to create the visualisation:

1. Place the dimension 'Reporting Month' into the columns section.
2. Add 'Average Delay Mins' and 'Previous Year Month Average Delay' to the rows section. This will likely create two separate graphs. To merge them into a single graph, right-click on the x-axis of the bottom graph and select Dual axis.
3. Right-click on the same axis again and choose the 'Synchronize Axis' option. This ensures that the bars look consistent across the graph.
4. Drag the measures that are present in the rows shelf into their respective Label section to display the exact labels.
5. Hide the axis situated on the right-hand side by right-clicking on it and deselecting 'Show Header'.

A graph with red and yellow lines

Description automatically generated

### Key findings from visualisation

* Seasonal Variation**:**

The data suggests a seasonal pattern in average delay minutes, with higher delays observed during summer months (June and July) and lower delays during winter months (November and February). This could be attributed to increased air travel demand during summer vacations and holidays, leading to congestion and potential delays.

* Year-over-Year Comparison**:**

Across corresponding months in 2022 and 2023, there are fluctuations in average delay minutes. Some months show increases in delay minutes from one year to the next (e.g., August), while others exhibit decreases (e.g., May). These variations indicate changes in delay performance over time and may warrant further analysis to understand underlying factors driving these shifts.

* Monthly Trends**:**

Certain months consistently exhibit higher average delay minutes compared to others. For instance, June and July consistently show higher delay minutes in both 2022 and 2023, suggesting recurring challenges during these months that may require targeted interventions to mitigate delays.

* Potential Areas for Improvement**:**

Months with consistently high delay minutes, such as June and July, highlight areas where operational improvements or infrastructure investments may be needed to enhance efficiency and reduce delays. Similarly, months with significant year-over-year increases in delay minutes signal areas requiring attention to prevent further deterioration in delay performance.

### Flight Cancellations: Top 5 Airports vs. Others

#### Aim:

The aim of this visualization is to identify the top 5 airports with the highest number of flight cancellations and compare them with the remaining airports. This visualization aims to provide insights into the distribution of flight cancellations across different airports.

#### Steps required to create the visualisation

1. Place the dimension 'Reporting Airport' in the rows section. Add 'Number Flights Cancelled' to the columns section.
2. Drag 'Sum of Number of Cancelled Flights' in the Label section to display the exact numbers.
3. Create a calculated field named ‘Rank By No. Flights Cancelled’ for ranking the 'Number Flight Cancelled' measure using this formula:

RANK(SUM([Number Flights Cancelled]), 'desc')

1. Create another calculated field by the name ‘Top 5 by No. of Flights Cancelled’ to filter the top 5 airports based on their rank, using this formula:

IF [Rank By No. of Flights Matched] <= 5 THEN "Top 5" ELSE "Others" END

1. Now drop the previously created field into the Colour to colour Top 5 Airports in the visualization.

A graph of flight cancellation

Description automatically generated

#### Key findings from visualisation

* Top 5 Airports with Highest Cancellations**:**

The visualization clearly identifies the top 5 airports with the highest number of flight cancellations: Heathrow, Gatwick, Manchester, Edinburgh, and London City.

These airports stand out with significantly higher numbers of cancellations compared to the other airports in the dataset.

* Distribution of Cancellations**:**

Majority of the airports fall into the "Others" category, indicating that they have lower numbers of flight cancellations compared to the top 5 airports.

This highlights the concentration of flight cancellations at a few major airports, possibly due to factors such as higher traffic volume, adverse weather conditions, or operational challenges.

### Flight Activity Insights

#### Aim:

The aim of this visualization is to analyse the distribution of flight activity across airports, with a focus on identifying the top 5 airports by the number of flights matched and comparing them with the remaining airports.

#### Steps required to create the visualisation:

1. Place the dimension 'Reporting Airport' in the rows section. Add 'Number Flights Matched' to the columns section.
2. Drag 'Sum of Add Number Flights Matched' in the Label section to display the exact numbers.
3. Create a calculated field named ‘Rank By No. of Flights Matched’ for ranking the 'Number Flight Cancelled' measure using this formula:

RANK(SUM([Number Flights Matched]), 'desc')

1. Create another calculated field by the name ‘Top 5 by No. of Flights’ to filter the top 5 airports based on their rank, using this formula:

IF [Rank By No. of Flights Matched] <= 5 THEN "Top 5" ELSE "Others" END

1. Now drop the previously created field into the Colour section to colour Top 5 Airports in the visualization.

A graph of flight prices

Description automatically generated with medium confidence

#### Key findings from visualisation:

* Top 5 Airports Dominance**:**

The visualization highlights the dominance of a select few airports, termed as the "Top 5," which handle significantly higher numbers of matched flights compared to the rest of the airports.

Heathrow, Gatwick, Manchester, Stansted, and Edinburgh emerge as the top-performing airports in terms of the number of flights matched.

* Distribution of Flight Activity**:**

Majority of the airports fall into the "Others" category, indicating they handle a lower volume of matched flights compared to the top 5 airports.

This disparity in flight activity distribution underscores the concentration of air traffic at a few major airports, possibly due to factors such as geographical location, infrastructure capacity, and airline preferences.

* Significance of Top 5 Airports**:**

The top 5 airports play a critical role in the aviation network, serving as major hubs for passenger and cargo transportation.

Their high volume of matched flights reflects their importance in facilitating connectivity and serving as key transit points for domestic and international travel.

* Opportunities for Analysis**:**

Further analysis can delve into the factors driving the success of the top 5 airports, such as their geographical location, airline partnerships, route networks, and operational efficiency.

Additionally, exploring the performance trends over time and comparing them with industry benchmarks can provide valuable insights for airport operators, airlines, and policymakers in optimizing infrastructure investments and enhancing overall air transport efficiency.

## Conclusion:

In conclusion, this assignment provided valuable insights into the process of building a data warehouse, designing a star schema, populating data, and creating visualizations using Tableau.

### What went well:

One of the highlights was the learning gained from previous assignments, which facilitated the creation of the star schema and loading of data. The familiarity with the process streamlined the task and made it relatively easy to execute. Additionally, the visualization task, although initially daunting, turned out to be an enjoyable learning experience.

### Challenges faced:

Despite the overall success of the assignment, certain challenges were encountered. Data preprocessing consumed a significant amount of time, suggesting the need for more efficient strategies in this area. Loading date data into the DATE\_DIM table posed difficulties, which required perseverance to overcome.

### Areas for improvement:

Given the opportunity to redo the assignment, there are several areas I would focus on. Firstly, I would aim to optimize the data preprocessing phase to minimize time spent on this task. Additionally, I would invest more time in familiarizing myself with Tableau and exploring its capabilities for both data processing and visualization. Enhancing my skills in data preprocessing and visualization techniques would contribute to more efficient and impactful outcomes in future assignments.

Overall, this assignment served as a valuable learning experience, reinforcing key concepts related to data warehousing, schema design, data manipulation, and visualization. Furthermore, creating visualizations using Tableau provided an immersive experience in translating raw data into meaningful insights, empowering effective decision-making. Through overcoming challenges and embracing new tools and techniques, I have gained valuable insights that will inform my approach to similar tasks in the future.