MSCA 31012 Data Engineering Platforms for Analytics
Final Proposal: 2017- 2022 Airline Delay Analysis
by Group 2

Kairan Zhong, Polaris Chen, Sunny Sun, Xiran Li, Yue Zhang

Ⅰ. Executive summary

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of January 2017 - July 2022 flight delays and cancellations. In addition, the DOT also tracks the quarterly operating balance statements for large and small US. air careers with different annual operating revenues. This project will create and analyze the structure of dimensional, relational and analytical data models based on the airline's flights' delay status and their airlines financial situation and determine their appropriateness.

Ⅱ. Business case and objective

Initial business statement:

- Why project is needed:
  Increase time-efficiency within airports across the United States. Companies want to use data and analytics platforms to support scalable, self-service business intelligence solutions to decrease airlines' delay rates, cut costs, and increase customer satisfaction.
- What stakeholders want to know:
  - Carrier: 1. Find out which airports have higher average delay rates. 2. Find out whether the cause of the delay is controllable(carrier management, security delay, etc) or uncontrollable(weather). 3. How to eliminate controllable delays.
  - Airport: 1. Find out which airlines are most likely to have delays and provide a longer time-estimation for specific airlines. 2. Find out whether weather is a key factor in causing delays and adjust its estimation time according to the likelihood of severe weather conditions. 3. Optimize air traffic control & take off time schedule after understanding the delay reasons for airlines
- Description of proposed analysis:
  - Modeling: build relational data modeling (conceptual/logical/physical), conduct data normalization, and build dimensional modeling
  - EDA: 1. Find out the delay rates among airports, airlines and airport & airlines. 2. Find out the main cause of delays among airports and airlines.
  - Visualizations: Build Tableau dashboards, featuring delay distributions in the U.S. and utilizing map, bar charts, line charts etc. to deliver insights.

Current business statement:
However, after data modeling, ETL process and data visualization, we figure out that there is no positive relationship between delay rates and carriers' operating costs. The evidence will be provided in the Insights part later.
Thus, instead of supporting business intelligence solutions for carriers and airports, **we shift the perspective to the passenger side.** We want to employ data engineering and data analytics tools to build a recommendation engine for passengers. To help them check which airports and carriers have higher average delay rates and the reason behind the delays. To provide them with travel advice that can effectively avoid the risk of delays.

Database Source:

- Airline_Delay_Cause.csv (https://www.kaggle.com/datasets/jawadkhattak/us-flight-delay-from-january-2017-july-2022)
- T_F41SCHEDULE_B1.csv (https://transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FKM&QO_fu146_anzr=Nv4%20Pn44vr4%20Sv0n0pvny)
- T_SCHEDULE_T3.csv (https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FKG&QO_fu146_anzr=Nv4%20Pn44vr4%20f7zzn4B)
- T_T100_MARKET_ALL_CARRIER (https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FMF&QO_fu146_anzr=Nv4%20Pn44vr45)

Database & Tools: MySQL, Python, Tableau

Ⅲ. Data Model

Data preparation/cleansing steps needed/proposed:

In addition to the *Flight Delay from Jan 2017 to July 2022* dataset from Kaggle, we also collected data from the Bureau of Transportation Statistics that help us identify the carrier's region, group, service class, and the airport type(whether it is domestic or international).

We have 4 datasets collected in total. Below are some key information in each dataset:

Data Set	Data Source	Attributes	Data Period	Data Size	Missing Values	Anomalies

Airline_Delay_Cause	Bureau of Transportation Statistics (https://www.kaggle.com/datasets/jawadkhattak/us-flight-delay-from-january-2017-july-2022)	Year; Month; Carrier; Airport; Delay count due to different reasons; Delay minutes due to different reasons	2017.1 - 2022.7	21 columns * 100K rows	158 rows of missing values	NA
T_T100_MARKET_ALL_CARRIER	Bureau of Transportation Statistics (https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FMF&QO_fu146_anzr=Nv4%20Pn44vr45)	Carrier ID, Carrier name, Region, Carrier group, Class, Data source	Latest Available Data: 2022.05	6 columns * 106K rows	NA	NA
T_SCHEDULE_T3	Bureau of Transportation Statistics (https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FKG&QO_fu146_anzr=Nv4%20Pn44vr4%20f7zzn4B)	Airport type, Origin airport ID, Origin airport code, Origin airport city name, Service class, Total departures performed, Total freight, Total mail	Latest Available Data: 2022.06	8 columns * 130K rows	868 missing values in the column “Total freight” and “Total mail”	NA
T_F41SCHEDULE_B1	Bureau of Transportation	Current cash flow; asset, flight	Latest Available Data:	8 columns * 96 rows	NA	There are duplicated rows within

	Statistics (https://transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FKM&QO_fu146_anzr=Nv4%20Pn44vr4%20Sv0n0pvny)	equipment of unique carriers	2022.6			the dataset

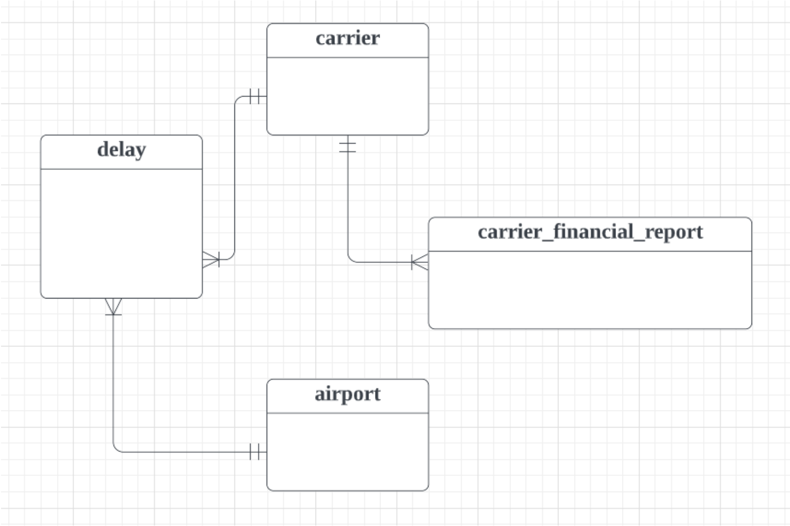
The datasets come from the ***Bureau of Transportation Statistics***, which is considered as an authoritative website that provides precise information regarding the flights and carriers. There are some missing values within the *Airline_Delay_Cause* dataset, but it only accounts for 0.1% of the total rows. In the *T_SCHEDULE_T3* dataset, the missing values are around 0.6% of the total dataset, and we have decided to remove that part as well. With that being said, we are removing the missing values given the overall high analytical maturity of the dataset. Besides, we found duplicated rows in the *T_F41SCHEDULE_B1* dataset, and we were able to remove those for future implementation.

Database Platform Considerations:

We aim to use MySQL as the main database platform for this project as it would be efficient for us to query information and data from multiple tables at once. Given that our dataset is relatively large, MySQL would be able to process large datasets in a fast and reliable manner. Besides, we would like to use Python for data cleaning purposes and Tableau to present data visualizations.

Conceptual Model:

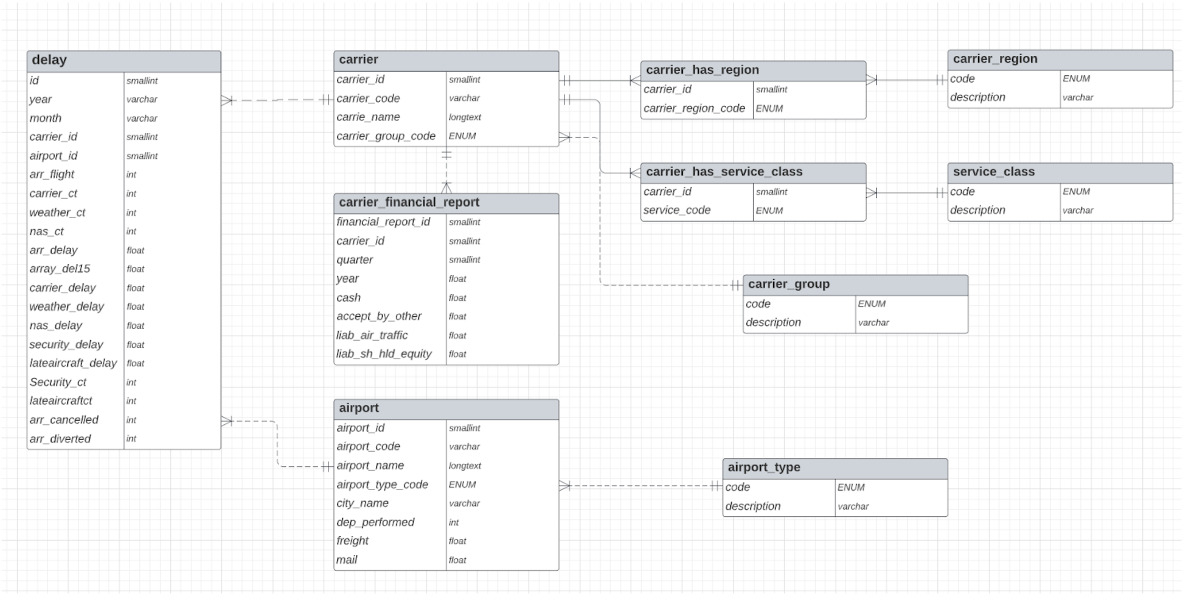
Here are the four major tables we are considering for the conceptual model.



The delay table will have information regarding the flight itself, the carrier and the departing airport, the count of delays caused by different reasons, as well as the total minutes of delays caused by different reasons calculated on a monthly basis. We will have table carrier and table airport that records all the carrier and airport information which could be linked to the delay table. Besides, we are creating a table called carrier_financial_report, having it linked to the carrier table, which records the financial data for each carrier, and we would like to see if frequent delays will actually lead to decrease in revenue because of customer dissatisfaction.

Logical Model:

Besides the 4 tables we intended to create for the conceptual model, we are bringing more tables that are linked to the carrier table and airport table, which could help explain some technical terms within carriers and airports, such as carrier group, airport type, service class, etc. In this way, we could have a better understanding of what each code corresponds to by adding descriptions.



Physical Model-Relational Models:

Normalization

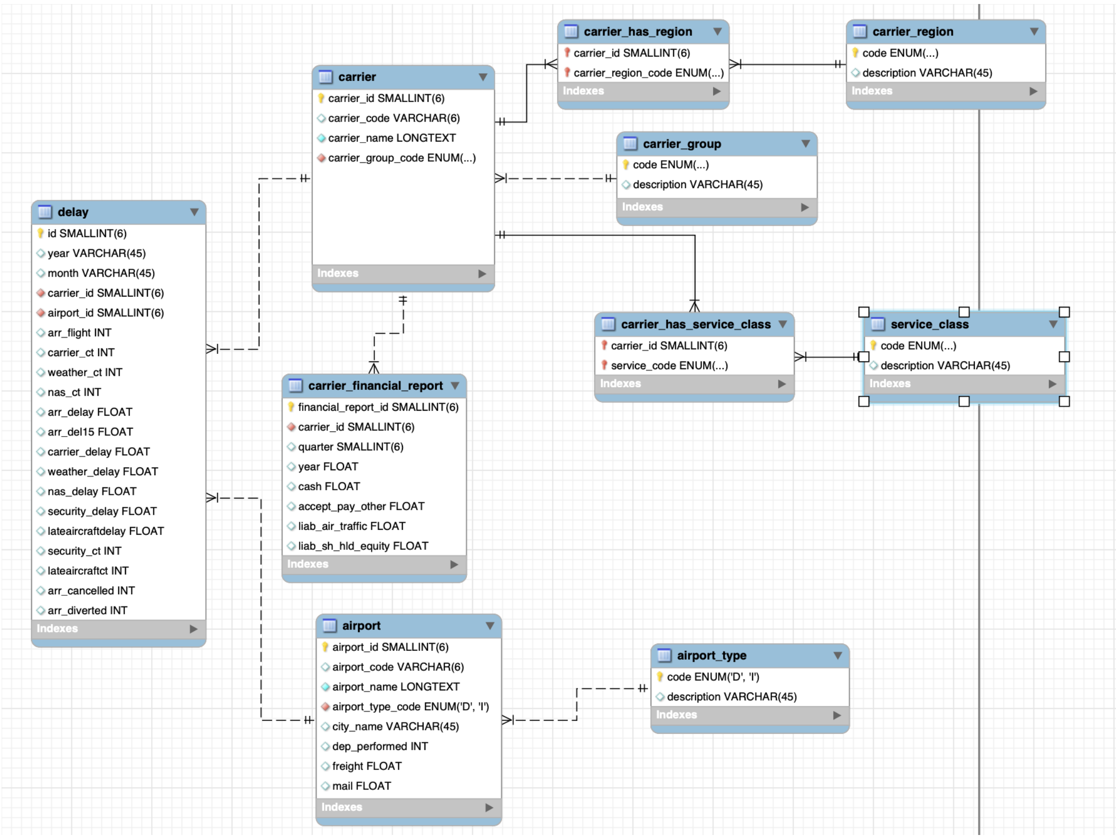
We normalized the four original tables into 10 3NF tables. The detailed information of 3NF tables are listed in Appendix.

Assumptions

Here are our assumptions for data normalization and modeling:

1. For each month, there is a unique delay id that records delay related information(counts of delays due to different reasons, minutes of delay in total due to different reasons) with respect to carriers and airports.
2. Each carrier has a unique carrier code.
3. Each airport has a unique airport code.
4. Each airport type has a unique type code.
5. Each financial report has a unique financial report id.
6. Each carrier region has a unique region code.
7. Each carrier group has a unique group code.
8. Each service class has a unique service code.
9. Each carrier can have more than one service class.
10. Multiple carriers can have the same service class.
11. Each carrier can only belong to one carrier group.
12. Multiple carriers can have the same carrier group.
13. Each carrier can belong to multiple regions.
14. Multiple carriers can have the same carrier region.
15. Each carrier can have multiple financial reports for different fiscal years.
16. Each airport can only be performed by only one airport type.
17. Multiple airports can have the same airport type.

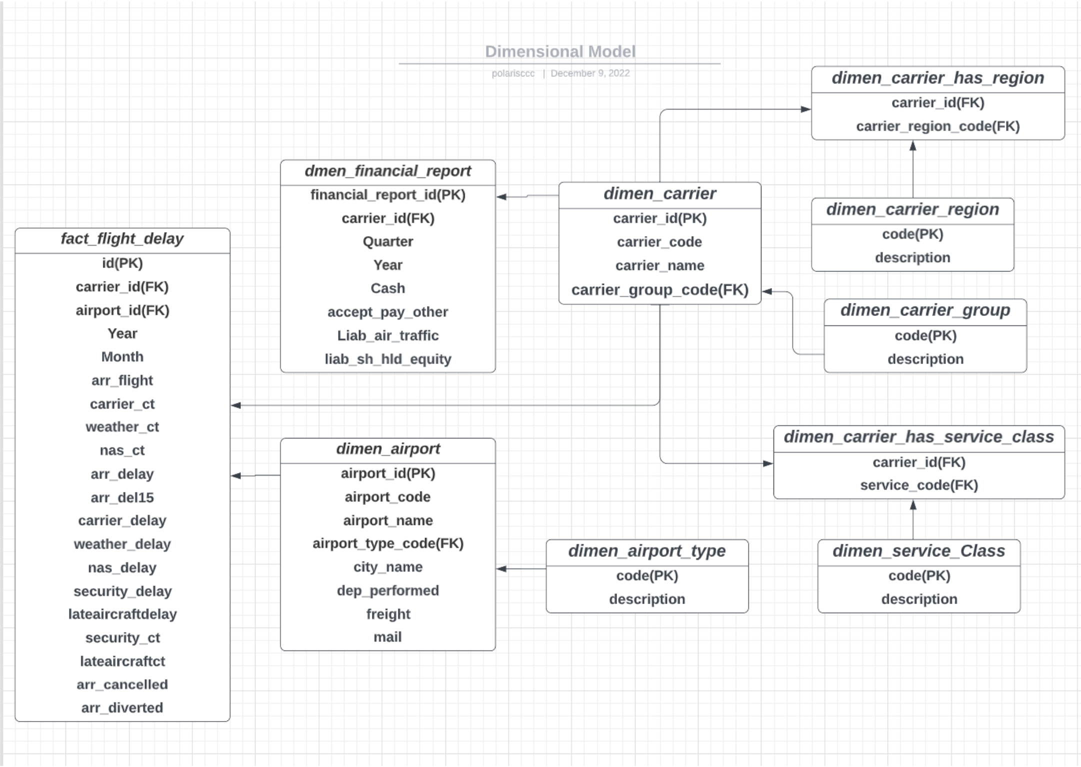
18. Each delay id can only contain one unique combination of one carrier and one airport at a certain month.
19. Each carrier and airport can have multiple delay records.



The table ‘delay’ includes delay-related information of different carriers and airports. The primary key here is a unique delay_id for each record and there are two foreign keys in this table. One is the carrier_id, which is a surrogate key we create to identify each unique carrier, connecting the delay table to the carrier table which basically includes some information about each carrier. And the other one is the airport_id, which is also a surrogate key, similarly connecting the delay table to the airport table.
We also have the carrier_financial_report table here, using a new dataset to see the financial reports of each carrier. Here one carrier may have multiple financial reports in different quarters. So it's a many to 1 relationship. And In this table, we use carrier_id as a foreign key and link it to the table carrier.

Physical Model - Dimensional Model:

In the dimensional model, the delay table is the fact table and others are all the dimensional tables.



Rationale of using relational model rather than dimensional model:
A relational database works best if we are representing data as a set of related tables. The relationships among tables allow us to link instances from one table to those in another (a flight number to its carrier & departing airport, for example) while storing the data about each of those (the flight number and the carrier) only once. That is useful if we want to change the details of one thing (carrier group) as it is stored in one place only. We don’t have to update every flight with that carrier. In this way, the tables are of a fixed structure.

Ⅳ. Data Profiling

Since our dataset is high in analytical maturity, N.A. values account for less than zero point one percent. To ensure that there are no duplicates or confusing carrier names across the delay dataset and financial statements, we conducted a cross-validation of the unique carrier identifiers and unique carrier names as shown in EDA FInal Project.ipynb.

Ⅴ. Methodology & various tools used in the process

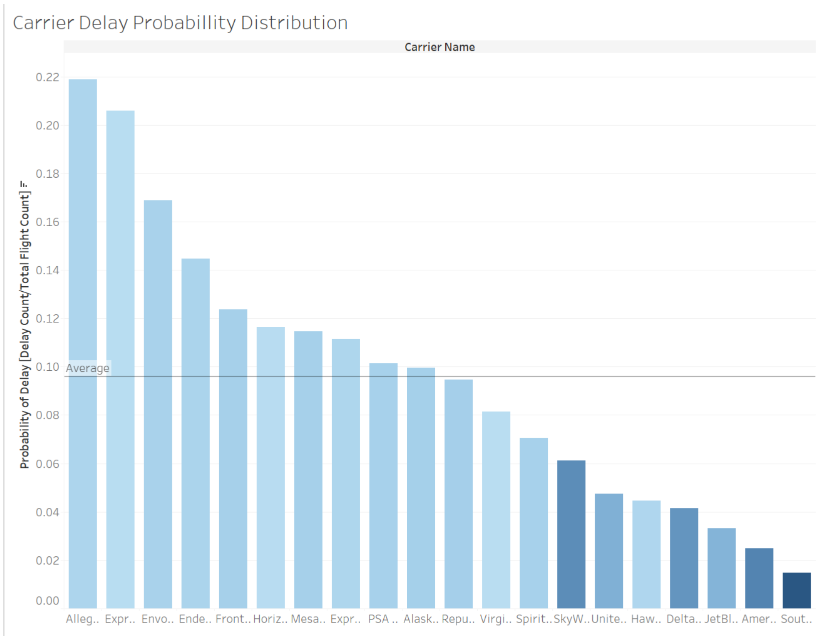
Overall speaking, we used Excel for initial data screening purposes when we were scrolling down the dataset and deciding which columns to use for the project. Moving forward, we utilized Excel to prepare the datasets, making sure that each dataset matches the corresponding table in our relational model. We also utilized Python for data cleaning purposes and did exploratory data analysis to better understand the datasets. Later on, we used MySQL to load the dataset and Tableau to create and present visualization charts.

Here are a few pros when we decided to use the tools above:

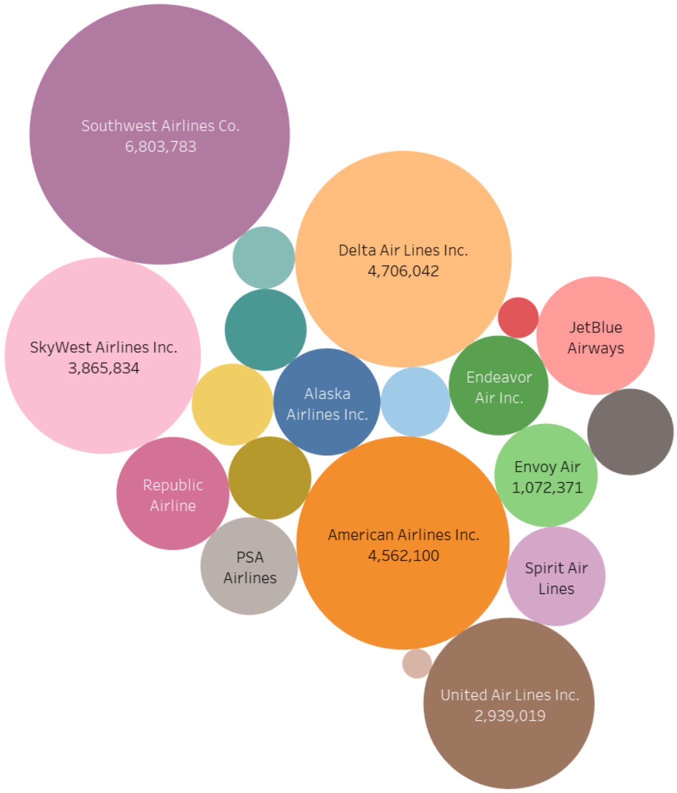
- Excel: handy to use, the easiest way to have a basic understanding of the dataset. It is also quite convenient to use Excel to view and create new worksheets that match the diagram.
- Python: could perform various calculations quickly and is able to process large datasets when doing data exploratory analysis and data cleaning.
- MySQL: also able to load large datasets within a short time, could be a good platform for future data queries.
- Tableau: we are able to present intuitive results by using visualization tools in Tableau.

VI. Insights

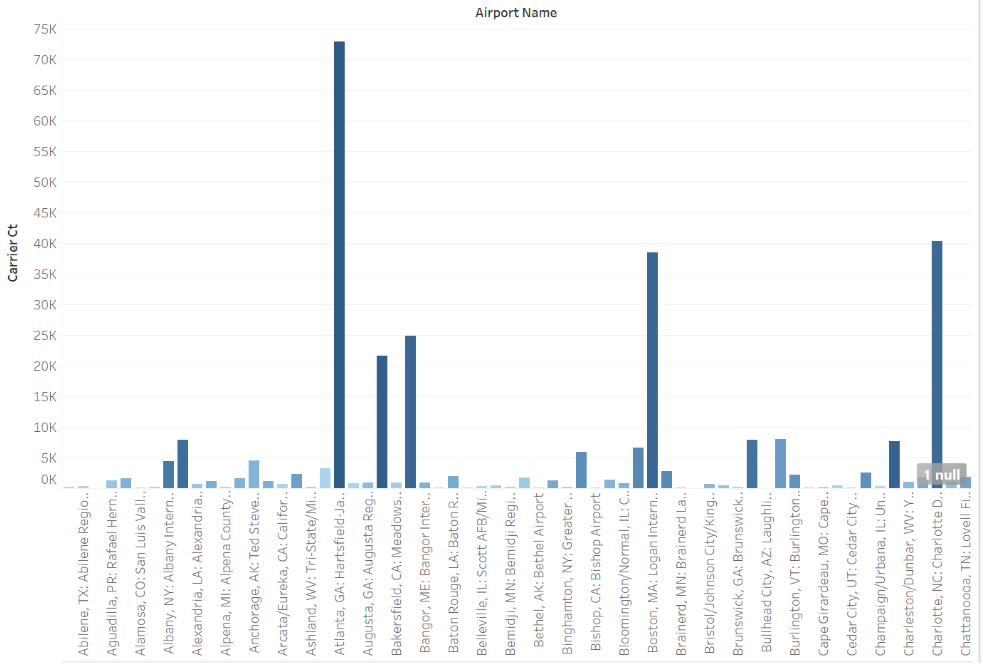
Dashboards:



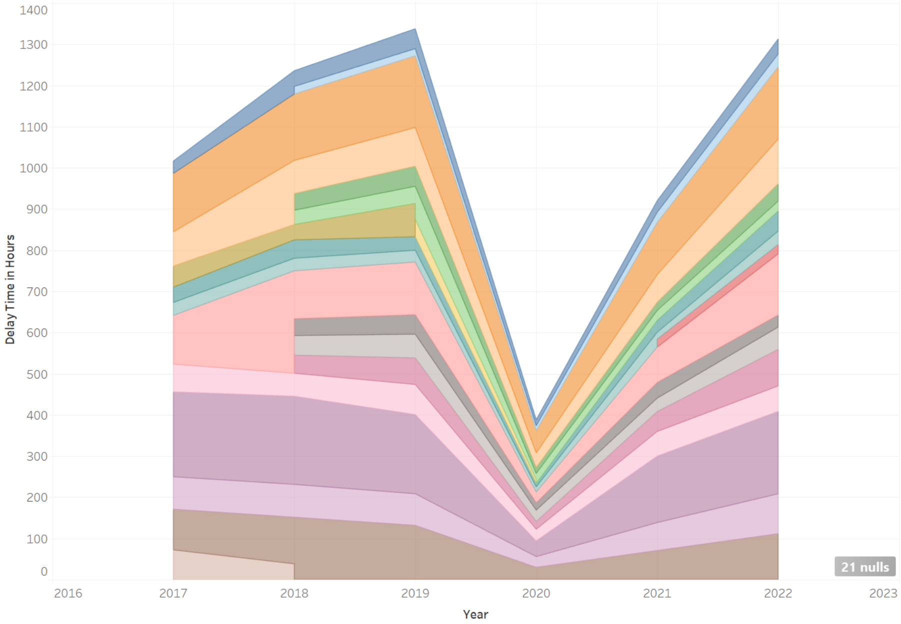
Total Flight Counts from 2017 - 2022



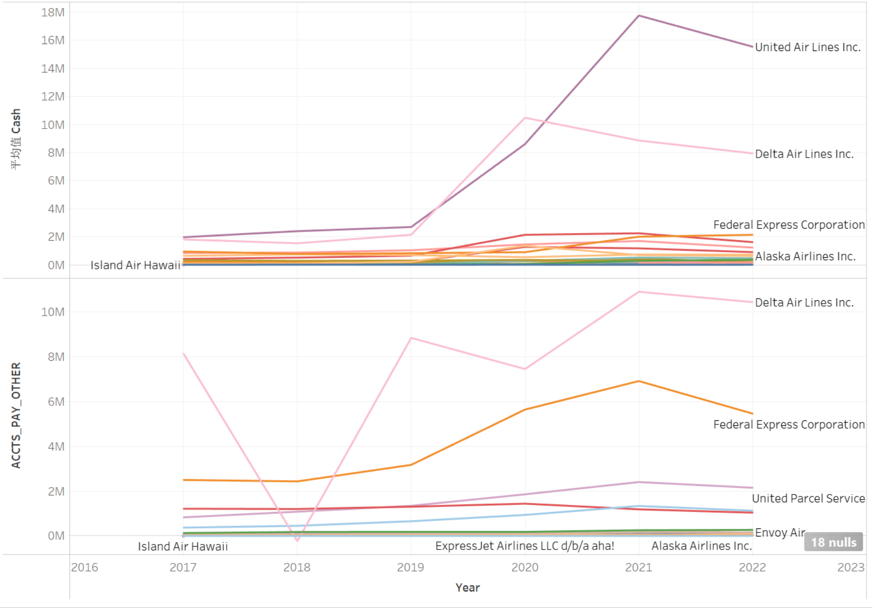
Total number of delays due to carriers at different airports



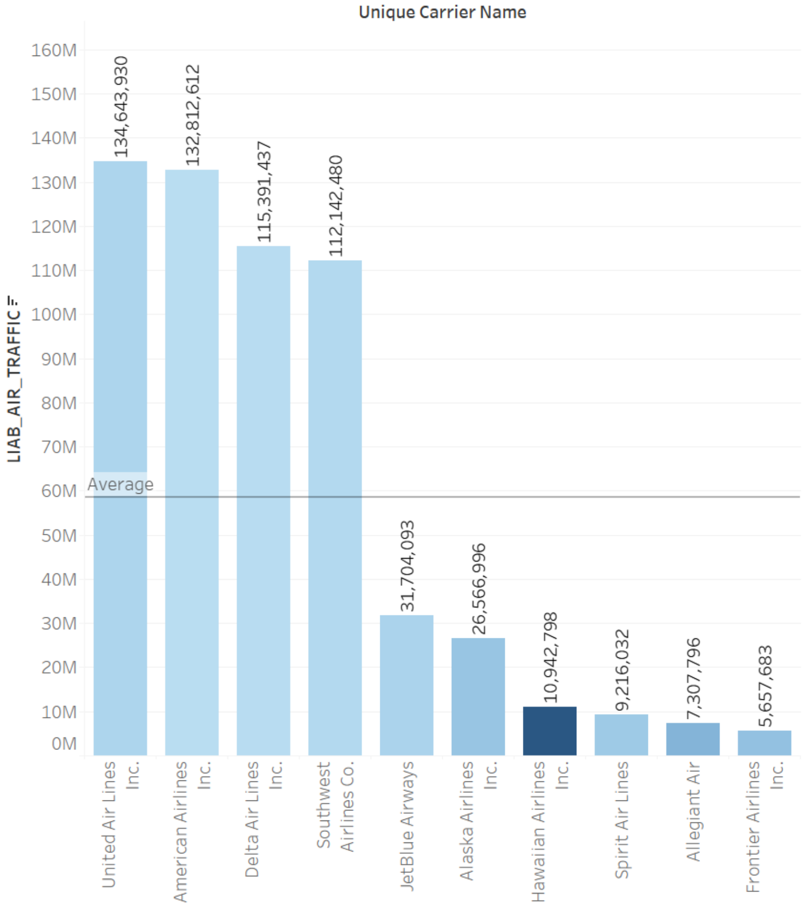
Average Delay Time per Month across Carriers



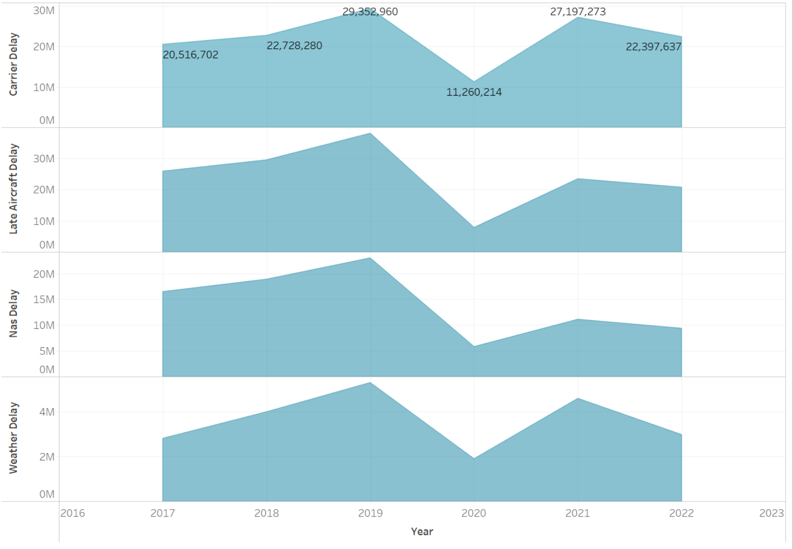
Average Cash/Account Payable across Carriers

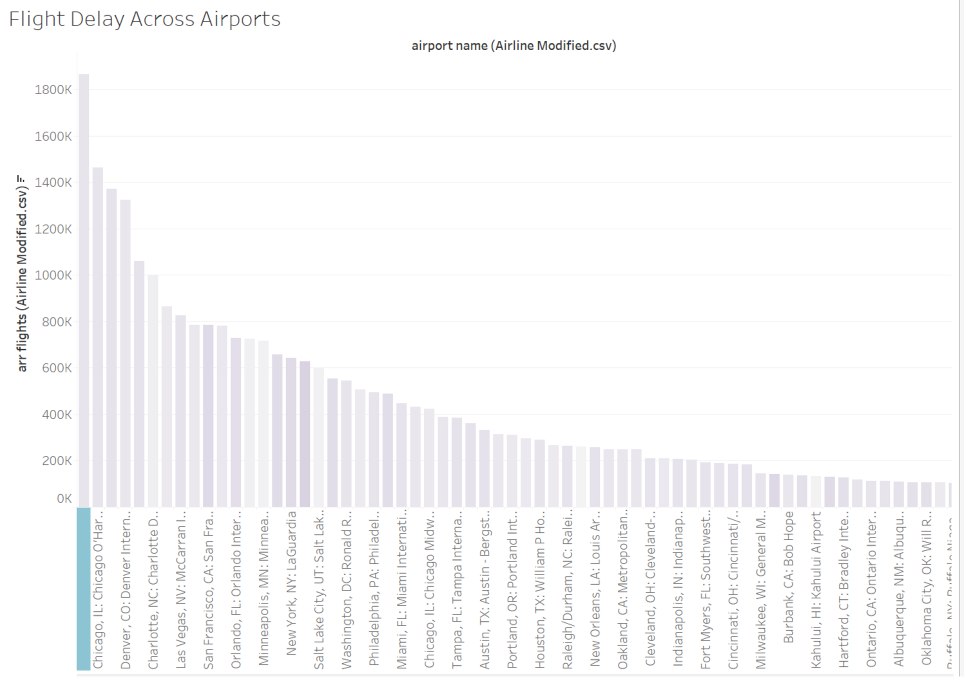


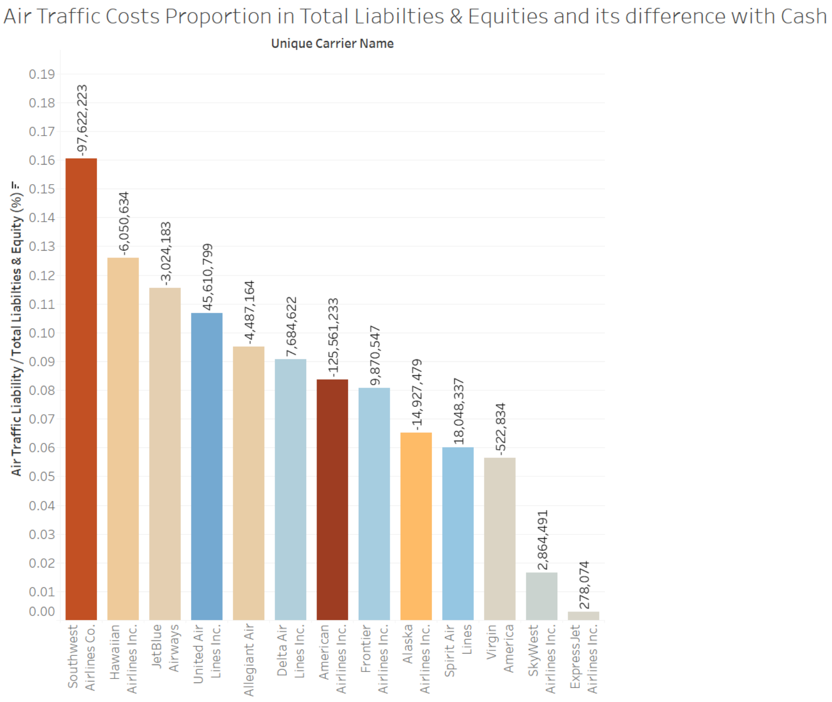
Air Traffic Operational Costs across Airlines (Above 3M)



Total delays due to different causes per year




Flight Delay Across Airports


Air Traffic Costs Proportion in Total Liabilties & Equities and its difference with Cash

Insights:

For insights, we utilized Tableau to create all visualizations. We found out that for major airline carriers like United Airlines, Southwest Airlines, American Airlines, and Delta Airlines, they all have high flight volume, low delay probabilities (number of delayed flights/total number of flights) and high cash/assets volume. On the other hand, smaller airlines like Allegiant Airline have low flight volume, high delay probabilities and low cash/assets volume. When analyzing the operational costs for each airline, we found that despite smaller airlines like Allegiant Airline having significantly higher delay probability and more delay time, their operational cost per flight is similar to that of a major airline. We suspect that this might be due to the lack of customer compensation or customer service provided by smaller airlines when having delays. Moreover, we also conduct a comparison analysis between operational costs and cash suspecting that bigger airlines would have more cash to ensure cash flow and future operations improvement. However, the comparison analysis showed the opposite. Some big airlines like Southwest Airlines and some small airlines like Allegiant Airlines all showed a deficit in cash after paying off their operational costs whereas other airlines like United Airlines and Frontier Airlines have over millions of excess cash after paying off their operational costs.

Ⅶ. Recommendations & Lesson learned

For recommendations, we would suggest customers to choose bigger airlines if they want to minimize the probability of delay and delay wait times. Overall, we believe that Southwest Airlines is the best airline to fly in.

A few lessons we have learned throughout this project:

1. Use surrogate key instead of letter key for easier data import.
2. Normalize data into 3NF before loading into tables to avoid confusing M-M relationships.
3. If there is a string variable, either use LONGTEXT() for safety, or make sure we know the maximum length so it will not exceed the capacity of VARCHAR().

Appendix

Data Description:

1. https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays
2. https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations

Tables for Data Modeling and Profiling:

Table 1: delay:

From Airline_Delay_Cause.csv:

- id (PK)
- year
- month
- carrier_id (FK)
- airport_id (FK)
- arr_flights: Number of flights arriving at airport
- arr_del15: Number of flights more than 15 minutes late
- carrier_ct: Number of flights delayed due to air carriers. (e.g. no crew)
- weather_ct: Number of flights due to weather.
- nas_ct: Number of flights delayed due to the National Aviation System (e.g. heavy air traffic).
- arr_delay:Total time (minutes) of delayed flight.
- carrier_delay: Total time (minutes) of delay due to air carrier
- weather_delay: Total time (minutes) of delay due to inclement weather.
- nas_delay: Total time (minutes) of delay due to the National Aviation System.
- security_delay: Total time (minutes) of delay as a result of a security issue.
- lateaircraftdelay: Total time (minutes) of delay flights as a result of a previous flight on the same airplane being late.
- security_ct: Number of flights canceled due to a security breach.
- lateaircraftct;Number of flights delayed as a result of another flight on the same aircraft delayed
- arr_cancelled: Number of canceled flights
- arr_diverted: Number of flights that were diverted

Table 2: carrier

- id (PK)

From Airline_Delay_Cause.csv:

- carrier (PK)
- carrier name

From T_T100_MARKET_ALL_CARRIER.csv:

- carrier_region: change to carrier_region_code (FK to table 4)
- carrier_group: change to carrier_group_code (FK to table 5)
- class: Service Class, change to sevice_class_code (FK to table 6)

- data_source: Source of Data (D = Domestic, I = International)

Table 3: airport

- airport_id (PK)

From airline_delay_cause.csv:

- airport_code
- airport_name

From T_SCHEDULE_T3.csv:

- code: Unique code for each airport
- airport_type: change to airport_type_code (FK to table 7)
- City_name
- dep_performed: (total departures performed)
- freight: Total Freight (tons)
- mail: Total Mail (tons)

Table 4: carrier_financial_report

- financial_report_id (PK)
- carrier_id (FK)
- quarter
- year

From T_F41SCHEDULE_B1.csv:

- cash: total cash
  calculate the following columns:
- ACCTS_PAY_OTHER: Total Short-Term Accounts Payables (Other)
- LIAB_AIR_TRAFFIC: Total Operating Liabilities on Air Traffic
- LIAB_SH_HLD_EQUITY: Total Liabilities and Stockholders' Equity

Table 5: carrier_has_service_class (used to connect table 2 & table 9)

- carrier_id (FK)

From T_T100_MARKET_ALL_CARRIER.csv:

- service_code: (FK)

Table 6: carrier_has_region (used to connect table 2 & table 7)

- carrier_id (FK)

From T_T100_MARKET_ALL_CARRIER.csv:

- carrier_region_code(FK)

Table 7: carrier_region:

Code	Description
A	Atlantic
D	Domestic

I	International
L	Latin America
P	Pacific
S	System

Table 8: carrier_group:

Code	Description
0	International Carriers
1	Regional Carriers (including Large, Medium, Commuter, Small Certified)
2	National Carriers
3	Major Carriers
7	Domestic Only - All Cargo Carriers

Table 9: service_class:

Code	Description
A	Scheduled First Class Passenger/ Cargo Service A
C	Scheduled Coach Passenger/ Cargo Service C
E	Scheduled Mixed First Class And Coach, Passenger/ Cargo Service E
F	Scheduled Passenger/ Cargo Service F
G	Scheduled All Cargo Service G
H	Humane Reason Unscheduled, Non-Revenue-Generating
K	Scheduled Service K (F+G)
L	Non-Scheduled Civilian Passenger/ Cargo Service L
N	Non-Scheduled Military Passenger/ Cargo Service N

P	Non-Scheduled Civilian All Cargo Service P
Q	Non-Scheduled Services (Other Than Charter) Q
R	Non-Scheduled Military All Cargo R
V	Non-Scheduled Service V (L+N+P+R) For U.S. Carrier And (L+P+Q) For Foreign Carrier
Z	All Service Z (K+V)

Table 9: airport_type

Code	Description
D	Domestic
I	International