Airline Flight Delays and Airport Performance

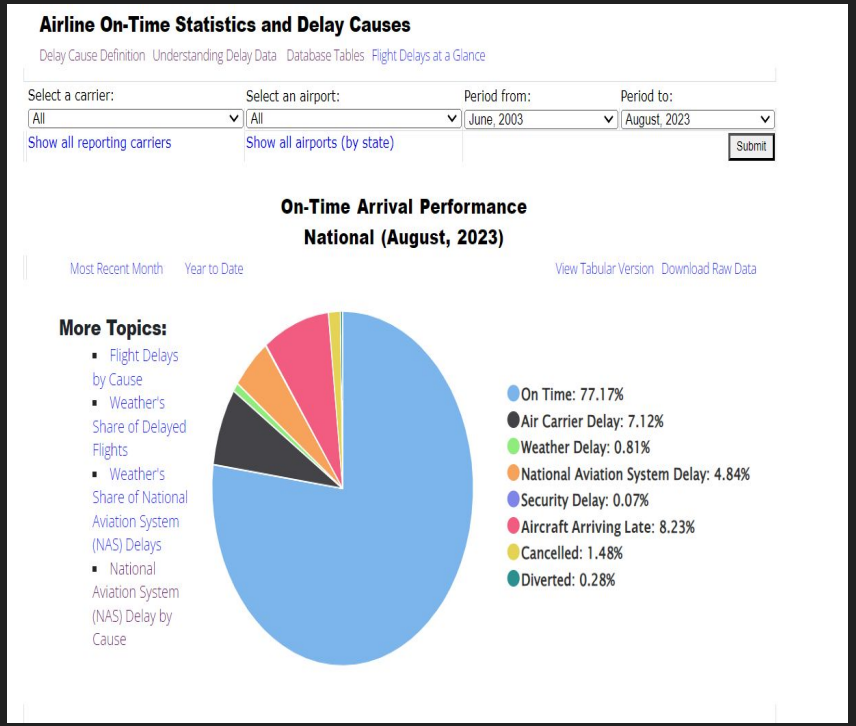
A look into flight delay reasons across USA

Contents

Understanding the data

Data Source

United States Department of Transportation: Bureau of Transportation Statistics – https://www.transtats.bts.gov/ot_delay/OT_DelayCause1.asp?20=E



Snapshot of the CSV file

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	year	month	carrier	carrier_na	airport	airport_na	arr_flights	arr_del15	carrier_ct	weather_c	nas_ct	security_c	late_aircra	arr_cancel	arr_divert	arr_delay	carrier_de	weather_c	nas_delay	security_d	late_aircraft_delay	
2	2023	8	9E	Endeavor	ABE	Allentown	89	13	2.25	1.6	3.16	0	5.99	2	1	1375	71	761	118	0	425	
3	2023	8	9E	Endeavor	ABY	Albany, GA	62	10	1.97	0.04	0.57	0	7.42	0	1	799	218	1	62	0	518	
4	2023	8	9E	Endeavor	AEX	Alexandria	62	10	2.73	1.18	1.8	0	4.28	1	0	766	56	188	78	0	444	
5	2023	8	9E	Endeavor	AGS	Augusta, G	66	12	3.69	2.27	4.47	0	1.57	1	1	1397	471	320	388	0	218	
6	2023	8	9E	Endeavor	ALB	Albany, NY	92	22	7.76	0	2.96	0	11.28	2	0	1530	628	0	134	0	768	
7	2023	8	9E	Endeavor	ATL	Atlanta, GA	1636	256	55.98	27.81	63.64	0	108.57	32	11	29768	9339	4557	4676	0	11196	
8	2023	8	9E	Endeavor	AUS	Austin, TX	75	12	5.62	0.97	4.41	0	1	0	0	843	535	170	111	0	27	
9	2023	8	9E	Endeavor	AVL	Asheville,	59	7	3.32	0	0.42	0	3.26	2	0	324	117	0	25	0	182	
10	2023	8	9E	Endeavor	AZO	Kalamazoo	62	13	6.53	0.94	3.54	0	1.99	0	0	707	470	77	87	0	73	
11	2023	8	9E	Endeavor	BDL	Hartford, (	30	4	0	0.82	0	0	3.18	1	0	1421	0	532	0	0	889	
12	2023	8	9E	Endeavor	BGM	Binghamto	58	10	2.78	0	3.18	0	4.03	1	0	1604	207	0	1049	0	348	
13	2023	8	9E	Endeavor	BGR	Bangor, M	124	13	8.42	1	0.5	0	3.08	3	0	1207	282	650	18	0	257	
14	2023	8	9E	Endeavor	BHM	Birmingha	84	17	4.11	0	4.24	0	8.65	2	2	1124	208	0	164	0	752	
15	2023	8	9E	Endeavor	BNA	Nashville,	166	25	6.02	2.91	11.4	0	4.68	2	0	1465	362	308	523	0	272	
16	2023	8	9E	Endeavor	BQK	Brunswick	62	14	7.46	0.2	3.1	0	3.24	2	2	2641	1238	184	771	0	448	
17	2023	8	9E	Endeavor	BTV	Burlington	147	30	11.85	0	11	0	7.16	3	0	1628	714	0	324	0	590	
18	2023	8	9E	Endeavor	BUF	Buffalo, NY	154	25	7.96	0	10.66	0	6.38	2	0	1065	289	0	448	0	328	
19	2023	8	9E	Endeavor	BWI	Baltimore,	62	13	1.92	0	3.82	0.2	7.05	2	0	900	89	0	203	28	580	
20	2023	8	9E	Endeavor	CAE	Columbia,	92	20	3.74	0	9.41	0	6.85	1	0	1375	398	0	448	0	529	
21	2023	8	9E	Endeavor	CHA	Chattanoo	119	17	5.23	3.1	1.96	0	6.71	1	0	1108	341	174	82	0	511	
22	2023	8	9E	Endeavor	CHO	Charlottes	139	17	4.17	2.6	6.67	0	3.55	2	1	891	226	118	348	0	199	
23	2023	8	9E	Endeavor	CHS	Charlestor	137	16	4.88	0.41	5.35	0	5.35	4	0	935	322	12	330	0	271	
24	2023	8	9E	Endeavor	CLE	Cleveland,	323	62	24.42	3.4	8.65	0	25.53	6	3	4601	1792	851	304	0	1654	
25	2023	8	9E	Endeavor	CLT	Charlotte,	232	38	9.94	1	13.46	0	13.6	9	1	2617	752	49	753	0	1063	

Understanding the columns

- Year and month (June 2003 - August 2023)
- Carrier code and Carrier name
- Airport code and Airport name
- Count of flights arrived
- Count of flights delayed; **flights arriving late 15 minutes or more are counted as official delays**
- Breakdown of count of flights delayed due to various *reasons*
- Count of flights cancelled
- Count of flights diverted
- Aggregated time (in minutes) by which the flights were delayed
- Breakdown of aggregated time (in minutes) by which the flights were delayed due to various *reasons*

Understanding the delay reasons

How are these categories defined?

- **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- **Extreme Weather:** Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- **Late-arriving aircraft:** A previous flight with same aircraft arrived late, causing the present flight to depart late.
- **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

Understanding the delay reasons

- Carrier delay (maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.) – **within Airline's control**
- Security delay (evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment, etc.) – **within Airport's control OR not Airline's fault**
- NAS delay (non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control) – **partially Airline and Airport's fault**
- Late aircraft delay (previous flight with same aircraft arrived late, causing the present flight to depart late) – **partially Airline and Airport's fault**
- Weather delay (tornado, blizzard, hurricane) – **neither Airline nor Airport's fault**

Data pre-processing

Checking for data quality

What are we looking for?

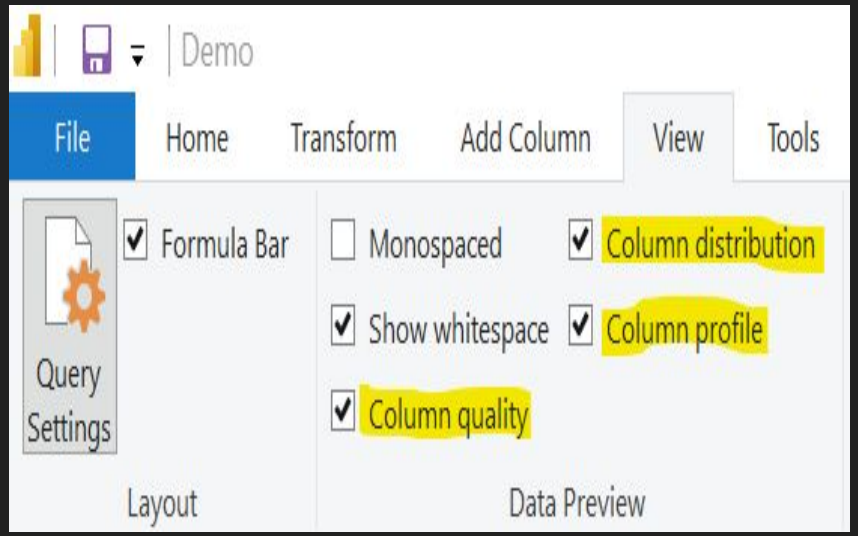
- Appropriate data types
- Missing data
- Redundant data
- Duplicated data
- Anything else that stands out

Checking for data quality (2)

Column distribution – count of distinct and unique values in every column

Column quality - ‘valid’, ‘erroneous’, or ‘empty’ entries in every column

Column profile - column statistics (count, min, max, average, standard deviation, etc.)



Checking for data quality (3)

- 21 years (2003 - 2023) and 12 months. Data type: *Whole Number*, instead of **Date**. We will create a calculated column using DAX later.
- 29 Carrier codes and 33 Carrier names. Some Airlines changed their names or got acquired. Example: ExpressJet Airlines LLC (not ExpressJet Airlines Inc. or Atlantic Southeast Airlines). The two columns were fixed to reflect that.
- 420 Airport codes however, 444 Airport names. The 24 Airport codes have both official and unofficial airport names. Example: Cape Cod Gateway and Barnstable Municipal-Boardman/Polando Field in Hyannis, MA.

Checking for data quality (4)

- For the remaining columns, there's less than 1% empty rows – there are 509 empty rows which can be removed.
- There are still 238 null values in the **arr_del15** (count of flights delayed) column. Those flights were either cancelled or diverted (as reflected in their respective columns) hence, that doesn't need to be fixed.
- The 5 columns reflecting the Delay reasons (carrier, security, NAS, late aircraft, weather) count are decimal numbers, instead of whole numbers. This is because a flight could be delayed for multiple reasons. Bureau of Transportation & Statistics accordingly divides those delays into the different categories/reasons hence, the decimal number. If you add all the 5 columns, it gives you a whole number equal to **arr_del15** column.

Star schema Dimensional Model

Dimensional Modeling

Problem

The four columns (carrier, carrier_name, airport, airport_name) have duplicated/redundant rows.

Resolution

Separate tables – “Carrier” and “Airport” such that each entry/row is unique (only happens once).

Dimensional Modeling (2)

What other columns might have similar problems in the future?

Year and *month* also have repeated entries.

What can we do today to avoid that?

Separate table – “Date” such that each entry/row is unique (only happens once).

Surrogate keys

A surrogate key is defined as a unique identifier for the records in a table. It does not have any business value or semantic meaning – it is only used for data analysis. It is not part of the application, and is invisible to the user.

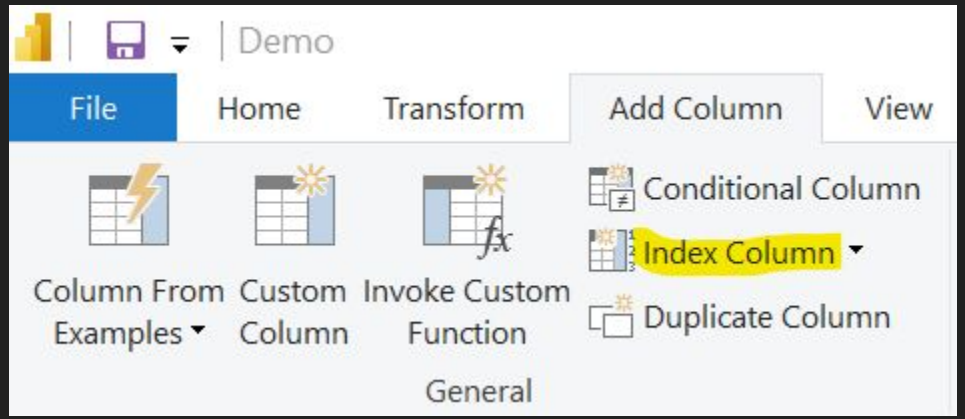
Characteristics

- It holds a unique value for all records.
- It is generated automatically.
- It can’t be modified by the user or the application.

Surrogate keys (2)

Advantages

- Makes queries run faster a.k.a improves Power BI performance
- Especially useful for Type 2 SCD (Slowly Changing Dimension)

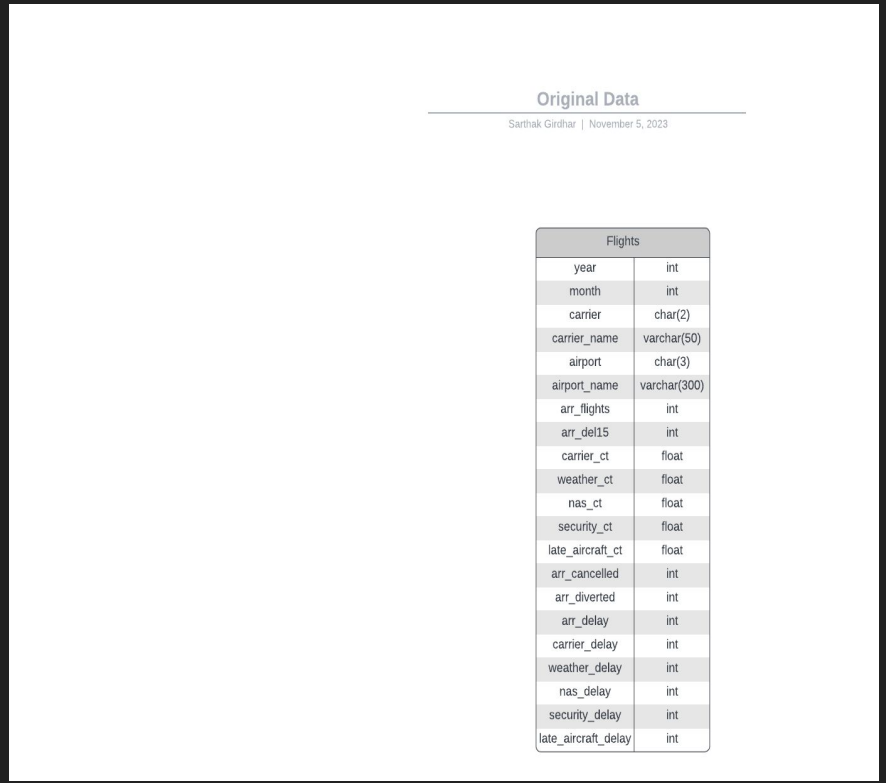


Type 1 SCD (Slowly Changing Dimension)

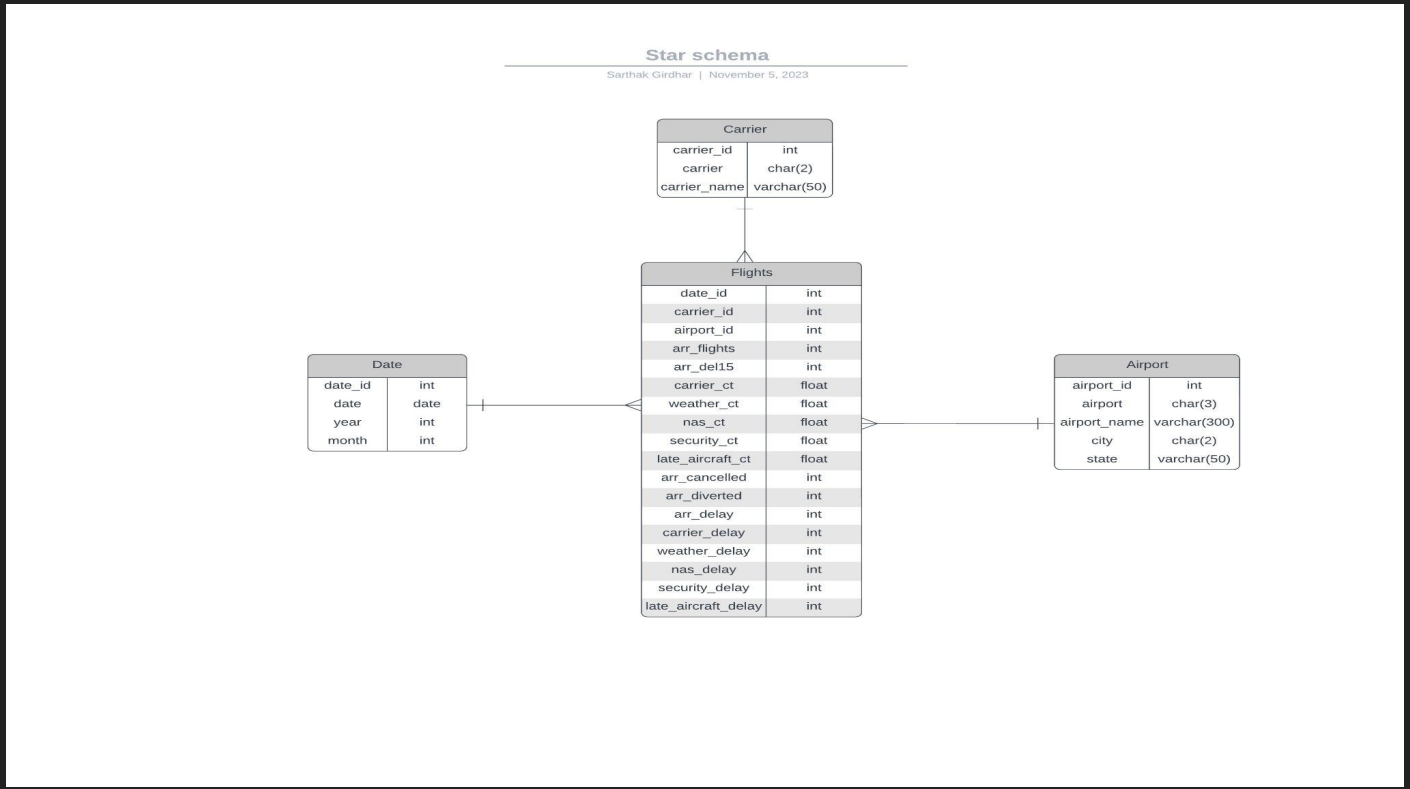
Since, our dataset has Type 1 SCD (old airline/airport names changing to new airline/airport names), we can just overwrite the old data.

We still introduce surrogate keys though, as it is a good database design practice.

Original data



Star schema Dimensional Model



Final Transformations & Managing Relationships

- 4 tables (Flights, Carrier, Airport, Date) with surrogate keys and redundant columns deleted from the *Flights* table.
- Extract city and state names from the **airport_name** column and add additional columns.

Relationships

- Date to Flights - One to Many
- Carrier to Flights - One to Many
- Airport to Flights - One to Many

Dashboards (Airline Flight Delays & Airport Performance)

Airline Flight Delay

- **Carrier Delay** is the only delay reason that is completely within an Airline/Carrier’s control.
- Scope of data: Jan 2013 - Aug 2023 (last 10 years).
- This is because not a lot of data exists for the first 10 years. Also, a lot of airlines that operated flights from 2003 to 2012 are no longer operational.

Airline Flight Delays (2)

- Taking into account i] operating carriers from Jan 2013 - Aug 2023, and ii] whether they report data, there are only 8 carriers:-
  - American Airlines Inc. (AA)
  - Delta Air Lines Inc. (DL)
  - Frontier Airlines Inc. (F9)
  - Hawaiian Airlines Inc. (HA)
  - JetBlue Airways (B6)
  - Southwest Airlines Co. (WN)
  - Spirit Air Lines (NK)
  - United Air Lines Inc. (UA)
- We create a new table in Power Query called **Carrier_Reporting** with columns ‘carrier_id’, ‘carrier’ and ‘carrier_name’. Finally, we also manage the relationship with the “Flights” table in the MODEL view.

Airline Flight Delays (3)

Metrics displayed

- Average carrier delay in minutes from 2013 to 2023.
- YoY % change for the 8 airlines.
- Count of delayed flights due to carrier VS other reasons (late aircraft, weather, NAS, security).

Airline Flight Delays (4)

Questions answered

- What is the average carrier delay for major US airlines in the past 10 years?
- What is the proportion of flights delayed due to airlines when compared to other reasons?

Questions raised

- When the number of flights operated in an year has remained approximately similar in the last 5 years (since 2018), why is the carrier delay (both in time and count) still increasing?

Airport Performance

- Scope of data: Jan 2013 - Aug 2023 (last 10 years).
- Less than 5% of the delays are attributed to **weather** and **security** (within Airport’s control) delay.

Airport Performance (2)

Metrics displayed

- Average arrival delay (in minutes) for the selected airport from 2013 to 2023.
- Percentage of delayed flights from 2013 to 2023 for the selected airport.
- Breakdown of flight delay reasons.
- Count of Total Flights, Delayed Flights and Delayed Flights % (till date as well as separately for each individual year).
- 2022 Average delay (in minutes).
- 2022 Airport Ranking (out of 373 airports).

Airport Performance (3)

Questions answered

- What is the average arrival delay and % delayed flights for US airports in the past 10 years?
- What is the breakdown of flight delay reasons for US airports?

Questions raised

- When we know (for most airports), over 65% of the delays happen due to *carrier* and *late aircraft*, can we increase the check-in time cut-off for the passengers? Another option is to start penalizing airlines for increased taxi time however, there's a possibility that the airlines might transfer that cost to the passengers.

DAX calculations (new measures, columns and tables)

1. Date [date] = FORMAT(DATE('Date'[year],'Date'[month],1),"MM-YYYY")
2. Flights [Average arr_delay] = DIVIDE (Flights[arr_delay], Flights[arr_del15], 0)
3. Flights [Average carrier_delay] = DIVIDE (Flights[carrier_delay], Flights[carrier_ct], 0)
4. Flights [Average weather_delay] = DIVIDE (Flights[weather_delay], Flights[weather_ct], 0)
5. Flights [Average nas_delay] = DIVIDE (Flights[nas_delay], Flights[nas_ct], 0)
6. Flights [Average security_delay] = DIVIDE (Flights[security_delay], Flights[security_ct], 0)
7. Flights [Average late_aircraft_delay] = DIVIDE (Flights[late_aircraft_delay], Flights[late_aircraft_ct], 0)
8. Flights [on_time flights] = Flights[arr_flights] - Flights[arr_del15]

DAX calculations (new measures, columns and tables)

```
9. carrier_delay YoY% = IF(
        ISFILTERED('Date'[date]),
        VAR __PREV_YEAR =
            CALCULATE(
                SUM('Flights'[Average carrier_delay]),
                DATEADD('Date'[date].[Date], -1, YEAR)
            )
        RETURN
            DIVIDE(SUM('Flights'[Average carrier_delay]) - __PREV_YEAR, __PREV_YEAR)
)
```

DAX calculations (new measures, columns and tables)

10. Flights [delayed_flights %] = DIVIDE (Flights[arr_del15], Flights[arr_flights], 0) * 100

11. Flights [non carrier_ct delayed_flights] = Flights[arr_del15] - Flights[carrier_ct]

12. Flights [carrier_ct delayed_flights %] = DIVIDE (Flights[carrier_ct], Flights[arr_del15], 0) * 100

13. Flights [non carrier_ct delayed_flights %] = DIVIDE (Flights[non carrier_ct delayed_flights], Flights[arr_del15], 0) * 100

DAX calculations (new measures, columns and tables)

14. Airport_Ranking =

```
CALCULATETABLE (
    SUMMARIZE (
        Flights,
        'Airport'[airport_id],
        'Airport'[airport_name],
        'Date'[date].[Year],
        "Average Delay", AVERAGE ( Flights[Average arr_delay] )
    ),
    'Date'[date].[Year] IN {2022}
)
```

15. Ranking = RANKX('Airport_Ranking', Airport_Ranking[Average Delay], ,ASC)

Skills demonstrated

Skills demonstrated

- Understanding dataset and reading supporting documentation.
- Checking for data quality and transformations in Power Query/M.
- Data Modeling (fact & dimension tables, star schema, surrogate keys, SCD).
- Data analysis - answering stakeholder’s questions via dashboards.
- DAX calculations to create new measures, tables, and columns.
- Visualizations (line chart, bar chart, donut chart, card visuals, text boxes, conditional sub-headings, action-based buttons, icons as buttons, use of tooltips, accessible theme, etc.)

Future scope

Future scope

- The data provided by Bureau of Transportation & Statistics is a time series. Therefore, **time series analysis** can be used to build forecasting models to predict and eventually, manage air travel delays in the future.
- **Incorporating new data sources** like, *routes* and *number of passengers flying* could reveal some new insights into the delay reasons.

References

- https://www.transtats.bts.gov/ot_delay/OT_DelayCause1.asp?20=E
- https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
- https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays#q10
- https://www.transtats.bts.gov/airports.asp
- https://learn.microsoft.com/en-us/power-bi/guidance/star-schema
- https://blog.devart.com/surrogate-key-in-sql.html
- https://learn.microsoft.com/en-us/dax/divide-function-dax
- https://hevodata.com/learn/rankx-power-bi/
- https://dax.guide/summarize/
- https://learn.microsoft.com/en-us/dax/summarize-function-dax
- https://www.sqlbi.com/articles/create-static-tables-in-dax-using-the-datatable-function/

Questions?