Project Proposal

Motivation

Flying on airplanes has become something that people do quite frequently, and at one point or another, we all have experienced having a flight be delayed or cancelled. These events are always very annoying and can cause a lot of issues. If there was a way to predict which flights could be delayed/cancelled before it actually happens, this would be extremely beneficial for people to ensure they choose flights that will stay on time. This would save people from a lot of unwanted stress as well as time, especially if they have time constraints (i.e. layover flight, holidays, emergencies, work, etc.).

Goals

The goal of this project is to explore how various factors correlate with flight delays across the United States in 2023. Using a comprehensive dataset that includes flight-level data, weather conditions, airline information, and aircraft specifications, we aim to uncover key patterns and predictors of delays.

Questions:

Part A: Can we use weather (airline, flight duration, flight distance, day of the week, departure airport) data to predict a flight was cancelled/delayed?
Part B: If delayed, can we predict for how long?

Dataset

The dataset includes:

1. Detailed departure and arrival delays for each commercial flight.
2. Weather data at the airport level, including temperature, precipitation, snow cover, air pressure, wind speed, and direction.
3. Aircraft manufacturer, model, and age information.
4. Airline carrier and airport metadata, including geolocation.

The main flight information is stored in US_flights_2023.csv, weather data in weather_meteo_by_airport.csv, and canceled or diverted flights are separated into Cancelled_Diverted_2023.csv for unbiased analysis.

Models

To analyze these data, we apply two modeling approaches: **Linear Regression** and **Decision Tree**. These models help us assess the importance of different features—such as weather patterns, aircraft age, and airline carrier—in predicting delays, and compare how well each algorithm performs on this multivariate structured data. For our decision tree we will use variables like precipitation, snow, wind speed, and location to predict whether or not a flight has been delayed or cancelled, we can use those same variables to predict the duration of the delay using linear regression.

Code to read data

```
cancelled_df=pd.read_csv("/Users/evanli/Desktop/451/Project/Cancelled_Diverted_2023.csv")
weather_df=pd.read_csv("/Users/evanli/Desktop/451/Project/weather_meteo_by_airport.csv")
flights_df=pd.read_csv("/Users/evanli/Desktop/451/Project/US_flights_2023.csv")
```

Dataset link

https://www.kaggle.com/datasets/bordanova/2023-us-civil-flights-delay-meteo-and-aircraft?select=US_flights_2023.csv