



# MIS41040 Business Decision Support System

## Tableau Report

Team 33

Bo-Tsun Chen (21205831)

Cheuk Nam Cyann Ng (21205368)

Chen Chen (21202636)

April 2022



UCD Michael Smurfit  
Graduate Business School

## Table of Content

<b>Introduction</b>	<b>3</b>
<b>Analysis of the Problem for Decision-Making</b>	<b>4</b>
Recommendation for passengers	4
Potential Solution for Airport Management	5
Information & Competitor Analysis for Airlines	6
<b>Data Cleaning</b>	<b>7</b>
<b>User Manual</b>	<b>12</b>
Customer	12
Airport Management	14
Airlines	16
<b>Appendix I</b>	<b>18</b>

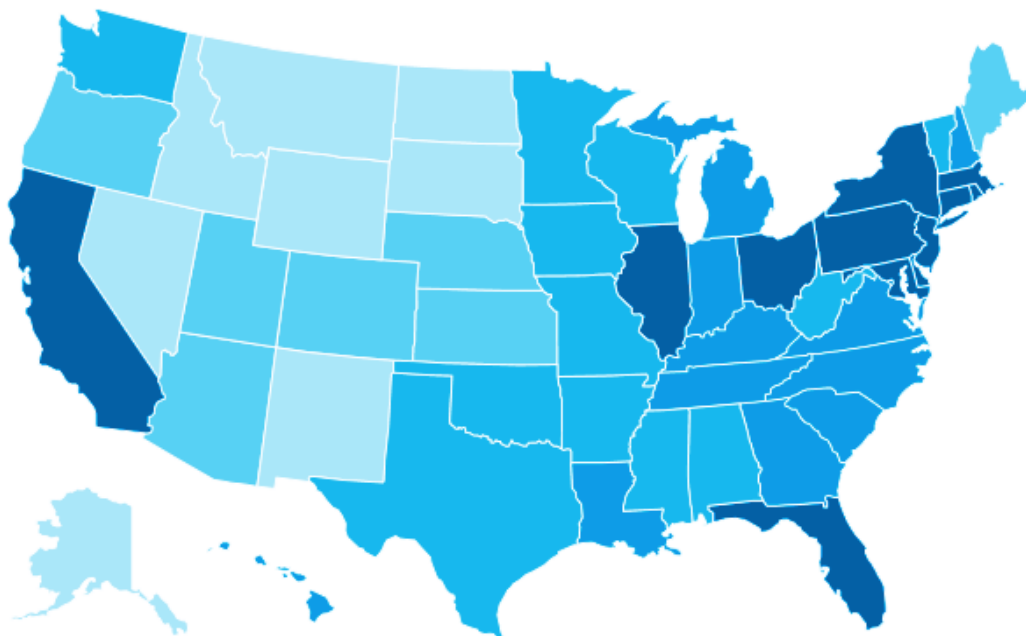
## Introduction

Aviation transportation plays an important role in the global economy. In order to narrow down the influence of flight delay in the aviation industry, the following three aspects were analysed: customer, airport and airline. And a decision support system is built to help them make wise decisions and improve their performances. The objective is to build a symbiotic system for passengers, airport management and airlines. All of them can benefit each if most of the controllable risk factors can be eliminated.

### Flight Delay in the USA 2015

The data source is from Kaggle (<https://www.kaggle.com/datasets/usdot/flight-delays>).

Delays can be classified into five categories from the explanation of the [Federal Aviation Administration](#) page: air carrier, extreme weather, National Aviation System (NAS), late-arriving aircraft, and security.



## Analysis of the Problem for Decision-Making

### Recommendation for passengers

#### Identification of the problem

When it comes to delay, frustration and desperation will overwhelm the joyful mood of passengers. Delays often do harm to one's schedule for both work or leisure. Hence, the following issue needs to be identified for a better transportation resolution.

#### Recommendations for passengers

1. Find the airport with the least delayed time in departure/arrival city.
2. Identify the airline with least departure delayed time (controllable factor).
3. Based on the seasonality, find the best period of time for travel.
4. Decide what time to depart on the designated day, avoiding the popular (busy) time.
5. Inform the potential delayed time, based on the probability of flight delayed and security delay in departure airport.

#### Descriptive analytics for passengers

Thanks to the historical data, the dashboard can inform the passenger, the airport and the airline with least delayed time of the certain route between two cities, and the time passenger should spare to avoid the queue and delay in the airport.

With the awareness of delay, passengers can find a better airport to depart or prepare enough time for a carefree travel.

## Potential Solution for Airport Management

### Identification of the problem

Delay strongly endangers the efficiency of the airports and passenger traffic. The problem of delays can be divided into two parts based on the definition of FAA, NAS reason by Air Traffic Control (ATC) system; and security reason by long queue in custom inspection systems, security breach and evacuation of a terminal.

The tableau dashboard aims to provide in-depth information to specify critical issues which can be improved by airport management. In terms of delayed flight, the parameters are assessed by the following four indicators:

1. Departure city/airport and arrival city/airport
2. Route(s)
3. Seasonality (month/day)
4. Different factors of delays.

### Empirical evidence for improvement

The parameters visualise the pattern with measurable attributes for each airport. The management of the airport is able to pin down the most severe drawback compared to the overall performance and start to execute remedial measures. Moreover, seasonality generates the concern of the demand for scalability and flexibility. Airports can have hundreds of flights in a single as well as a few, both tangible and intangible resources can be allocated efficiently.

## Information & Competitor Analysis for Airlines

Identify the competency amongst all airlines

In terms of a competitive airline, outperforming other airlines is the priority. In order to achieve the target, the management roles need to identify the following metrics meet the criteria:

1. Keep number of delayed flights as few as possible
2. Keep average flight delay time as low as possible.
3. analyse which airports have the largest number of flights delay and longest average flights delay time
  - a. type of delay (departure/arrival)
  - b. Major factors leading to the delay.
4. Seasonality of delay by month.
5. Pattern of delay by week/day.
6. Minimise the controllable factors of delay (air carrier delay, late aircraft delay).

Provide insightful recommendations for airlines

Airlines can develop their strategy based on the recommendations with measurable benchmarks amongst all airlines. The analysis provides unique information for selected airlines according to the route, city or airport. Airlines can know their position within different routes, airports and month/day/time.

## Data Cleaning

Code is available in appendix I.

### *Introduction*

To prepare and clean the dataset in the following, Python is used to support the step of data cleaning, two of the libraries, *pandas* and *numpy*, were imported to assist data cleaning in Python.

In the beginning of the data cleaning phase, four files related to the project were used.

1. Flights.csv
2. Airlines.csv
3. Airports.csv
4. Mapping\_Data\_dictionary.xlsx

In an attempt to make the access of dataset convenient, the related tables would be aggregated into one dataset. With the help of a single dataset, the connection between tableau and database.

We have done seven steps in data cleaning:

1. Drop unrelated columns (minor details)
2. Find and drop duplicates
3. Detect & deal with outliers and missing values
4. Transform Nan data into 0
5. Map values
6. Connect tables
7. Rename columns.

The details will be explained in the next section.

### *Seven phases of data cleaning in detail*

The details are explained in the following:

1. Drop columns: A few of the columns are not necessary in building the decision support system. For example, 'TAIL\_NUMBER', 'TAXI\_OUT', 'WHEELS\_OFF', 'SCHEDULED\_TIME', 'ELAPSED\_TIME', 'WHEELS\_ON', 'TAXI\_IN', 'DIVERTED' in "flights.csv" file, so the column above were dropped.. In addition, two columns in the final joined table, 'COUNTRY', 'IATA\_CODE', were dropped as well.
2. Duplicates: There are no duplicates, thus we did not work on this phase.
3. Outliers: Several outliers were detected by drawing the boxplots. With the calculation of IQR, the outliers were detected as the following table.

The statistical range of the delay time (departure/arrival) duration is extremely large, roughly 2000 minutes. Yet the pattern of the distribution makes the outliers less obvious. Based on the real-life scenario, some of the flight delays could result from uncontrollable events or weather, we should not exclude the type of data.

4. Missing values: The "nan" values were replaced by 0 in the following columns:

- i. 'AIR\_SYSTEM\_DELAY'
- ii. 'SECURITY\_DELAY'
- iii. 'AIRLINE\_DELAY'
- iv. 'LATE\_AIRCRAFT\_DELAY'
- v. 'WEATHER\_DELAY'

Aside from the five types of delays above, the missing values amongst other category are in appendix I:

490775 rows of missing values were identified in column 'LONGITUDE\_destination'. In order to draw the correct map, we decided to drop these rows with missing values. This is because roughly 5.8 million records in the dataset, and 500,000 records seem to be relatively small factors. Meanwhile, most of the flights were not cancelled (5,729,195). Similarly, the cancelled flights were executed by the same approach. For columns 'DEPARTURE\_TIME', 'ARRIVAL\_DELAY' and 'LATITUDE\_origin', only a few rows of missing values were found, the rows of these rare values were dropped.



5. Transform data types: The two columns from “flights.csv”, which were 'ORIGIN\_AIRPORT' & 'DESTINATION\_AIRPORT', including mixed data types. Thus, the data type were transformed into string value.

The column 'ORIGIN\_AIRPORT', from "Mapping\_Data\_Dictionary.xlsx", was transformed to string in order to match with the column in “flights.csv”; while the values in the following four columns from “flights.csv”, 'SCHEDULED\_DEPARTURE', 'DEPARTURE\_TIME', 'SCHEDULED\_ARRIVAL', 'ARRIVAL\_TIME', were converted to datetime type.

6. Create connection amongst csv files: The “flights.csv” file was left joined with the “airlines.csv” file on 'AIRLINE' and 'IATA\_CODE'. The joint table then was left joined with the “airports.csv” file twice on 'ORIGIN\_AIRPORT' & 'IATA\_CODE' and 'DESTINATION\_AIRPORT' & 'IATA\_CODE' respectively in order to get the coordinates of both the origin airports and the destination airports.
7. Mapping: The origin airport id was mapped into IATA code for “flights.csv” by using the information from "Mapping\_Data\_Dictionary.xlsx". The series of numbers 1 to 7, were mapped to Sunday-Saturday, in the field 'DAY\_OF\_WEEK' from “flights.csv”, with the approach of mapping, the data can be told simply.
8. Rename columns: Columns or fields were renamed for operating purposes, to keep everything in order and make everyone easy to understand. For example, the field name “airline long” stands for the full name of airlines.

### *Summary*

Data cleaning could be time-consuming and effort-taking, but it's essential for data analytics and building up a better understanding of data . The knowledge can only be accessible if we realise the veracity, variety and value in it.

# User Manual

## Customer

### Introduction

The customer user manual aims to provide helpful information for passengers to decide the best solution for their flying experience. The intuitive dashboard has the following measurement for the potential delay of each route; passengers can simply find out the best way to travel in terms of an estimated delay.

The dashboard tells:

1. Available airports based on your departure/arrival city
2. Recommended time to fly during each day
3. Recommended day to fly during selected duration
4. Recommended airline to take (least delay time)
5. Estimated security delay passengers should spare in case of queue
6. All routes between (selected/available) airports
7. The seasonality of average minutes delayed of departure/arrival

### How to access the dashboard

Passengers are always concerned about the most efficient way to travel. The customer dashboard allows passengers to enter their preference step by step:

1. Enter the name of the city or the airport the passenger would like to take off/land.
  - a. Can simply select one airport and the dashboard can provide viable options.
  - b. If the options (city/airport) don't show, it means the route is not feasible.
2. Enter the passenger's preference of airlines.
  - a. Only the operating airlines are shown.
3. Select the expected travel period.
  - a. Enter the day to leave.
  - b. Enter the day to return.

### Insight from the dashboard

Upon receiving the inquiry, the dashboard will demonstrate the historical data for passengers to check as the followings:

1. The number of delayed departure flights from where the passenger wants to depart, and the percentage of flights delayed with estimated time.
2. Average security delay the passenger should expect.
3. Route(s) via available city or airport on the map.
4. Airlines with relatively less average delay time.
5. Which day of week has the lowest average delayed time.
6. What time of day has the lowest average delayed time.

## Airport Management

### Introduction

The airport user dashboard aims to provide penetrative information for airport administrations to improve performance of the airport system. The intuitive dashboard has the following measurement to help them decide which aspects they performed while benchmarking with competitors.

The dashboard tells:

1. Top 5 airports have the largest number of (departure/arrival) delayed flights
2. The total number of departure/arrival delays of the airport
3. The average minutes of departure/arrival delay of the airport
4. The percentage of delayed departures (arrival delays)
5. Five factors benchmarking of selected airport with overall performance
6. Which airline has the largest number of each delay causes
7. The seasonality of number of flights delayed in departure/arrival

### How to access the dashboard

Airport administrations are always concerned about putting themselves into best performances. Through the following step, airport management can easily access the information via dashboard:

1. Step by step narrow down
  - a. Locate departure city or/and airport
  - b. Locate arrival city or/and airport
  - c. Select the range of date
  - d. Choose airline to compare overall delay statistic
2. Enter Flight Number directly into the dropdown menu
  - a. Can simply select Flight number. (e.g. AA1298)
  - b. Range certain time period

### Insight from the dashboard

Upon receiving the inquiry, the dashboard will demonstrate the historical data for airport administratives to check as the following.

1. Number of departure/arrival flights delayed in selected airports.
  - a. Shown in map.
  - b. Blue point(s) tell the location of the departure airport(s).
  - c. Orange point(s) tell the coordinate of the arrival airport(s).
2. The percentage of ontime/delayed flights in selected airports.
  - a. Shown in pie chart.
3. Average delayed time on departure/arrival.
4. Delay analysis (comparison):
  - a. Average delayed minutes according to five factors of all airports.
  - b. Average delayed minutes according to five factors of the selected airport.
5. Top 5 airports with departure/arrival delays
6. Most departure delayed airlines compared to most arrival airlines.

## Airlines

### Introduction

The airline user manual aims to provide helpful information for airline companies to compare themselves with their competitors. The intuitive dashboard has the following measurement to determine their historical performance and identify the comparative advantages.

The dashboard tells:

1. Which airline has the least number of delayed flights
2. Which airline has the lowest average departure/arrival delay time
3. Which month, day of week and time of day have the largest number of flights and longest average departure/arrival delay time
4. Which airline has the largest number of each delay causes
5. Five factors benchmarking of selected airline with overall performance
6. The seasonality of number of flights delayed in departure/arrival

### How to access the dashboard

Airline management are sensibly aware of the causes that led to unexpected delays. The dashboard can input the operating airline through the dashboard via different options:

- a. Step by step narrow down
  - i. Locate departure city or/and airport
  - ii. Locate arrival city or/and airport
  - iii. Select the range of date
  - iv. Choose airline to compare overall delay statistic
- b. Enter Flight Number directly into the dropdown menu
  - i. Can simply select Flight number. (e.g. AA1298)
  - ii. Range certain time period

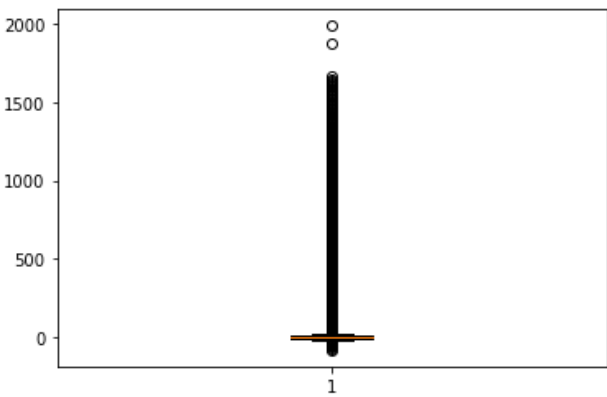
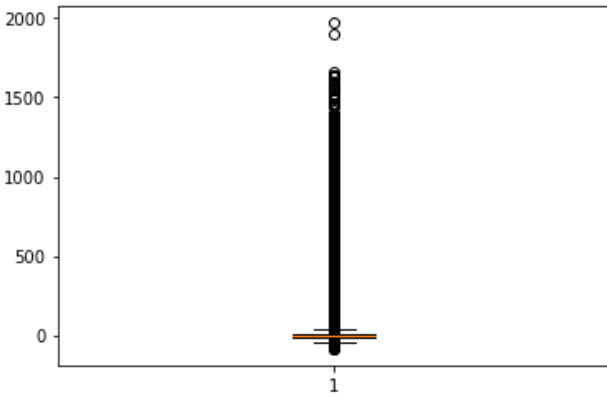
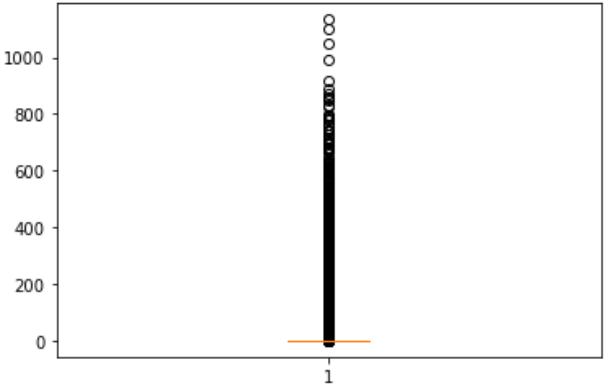
### Insight from the dashboard

Upon receiving the inquiry, the dashboard will demonstrate the historical data for airlines to check as the followings for the designated flight route(s):

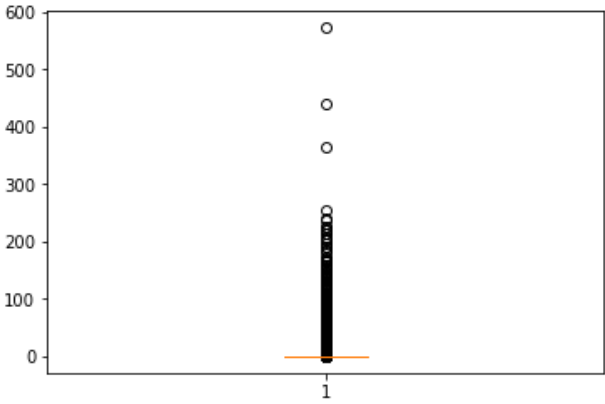
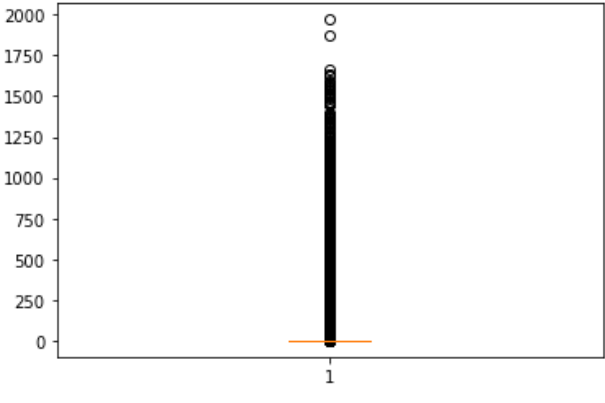
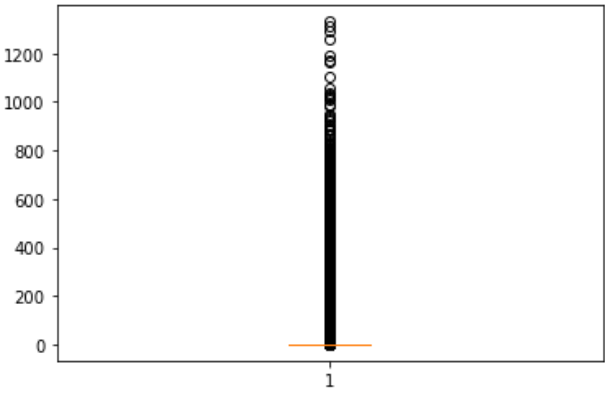
1. Number of departure/arrival flights delayed.
  - a. Shown in map.
  - b. Blue plane icon(s) tell the location of the departure airport(s).
  - c. Orange point(s) tell the coordinate of the arrival airport(s).
2. The percentage of ontime/delayed flights.
  - a. Shown in pie chart.
3. Average delayed time on departure/arrival.
4. Delay analysis (comparison):
  - a. Average delayed minutes according to five factors of all airlines.
  - b. Average delayed minutes according to five factors of the selected airline.
5. Types of delay graph elaborates the total delayed time and delayed time of each factor in every airline available.
  - a. The size of the circle stands for the total delayed time, the smallest point means no delay..
  - b. The y-intercept of the circles stands for the value of delayed minutes.
  - c. If the circle is on the line (or below the line), the factors didn't contribute to the delay.

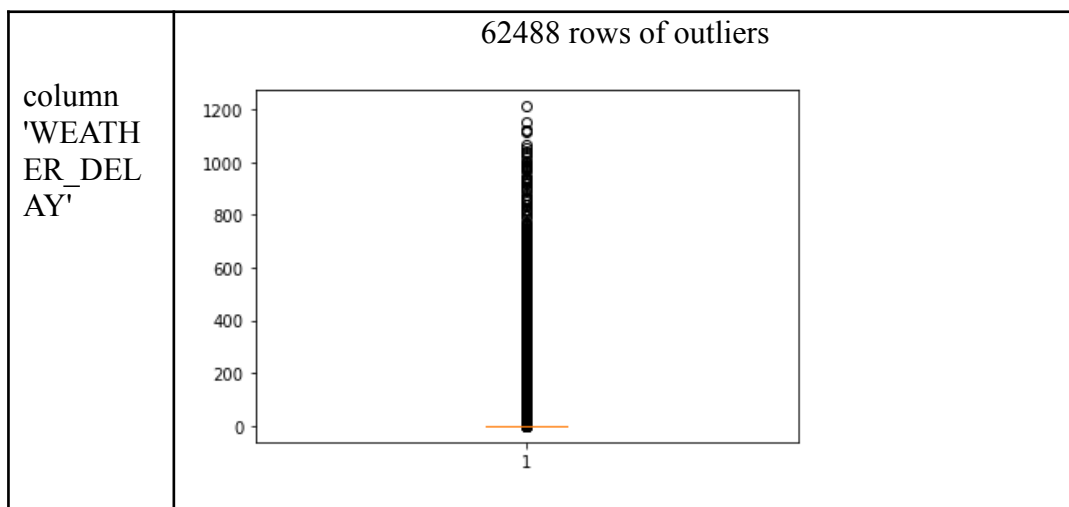
## Appendix

### 1. Table of outliers

column 'DEPART URE_DE LAY'	<p>651567 rows of outliers</p> 
column 'ARRIVA L_DELA Y'	<p>485232 rows of outliers</p> 
column 'AIR_SYS TEM_DE LAY'	<p>531928 rows of outliers</p> 
column 'SECURI TY_DEL AY'	<p>3337 rows of outliers</p>



	 <p>A box plot for the column 'AIRLINE_DELAY'. The y-axis ranges from 0 to 600. The plot shows a dense cluster of data points at the bottom, with a few outliers extending upwards to nearly 600. An orange horizontal line is drawn at the bottom of the plot area.</p>
column 'AIRLINE _DELAY'	<p>537885 rows of outliers</p>  <p>A box plot for the column 'LATE_AIRCRAFT_DELAY'. The y-axis ranges from 0 to 2000. The plot shows a dense cluster of data points at the bottom, with a few outliers extending upwards to nearly 2000. An orange horizontal line is drawn at the bottom of the plot area.</p>
column 'LATE_AI RCRAFT _DELAY'	<p>528297 rows of outliers</p>  <p>A box plot for the column 'LATE_AIRCRAFT_DELAY'. The y-axis ranges from 0 to 1200. The plot shows a dense cluster of data points at the bottom, with a few outliers extending upwards to nearly 1200. An orange horizontal line is drawn at the bottom of the plot area.</p>



## 2. Table of missing values

YEAR	0
MONTH	0
DAY	0
DAY OF WEEK	0
AIRLINE	0
FLIGHT NUMBER	0
ORIGIN AIRPORT	0
DESTINATION AIRPORT	0
SCHEDULED DEPARTURE	0
DEPARTURE TIME	86153
DEPARTURE DELAY	86153
AIR TIME	105071
DISTANCE	0
SCHEDULED ARRIVAL	0
ARRIVAL TIME	92513
ARRIVAL DELAY	105071
DIVERTED	0
CANCELLED	0
CANCELLATION REASON	5729195
AIR SYSTEM DELAY	0
SECURITY DELAY	0
AIRLINE DELAY	0
LATE AIRCRAFT DELAY	0
WEATHER DELAY	0
IATA CODE airline	0
AIRLINE long	0
AIRPORT origin long	0
CITY origin	0
STATE origin	0
LATITUDE origin	5008
LONGITUDE origin	5008
AIRPORT destination long	486165
CITY destination	486165
STATE destination	486165
LATITUDE destination	490775
LONGITUDE destination	490775

### 3. Python code of data cleaning

```
import pandas as pd
import numpy as np
import datetime
import matplotlib.pyplot as plt

fig, ax = plt.subplots()

pd.set_option('display.max_columns', None)

path = r"D:/UCD/Trimester_3/MIS41040/Tableau_Report/"

filename1 = "flights.csv"
filename2 = "airlines.csv"
filename3 = "airports.csv"
filename4 = "Mapping_Data_Dictionary.xlsx"

flights = pd.read_csv(path + filename1, dtype = {'SCHEDULED_DEPARTURE': 'str', 'DEPARTURE_TIME': 'str',
'SCHEDULED_ARRIVAL': 'str', 'ARRIVAL_TIME': 'str'})
airlines = pd.read_csv(path + filename2)
airports = pd.read_csv(path + filename3)
mapping = pd.read_excel(path + filename4, sheet_name = 'Mapping', usecols = [2, 3])

flights.drop(['TAIL_NUMBER', 'TAXI_OUT', 'WHEELS_OFF', 'SCHEDULED_TIME', \
'ELAPSED_TIME', 'WHEELS_ON', 'TAXI_IN', 'DIVERTED'], axis = 1, inplace = True)

flights['ORIGIN_AIRPORT'] = flights['ORIGIN_AIRPORT'].astype('str')
flights['DESTINATION_AIRPORT'] = flights['DESTINATION_AIRPORT'].astype('str')
mapping['ORIGIN_AIRPORT'] = mapping['ORIGIN_AIRPORT'].astype('str')

flights.dropna(subset = ['DEPARTURE_TIME', 'ARRIVAL_TIME'], inplace = True)
flights['DEPARTURE_TIME'].replace("2400", "0000", inplace = True)
flights['SCHEDULED_ARRIVAL'].replace("2400", "0000", inplace = True)
flights['ARRIVAL_TIME'].replace("2400", "0000", inplace = True)

flights['SCHEDULED_DEPARTURE'] = flights['SCHEDULED_DEPARTURE'].apply(lambda x:
datetime.datetime.strptime(x, "%H%M").time())
flights['DEPARTURE_TIME'] = flights['DEPARTURE_TIME'].apply(lambda x: datetime.datetime.strptime(x,
"%H%M").time())
flights['SCHEDULED_ARRIVAL'] = flights['SCHEDULED_ARRIVAL'].apply(lambda x: datetime.datetime.strptime(x,
"%H%M").time())
flights['ARRIVAL_TIME'] = flights['ARRIVAL_TIME'].apply(lambda x: datetime.datetime.strptime(x, "%H%M").time())

mapping = dict(mapping.values)
flights['ORIGIN_AIRPORT'].replace(mapping, inplace = True)
data = flights.merge(airlines, how = 'left', left_on = 'AIRLINE', right_on = 'IATA_CODE', suffixes=(None, '_long'))
data = data.merge(airports, how = 'left', left_on = 'ORIGIN_AIRPORT', right_on = 'IATA_CODE')

data.rename(columns={'AIRPORT': 'AIRPORT_origin_long', \
'IATA_CODE_x': 'IATA_CODE_airline', \
'CITY': 'CITY_origin', 'STATE': 'STATE_origin', \
'LATITUDE': 'LATITUDE_origin', \
'LONGITUDE': 'LONGITUDE_origin', \
'IATA_CODE_y': 'IATA_CODE_airport'}, inplace = True)

data = data.merge(airports, how = 'left', left_on = 'DESTINATION_AIRPORT', right_on = 'IATA_CODE')

data.rename(columns={'AIRPORT': 'AIRPORT_destination_long', \
'CITY': 'CITY_destination', 'STATE': 'STATE_destination', \
'LATITUDE': 'LATITUDE_destination', \
```

```
data.drop(['LONGITUDE','LONGITUDE_destination'], inplace = True)

data.drop(['IATA_CODE', 'COUNTRY_y', 'COUNTRY_x', 'IATA_CODE_airport'], axis = 1, inplace = True)

data['AIR_SYSTEM_DELAY'].replace(np.nan, 0, inplace = True)
data['SECURITY_DELAY'].replace(np.nan, 0, inplace = True)
data['AIRLINE_DELAY'].replace(np.nan, 0, inplace = True)
data['LATE_AIRCRAFT_DELAY'].replace(np.nan, 0, inplace = True)
data['WEATHER_DELAY'].replace(np.nan, 0, inplace = True)

data.dropna(subset = ['LONGITUDE_destination', 'DEPARTURE_TIME', 'ARRIVAL_DELAY', 'LATITUDE_origin'],
inplace = True)
data = data[data['CANCELLED']==0]
data.drop(['CANCELLED', 'CANCELLATION_REASON'], axis = 1, inplace = True)
data['DAY_OF_WEEK'] = data['DAY_OF_WEEK'].map({1:"Sunday", 2:"Monday", 3:"Tuesday", 4:"Wednesday",
5:"Thursday", 6:"Friday", 7:"Saturday"})

data.to_csv(path + 'flights_delay_cleaned.csv', encoding = 'utf-8')
```