ECE9017 Advanced Databases

Final Project Report

Project title:

2019 U.S. aviation industry analysis

Group name:

J.A.R.V.I.S.

Team member:

Haoran Ding, 251099437, hding49@uwo.ca
Jiaxin Zhao, 251105013, jzhao537@uwo.ca
Jingchao Xu, 250835833, jxu443@uwo.ca
Zhaokai Sun, 251094021, zsun323@uwo.ca

1. Database Introduction

This database is the 2019 U.S. Air Carrier Statistics (Form 41 Traffic) downloaded from the official website of the United States Department of Transportation. It includes monthly data reported by certificated U.S. air carriers on passengers, freight and mail transported. It also includes aircraft type, service class, available capacity and seats, and aircraft hours ramp-to-ramp and airborne.

The original data consists of 45 columns (corresponding to 45 attributes), with a total of 391009 rows of data. However, there is only one table in the original data, and the form of the data does not meet the specifications of the third normal form. So the original data was manually sorted out, and one table of the original data was split into six tables, each of which met the requirements of the third normal form.

The data contents in the six tables after splitting are as follows:

(1) Summary:

Record flight operation information, such as the number of seats, number of passengers, airborne time.

(2) Carrier:

Record airline information, such as airline ID, airline name.

(3) Airport:

Record the information of the airport, such as the IATA code of the airport and the name of the city.

(4) State:

Record the information of the state and the code of the state where the airport is located.

(5) Aircraft:

Record aircraft mechanical information, such as aircraft model and hardware configuration.

(6) Flightdate:

Record the flight time information, such as the quarter and month of the flight.

For the names and meanings of all the columns in the database, please see the appendix at the end of this report.

2. Project idea

The idea of this project is to analyze the operating data of airlines, especially the profitability of each route, so as to increase the profits of airlines. As a typical industry with heavy assets and weak profits, the profitability of each route is very important to airlines. And whether a route is profitable generally depends on the following three factors:

(1) Passenger transportation efficiency:

Refers to the ratio of the actual number of passengers to the total number of seats that the aircraft can provide during a flight. The larger the ratio, the more seats are sold and the higher the airline's revenue.

(2) Freight transportation efficiency:

Refers to the ratio of the actual weight of the freight transported (the sum of the weight of the cargo and the weight of the mail) to the available payload of the aircraft in one flight. The larger the ratio, the aircraft's cargo transportation revenue is higher.

(3) Flight efficiency:

Refers to the airborne time in the air and the time from the departure ramp to the landing ramp.

Since the aircraft began to taxi to the departure ramp, it no longer accepts the power supply from the airport, and uses its own onboard engine to generate electricity. Obviously, from taxiing to the departure ramp, the faster the take off, the higher the fuel efficiency of the aircraft.

3. Project Objectives

Split the original data into multiple tables that meet the third normal form specification. And based on these tables, build a new database, including the connection of the primary key and the foreign key, appropriate indexes, and so on.

In addition, according to the new database, create a data mart and build an ETL process. Realize the above project idea in the fact table.

Try to use SSAS to build a multi-dimensional cube.

Finally, combined with the calculated three operational efficiency data in the fact table, try to use the machine learning algorithm to predict the future passenger demand and freight demand of the airline, thereby helping the airline to allocate aircraft and crew members more reasonably.

4. Project deliverables

This project is expected to have the following deliverables:

(1) A database conforming to the third normal form.

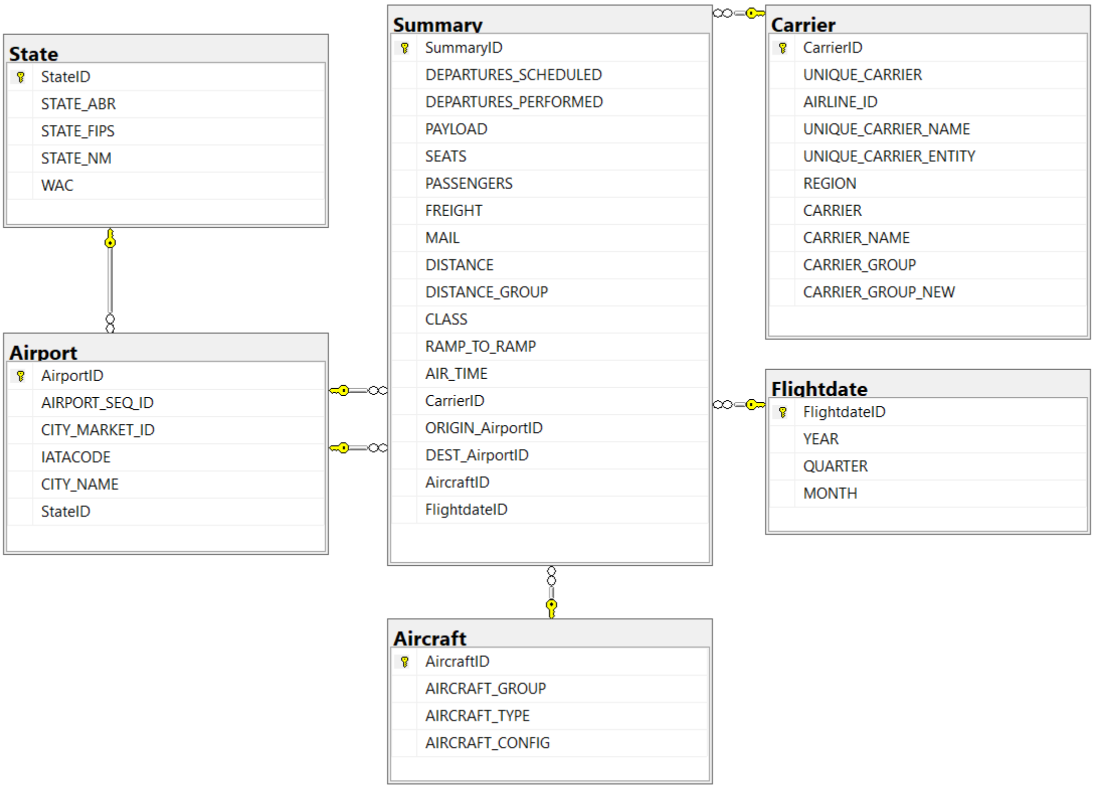
(2) Data marts and ETL processes implemented using stored procedures.

(3) The ETL process implemented using SSIS packages.

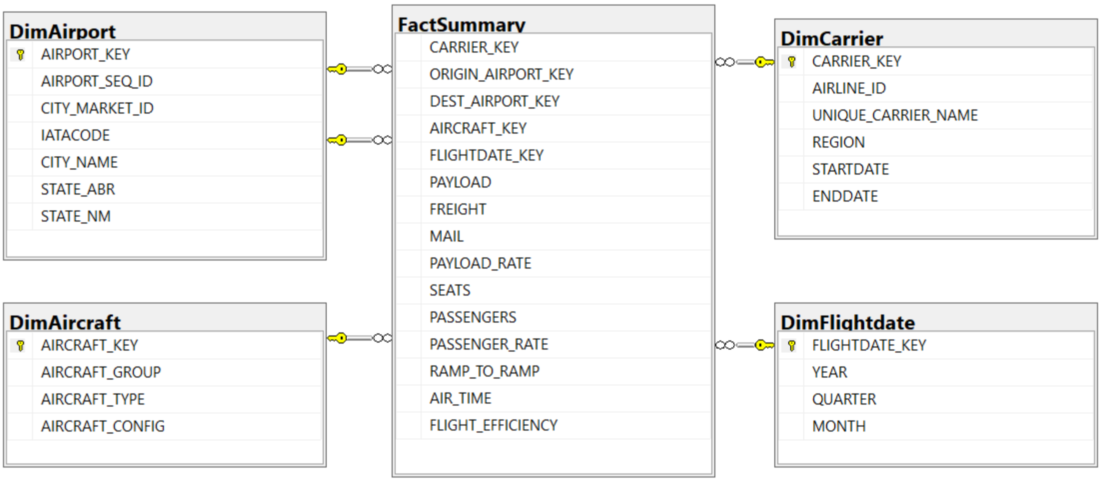
(4) SSAS multi-dimensional cube.

(5) The prediction results (accuracy) of machine learning algorithms.

5. Database ER diagram

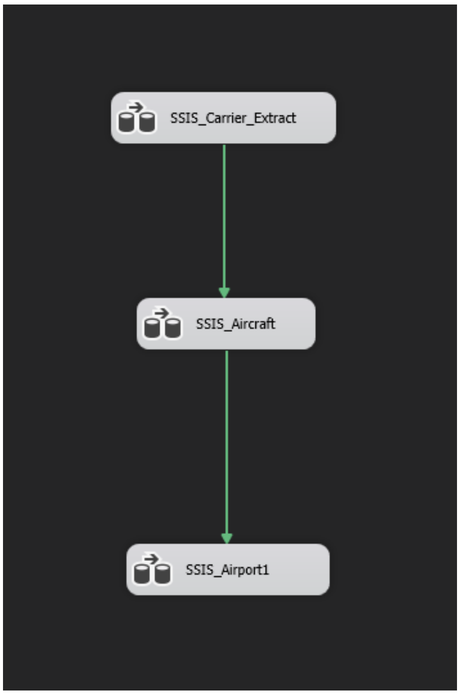


6. Data mart schema diagram

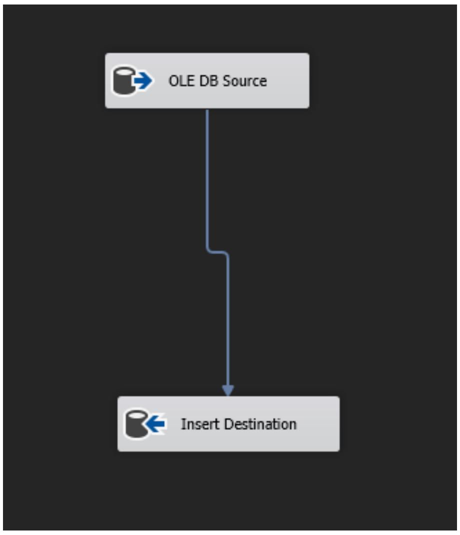


7. SSIS diagram

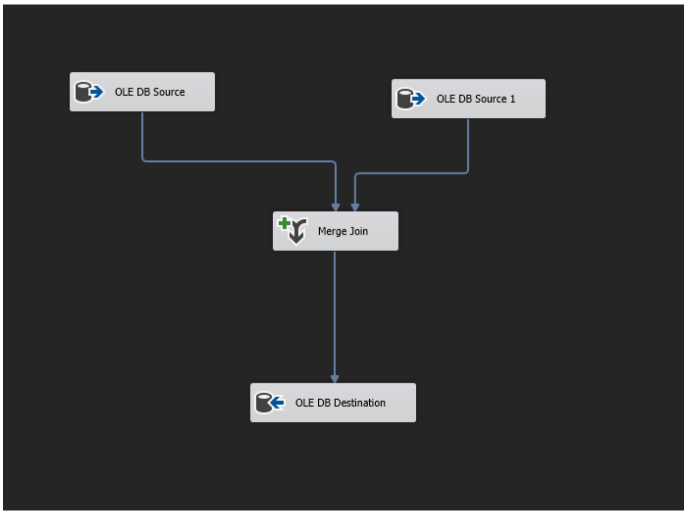
(1) Control flow



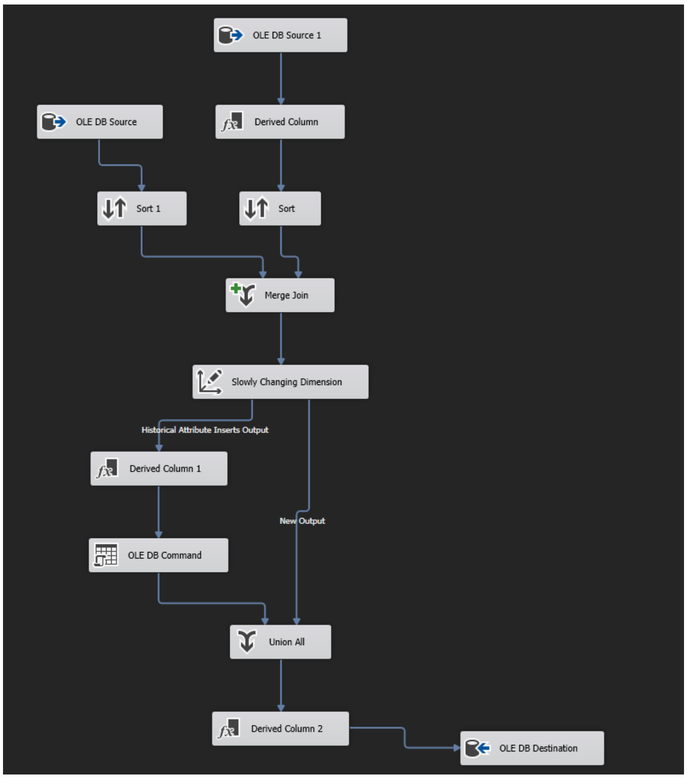
(2) SSIS_Aircraft



(3) SSIS_Airport

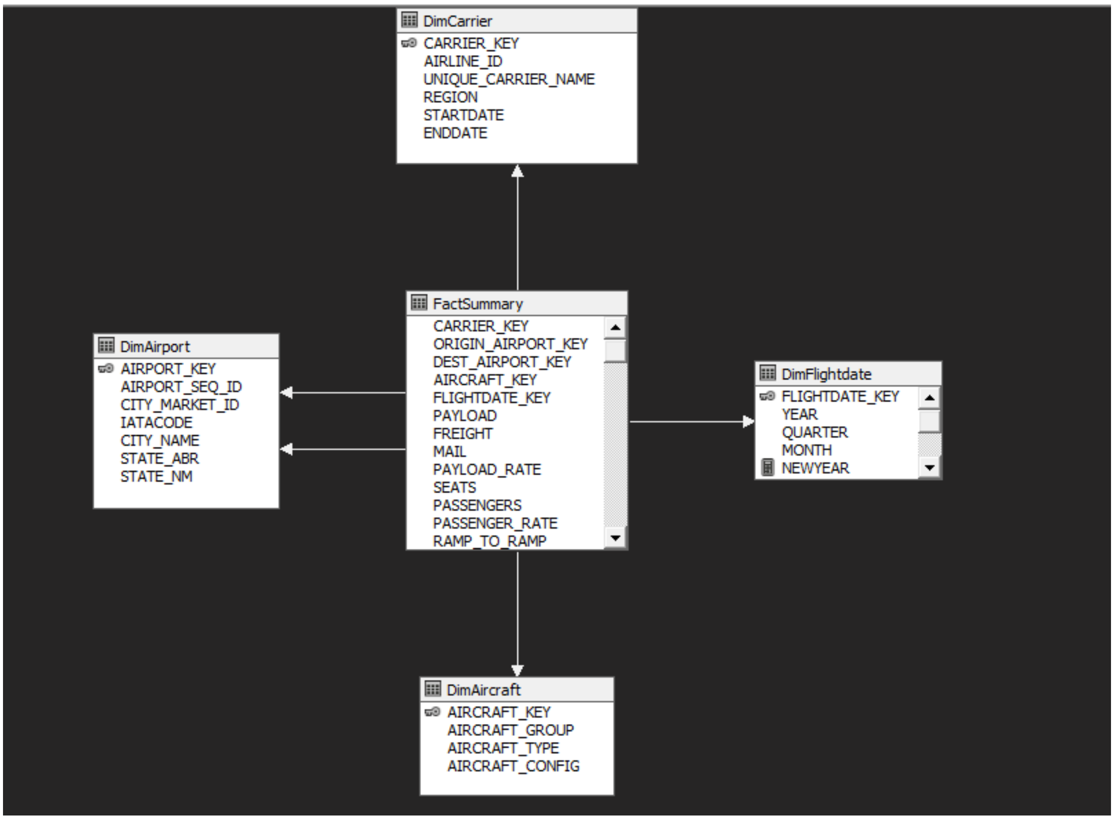


(4) SSIS_Carrier



8. SSAS diagram

(1) Data source views



DimCarrier
CARRIER_KEY
AIRLINE_ID
UNIQUE_CARRIER_NAME
REGION
STARTDATE
ENDDATE

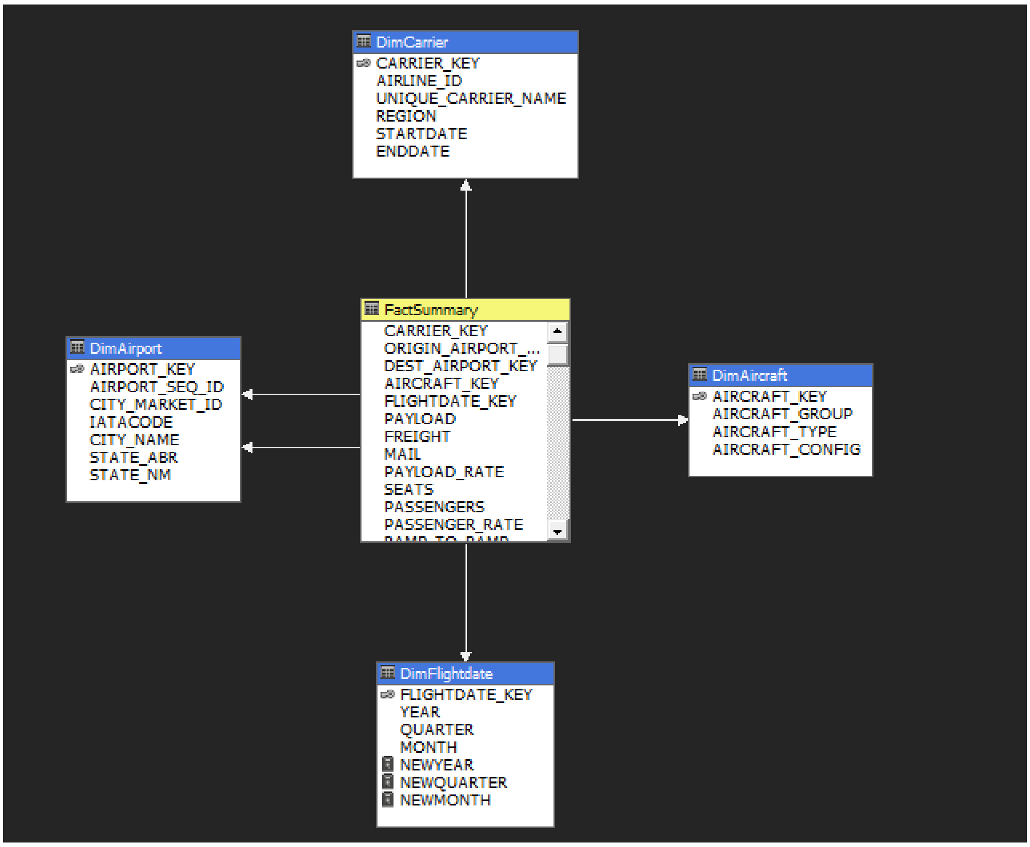
FactSummary
CARRIER_KEY
ORIGIN_AIRPORT_KEY
DEST_AIRPORT_KEY
AIRCRAFT_KEY
FLIGHTDATE_KEY
PAYLOAD
FREIGHT
MAIL
PAYLOAD_RATE
SEATS
PASSENGERS
PASSENGER_RATE
RAMP_TO_RAMP

DimAirport
AIRPORT_KEY
AIRPORT_SEQ_ID
CITY_MARKET_ID
IATACODE
CITY_NAME
STATE_ABR
STATE_NM

DimFlightdate
FLIGHTDATE_KEY
YEAR
QUARTER
MONTH
NEWYEAR

DimAircraft
AIRCRAFT_KEY
AIRCRAFT_GROUP
AIRCRAFT_TYPE
AIRCRAFT_CONFIG

(2) Cube



DimCarrier
CARRIER_KEY
AIRLINE_ID
UNIQUE_CARRIER_NAME
REGION
STARTDATE
ENDDATE

FactSummary
CARRIER_KEY
ORIGIN_AIRPORT_...
DEST_AIRPORT_KEY
AIRCRAFT_KEY
FLIGHTDATE_KEY
PAYLOAD
FREIGHT
MAIL
PAYLOAD_RATE
SEATS
PASSENGERS
PASSENGER_RATE
RAMP_TO_RAMP

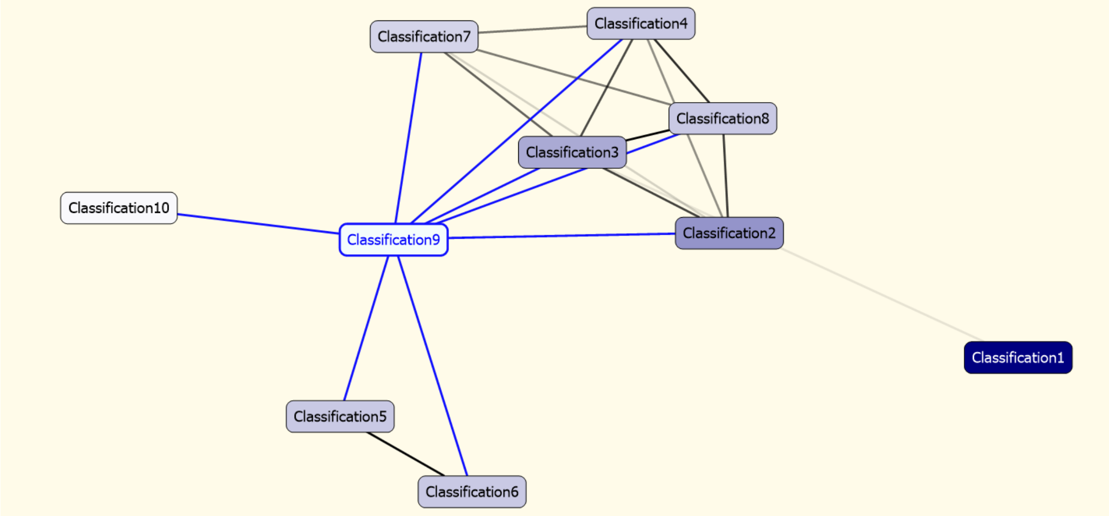
DimAirport
AIRPORT_KEY
AIRPORT_SEQ_ID
CITY_MARKET_ID
IATACODE
CITY_NAME
STATE_ABR
STATE_NM

DimAircraft
AIRCRAFT_KEY
AIRCRAFT_GROUP
AIRCRAFT_TYPE
AIRCRAFT_CONFIG

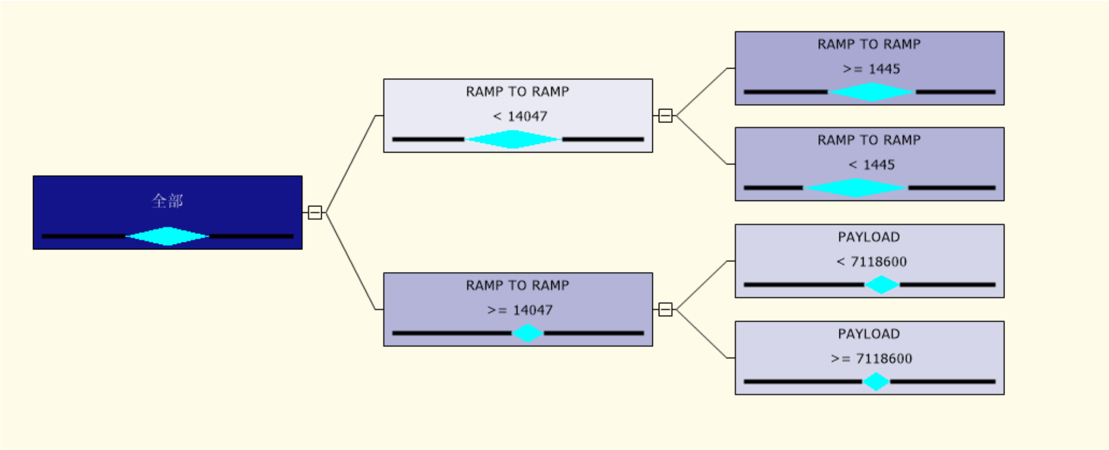
DimFlightdate
FLIGHTDATE_KEY
YEAR
QUARTER
MONTH
NEWYEAR
NEWQUARTER
NEWMONTH

9. Data Mining (bonus part)

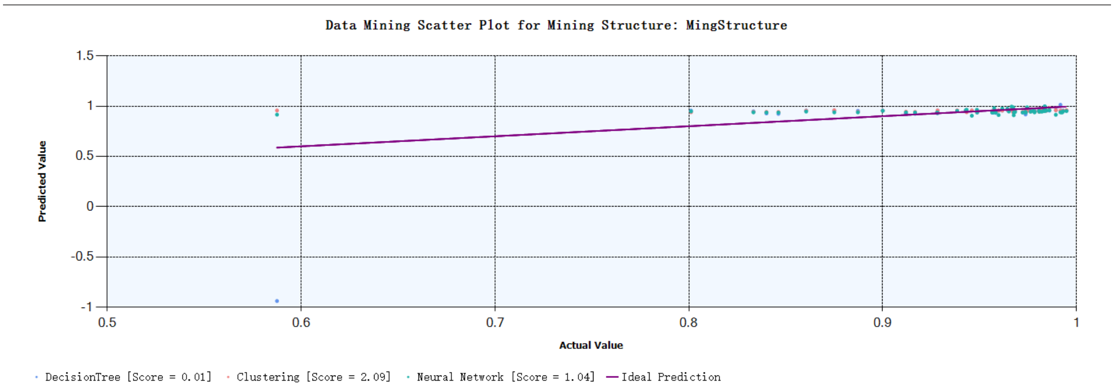
(1) Clustering



(2) Decision tree



(3) Prediction



Appendix

	Field Name	Description
Summary	SummaryID	PK
	DEPARTURES_SCHEDULED	Departures Scheduled
	DEPARTURES_PERFORMED	Departures Performed
	PAYLOAD	Available Payload (pounds)
	SEATS	Available Seats
	PASSENGERS	Non-Stop Segment Passengers Transported
	FREIGHT	Non-Stop Segment Freight Transported (pounds)
	MAIL	Non-Stop Segment Mail Transported (pounds)
	DISTANCE	Distance between airports (miles)
	DISTANCE_GROUP	Distance Intervals, every 500 Miles, for Flight Segment
	CLASS	Service Class
	RAMP_TO_RAMP	Ramp to Ramp Time (minutes)
	AIR_TIME	Airborne Time (minutes)
	CarrierID	FK
	ORIGIN_AirportID	FK
	DEST_AirportID	FK
	AircraftID	FK
	FlightdateID	FK
Carrier	CarrierID	PK
	UNIQUE_CARRIER	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
	AIRLINE_ID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
	UNIQUE_CARRIER_NAME	Unique Carrier Name. When the same name has been used by multiple carriers, a numeric suffix is used for earlier users, for example, Air Caribbean, Air Caribbean (1).
	UNIQUE_CARRIER_ENTITY	Unique Entity for a Carrier's Operation Region.
	REGION	Carrier's Operation Region. Carriers Report Data by Operation Region
	CARRIER	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
	CARRIER_NAME	Carrier Name
	CARRIER_GROUP	Carrier Group Code. Used in Legacy Analysis
	CARRIER_GROUP_NEW	Carrier Group New
Airport	AirportID	PK

	AIRPORT_SEQ_ID	Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
	CITY_MARKET_ID	City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
	IATACODE	Airport IATA code
	CITY_NAME	Airport cirt name
	StateID	FK
State	StateID	PK
	STATE_ABR	State Code
	STATE_FIPS	State FIPS (U.S. Federal Information Processing Standard Codes)
	STATE_NM	Airport state name
	WAC	Airport, World Area Code
Aircraft	AircraftID	PK
	AIRCRAFT_GROUP	Aircraft Group
	AIRCRAFT_TYPE	Aircraft Type
	AIRCRAFT_CONFIG	Aircraft Configuration
Flightdate	FlightdateID	PK
	YEAR	Flight year
	QUARTER	Flight quarter
	MONTH	Flight month