# Flight Punctuality Data Warehouse Report

# Star Schema Data Warehouse

## Star Schema Design

Below we have the star schema design which was devised for the “Flights Punctuality” excel spreadsheets.

A screenshot of a computer

Description automatically generated with medium confidence

### Dimension Tables

The dimension tables will contain the descriptive elements of the data. The first dimension table is the “Airline” table, which will contain the information on the different airlines. For this the column found in the excel spreadsheet with the name “airline\_name” will be used to show the name of the airline, as well as working as a primary key. General information about the airline will also be stored here, including the email, telephone number and the location of the headquarters for said airline.

The “Airport” dimension table will contain the information on the different airports, where similarly to the “Airline” table, it will use the “reporting\_airport” column in the table for the name of the airport, as well as being the primary key. This table will also include the general information of email, telephone number and location for the airport.

The “Reporting Period” table will be used to identify the month and year in which the report took place, so that the data is time stamped within the data warehouse. This will include the “year” and “month” as well as the “reporting\_period” which will be in the same format as the “reporting\_period” in the excel spreadsheet.

The “Destination Country” table will consist of the name of the country which the airplane will land, which is found in the “origin\_destination\_country” which is also the primary key. Extra information, such as the capital of the country, the time zone it operates in, and the predominant language spoken in said country will be listed.

The “Destination Location” will contain the “origin\_destination” which is the location in which the plane will land. The “origin\_destination” will also work as a primary key. Extra information, such as the region, population size, as well as the longitude and latitude will be saved in this table.

Finally, the “Flight Type” table will contain the “scheduled\_charter” as a primary key, which will tell the user whether the flight is scheduled or chartered, which is represented by a “C” or an “S”. The description as well as the name of the flight type will also be stored to explain to the user what the “C” and “S” represents.

### Fact Table

The fact table titled “Airport Punctuality” will contain all the data that needs to be measured for analysis and reports. It will contain all the dimension tables primary keys as foreign keys to aid in uniquely identifying the aggregated entries. The table will contain all the numerical data that will be analysed, such as the percentage of late and early flights, cancelled flights and so on. All measurable data which was in the excel spreadsheet will be present within this table for analysis purposes. This was done since each data column is of value for analysis.

### Design Choices

* Opted to use the same names for the attributes which were used within the excel spreadsheets for consistency, despite personally disliking some of the naming conventions, such as “origin\_destination\_country” and “origin\_destination”. This was done for consistency.
* The “airline\_name” and “reporting\_airport”, as well as other dimension table attributes are more than capable to be primary keys for their table, therefore the decision to keep them as the primary keys instead of creating a new unique ID was made, which also makes loading the data directly into Apex Oracle easier, since a unique identifier would not have to be generated or made for each column.
* Many of the dimension tables could have connected to one another, such as the “Airline” table connecting to the “Airport” table, or “Destination Country” connecting to “Destination Location”, however, the decision to connect them all directly to the fact table was made so that there would not be a complex relationship structure leading to longer queries, and the star schema design would still function the same either way, therefore, making all the dimension tables connect directly to the fact table should be the better option.
* Included extra information about the tables, such as the email, addresses and telephone numbers of the airport and airline, as well as the information about the origin destination and origin destination country so that the data warehouse has dimension tables which provide context and background information that are not just the primary key, whereas the fact table will record the data that needs to be stored for analysis. This will give longevity to the data warehouse since future users will know what certain pieces of data in the fact table means, such as “S” and “C” for scheduled and chartered. This will aid in understandability and readability for future users.

# CREATE Table Statements for Star Schema

## Airline Table

CREATE TABLE "AIRLINE"

( "AIRLINE\_NAME" VARCHAR2(100 CHAR),

"AIRLINE\_EMAIL" VARCHAR2(50),

"AIRLINE\_TELEPHONE\_NUMBER" VARCHAR2(50),

"AIRLINE\_HEADQUARTERS\_LOCATION" VARCHAR2(50),

CONSTRAINT "AIRLINE\_PK" PRIMARY KEY ("AIRLINE\_NAME")

USING INDEX ENABLE

) ;

## Airport Table

CREATE TABLE "AIRPORT"

( "REPORTING\_AIRPORT" VARCHAR2(100 CHAR) NOT NULL ENABLE,

"AIRPORT\_EMAIL" VARCHAR2(50),

"AIRPORT\_TELEPHONE\_NUMBER" VARCHAR2(50),

"AIRPORT\_LOCATION" VARCHAR2(50),

CONSTRAINT "AIRPORT\_PK" PRIMARY KEY ("REPORTING\_AIRPORT")

USING INDEX ENABLE

) ;

## Reporting Period Table

CREATE TABLE "REPORTING\_PERIOD"

( "REPORTING\_PERIOD" NUMBER GENERATED BY DEFAULT ON NULL AS IDENTITY MINVALUE 1 MAXVALUE 9999999999999999999999999999 INCREMENT BY 1 START WITH 1 CACHE 20 NOORDER NOCYCLE NOKEEP NOSCALE NOT NULL ENABLE,

"YEAR" VARCHAR2(4 CHAR) NOT NULL ENABLE,

"MONTH" VARCHAR2(20 CHAR) NOT NULL ENABLE,

CONSTRAINT "REPORTING\_PERIOD\_PK" PRIMARY KEY ("REPORTING\_PERIOD")

USING INDEX ENABLE

) ;

## Destination Country Table

CREATE TABLE "DESTINATION\_COUNTRY"

( "ORIGIN\_DESTINATION\_COUNTRY" VARCHAR2(50 CHAR),

"CAPITAL" VARCHAR2(50),

"TIME\_ZONE" VARCHAR2(50),

"LANGUAGE" VARCHAR2(50),

CONSTRAINT "DESTINATION\_COUNTRY\_PK" PRIMARY KEY ("ORIGIN\_DESTINATION\_COUNTRY")

USING INDEX ENABLE

) ;

## Destination Location Table

CREATE TABLE "DESTINATION\_LOCATION"

( "ORIGIN\_DESTINATION" VARCHAR2(255 CHAR),

"REGION" VARCHAR2(50),

"POPULATION\_SIZE" NUMBER(10,0),

"LONGITUDE" VARCHAR2(50),

"LATITUDE" VARCHAR2(50),

CONSTRAINT "DESTINATION\_LOCATION\_PK" PRIMARY KEY ("ORIGIN\_DESTINATION")

USING INDEX ENABLE

) ;

## Flight Type Table

CREATE TABLE "FLIGHT\_TYPE"

( "SCHEDULED\_CHARTER" VARCHAR2(1 CHAR) NOT NULL ENABLE,

"FLIGHT\_TYPE" VARCHAR2(50 CHAR) NOT NULL ENABLE,

"FLIGHT\_TYPE\_DESCRIPTION" VARCHAR2(255 CHAR) NOT NULL ENABLE,

CONSTRAINT "FLIGHT\_TYPE\_PK" PRIMARY KEY ("SCHEDULED\_CHARTER")

USING INDEX ENABLE

) ;

## Airport Punctuality Table

CREATE TABLE "AIRPORT\_PUNCTUALITY"

( "AIRPORT\_PUNCTUALITY\_ID" NUMBER GENERATED BY DEFAULT ON NULL AS IDENTITY MINVALUE 1 MAXVALUE 9999999999999999999999999999 INCREMENT BY 1 START WITH 1 CACHE 20 NOORDER NOCYCLE NOKEEP NOSCALE NOT NULL ENABLE,

"REPORTING\_PERIOD" NUMBER,

"REPORTING\_AIRPORT" VARCHAR2(50),

"ORIGIN\_DESTINATION\_COUNTRY" VARCHAR2(50),

"ORIGIN\_DESTINATION" VARCHAR2(255),

"AIRLINE\_NAME" VARCHAR2(100),

"SCHEDULED\_CHARTER" VARCHAR2(1),

"NUMBER\_FLIGHTS\_MATCHED" NUMBER,

"ACTUAL\_FLIGHTS\_UNMATCHED" NUMBER,

"NUMBER\_FLIGHTS\_CANCELLED" NUMBER,

"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,

"FLIGHTS\_UNMATCHED\_PERCENT" NUMBER,

"FLIGHTS\_CANCELLED\_PERCENT" NUMBER,

"AVERAGE\_DELAY\_MINS" NUMBER,

"PREVIOUS\_YEAR\_MONTH\_FLIGHTS\_MATCHED" NUMBER,

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

"PREVIOUS\_YEAR\_MONTH\_AVERAGE\_DELAY" NUMBER,

PRIMARY KEY ("AIRPORT\_PUNCTUALITY\_ID")

USING INDEX ENABLE

) ;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRLINE\_PUNCTUALITY\_AIRLINE\_NAME" FOREIGN KEY ("AIRLINE\_NAME")

REFERENCES "AIRLINE" ("AIRLINE\_NAME") ENABLE;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRLINE\_PUNCTUALITY\_REPORTING\_AIRPORT" FOREIGN KEY ("REPORTING\_AIRPORT")

REFERENCES "AIRPORT" ("REPORTING\_AIRPORT") ENABLE;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRPORT\_PUNCTUALITY\_DESTINATION\_COUNTRY\_FK" FOREIGN KEY ("ORIGIN\_DESTINATION\_COUNTRY")

REFERENCES "DESTINATION\_COUNTRY" ("ORIGIN\_DESTINATION\_COUNTRY") ENABLE;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRPORT\_PUNCTUALITY\_DESTINATION\_LOCATION\_FK" FOREIGN KEY ("ORIGIN\_DESTINATION")

REFERENCES "DESTINATION\_LOCATION" ("ORIGIN\_DESTINATION") ENABLE;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRPORT\_PUNCTUALITY\_FLIGHT\_TYPE\_FK" FOREIGN KEY ("SCHEDULED\_CHARTER")

REFERENCES "FLIGHT\_TYPE" ("SCHEDULED\_CHARTER") ENABLE;

ALTER TABLE "AIRPORT\_PUNCTUALITY" ADD CONSTRAINT "AIRPORT\_PUNCTUALITY\_REPORTING\_PERIOD\_FK" FOREIGN KEY ("REPORTING\_PERIOD")

REFERENCES "REPORTING\_PERIOD" ("REPORTING\_PERIOD") ENABLE;

# Discussion of Database Creation & Population

## Excel Spreadsheets Data Preparation

Not much of the data had to be prepped to create and populate the data warehouse since the excel spreadsheets already contained large amounts of aggregated data which was time stamped by the “reporting\_period”, showing that the data was well prepared prior. However, there was a column which seemed redundant in the excel spreadsheets, which is the “run\_date” which is only used to display the date and time of when the report was run, therefore, it has no real value within our data warehouse, so the column was deleted in each spreadsheet.

For further preparation, all the excel spreadsheets were put together into a singular excel workbook, with each month being a different sheet within the workbook, allowing for ease of access to the data, and allowing for a fast import of data into Oracle Apex and Tableau for the visualisations that are needed later.

## Creating & Populating the Airport Punctuality Table

The first table which was created, was the fact table of “Airport\_Punctuality”. This was done since it is the table which contains all the data and links all the dimension tables, therefore, it would be easy to load in the data directly from the excel spreadsheet into Apex Oracle, since this table will contain everything. Below, we can see that the table was created alongside loading in the data. The sheet which corresponded to the month, January, was loaded in, where we can see all the corresponding columns were created and loaded in.A screenshot of a computer

Description automatically generated

The message confirming the table was created and all the data from the month of January was loaded into the table. This process had to then be repeated for all months, since each month had a separate worksheet.A green tick on a black background

Description automatically generated with low confidence

## Airline Table

The airline table was then manually created, where the same load process was used to load in the data of all the airline names. However, since this table was already created, we were able to just upload the data from the “airline\_name” column. Since the “airline\_name” was set as the primary key, this allowed for all duplicates to not be added when loading in the airline names into the newly created airline table, where Apex Oracle would automatically reject them. Just like the fact table, the loading process was done for each worksheet for every single month.

A screenshot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generated with low confidence

## Airport Table

The same approach was also done for the “Airport” table, where all worksheets were loaded into the table, where only the “reporting\_airport” column was accepted. There are no duplicates since the “reporting\_airport” is the primary key.

A screenshot of a computer

Description automatically generated

## Destination Country & Destination Location Tables

Both the “Destination\_Country” and “Destination\_Location” tables had the same approach as the prior tables, where due to the sheer size of data in the excel spreadsheets required for the tables, they were loaded in, where only data from the “origin\_destination\_county” columns of the spreadsheets were loaded in. A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence

## Reporting Period Table

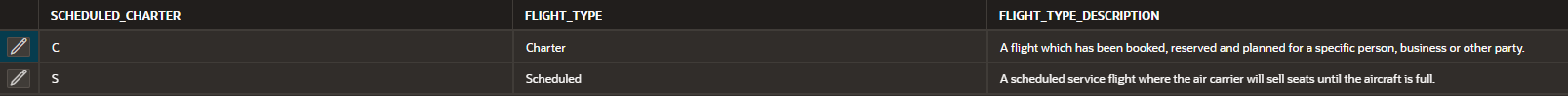
The reporting period table did not have the data automatically loaded into it, since it is a relatively small table that is representing the reporting period through a code that combines the year and month, therefore, manually typing out the code to be linked to the fact table later was possible. The “year” and “month” columns were also manually inputted.

A screenshot of a computer

Description automatically generated with medium confidence

## Flight Type Table

Like that of the “Reporting Period” table, this table was manually inputted since it only contains two columns. The “flight\_type” and “flight\_type\_description” were also manually inputted to be able to describe what the “C” and “S” would stand for in the fact table.



## Constraints for The Fact Table

After the creation of all the tables, it was imperative to ensure that the tables all linked together for the data warehouse to perform queries as intended. This was done by making sure all the dimension table primary keys matched their corresponding columns in the fact table, which would then be assigned as a foreign key. To do this, all primary keys had to have the same variables as the target column which will be the foreign key, as well as the same name, which was already all planned within the star schema, making the creation process easier.

A screenshot of a computer

Description automatically generated with medium confidence

From the screenshot above, we can see that there are six foreign keys within the “Airport\_Punctuality” fact table, showing that each dimension table has been successfully connected with the fact table, showing that the data warehouse should operate efficiently and effectively, where queries will run fast and optimally.

# OLAP Cubes Benefits to The Airline Industry

## What is An OLAP Cube?

An ‘Online Analytical Processing’ (OLAP) cube is a multidimensional database that can be used for fast analysis of data for queries to be run. Once a cube has been made by multiple dimensions, the rendering of the cube may take some time, however, once it has been rendered, the data, which a company may be looking for to aid in solving a business problem, can be extracted fast by ‘slicing’ the cube and getting the exact data which is required based on the dimensions needed, such as slicing the cube based off of time, geography and more. An OLAP cube is much better for business operations than a regular relational database, since it allows for instantaneous analysis of large pieces of data.

## Flight Operation Cube

The airline industry in general could benefit greatly from different OLAP cubes that could aid in different aspects, for example, we can start with a ‘Flight Operations’ cube which is in theme with the report so far, since the data warehouse which we have created revolves around flight operations and punctuality. The ‘Flight Operations’ cube may contain the different schedules of each aircraft, the schedule of the crew members, as well as maintenance schedules and other types of schedules revolving around the flights, such as how punctual the flight was at arriving at its destination. By having this cube, the airline industry can strategically analyse the data to find methods and different means to optimise the operations of their aircrafts, so that they can be on time and ready to improve customer satisfaction.

## Customer Cube

Another cube which could be created to aid the airline industry could be a ‘Customer’ cube which could contain different pieces of data about the customers, such as their demographic (gender, age etc), flight purchasing behaviour (purchase history and frequency of purchases) and potentially customer satisfaction, measured through surveys and feedback. Depending on how specific the cube can be, we can also include marketing, where information about how well a marketing campaign performed, such as a television advertisement or poster, stores the customer response rate to it. This cube would be pivotal for the airline industry since it allows for business operations to be created around the customer, where the airline can better understand the customers’ reaction to their services and advertisements and make changes accordingly based on their needs, such as investing more into certain advertisements that draw in more customers. It can also be used to improve business operations by catering towards certain demographics based on the demographic research, such as reviewing the feedback from elderly people and making changes to their flights and scheduling accordingly, such as making flights more comfortable. The customer cube is very effective for managing demographics since we can go in depth with the different demographics, since we can slice the cube into various amounts, for example, we can slice the cube in a way where we can run queries for elderly customers from a specific country, or we can slice it to just be elderly people in general, therefore we can determine how in depth we want to go in our analytical dive into customer satisfaction, so that the business operations can be as detailed as possible.

## Environmental Cube

The airline industry may also need a cube that would be used to measure their environmental impact, this cube can contain data on the fuel consumption of a plane, carbon emissions, amount of waste generated and more. From this cube, the CAA can use it to see which airlines are meeting environmental performance goals, and to make changes accordingly to those that aren’t, by putting in place different means to which these aircrafts can improve, which can aid in cost saving opportunities and aid in business operation decision making.

## Revenue Management Cube

In conjunction with the customer cube mentioned prior, a cube based around revenue management can be created to analyse the demand for different seats, the price of said seat and its availability. This cube is effective since these data points can be sliced based on the different airlines, aircrafts, location, time of day and more which may affect when a customer may pay for a specific seat, when a specific seat is available and when certain seats are in higher demand. This will allow for the airline industry to optimise their business operations to create more revenue by making these seats more readily available based on their findings, for example, for a specific flight at a specific time to a specific location, the first-class seats may be less available, therefore, they can potentially design their aircraft journeys to have more available to maximise their profits.

## Safety Regulations Cube

One final cube which could be created to aid the airline industry, which majority of industries should implement, is the “Safety Regulations” cube. For the airline industry, this cube may contain data on the safety incidents that occurred on flights, which will include details on what happened, the location, time of day, airline, which aircraft it occurred on and the severity of the incident. This cube can also include any risk management steps taken where it is recorded within the cube the type of risk assessment taken, the results of said assessment and any further risk mitigations put in place. This cube can also include more factors, such as data on the airline regulations. This cube would be very important for the CAA and airline industry since it allows for the data to be analysed in a way to see if the safety of the flights have improved over time, and the cube also makes it possible to slice the data in a way to see which incidents may occur more frequently, as well as how severe they are, so through the analysis of this data, the CAA can put more regulations and restrictions in place to improve customer and employee safety.

## Conclusion

As showcased by the potential cubes that could be created by the airline industry, shown above, by using OLAP cubes, the CAA can vastly improve their business operations due to OLAP cubes allowing for the fast analysis of data in comparison to a regular relational database. Therefore, informed decisions can be made, which could lead to cost-saving opportunities, improved satisfaction for both the customer and employees, as well as making the airline industry more profitable.

# Tableau Visualisations

## Visualisation 1 – Line Graphs of Cancellations: A picture containing text, line, handwriting, plot Description automatically generated

### Aim

The aim of the visualisation listed above is to check if the cancellations of flights across the United Kingdom is affected by the month, and to see if the cancellations are different depending on the location of the airport within the United Kingdom. For this example, a dashboard was created comparing three bar charts, the first showing the cancellations for all months for every single airport within the United Kingdom (red graph), the second showing the cancellations for Heathrow for all months (green graph) and finally the cancellations for Aberdeen for all months (blue graph). Heathrow will be used to measure the London flights in England (further south), where as Aberdeen will be used to measure the Scotland flights (further north).

### Creative Steps

#### Reporting Period into Date Variable

The first step in creating this visualisation was to change the “Reporting Period” into a Date so that we can create line graphs, since a line graph requires a date or time. The date was then formatted to only display the month and year for ease of reading for the stakeholder.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

#### Columns & Rows

To create the graphs, the column section would contain the dimension “Reporting Period” so that we can compare the months to another factor, and in this case, it would be the measure of the “Number Flights Cancelled”, where we would then use the SUM measure to total up all the cancellations for each month accordingly.

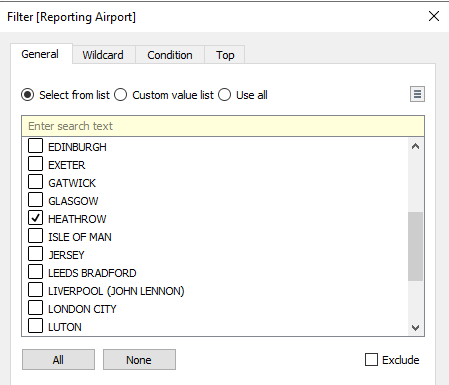
A picture containing text, screenshot, font, number

Description automatically generated

#### Filters

Depending on the sheet, we will use the dimension of “Reporting Airport” as a filter. For the red graph, we included all airports, for the green graph we only ticked “Heathrow” as the airport to measure, and finally for the blue graph we only ticked “Aberdeen” as the airport to measure.

A picture containing text, screenshot, font, design

Description automatically generated

#### Combining Graphs into a Dashboard

The three sheets were then combined to make a dashboard where the visualisation is easy to observe and understand.A picture containing text, screenshot, font, number

Description automatically generated

### Key Findings

From this visualisation, we can observe that most cancellations from all airports happened within the winter months, such as December and January. This is most likely due to weather conditions causing the cancellations, such as snow and ice, where many places within United Kingdom may not have measures in place to continue a flight schedule in snowy or icy weather. For this reason, the Heathrow and Aberdeen graphs were shown alongside one another with the all airports graph.

The Heathrow December calculations totalled up to “449” out of the total December cancellations of “1,463”, which may seem like many, however, Heathrow is also the most popular airport in the United Kingdom, therefore it would make up the majority of results.

Aberdeen on the other hand has consistent cancellations across multiple months, where December is still the highest at “121”, however, May and October are not too far behind at “115” and “89”, showing that it may not be a seasonal/weather issue, since January is also one of the lowest values at only “16” cancellations.

From this, we can see that places that are usually colder, such as Aberdeen in Scotland, may have better protocols in place to ensure that flights are not cancelled due to poor weather, since they may be more used to the weather and can have safeguards in place. Within London based airports, such as Heathrow, we can assume that they may not be as well equipped to handle the bad weather during the winter months, since we can see that the cancellations seem to mainly happen during the winter months, with January and December having big spikes in comparison to the rest of the months, whereas with Aberdeen, the cancellations seem to also spike in April.

These cancellations can also be due to the holidays as well as different weather conditions, since the spikes happening in December could be due to Christmas, where many flights may get cancelled, due to lack of staff, maintenance issues or overbooking. This can be further supported by the Aberdeen airport having a spike in April, which is during the Easter holidays.

From this visualisation, we can assume that more measures should be put in place to ensure that cancellations cannot occur as frequently during the winter months since the airline industry can be losing out on a lot of profit due to said cancellations. More staff can be hired for the winter months to ensure that over booked flights may still operate, or to even aid in safety measures for the flights in the bad weather.

## Visualisation 2 – Side-By-Side Bar of Airline Delays: A picture containing text, screenshot, font, plot Description automatically generated

### Aim

The aim of this visualisation is to see whether the most popular airlines within the United Kingdom were able to improve their performance by minimising the overall average time their flights were delayed, by comparing the year 2021 with the previous year, 2020.

### Creative Steps

#### Columns & Rows

The first step in creating this visualisation was to add the “Airline Name” as a column. Then, the “Measure Names” and “Measure Values” (which will be defined below) were used as the remaining column and row.

A picture containing text, screenshot, font, number

Description automatically generated

#### Filters & Measure Values

The measures which need to be used are the “Average Delay Mins” and “Previous Year Month Average Delay” which will be used to compare the delays in 2021 and 2020. To ensure that the values are correct and consistent, we must change these measures into an AVERAGE instead of a SUM, otherwise all the values will be totalled up instead of having an average delay in minutes. The colours red and blue were chosen to represent the two bars within the side-by-side bar, where red represents “Previous Year Month Average Delay” and blue represents “Average Delay Mins”.

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

The airlines were then filtered, where the only ones that will be shown in this visualisation are the most popular airlines within the United Kingdom (found via external research).

A screenshot of a computer

Description automatically generated with low confidence

### Key Findings

From this visualisation, we can see that many of the airlines ended up improving on their delays, where five out of the thirteen made clear improvements. However, this shows that most of the airlines somehow failed to improve on their average delay in minutes from 2020 to 2021.

The first reason which may have led to this is the airlines not putting more preventative measures in place to improve the likeliness of their planes being earlier. However, it is imperative to keep in mind that we are comparing the years 2021 and 2020, where in 2020, the COVID-19 pandemic had begun, which had led to less flights due to the whole world being put into lockdown, this means that the sample size of 2020 (the previous year) is much smaller than that of 2021, therefore, this must be considered when comparing the delays in minutes. However, for future references, it is important for the airlines to find ways to reduce their delays, so that the next year can be a good comparative year (if there is no global issues or pandemics).

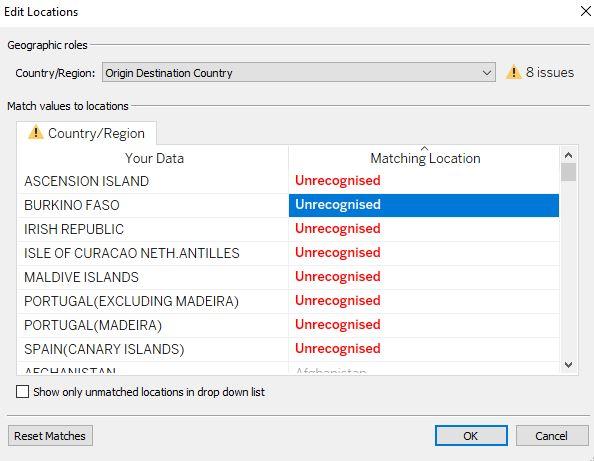
## Visualisation 3 – Heat Maps of Cancellations, Late and Early Flights: A screenshot of a computer screen Description automatically generated with low confidence

### Aim

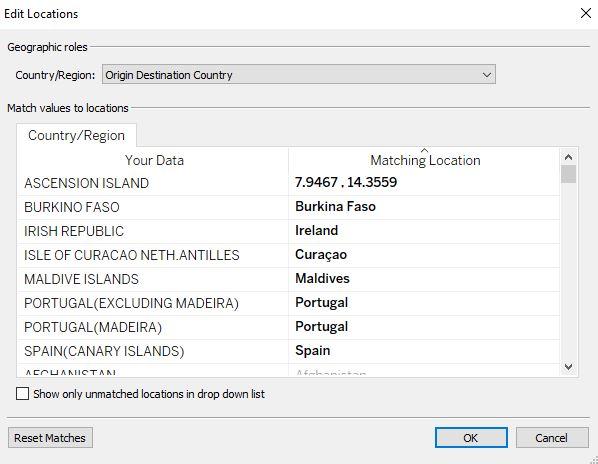
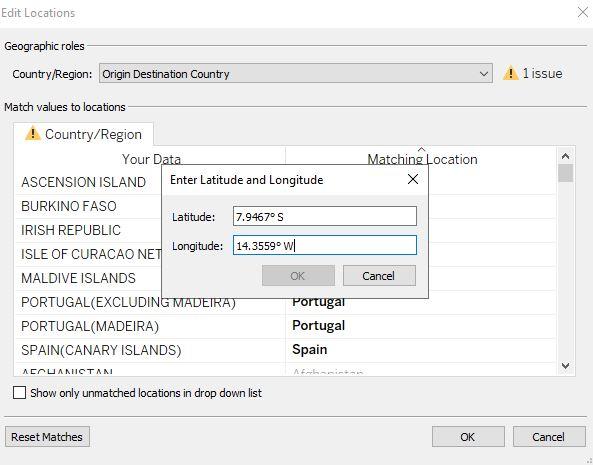
The aim of this visualisation is to see that based on the destination country which the airplane will land, if the plane will have a higher chance of being cancelled, late or early/on-time.

### Creative Steps

#### Fixing Geo-Locations

When selecting the map pre-set after dragging in the “Origin Destination Country”, which by default had a geographic role of “Country/Region”, allowing for us to make a map out of it, the error message of eight locations within the data set needing to be renamed to match a location popped up.

From this screen, the locations from the data set were then adjusted to have a matching location by selecting the corresponding countries from the drop-down menu. However, the “Ascension Island” did not have a location within the drop-down menu, therefore a google search of the longitude and latitude was required to set a precise location for the “Ascension Island”.



#### Colour Markings

For the map of the average of flights cancelled, the colour red was chosen to represent the heat points on the map, where the colour is based off the percent of the flights cancelled, where the value was set to an AVERAGE instead of the SUM. For the early flights map, the colour green was used where the AVERAGE of the “Flights Cancelled Percent” was used to measure how many flights were cancelled. Finally, the colour blue was used for the late flights where the flights were late between thirty-one and sixty-one minutes, where the AVERAGE was used to measure said late flights.

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

### Key Findings

From the flight cancellations map, we can see that Russia has many flights cancelled when compared to other countries. Uganda has a 100% cancellation, however, the sample size of flights to Uganda is extremely small in comparison to Russia. The reasons as to why the cancellations in Russia may be so high from the United Kingdom airports, could be due to poor weather conditions since Russia is a very cold and snowy country, where the airlines may not be well versed and equipped for the journey through the Russian tundra. It could also be due to the rising political tensions in 2021 which eventually led to the Russian invasion of Ukraine in early 2022. Afghanistan and Iraq have two of the highest percentages at 33.3% and 31.8% respectively. This could be due to on going conflicts and political instability which leads to the flights being cancelled, as well as limited airport infrastructure. China also had a high cancellation percentage of 13%, which could be due to the COVID-19 pandemic originating from Wuhan, China. Argentina seems to also have a large percentage (14.3%), which could be to the political structure or heavy COVID-19 restrictions.

From the map of very late flights (31 to 60 minutes), we can see that the countries with many late flights were within the Asia region, such as Pakistan with 11.5% and India with 7.7%. Due to the sheer amount of people living in India, there may be a lot of air traffic congestion which leads to flights being extremely delayed in comparison to other countries. Despite Russia having many cancellations, it did not seem to have many late flights, where only 2.6% of the flights to Russia were 31 to 60 minutes late. Also, despite the sample sizes for the USA, Canada and the United Kingdom being large, they all have a relatively low percentage of very late flights, showing that preventative measures have been put in place to deny flights being extremely late in these countries. The airlines can potentially put these measures in place for some of the Asian countries like India and Pakistan to potentially reduce the amount of extremely late flights, to improve customer satisfaction and make more profit in the future.

Finally, the map of the early flights shows Greenland to be very efficient, since 50% of the flights to Greenland are early, however, this could be due to the small sample size of the flights to Greenland. Ethiopia and the Ivory Coast also have high percentages at 41.8% and 50% respectively, however, once again their sample size is smaller than the other countries. Many of the European countries have high percentages, such as the United Kingdom with 29%, which has a large sample size and yet still has a large percentage of flights which are early, showing effective measures are in place. The United States is similar, with 25% of flights being early with a large sample size, showing efficient measures are also in place.

Throughout the three maps, we can see that most countries seem to have flights that are arriving earlier than they are arriving extremely late, showing that the airlines are operating efficiently, however, they could be much better since many countries still have a relatively high percentage of late flights, such as Pakistan with 11.5%.