

Big Data and Business Intelligence

Module Code: CIS4008-N-BF1-2023

Module Leader: Mansha Nawaz

Topic: Analysis of US Domestic Flight Delays and Cancellations in 2023.

Power BI Technical document

Name: Konda Reddy Thokala

Student number: C2994587

Submission date: 10 Jan 2024



MSc Data Science
School of Computing, Engineering & Digital Technologies
Teesside University
Middlesbrough
Tees Valley, TS1 3BX
United Kingdom

Table of Contents

Power BI Technical document	2
1. Overview of Dataset	2
1.1 Introduction.....	2
1.2 Data Set Source and Description.....	2
2. Data preparation steps:	4
2.1 Data import:	4
2.2 Data Cleaning:	6
2.3 Data preprocessing:	6
2.4 DAX formulas:	7
3. Data modelling:	15
4. Dashboard:	17
4.1 US Department of transportation logo.....	17
4.2 Dashboards	18
5. Reference:	21
6. Self-Assessment:	21

Power BI Technical document

1. Overview of Dataset

1.1 Introduction

In 2023, the domestic air travel landscape in the United States faced several challenges and fluctuations, particularly in terms of flight delays and cancellations. The aviation industry, which is an important component of the nation's transportation system, experienced disruptions that impacted both passengers and airline operations. This overview provides an overview of the key factors and trends that contributed to flight delays and cancellations during the specified period.

By visually representing patterns, trends, and insights from the dataset, Power BI enables analysis of US domestic flight delays and cancellations in 2023, facilitating informed decision-making and strategic improvements.

1.2 Data Set Source and Description

Dataset source: <https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023>

The Bureau of Transportation Statistics of the United States Department of Transportation (DOT) monitors the on-time performance of domestic flights operated by large air carriers. The monthly Air Travel Consumer Report contains summary information on the number of on-time, delayed, cancelled, and diverted flights.

Retrieved in November 2023 using the [On-Time: Reporting Carrier On-Time Performance \(1987-present\)](#) application.

The source data was downloaded in monthly subsets and then joined by year. The most recent data for 2023 is from August. csvkit, Python, and Excel were used for data consolidation, transformation, wrangling, variable selection, and label updates.

To visualize and find the patterns, 2023 data is taken from the complete dataset, Hence the data is of 2023 from January to August.

The dataset contains 33 columns and 4,545,422 rows in total.

Header	Data Type	Description
FL_DATE	object	Flight Date (yyyymmdd)
AIRLINE_CODE	object	Unique Carrier Code.
DOT_CODE	int64	An identification number assigned by US DOT to identify a unique airline (carrier).
FL_NUMBER	int64	Flight Number
ORIGIN	object	Origin Airport
ORIGIN_CITY	object	Origin Airport, City Name
DEST	object	Destination Airport
DEST_CITY	object	Destination Airport, City Name

CRS_DEP_TIME	int64	CRS Departure Time (local time: hhmm)
DEP_TIME	float64	Actual Departure Time (local time: hhmm)
DEP_DELAY	float64	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
TAXI_OUT	float64	Taxi Out Time, in Minutes
WHEELS_OFF	float64	Wheels Off Time (local time: hhmm)
WHEELS_ON	float64	Wheels On Time (local time: hhmm)
TAXI_IN	float64	Taxi In Time, in Minutes
CRS_ARR_TIME	int64	CRS Arrival Time (local time: hhmm)
ARR_TIME	float64	Actual Arrival Time (local time: hhmm)
ARR_DELAY	float64	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
CANCELLED	float64	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	object	Specifies The Reason For Cancellation
DIVERTED	float64	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	float64	CRS Elapsed Time of Flight, in Minutes
ELAPSED_TIME	float64	Elapsed Time of Flight, in Minutes
AIR_TIME	float64	Flight Time, in Minutes
DISTANCE	float64	Distance between airports (miles)
DELAY_DUE_CARRIER	float64	Carrier Delay, in Minutes
DELAY_DUE_WEATHER	float64	Weather Delay, in Minutes
DELAY_DUE_NAS	float64	National Air System Delay, in Minutes
DELAY_DUE_SECURITY	float64	Security Delay, in Minutes
DELAY_DUE_LATE_AIRCRAFT	float64	Late Aircraft Delay, in Minutes
FL_YEAR	int64	Flight Date (yyyy)
FL_MONTH	int64	Flight Date (mm)
FL_DAY	int64	Flight Date (dd)

There are another supporting datasets for the full name of airlines based on codes in the main dataset and for Cancellation codes.

AIRLINE_CODE_DICTIONARY table is having 2 columns and 1729 rows.

Header	Data Type	Description
Code	Object	Unique Carrier Code.
Description	Object	Complete name of the carrier

cancellation codes table is having 2 columns and 5 columns.

Header	Data Type	Description
Code	Object	Unique cancellation code
Description	Object	Meaning of the code as per DOT

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	FL_DATE	AIRLINE_CD	DOT_CODE	FL_NUM	ORIGIN	ORIGIN_CITY	DEST	DEST_CITY	CRS_DEP	DEP_TIME	DEP_DELAY	TAXI_OUT	WHEELS_O	WHEELS_O	TAXI_IN	CRS_ARR	TARR_TIME	ARR_DELAY	CANCELLED	CANCELLATION
2	04/06/2023	9E		20363	4628	DSM	Des Moines	LGA	New York	1033	1053	20	17	1110	1424	15	1415	1439	24	0
3	01/06/2023	9E		20363	4628	ROC	Rochester	LGA	New York	1001	956	-5	15	1011	1100	16	1119	1116	-3	0
4	02/06/2023	9E		20363	4628	ROC	Rochester	LGA	New York	1001	956	-5	9	1005	1051	11	1119	1102	-17	0
5	01/06/2023	9E		20363	4629	ITH	Ithaca/Cort	JFK	New York	1452	1447	-5	11	1458	1544	12	1559	1556	-3	0
6	02/06/2023	9E		20363	4629	ITH	Ithaca/Cort	JFK	New York	1452	1448	-4	13	1501	1547	7	1559	1554	-5	0
7	03/06/2023	9E		20363	4629	ITH	Ithaca/Cort	JFK	New York	1452	1447	-5	9	1456	1543	24	1559	1607	8	0
8	04/06/2023	9E		20363	4629	ITH	Ithaca/Cort	JFK	New York	1453	1443	-10	12	1455	1538	28	1600	1606	6	0
9	01/06/2023	9E		20363	4629	JFK	New York	1ITH	Ithaca/Cort	1255	1243	-12	28	1311	1347	3	1404	1350	-14	0
10	02/06/2023	9E		20363	4629	JFK	New York	1ITH	Ithaca/Cort	1255	1253	-2	30	1323	1403	3	1404	1406	2	0
11	03/06/2023	9E		20363	4629	JFK	New York	1ITH	Ithaca/Cort	1255	1250	-5	20	1310	1349	2	1404	1351	-13	0

Figure 1: Main Dataset

2. Data preparation steps:

2.1 Data import:

The first step to start the data preparation is by importing the data into Power BI application.

To import the data, we need to know the format of the dataset. As the main dataset format is CSV, we select Get data > Text/CSV as showing in figure 2.

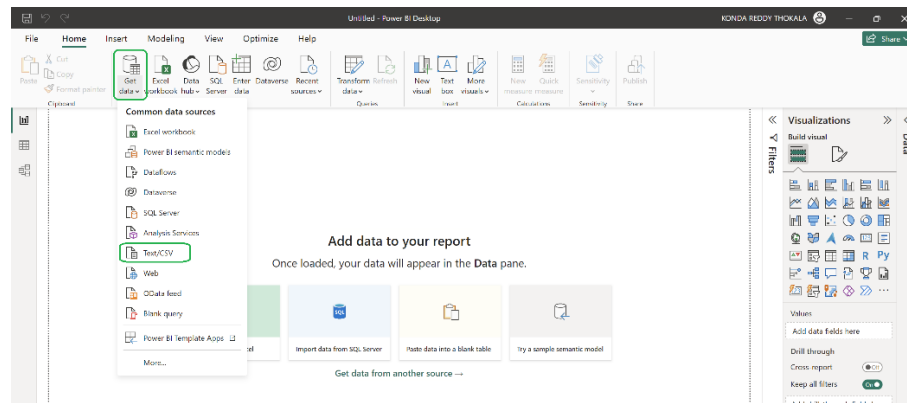


Figure 2: Data import Get Data.

Now select the data file from the system and click open.

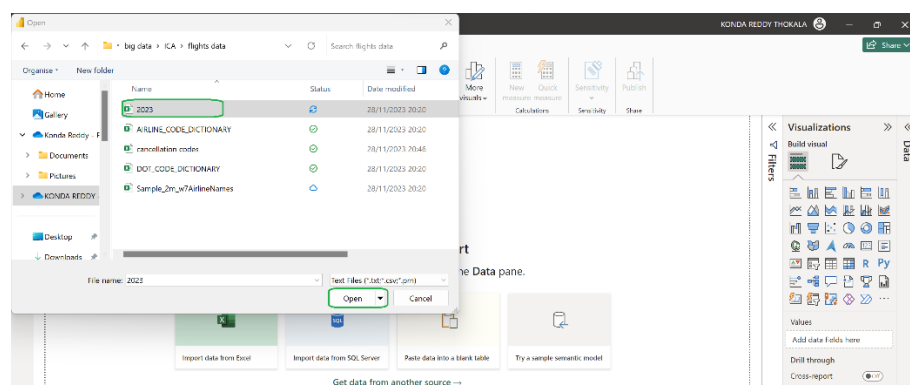


Figure 3: Data import: Dataset file upload

Once the data is available in the power BI, then we can transform the data if we need to perform data cleaning or else, we can load the data directly into Power BI as showing in the figure 4

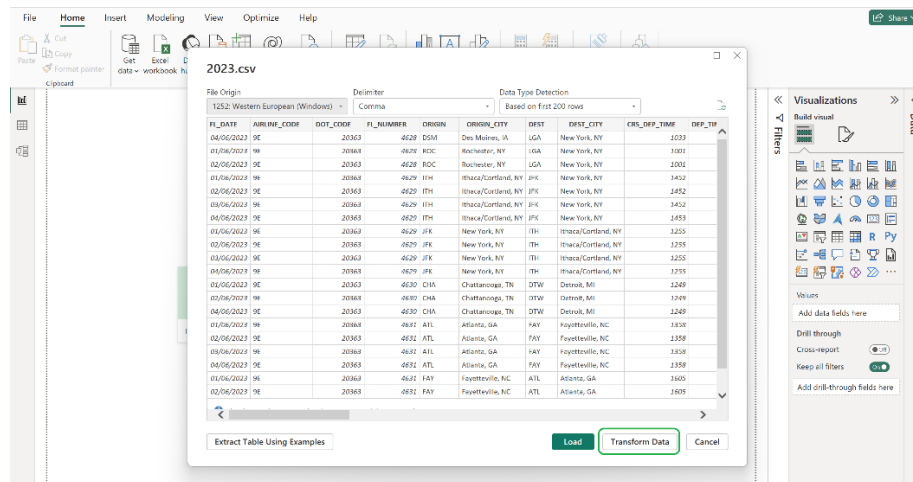


Figure 4: data import: Data transform

Now import the secondary datasets by following the same steps as the main dataset.

➤ Importing AIRLINE_CODE_DICTIONARY table

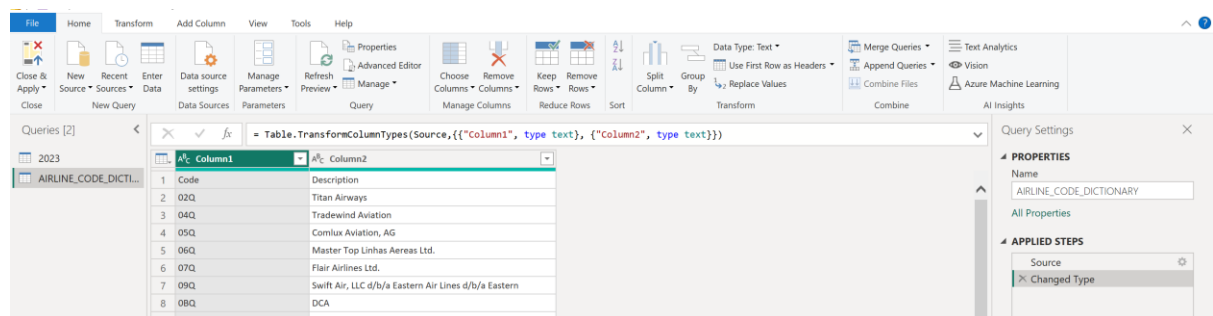


Figure 5: Data import, supporting file 1

➤ Importing Cancellation codes table and make the first line as header.

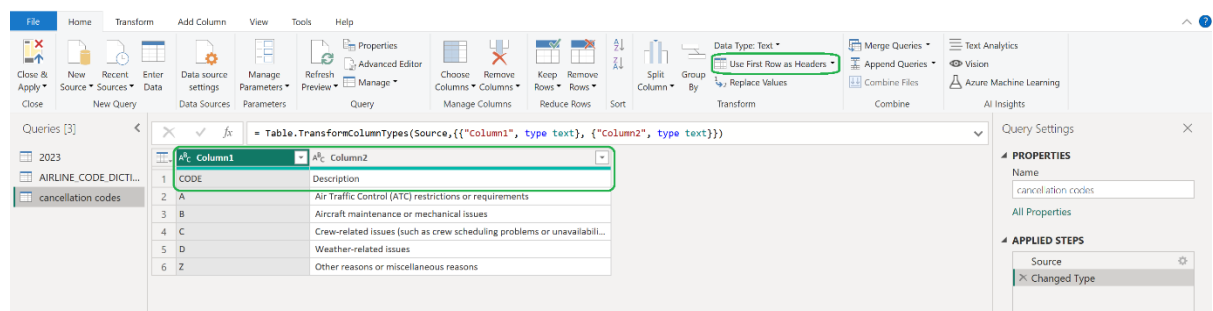


Figure 6: Data import, supporting file 2

Final data:

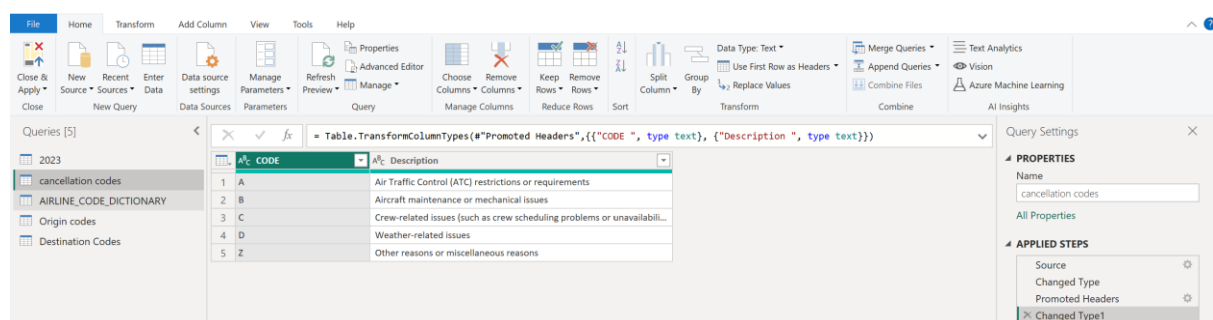


Figure 7: Data import, Final data tables.

After importing both the datasets, we can now do the necessary cleaning and processing in transform page as showing in figure 5 to visualize the data.

2.2 Data Cleaning:

As the data is from Kaggle, it is already cleansed. Hence performing basic cleaning operations to overcome if any mismatches.

- Removed duplicates from the datasets.
- Replaced null values in "DELAY_DUE_CARRIER", "DELAY_DUE_WEATHER", "DELAY_DUE_NAS", "DELAY_DUE_SECURITY", "DELAY_DUE_LATE_AIRCRAFT" with zeros.

Below is the M CODE for the same.



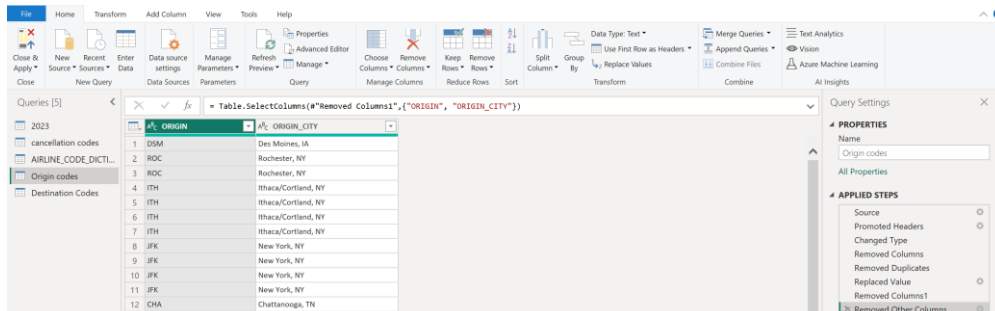
```

let
    Source = Csv.Document(File.Contents("C:\Users\konda\OneDrive - Teesside University\big data\ICA\flights data\2023.csv"),[Delimiter=","],
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{{"FL_DATE", type date}, {"AIRLINE_CODE", type text}, {"DOT_CODE", Int64}, {"DELAY_DUE_CARRIER", type text}, {"DELAY_DUE_WEATHER", type text}, {"DELAY_DUE_NAS", type text}, {"DELAY_DUE_SECURITY", type text}, {"DELAY_DUE_LATE_AIRCRAFT", type text}}}},
    #"Removed Duplicates" = Table.Distinct(#"Changed Type"),
    #"Replaced Value" = Table.ReplaceValue(#"Removed Duplicates",null,0,Replacer.ReplaceValue,{"DELAY_DUE_CARRIER", "DELAY_DUE_WEATHER", "DELAY_DUE_NAS", "DELAY_DUE_SECURITY", "DELAY_DUE_LATE_AIRCRAFT"}),
    in
        #"Replaced Value"
  
```

Figure 8: Main dataset M code.

2.3 Data preprocessing:

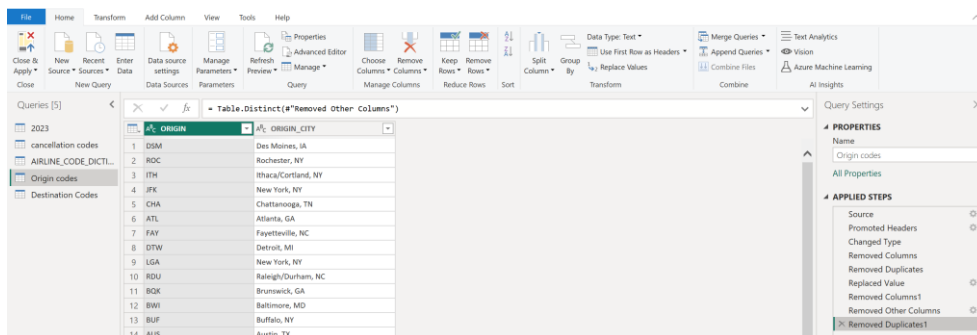
In data preprocessing, created a duplicate of 2023 dataset and removed all other columns except "ORIGIN", "ORIGIN_CITY" and created a new table named "Origin codes".



ORIGIN	ORIGIN_CITY
1 DSM	Des Moines, IA
2 ROC	Rochester, NY
3 ROC	Rochester, NY
4 ITH	Ithaca/Cortland, NY
5 ITH	Ithaca/Cortland, NY
6 ITH	Ithaca/Cortland, NY
7 ITH	Ithaca/Cortland, NY
8 JFK	New York, NY
9 JFK	New York, NY
10 JFK	New York, NY
11 JFK	New York, NY
12 CHA	Chattanooga, TN

Figure 9: Origin table creation

Now removed the duplicate from the table "Origin codes".



ORIGIN	ORIGIN_CITY
1 DSM	Des Moines, IA
2 ROC	Rochester, NY
3 ITH	Ithaca/Cortland, NY
4 JFK	New York, NY
5 CHA	Chattanooga, TN
6 ATL	Atlanta, GA
7 FAY	Fayetteville, NC
8 DTW	Detroit, MI
9 LGA	New York, NY
10 IND	Indianapolis, IN
11 BOK	Birmingham, AL
12 BWI	Baltimore, MD
13 BUF	Buffalo, NY
14 AUS	Austin, TX

Figure 10: Origin table duplicates deletion.

The M code of Origin codes is:



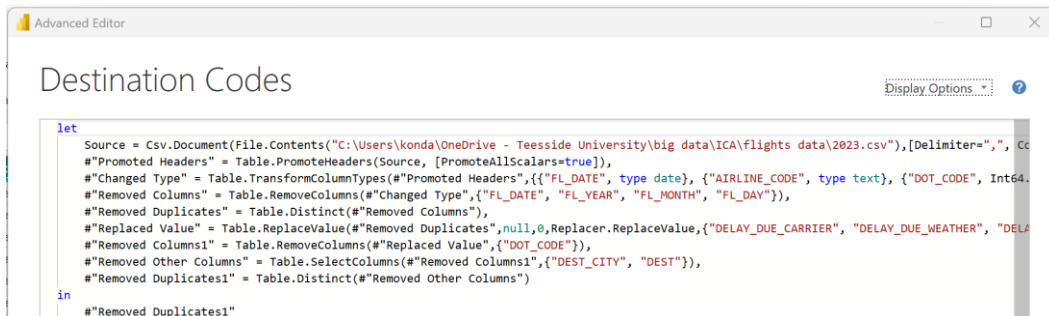
```

let
    Source = Csv.Document(File.Contents("C:\Users\konda\OneDrive - Teesside University\big data\ICA\flights data\2023.csv"),[Delimiter=","], Cc
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"FL_DATE", type date}, {"AIRLINE_CODE", type text}, {"DOT_CODE", Int64.
    #"Removed Columns" = Table.RemoveColumns(#"Changed Type",{"FL_DATE", "FL_YEAR", "FL_MONTH", "FL_DAY"}),
    #"Removed Duplicates" = Table.Distinct(#"Removed Columns"),
    #"Replaced Value" = Table.ReplaceValue(#"Removed Duplicates",null,0,Replacer.ReplaceValue,{"DELAY_DUE_CARRIER", "DELAY_DUE_WEATHER", "DELA
    #"Removed Columns1" = Table.RemoveColumns(#"Replaced Value",{"DOT_CODE"}),
    #"Removed Other Columns" = Table.SelectColumns(#"Removed Columns1",{"ORIGIN", "ORIGIN_CITY"}),
    #"Removed Duplicates1" = Table.Distinct(#"Removed Other Columns")
in
    #"Removed Duplicates1"
  
```

Figure 11: Origin table M Code

Similarly created a new table named "Destination Codes" by duplicating the main dataset i.e., 2023 and deleted all the columns except "DEST_CITY", "DEST" and removing the duplicates.

The M code for Destination codes is:



```

let
    Source = Csv.Document(File.Contents("C:\Users\konda\OneDrive - Teesside University\big data\ICA\flights data\2023.csv"),[Delimiter=","], Cc
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"FL_DATE", type date}, {"AIRLINE_CODE", type text}, {"DOT_CODE", Int64.
    #"Removed Columns" = Table.RemoveColumns(#"Changed Type",{"FL_DATE", "FL_YEAR", "FL_MONTH", "FL_DAY"}),
    #"Removed Duplicates" = Table.Distinct(#"Removed Columns"),
    #"Replaced Value" = Table.ReplaceValue(#"Removed Duplicates",null,0,Replacer.ReplaceValue,{"DELAY_DUE_CARRIER", "DELAY_DUE_WEATHER", "DELA
    #"Removed Columns1" = Table.RemoveColumns(#"Replaced Value",{"DOT_CODE"}),
    #"Removed Other Columns" = Table.SelectColumns(#"Removed Columns1",{"DEST_CITY", "DEST"}),
    #"Removed Duplicates1" = Table.Distinct(#"Removed Other Columns")
in
    #"Removed Duplicates1"
  
```

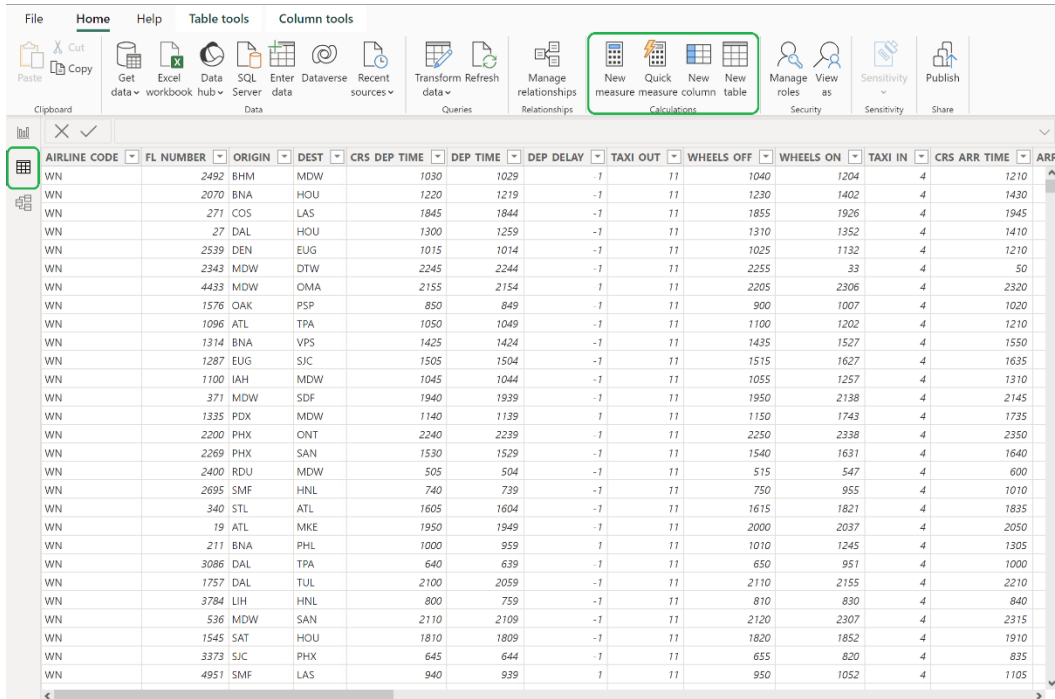
Figure 12: Destination Codes tables M code

Once all the necessary tables are created, then click on close and apply to load the data into PowerBI data pane to visualize the data.

2.4 DAX formulas:

DAX (Data Analysis Expressions) is a formula language used in Microsoft tools such as Power BI, Excel, and others. It is used to manipulate data, perform calculations, and create custom measures. DAX improves data modelling and analysis by providing powerful aggregation and transformation functions.

We can use DAX formulas in Power BI in table view mode and using DAX we can create measures, columns and tables as showing in the figure 13



AIRLINE CODE	FL NUMBER	ORIGIN	DEST	CRS DEP TIME	DEP TIME	DEP DELAY	TAXI OUT	WHEELS OFF	WHEELS ON	TAXI IN	CRS ARR TIME	ARR
WN	2492	BHM	MDW	1030	1029	-1	11	1040	1204	4	1210	
WN	2070	BNA	HOU	1220	1219	-1	11	1230	1402	4	1430	
WN	271	COS	LAS	1845	1844	-1	11	1855	1926	4	1945	
WN	27	DAL	HOU	1300	1259	-1	11	1310	1352	4	1410	
WN	2539	DEN	EUG	1015	1014	-1	11	1025	1132	4	1210	
WN	2343	MDW	DTW	2245	2244	-1	11	2255	33	4	50	
WN	4433	MDW	OMA	2155	2154	-1	11	2205	2306	4	2320	
WN	1576	OAK	PSP	850	849	-1	11	900	1007	4	1020	
WN	1096	ATL	TPA	1050	1049	-1	11	1100	1202	4	1210	
WN	1314	BNA	VPS	1425	1424	-1	11	1435	1527	4	1550	
WN	1287	EUG	SJC	1505	1504	-1	11	1515	1627	4	1635	
WN	1100	IAH	MDW	1045	1044	-1	11	1055	1257	4	1310	
WN	371	MDW	SDF	1940	1939	-1	11	1950	2138	4	2145	
WN	1335	PDX	MDW	1140	1139	-1	11	1150	1743	4	1735	
WN	2200	PHX	ONT	2240	2239	-1	11	2250	2338	4	2350	
WN	2269	PHX	SAN	1530	1529	-1	11	1540	1631	4	1640	
WN	2400	RDU	MDW	505	504	-1	11	515	547	4	600	
WN	2695	SMF	HNL	740	739	-1	11	750	955	4	1010	
WN	340	STL	ATL	1605	1604	-1	11	1615	1821	4	1835	
WN	19	ATL	MKE	1950	1949	-1	11	2000	2037	4	2050	
WN	211	BNA	PHL	1000	959	-1	11	1010	1245	4	1305	
WN	3086	DAL	TPA	640	639	-1	11	650	951	4	1000	
WN	1757	DAL	TUL	2100	2059	-1	11	2110	2155	4	2210	
WN	3784	LIH	HNL	800	759	-1	11	810	830	4	840	
WN	536	MDW	SAN	2110	2109	-1	11	2120	2307	4	2315	
WN	1545	SAT	HOU	1810	1809	-1	11	1820	1852	4	1910	
WN	3373	SJC	PHX	645	644	-1	11	655	820	4	835	
WN	4951	SMF	LAS	940	939	-1	11	950	1052	4	1105	

Figure 13: Data View tab.

2.4.1 Airline Rank column:

- Created new column using DAX formula to get the rank of the airline companies based on total number of flights handled by the carrier.
- Below is the DAX code for the same:

```
AirlinesRank =
RANKX(
    ALL('2023'[AIRLINE_CODE]),
    CALCULATE(
        COUNTROWS('2023'),
        ALLEXCEPT('2023', '2023'[AIRLINE_CODE])
    ),
    ,
    DESC
)
```

- This DAX code generates the calculated column 'AirlinesRank' in the table or model '2023.' It ranks airlines in descending order based on the number of flights handled. The code uses the RANKX function, which considers all unique airline codes and calculates counts for each airline separately. The ALLEXCEPT function ensures that the count is evaluated for each airline, and DESC indicates the descending ranking order. This column can be used to compare and visualize airlines' relative performance in handling flights in 2023.

2.4.2 AvgArrivalDelay

- Created a new measure using DAX formula to get the Average Arrival Delay
- Below is the DAX code for the same:

AvgArrivalDelay = `AVERAGE('2023'[ARR_DELAY])`

- The average arrival delay for flights in the '2023' table is calculated using this DAX code. It generates a measure called 'AvgArrivalDelay' by employing the AVERAGE function, which computes the mean of the 'ARR_DELAY' column, which represents the average time that flights arrived later than scheduled. This metric provides an overall metric for analysing the arrival timeliness in the specified dataset.

2.4.3 AvgDepartureDelay

- Created a new measure using DAX formula to get the Average Departure Delay
- Below is the DAX code for the same:

AvgDepartureDelay = `AVERAGE('2023'[DEP_DELAY])`

- The average Departure delay for flights in the '2023' table is calculated using this DAX code. It generates a measure called 'AvgDepartureDelay' by employing the AVERAGE function, which computes the mean of the 'DEP_DELAY' column, which represents the average time that flights arrived later than scheduled. This metric provides an overall metric for analysing the arrival timeliness in the specified dataset.

2.4.4 CancellationRate

- Created new measure using DAX formula to get the cancellation percentage in overall flight count.
- Below is the DAX code for the same:

CancellationRate = `DIVIDE(COUNTROWS(FILTER('2023', '2023'[CANCELLED] = 1)), COUNTROWS('2023'))`

- This DAX code creates a 'CancellationRate' measure into the '2023' table. The cancellation rate is calculated by dividing the number of rows where 'CANCELLED' is 1 (indicating a cancelled flight) by the total number of rows in the entire '2023' table. The DIVIDE function deals with division by zero errors. The resulting measure represents the proportion of flights in the given dataset that were cancelled, providing insight into the overall cancellation rate for the specified period.

2.4.5 Cancelled flights.

- Created new measure using DAX formula to get the cancellation flights count.
- Below is the DAX code for the same:

Cancelled flights = `COUNTROWS(FILTER('2023', '2023'[CANCELLED] = 1))`

- The DAX code generates a 'Cancelled flights' measure in the '2023' table. It counts the number of rows with the 'CANCELLED' column set to 1, indicating cancelled flights. This metric returns the total number of cancelled flights in the specified dataset, providing a quantitative measure of the total number of flight cancellations during the specified period.

2.4.6 DelayDueToCarrierPercentage

- Created new measure using DAX formula to get the delayed percentage due to Carrier.
- Below is the DAX code for the same:

```
DelayDueToCarrierPercentage = DIVIDE(SUM('2023'[DELAY_DUE_CARRIER]),  
SUM('2023'[ARR_DELAY]))
```

- The DAX code 'DelayDueToCarrierPercentage' calculates the percentage of total delay that can be attributed to the carrier for flights in the '2023' table. It accomplishes this by dividing the total number of delays caused by the carrier ('DELAY_DUE_CARRIER') by the total number of arrival delays ('ARR_DELAY'). To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of total delays that can be attributed to carrier-related issues during the specified time period.

2.4.7 DelayDueToLateAircraftPercentage

- Created new measure using DAX formula to get the delayed percentage due to Aircraft.
- Below is the DAX code for the same:

```
DelayDueToNasPercentage =  
DIVIDE(SUM('2023'[DELAY_DUE_LATE_AIRCRAFT]), SUM('2023'[ARR_DELAY]))
```

- The DAX code 'DelayDueToAircraftPercentage' calculates the percentage of total delay that can be attributed to the Aircraft for flights in the '2023' table. It accomplishes this by dividing the total number of delays caused by the carrier ('DELAY_DUE_LATE_AIRCRAFT ') by the total number of arrival delays ('ARR_DELAY'). To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of total delays that can be attributed to aircraft-related issues during the specified time period.

2.4.8 DelayDueToNasPercentage

- Created new measure using DAX formula to get the delayed percentage due to Nas.
- Below is the DAX code for the same:

```
DelayDueToNasPercentage = DIVIDE(SUM('2023'[DELAY_DUE_NAS]),  
SUM('2023'[ARR_DELAY]))
```

- The DAX code 'DelayDueToNasPercentage' calculates the percentage of total delay that can be attributed to the Nas for flights in the '2023' table. It accomplishes this by dividing the total number of delays caused by the carrier ('DELAY_DUE_NAS') by the total number of arrival delays ('ARR_DELAY'). To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of total delays that can be attributed to Nas related issues during the specified time period.

2.4.9 DelayDueToSecurityPercentage

- Created new measure using DAX formula to get the delayed percentage due to Security.
- Below is the DAX code for the same:

```
DelayDueToSecurityPercentage =  
DIVIDE(SUM('2023'[DELAY_DUE_SECURITY]), SUM('2023'[ARR_DELAY]))
```

- The DAX code 'DelayDueToSecurityPercentage' calculates the percentage of total delay that can be attributed to the security for flights in the '2023' table. It accomplishes this by dividing the total number of delays caused by the carrier ('DELAY_DUE_SECURITY') by the total number of arrival delays ('ARR_DELAY'). To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of total delays that can be attributed to security related issues during the specified time period.

2.4.10 DelayDueTo WeatherPercentage

- Created new measure using DAX formula to get the delayed percentage due to Weather.
- Below is the DAX code for the same:

```
DelayDueToWeatherPercentage = DIVIDE(SUM('2023'[DELAY_DUE_WEATHER]),  
SUM('2023'[ARR_DELAY]))
```

- The DAX code 'DelayDueToWeatherPercentage' calculates the percentage of total delay that can be attributed to the weather for flights in the '2023' table. It accomplishes this by dividing the total number of delays caused by the carrier ('DELAY_DUE_WEATHER') by the total number of arrival delays ('ARR_DELAY'). To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of total delays that can be attributed to weather related issues during the specified time period.

2.4.11 FlightsDivertedPercentage

- Created new measure using DAX formula to get the flight diverted percentage from the overall percentage.
- Below is the DAX code for the same:

```
FlightsDivertedPercentage = DIVIDE(COUNTROWS(FILTER('2023',  
'2023'[DIVERTED] = 1)), COUNTROWS('2023'))
```

- In the '2023' table, the DAX code 'FlightsDivertedPercentage' calculates the percentage of flights that were diverted. It accomplishes this by dividing the number of rows in the '2023' table where the 'DIVERTED' column is equal to 1 (indicating a diverted flight) by the total number of rows in the '2023' table. The DIVIDE function deals with division by zero errors. This metric indicates the proportion of flights that experienced diversions during the specified time period.

2.4.12 OnTimeDeparturesPercentage

- Created new measure using DAX formula to get the flight percentage which are having departures ontime.
- Below is the DAX code for the same:

```
OnTimeDeparturesPercentage = DIVIDE(COUNTROWS(FILTER('2023',
'2023'[DEP_DELAY] <= 0)), COUNTROWS('2023'))
```

- In the '2023' table, the DAX code 'OnTimeDeparturesPercentage' calculates the percentage of on-time departures. It accomplishes this by counting the rows in the '2023' table where the 'DEP_DELAY' column (departure delay) is less than or equal to 0 (indicating an on-time departure) and then dividing that total by the total number of rows in the '2023' table. To handle potential division by zero errors, the DIVIDE function is used. This metric provides information on the percentage of flights that departed on time or with no delay during the specified time period.

2.4.13 Total Diverted Flights

- Created new measure using DAX formula to get the total count of diverted flights.
- Below is the DAX code for the same:

```
Total Diverted Flights = COUNTROWS(FILTER('2023', '2023'[DIVERTED] = 1))
```

- The DAX code 'Total Diverted Flights' generates a measure in the '2023' table that counts the number of rows with the 'DIVERTED' column set to 1. This figure represents the total number of flights that were diverted during the time period specified in the dataset. It provides a numerical measure of the magnitude of flight diversions in the given context.

2.4.14 TotalAirTime

- Created new measure using DAX formula to get the total airtime of the flights.
- Below is the DAX code for the same:

```
TotalAirTime = SUM('2023'[AIR_TIME])
```

- The DAX code 'TotalAirTime' creates a measure in the '2023' table by adding the 'AIR_TIME' column sums. This metric represents the total amount of time spent in the air for all flights in the specified dataset (2023). It provides a total metric of airborne duration for all flights during the specified time period.

2.4.15 TotalDistance

- Created new measure using DAX formula to get the total distance of the flights.
- Below is the DAX code for the same:

```
TotalDistance = SUM('2023'[DISTANCE])
```

- The DAX code 'TotalDistance' creates a measure in the '2023' table by adding the 'DISTANCE' column sums. This metric represents the total distance travelled by all flights in the given dataset ('2023'). It calculates the total travel distance for all flights during the specified time period.

2.4.16 UniqueDestinationsByAirline

- Created new measure using DAX formula to get the count of unique destinations based on airlines.
- Below is the DAX code for the same:

```
UniqueDestinationsByAirline = COUNTROWS( VALUES('2023'[DEST]))
```

- The DAX code 'UniqueDestinationsByAirline' generates a measure in the '2023' table that counts the number of distinct destinations visited by flights in the dataset. The VALUES function is applied to the 'DEST' column, which contains the destination airport codes. This metric represents the variety of destinations served by airlines during the specified time period.

2.4.17 Airline with Most Diverted Flights

- Created new measure using DAX formula to get the most diverted flights based on airlines.
- Below is the DAX code for the same:

```
Airline with Most Diverted Flights =
    CALCULATE(
        MAXX(
            VALUES('2023'[AIRLINE_CODE]),
            [Total Diverted Flights]
        )
    )
```

- The DAX code 'Airline with the Most Diverted Flights' computes the airline code with the most diverted flights in the '2023' dataset. It evaluates the expression within the context of the current filter context using the CALCULATE function, MAXX to iterate over unique airline codes, and references the previously created measure '[Total Diverted Flights]' to find the maximum value. This metric identifies the airline that had the most flight diversions during the specified time period.

2.4.18 TotalCancellations

- Created new measure using DAX formula to get the total number of cancelled flights based on Airlines.
- Below is the DAX code for the same:

```
TotalCancellations = CALCULATE(SUM('2023'[CANCELLED]),
    ALL('2023'[AIRLINE_CODE]))
```

- In the '2023' table, this DAX formula computes the total number of cancellations. It sums the 'CANCELLED' column with the CALCULATE function, and ALL('2023'[AIRLINE_CODE]) removes the filter on the 'AIRLINE_CODE' column, considering all airline codes for the calculation.

2.4.19 Top5Airports

- Created new table using DAX formula to get the list of airports based on number of flights handled by the airport.
- Below is the DAX code for the same:

```
Top5Airports =
SUMMARIZE(
    '2023',
    '2023'[ORIGIN],
    "NumberofFlights", CALCULATE(COUNTROWS('2023'))
)
```

- The DAX code 'Top5Airports' generates a table in the '2023' dataset by using the SUMMARIZE function. It summarizes the data based on the 'ORIGIN' column, counting the number of flights from each unique origin airport with the help of CALCULATE and COUNTROWS. This table lists the top airports based on the number of flights during the specified time period.

2.4.20 TopAirportsByArrivalDelay

- Created new table using DAX formula to get the list of airlines based on arrival delay.
- Below is the DAX code for the same:

```
TopAirportsByArrivalDelay =
SUMMARIZE(
    '2023',
    '2023'[DEST],
    "AverageArrivalDelay", AVERAGE('2023'[ARR_DELAY])
)
```

- The DAX code 'TopAirportsByArrivalDelay' creates a table in the '2023' dataset using the SUMMARIZE function. It summarizes the data based on the 'DEST' column (destination), calculating the average arrival delay for each unique destination airport using the AVERAGE function on the 'ARR_DELAY' column. This table shows the average arrival delays at various airports over the specified time period.

2.4.21 TopAirportsByDepartureDelay

- Created new table using DAX formula to get the list of airlines based on departure delay.
- Below is the DAX code for the same:

```
TopAirportsByDepartureDelay =
SUMMARIZE(
    '2023',
    '2023'[ORIGIN],
    "AverageDepartureDelay", AVERAGE('2023'[DEP_DELAY])
)
```

- The DAX code 'TopAirportsByDepartureDelay' creates a table in the '2023' dataset using the SUMMARIZE function. It summarizes the data based on the column 'ORIGIN' (origin), calculating the average departure delay for each unique origin airport using the AVERAGE function on the column 'DEP_DELAY'. This table shows the average departure delays at various airports over the specified time period.

2.4.22 Top10 diverted by destination.

- Created new table using DAX formula to get the list of Destinations based on the number of diversions.
- Below is the DAX code for the same:

```
Top10 Diverted by destinations =
TOPN(10,
    SUMMARIZE('2023', 'Destination Codes'[DEST_CITY],
    '2023'[DEST], "TotalDiverted", [Total Diverted Flights]),
    [TotalDiverted],
    DESC
)
```

- This DAX formula creates a table called 'Top10 Diverted by Destinations' by selecting the top ten destinations in the '2023' table based on the total number of diverted flights.

2.4.23 Destination Airport

- Created new column in the table "Top 10 diverted by destinations" to combine the Destination city and destination airport code for visualization.
- Below is the DAX code for the same:

```
Destination Airport = 'Top10 Diverted by
destinations'[DEST_CITY] & " (" & 'Top10 Diverted by
destinations'[DEST] & ")"
```

- This DAX formula adds a "Destination Airport" column to the 'Top10 Diverted by Destinations' table by combining the 'DEST_CITY' and 'DEST' columns, which are formatted as "DEST_CITY (DEST)".

3. Data modelling:

Data modelling is the process of organizing and structuring data to allow for effective analysis and visualization. Defining relationships, creating hierarchies, and optimizing data

structures for databases or analytical tools are all part of it. A well-designed data model improves data integrity, simplifies querying, and provides meaningful insights, allowing for better decision-making and comprehension of the underlying information in a dataset or database.

From the dataset '2023', we have created tables to optimize the data and get the insights for efficiently.

The schema used to optimize the dataset is snowflake schema figure 14. The snowflake schema is a database modelling technique that normalizes a normalized star schema by dividing its dimension tables into multiple related tables. This results in a snowflake-like hierarchical structure. While it reduces redundancy and improves data integrity, increased table joins can lead to more complex queries. Snowflake schemas are popular in data warehouses.

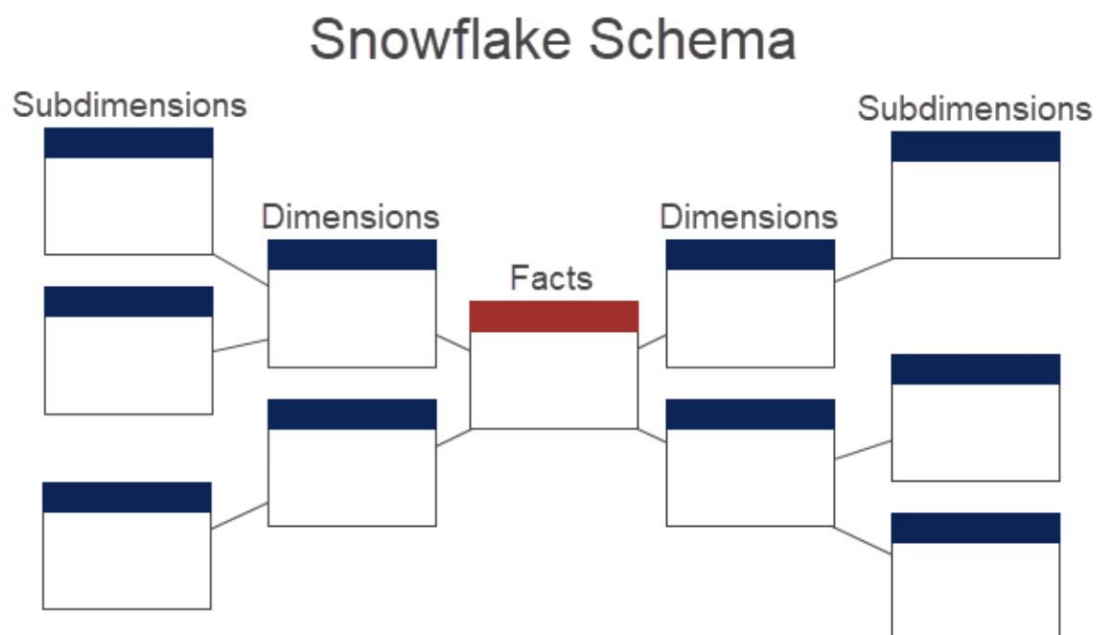


Figure 14: Snowflake Schema.

To create the snowflake schema, below tables are created.

Fact table:

- 2023

Dimension tables:

- AIRLINE_CODE_DICTIONARY
- cancellation codes
- Destination Codes
- Origin codes
- Top5Airports
- TopAirportsByArrivalDelay
- TopAirportsByDepartureDelay
- Top 10 Diverted by Destinations

The snowflake schema created based on the fact table and dimension tables, below is the star schema in figure 15.

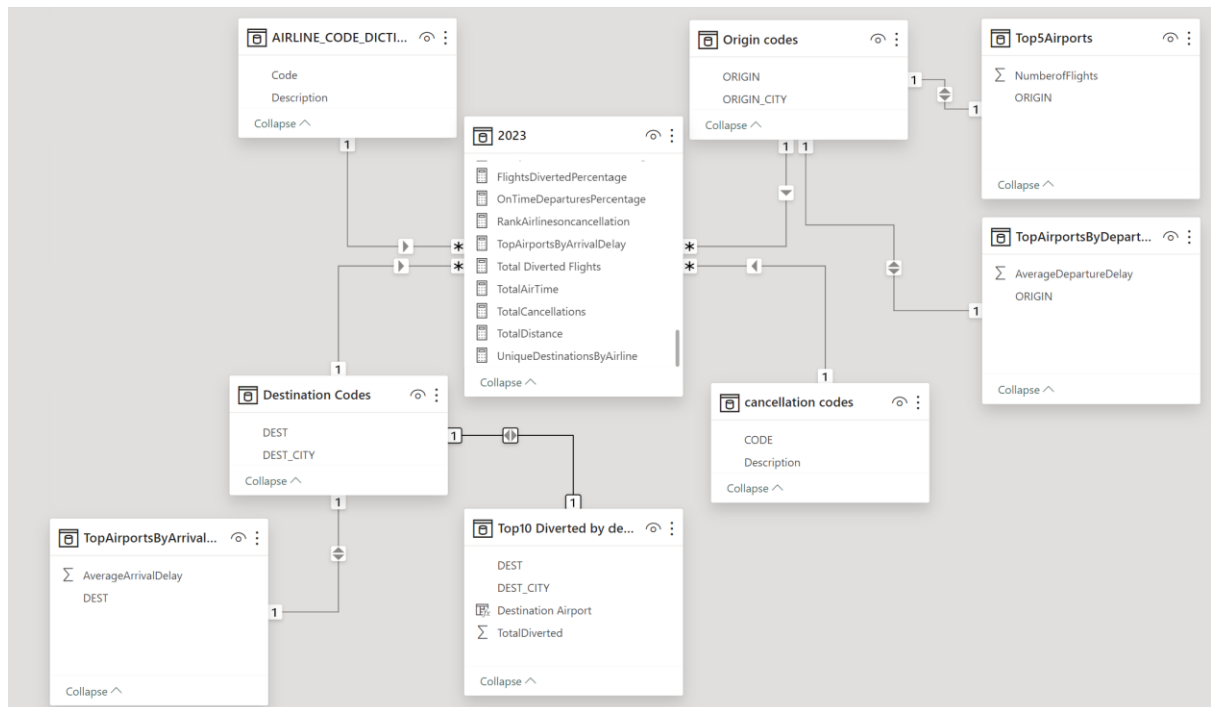


Figure 15: Snowflake schema in Power BI.

4. Dashboard:

The Dashboard created based on the dataset is divided into 4 categories, they are Overview of the data, cancelled flight analysis, delayed flight analysis and Diverted flight analysis. And the colour schema used to create the dashboard is based on the US Department of Transportation logo.

4.1 US Department of transportation logo.



Figure 16: US DOT Logo

The based colours in the logo are blue and white, hence all the dashboards background and the graphs backgrounds are based on blue colour and all the Alpha numeric values are in white colour.

Colour Hexa code:

Background: #151138

Font: #E6E6E6

4.2 Dashboards

4.2.1 Overview:

The overview dashboard gives the high-level information of the dataset.

The business questions answered using the overview dashboard are:

- The basic information on dataset and the source information
- What is the total distance travelled in miles?
- What is the total airtime of the flights in mins?
- Leaderboard of the carriers based on the flights.
- Total number of months in the dataset.
- Total number of carriers operated in that time period.
- Total number of destinations operated in that time period
- Total number of flights operated in that time period
- Top 5 airports based on number of flights handled.
- Top 5 airlines based on unique destinations handled.

Figure 17 shows the overview dashboard

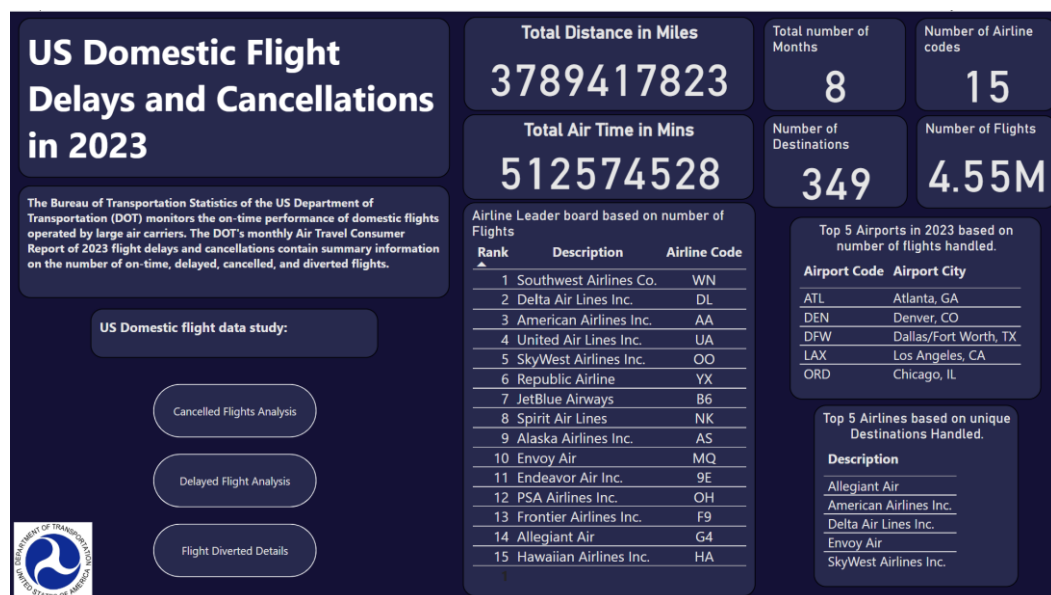


Figure 17: Overview dashboard

4.2.2 Delayed Flight Analysis

The delayed flights analysis is created to study about the flight delayed data. It helps to understand the reasons, and which month is affected the most, etc.

The business questions answered using the delayed flights analysis dashboard are:

- What is the average arrival delay?
- What is the average departure delay?
- What is the ontime departure percentage?
- What are major factors which are involved in the flight delays?

- Which month is having highest and lowest average arrival delay and departure delay?
- Top 5 origin cities based on average departure delay.
- Study all the above mentioned paraments based on flight carriers.

Figure 18 shows the delayed flight analysis dashboard



Figure 18: Delayed Flight Analysis dashboard.

4.2.3 Cancelled Flights Analysis

The cancelled flights analysis is created to study about the flight cancelled data. It helps to understand the cancellation reasons, and which month is affected the most, etc.

The business questions answered using the cancelled flights analysis dashboard are:

- What is the percentage of flights cancelled in that time period?
- What the overall count of the cancelled flights?
- The major reasons for the flight cancellations?
- Top 8 airlines which are having highest number of cancellations?
- Which month is the having highest and lowest number of cancellations?
- Study all the above mentioned paraments based on flight carriers.

Figure 19 shows the cancelled flights analysis dashboard

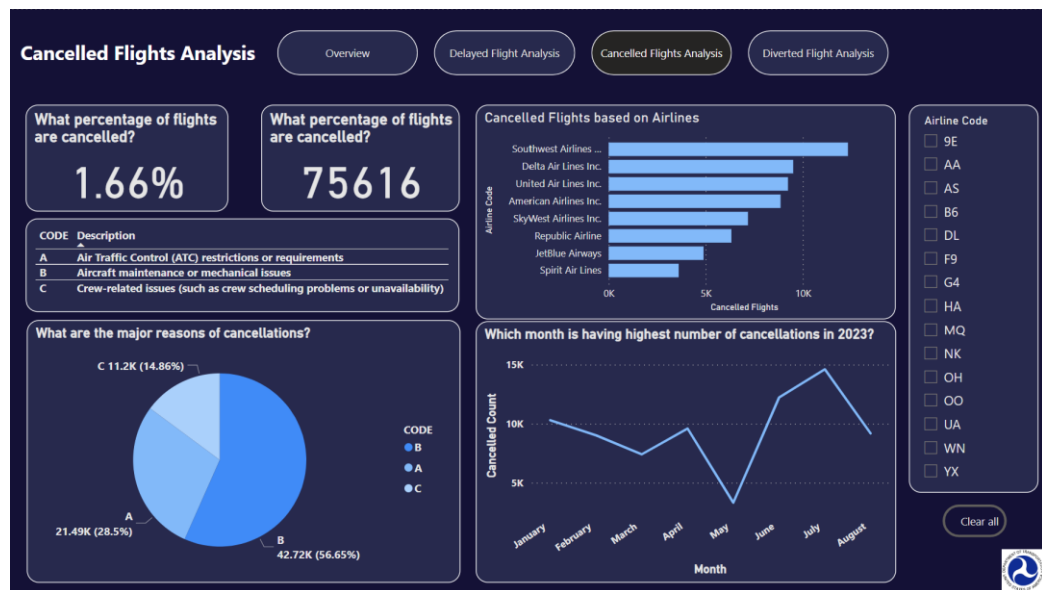


Figure 19: Cancelled Flight Analysis dashboard

4.2.4 Diverted Flight Analysis

The diverted flights analysis is created to study about the flight diverted data. It helps to understand month is affected the most, highest number of flights diverted based on airlines, etc.

The business questions answered using the diverted flights analysis dashboard are:

- What is the percentage of diverted flights?
- What is the total number of diverted flights?
- What is the average departure delay for diverted flights in mins?
- Which month is having highest and lowest diverted flights?
- Which airlines are highest having diverted flights?
- Study all the above mentioned parameters based on flight carriers.

Figure 20 shows the diverted flight analysis dashboard:

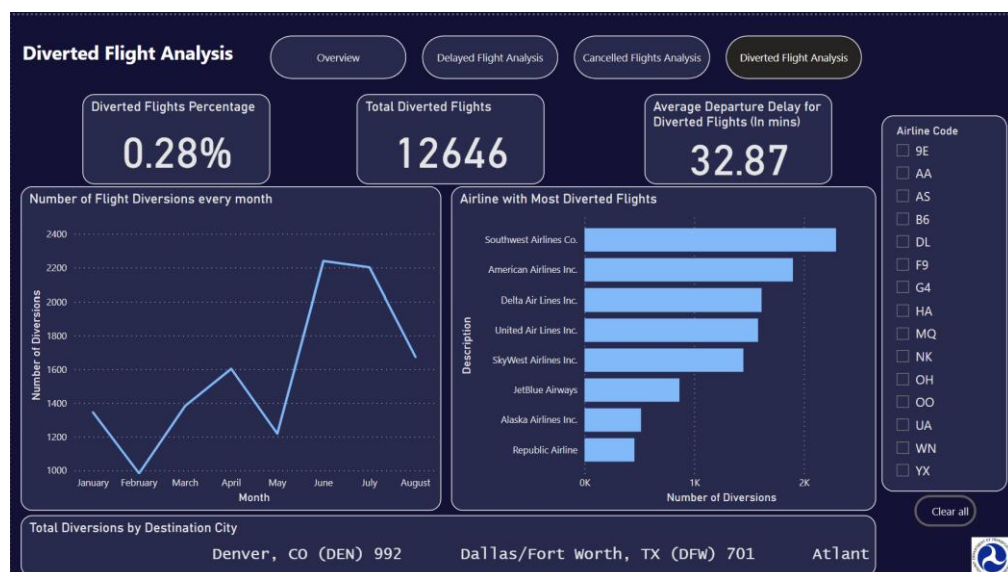


Figure 20: Diverted Flight Analysis dashboard.

5. Reference:

- US Department of Transportation:
https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr
- Dataset source: <https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023>
- Snowflake Schema Reference: <https://phoenixnap.com/kb/star-vs-snowflake-schema>

6. Self-Assessment:

Report Section	Description	Grade
Report Structure	The report is well-written, and it contains all the relevant sections	95
Data Pre-processing and Data Modelling	Many pre-processing steps have been applied. The data model is well-structured	98
Dax and M language	Both DAX and M Language have been extensively used in the report	95
Dashboard Design	The dashboard contains a variety of charts, including advanced ones not covered in the module.	85
Average		93.25