LAB 2 Report

Team Details:

Name: Yiyang Jia (ID: 3882757155)
Rongzi Xie (ID: 9100096985)
Yihao Wang (ID: 4755190842)

Team Domain of data:

Our team will focus on the U.S. domestic air travel domain, with an emphasis on airline on-time performance(this is my own domain right now which will be part of the team focus). We chose this domain because high-quality data are publicly available (e.g., DOT/BTS feeds and curated Kaggle files), and flight punctuality is a significant topic around the world with its high volume and demands.

We aim to analyze on-time percentages across carriers, airports, routes, and seasons to understand when and where delays are most likely to occur. In addition to summary statistics, we plan to explore drivers of delay—such as weather, airport congestion, and scheduling buffers—and compare performance differences among airlines and hubs. Insights from this work can inform trip planning, operational decisions, and broader discussions about travel reliability and cost.

Dataset:

CSV source (Link):

https://www.kaggle.com/datasets/shubhamsingh42/flight-delay-dataset-2018-2024/data

This is a kaggle dataset I found for the flight-delay information of US commercial flights from 2018 to 2024. Each row corresponds to a flight leg and includes typical fields such as date/time stamps, carrier and airport codes, scheduled vs. actual departure/arrival, and delay minutes. For my computing metrics like **on-time percentage** (arrival within 15 minutes of schedule), this dataset is quite a fit. The reasons I choose this dataset are that first it contains a very large sample of dataset which will generate meaningful outputs and its time period(2018-2024) is a valuable time trend for analysis which provides stats both before and after covid time. It's a dataset from kaggle which is a very authoritative website.

Website source (Link):

https://www.flightradar24.com/blog/

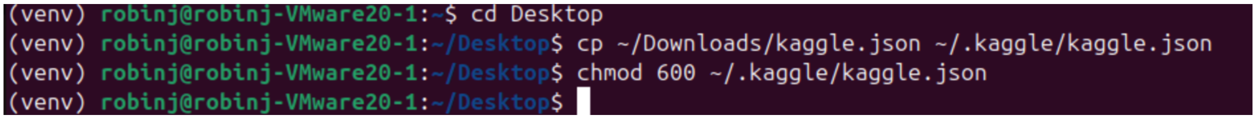
Flightradar is a very useful website and app I’ve used a lot for the flight information, and it’s also an authoritative source for flight information and this website publishes readable, data-driven posts on aviation operations and posts explains why delays spike (storms, strikes, holidays), so I can link patterns in my dataset to real events. In addition, its maps/charts help communicate findings to non-technical readers.

PDF source (Link):

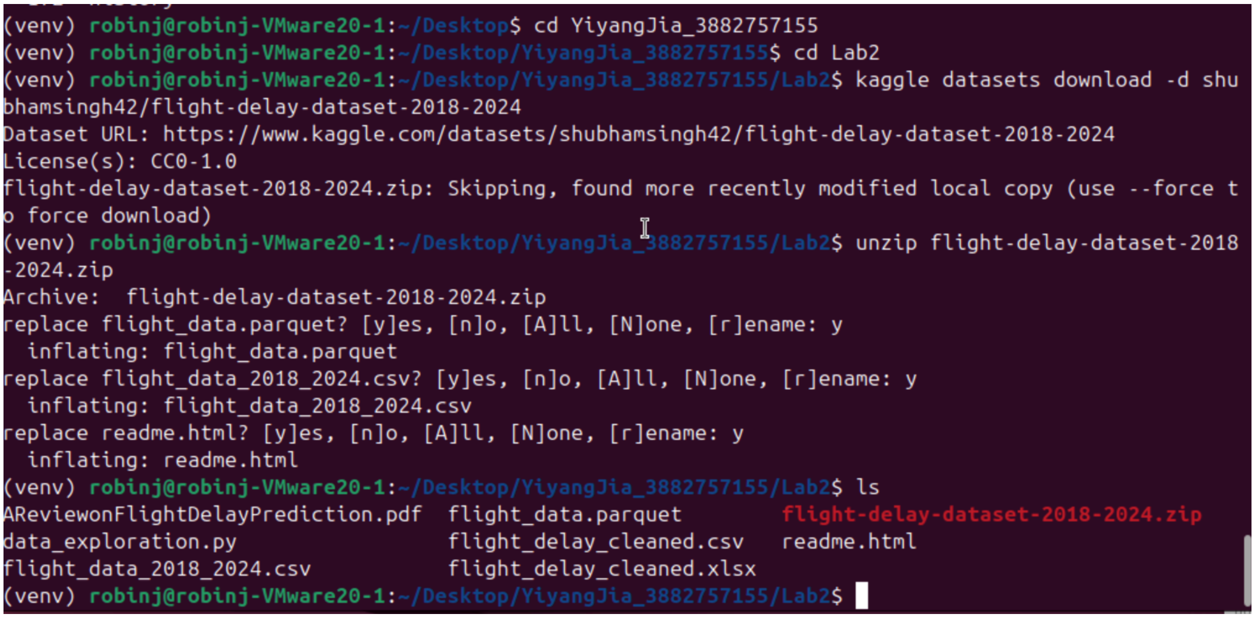
https://www.researchgate.net/publication/315382748_A_Review_on_Flight_Delay_Prediction

It’s a structured literature review that defines the flight-delay prediction landscape and proposes a taxonomy backed by a systematic mapping of prior work, and summarizes widely used data sources (U.S. DOT/FAA/BTS; Eurocontrol; NOAA weather) and discusses attributes/dimensions used in models. The reason why I chose this is because it’s a research summary about the prediction on flight delay and the variety of ML models to do the predictions which quite satisfy my major and the learning of this course. Also, it provides a timeline and trends for a credible Related-Work section which I believe will help our project a lot.

Kaggle API implementation:



```
(venv) robinj@robinj-VMware20-1:~$ cd Desktop
(venv) robinj@robinj-VMware20-1:~/Desktop$ cp ~/Downloads/kaggle.json ~/.kaggle/kaggle.json
(venv) robinj@robinj-VMware20-1:~/Desktop$ chmod 600 ~/.kaggle/kaggle.json
(venv) robinj@robinj-VMware20-1:~/Desktop$ 
```



```
(venv) robinj@robinj-VMware20-1:~/Desktop$ cd YiyangJia_3882757155
(venv) robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155$ cd Lab2
(venv) robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/Lab2$ kaggle datasets download -d shu
bhamsingh42/flight-delay-dataset-2018-2024
Dataset URL: https://www.kaggle.com/datasets/shubhamsingh42/flight-delay-dataset-2018-2024
License(s): CC0-1.0
flight-delay-dataset-2018-2024.zip: Skipping, found more recently modified local copy (use --force t
o force download)
(venv) robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/Lab2$ unzip flight-delay-dataset-2018
-2024.zip
Archive:  flight-delay-dataset-2018-2024.zip
replace flight_data.parquet? [y]es, [n]o, [A]ll, [N]one, [r]ename: y
  inflating: flight_data.parquet
replace flight_data_2018_2024.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: y
  inflating: flight_data_2018_2024.csv
replace readme.html? [y]es, [n]o, [A]ll, [N]one, [r]ename: y
  inflating: readme.html
(venv) robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/Lab2$ ls
AReviewonFlightDelayPrediction.pdf  flight_data.parquet        flight-delay-dataset-2018-2024.zip
data_exploration.py                 flight_delay_cleaned.csv   readme.html
flight_data_2018_2024.csv           flight_delay_cleaned.xlsx
(venv) robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/Lab2$
```

Before, I created my kaggle account and created a new token then downloaded the kaggle.json in my VMfusion downloads then gave the permission for using the kaggle API to download the zip files of the dataset. On the second image, just use kaggle(I installed before) to download the flight-delay dataset I used for this lab and ls show it.