UK Airport Punctuality Analysis – Data Warehouse Desing and Visual Insights (2023-2024)

**Introduction**

This report outlines the process of designing a star schema data warehouse for UK airport punctuality data. The dataset covers flights across 2023 and 2024 and includes information on delays, cancellations, airlines, and destinations.

The project involved preparing the data, creating and populating SQL tables in Oracle, and producing visualisations in Tableau to identify trends and insights.

**Star Schema**

The star schema was designed to optimise the performance of analytical queries and reporting. It consists of one central fact table, *FlightPunctualityFact,* and five surrounding dimension tables:

*AirportDim* – Stores reporting airport names.

*RouteDim* – Contains the destination and country for each flight route.

*AirlineDim* – Lists airline names.

*FlightTypeDim* – Identifies whether a flight is scheduled or chartered.

*DateDim* – Breaks down flight data by year, month number, and month name.

Each dimension table uses a surrogate primary key (e.g., *airport\_id*), which links to foreign keys in the fact table. This simplifies queries by reducing duplications and supporting flexibility if descriptive fields change (e.g. airline renames).

A diagram of a program

AI-generated content may be incorrect.

**SQL Table Creation**

All tables were created using Oracle SQL with appropriate data types, primary keys (PK), and foreign key (FK) constraints to maintain referential integrity. The *FlightPunctualityFact* table stores all measurable data, while each dimension table holds descriptive attributes used for filtering and grouping.

**AirportDim**

CREATE TABLE AirportDim (

airport\_id INT PRIMARY KEY,

reporting\_airport VARCHAR2(100)

);

A screenshot of a computer

AI-generated content may be incorrect.

**AirlineDim**

CREATE TABLE AirlineDim (

airline\_id INT PRIMARY KEY,

airline\_name VARCHAR2(100)

);

A screenshot of a computer

AI-generated content may be incorrect.

**RouteDim**

CREATE TABLE RouteDim (

route\_id INT PRIMARY KEY,

origin\_destination VARCHAR2(100),

origin\_destination\_country VARCHAR2(100)

);

A screenshot of a computer

AI-generated content may be incorrect.

**FlightTypeDim**

CREATE TABLE FlightTypeDim (

flight\_type\_id INT PRIMARY KEY,

scheduled\_charter VARCHAR2(1)

);

A screenshot of a computer

AI-generated content may be incorrect.

**DateDim**

CREATE TABLE DateDim (

date\_id INT PRIMARY KEY,

year INT,

month INT,

month\_name VARCHAR2(20)

);

A screenshot of a computer

AI-generated content may be incorrect.

**FlightPunctualityFact**

CREATE TABLE FlightPunctualityFact (

fact\_id INT PRIMARY KEY,

airport\_id INT,

route\_id INT,

airline\_id INT,

flight\_type\_id INT,

date\_id INT,

number\_flights\_matched INT,

number\_flights\_cancelled INT,

flights\_cancelled\_percent NUMBER(6,2),

flights\_more\_than\_15\_minutes\_early\_percent NUMBER(6,2),

flights\_15\_minutes\_early\_to\_1\_minute\_early\_percent NUMBER(6,2),

flights\_0\_to\_15\_minutes\_late\_percent NUMBER(6,2),

flights\_between\_16\_and\_30\_minutes\_late\_percent NUMBER(6,2),

flights\_between\_31\_and\_60\_minutes\_late\_percent NUMBER(6,2),

flights\_between\_61\_and\_120\_minutes\_late\_percent NUMBER(6,2),

flights\_between\_121\_and\_180\_minutes\_late\_percent NUMBER(6,2),

flights\_between\_181\_and\_360\_minutes\_late\_percent NUMBER(6,2),

flights\_more\_than\_360\_minutes\_late\_percent NUMBER(6,2),

average\_delay\_mins NUMBER(6,2),

FOREIGN KEY (airport\_id) REFERENCES AirportDim(airport\_id),

FOREIGN KEY (route\_id) REFERENCES RouteDim(route\_id),

FOREIGN KEY (airline\_id) REFERENCES AirlineDim(airline\_id),

FOREIGN KEY (flight\_type\_id) REFERENCES FlightTypeDim(flight\_type\_id),

FOREIGN KEY (date\_id) REFERENCES DateDim(date\_id)

);

A screenshot of a computer

AI-generated content may be incorrect.

All data types were chosen based on expected value ranges and to ensure compatibility with Tableau during export. Numerical fields were assigned precision using NUMBER(6,2) to avoid overflow errors during insertion.

**Data Preparation and Loading**

**Initial Data Cleaning**

* The original dataset consisted of 24 monthly CSV files.
* These files were manually merged in Excel into a single master dataset.
* The following columns were removed due to being incomplete or inconsistent across years:
  + *run\_date*
  + *actual\_flights\_unmatched*
  + *flights\_unmatched\_percent*
  + *previous\_year\_month\_flights\_matched*

**Preprocessing in Python**

The dataset was further prepared in Python using a Jupyter Notebook

Key steps performed:

* Loaded the merged Excel dataset into a pandas DataFrame
* Split the *reporting\_period* column into *year, month,* and *month\_name*:
  + This was essential to simplify the creation of the *DateDim* table in SQL
  + The year and month were extracted using slicing, and month numbers were mapped to month names
  + The original *reporting\_period* column was then dropped
* Cleaned and standardised all text fields:
  + Stripped whitespace, converted all values to uppercase for consistency
  + This reduced the risk of duplication and join mismatches caused by variations in case or trailing spaces
* Validated Numeric Values
  + Checked for negative values in all numeric columns; none were found
  + Checked all percentage columns for invalid values above 100%, none were found
  + Confirmed the absence of fully duplicated rows
* Final Verification:
  + Printed the shape of the cleaned dataset to ensure no data was lost
  + Saved the cleaned version as *Cleaned\_Master\_Dataset\_Backup.csv* for use in Oracle and Tableau

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**Uploading and Staging in Oracle**

* The cleaned CSV was uploaded to Oracle APEX via SQL Workshop > Data Workshop.
* A new staging table named *Staging\_Master* was automatically created to hold the raw data and mirror the structure of the cleaned dataset
* This table was used to populate all dimension and fact tables.

A screenshot of a computer

AI-generated content may be incorrect.

**Populating Dimension Tables**

Each dimension table was filled using SELECT DISTINCT queries from the staging table.

ROWNUM was used to assign surrogate keys for each unique entry.

Implemented Code:

INSERT INTO AirportDim (airport\_id, reporting\_airport)

SELECT ROWNUM, reporting\_airport

FROM (SELECT DISTINCT reporting\_airport FROM Staging\_Master);

This method was repeated for:

* *AirlineDim*
* *RouteDim*
* *FlightTypeDim*
* *DateDim*

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Populating the Fact Table**

* The *FlightPunctualityFact* table was populated using JOIN queries to link the staging table to each dimension table.
* This replaced descriptive values with their corresponding surrogate keys.

Implemented Code:

INSERT INTO FlightPunctualityFact (

fact\_id,

airport\_id,

route\_id,

airline\_id,

flight\_type\_id,

date\_id,

number\_flights\_matched,

number\_flights\_cancelled,

flights\_cancelled\_percent,

flights\_more\_than\_15\_minutes\_early\_percent,

flights\_15\_minutes\_early\_to\_1\_minute\_early\_percent,

flights\_0\_to\_15\_minutes\_late\_percent,

flights\_between\_16\_and\_30\_minutes\_late\_percent,

flights\_between\_31\_and\_60\_minutes\_late\_percent,

flights\_between\_61\_and\_120\_minutes\_late\_percent,

flights\_between\_121\_and\_180\_minutes\_late\_percent,

flights\_between\_181\_and\_360\_minutes\_late\_percent,

flights\_more\_than\_360\_minutes\_late\_percent,

average\_delay\_mins

)

SELECT

ROWNUM,

a.airport\_id,

r.route\_id,

al.airline\_id,

ft.flight\_type\_id,

d.date\_id,

sm.number\_flights\_matched,

sm.number\_flights\_cancelled,

sm.flights\_cancelled\_percent,

sm.flights\_more\_than\_15\_minutes\_early\_percent,

sm.flights\_15\_minutes\_early\_to\_1\_minute\_early\_percent,

sm.flights\_0\_to\_15\_minutes\_late\_percent,

sm.flights\_between\_16\_and\_30\_minutes\_late\_percent,

sm.flights\_between\_31\_and\_60\_minutes\_late\_percent,

sm.flights\_between\_61\_and\_120\_minutes\_late\_percent,

sm.flights\_between\_121\_and\_180\_minutes\_late\_percent,

sm.flights\_between\_181\_and\_360\_minutes\_late\_percent,

sm.flights\_more\_than\_360\_minutes\_late\_percent,

sm.average\_delay\_mins

FROM Staging\_Master sm

JOIN AirportDim a ON sm.reporting\_airport = a.reporting\_airport

JOIN RouteDim r ON sm.origin\_destination = r.origin\_destination

AND sm.origin\_destination\_country = r.origin\_destination\_country

JOIN AirlineDim al ON sm.airline\_name = al.airline\_name

JOIN FlightTypeDim ft ON sm.scheduled\_charter = ft.scheduled\_charter

JOIN DateDim d ON sm.year = d.year AND sm.month = d.month;

A screenshot of a computer

AI-generated content may be incorrect.

**Tableau Visualisations**

This section includes visual insights created using Tableau, based on the cleaned and structured dataset. Each visualisation is explained with its aim, creation steps, effectiveness, and key findings.

**Visualisation 1:** Airline Flight Volume – Top 10

**Aim**: To compare the top 10 airlines based on their total number of matched flights to highlight the scale of operations for each carrier.

A screenshot of a graph

AI-generated content may be incorrect.

**Steps Taken in Tableau:**

* Dragged *airline\_name* to Label and *number\_flights\_matched* to Size and Colour
* Selected Packed Bubbles from the Show Me panel
* Applied a filter to include the top 10 airlines
* Adjusted colour gradient and label display to emphasise scale differences among airlines

**Effectiveness and Presentation:**

This visualisation provides a clear representation of airline flight volumes. Larger bubbles correspond to airlines with higher matched flight counts, while colour reinforces the visual hierarchy. The use of tooltips compensates for labels not appearing on smaller bubbles.

**Key Findings:**

* EasyJet had the highest matched flight count, following by Ryanair and British Airways, each with over half a million flights
* The visual highlights a strong concentration of flight activity among a small number of dominant carriers

**Visualisation 2:** Distribution of Delays by Time Category (% of Flights)

**Aim**: To understand how flight delays are distributed across standard time buckets, helping assess the severity of delays in the dataset.

A graph with numbers and dots

AI-generated content may be incorrect.

**Steps taken in Tableau:**

* Selected all relevant delay percentage fields (0-15 min, 16-30 min, etc.)
* Used *Measure Name* and *Measure Values* to create a horizontal bar chart
* Aggregated all fields using AVG to reflect average delay percentages
* Applied a custom colour gradient to reflect delay severity

**Effectiveness and Presentation:**

The chart clearly shows that the vast majority of delays fall within the 0-30 minute range, with visual emphasis on severity using colour. Cleaning up the category names and axis labels made the chart easier to interpret.

**Key Findings:**

* Approximately 39% of delayed flights were within the first 30 minutes
* Severe delays (over 2 hours) were rare
* The dataset reflects an efficient network with minimal long delays

**Visualisation 3:** Average Delay Over Time – Top 5 Airlines

**Aim:** To track delay performance changes over time for the five most active airlines, highlighting seasonal peaks, operational trends and performance differences.

A graph of different colored lines

AI-generated content may be incorrect.

**Steps Taken in Tableau:**

* Created a calculated field *MonthYear\_Date* using:
  + MAKEDATE([year], [month], 1)
* Dragged *MonthYear\_Date* to Columns and *average\_delay\_mins* to Rows
* Added *airline\_name* to both Colour and Filter
* Applied a top 5 filter on *airline\_name* using *SUM(number\_flights\_matched)* to focus on the most active airlines

**Effectiveness and Presentation:**

The line chart reveals trends in airline delay performance over a two-year period. With colour-coded lines and a clear time progression, differences in peak periods and recovery phases are easy to observe. Filtering to the top 5 airlines keeps the chart focused and prevents visual clutter.

**Key Findings:**

* EasyJet and Ryanair consistently had the highest total delay minutes, with clear seasonal peaks in July 2023 and June-August 2024.
* Jet2 followed similar peaks, with lower overall volumes
* British Airways maintained a moderate but stable delay profile throughout, while Loganair had the lowest delay volumes among the top 5.
* Peaks appear to align with summer travel months, indicating predictable seasonal surges in delay incidents

**Visualisation 4:** Top 10 Routes by Average Cancellation Rate

**Aim:** To identify which flight routes have the highest average cancellation percentages, highlighting connections that regularly experience disruption.

A graph of a number of flights

AI-generated content may be incorrect.

**Steps Taken in Tableau:**

* Created a calculated field name *Route* combining *origin\_destination* and *origin\_destination\_country*
* Dragged *Route* to Rows and *flights\_cancelled\_percent* to Columns
* Changed aggregation to Average to reflect average cancellations per data record.
* Applied a Top 10 filter to the *Route* field based on *AVG(flights\_cancelled\_percent)*
* Sorted the chart in descending order
* Added *AVG(flights\_cancelled\_percent)* to both Label and Colour for clarity

**Effectiveness and Presentation:**

This chart highlights which routes are, on average, most affected by cancellations over time. Using average values instead of totals makes the insight more comparable across routes with different flight volumes. The Top 10 view keeps the chart concise and easy to interpret, while colour and labels enhance visibility.

**Key Findings:**

* Barra had the highest average cancellations, with 15.08 flights cancelled per period, followed by Tiree and Wick John O Groats, averaging 13.46 and 12.00 cancellations.
* The top 5 routes are remote or island destinations within the UK, indicating frequent disruption on smaller or weather-sensitive routes.
* Heathrow appears in the bottom of the top 10 with an average of 7.40 cancellations, suggesting that even high-volume routes face persistent cancellation rates