Data Exploration Insight

In this project, we will use 4 dataset: flights, airports, planes, carriers. Flights will be the main dataset and it will refer the data in other three tables

Airports

Airport dataset has 3000+ records. Column ‘iata’ is the single user key column, which will be used to uniquely identify an airport. In our exploration, we found:

1. All ‘Iata’ column values are unique
2. No column has any null value

Overall, the ‘Airport’ data is good and doesn’t need much cleaning.

Planes

Plane dataset has 5029 records.. Column ‘tailnum’ is the user key column. In our exploration, we also found:

1. ‘‘tailnum’’ column values are unique
2. 549 records have a lot of null value and are bad records. They need to be removed.
3. We also calculate value_count for some of the columns

Overall, the plane data is in good shape but we will need to remove some bad/empty records.

Carriers

Carrier dataset has 1491 records and only two columns: 'Code', 'Description’. Column ‘Code’ is the user key column. We found:

1. ‘Code’’ column values are unique
2. None of the columns has a null value

Overall, the Carrier data is in good and doesn’t need much cleansing

Flights

We are looking at the flights of year-of-2008 data, It has 2389217 records and 29 columns. We mainly checked null values and referential integrity

1. Null values

For data quality, we checked for each column, the number of records with a null value.

```
Year:0
Month:0
DayofMonth:0
DayOfWeek:0
DepTime:64442
CRSDepTime:0
ArrTime:70096
CRSArrTime:0
UniqueCarrier:0
FlightNum:0
TailNum:42452
ActualElapsedTime:70096
CRSElapsedTime:407
AirTime:70096
ArrDelay:70096
DepDelay:64442
Origin:0
Dest:0
Distance:0
TaxiIn:70096
TaxiOut:64442
Cancelled:0
CancellationCode:2324775
Diverted:0
CarrierDelay:1804634
WeatherDelay:1804634
NASDelay:1804634
SecurityDelay:1804634
LateAircraftDelay:1804634
Elapsed:120.030267
```

We can see the delayed columns have a lot of null values. This makes sense as many flights didn't have any delay

We also see for 4 column, Year, Month, DayofMonth and FlightNum, they don't have null values. As these 4 columns can be combined used to uniquely identify a flight, so we can see none of the user key values is t empty, which is really good.

2. Referential integrity check

As flights data will refer data from other dataset, we also check the dependency between the datasets. The below table shows the number of records of flight dataset that don't not match their referred tables. The results are pretty good.

Parent DataSet	Result
Airports	For 'Origin' and 'Dest' of Flights, all the values are matching "iata" column values from Airports
Planes	For 'TailNum', there are 84499 (3.6%) records that are not matching 'tailnum' column of Planes dataset. 151 distinct values.
Carrier	For 'UniqueCarrier' column, all the values are matching "Code" column values of carrier data set.

Reference 1: columns and schema of each dataset

Flight	root |-- Year: integer (nullable = true) |-- Month: integer (nullable = true) |-- DayofMonth: integer (nullable = true) |-- DayOfWeek: integer (nullable = true) |-- DepTime: integer (nullable = true) |-- CRSDepTime: integer (nullable = true) |-- ArrTime: integer (nullable = true) |-- CRSArrTime: integer (nullable = true) |-- UniqueCarrier: string (nullable = true) |-- FlightNum: string (nullable = true) |-- TailNum: string (nullable = true) |-- ActualElapsedTime: integer (nullable = true) |-- CRSElapsedTime: integer (nullable = true) |-- AirTime: integer (nullable = true) |-- ArrDelay: integer (nullable = true) |-- DepDelay: integer (nullable = true) |-- Origin: string (nullable = true) |-- Dest: string (nullable = true) |-- Distance: integer (nullable = true) |-- TaxiIn: integer (nullable = true) |-- TaxiOut: integer (nullable = true) |-- Cancelled: integer (nullable = true) |-- CancellationCode: string (nullable = true) |-- Diverted: string (nullable = true) |-- CarrierDelay: integer (nullable = true) |-- WeatherDelay: integer (nullable = true) |-- NASDelay: integer (nullable = true) |-- SecurityDelay: integer (nullable = true)

	|-- LateAircraftDelay: integer (nullable = true)
Carrier	root |-- Code: string (nullable = true) |-- Description: string (nullable = true)
Plane	root |-- tailnum: string (nullable = true) |-- type: string (nullable = true) |-- manufacturer: string (nullable = true) |-- issue_date: string (nullable = true) |-- model: string (nullable = true) |-- status: string (nullable = true) |-- aircraft_type: string (nullable = true) |-- engine_type: string (nullable = true) |-- year: string (nullable = true) |-- ROW_ID: long (nullable = false)
Airport	root |-- iata: string (nullable = true) |-- airport: string (nullable = true) |-- city: string (nullable = true) |-- state: string (nullable = true) |-- country: string (nullable = true) |-- lat: double (nullable = true) |-- long: double (nullable = true)