Predicting Flight Delays to Improve Airport Operations

Team 4-1 | April 2, 2025

Mohamed Bakr

Shruti Gupta

Erica Landreth

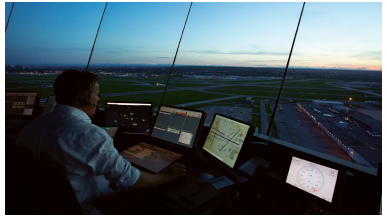
Danielle Yoseloff

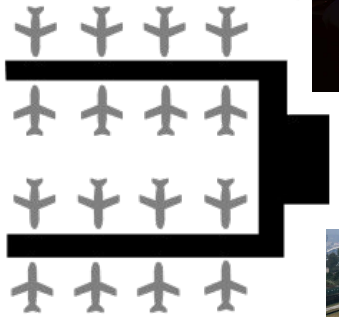
Outline

- Problem statement
- Data preparation
  - Input sources
  - Joining and cleaning
- Feature development
  - Seasonality
  - Recency and frequency
- Modeling
  - Baseline pipeline
  - Preliminary results
- Next steps

Problem Statement

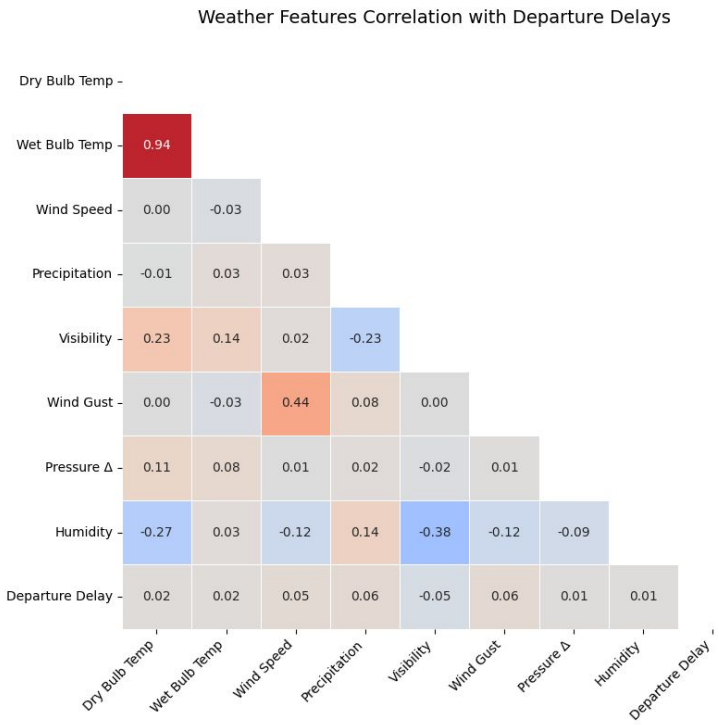
- Efficient airport operations require effective resource management
  - Personnel
  - Space
  - Scheduling
- Balancing resources heavily dependent on flight schedules
  - Unanticipated delays lead to inefficiencies–often costly
- **Modeling objective:** predict whether a flight's departure will be delayed, two hours before its scheduled departure
  - Delayed: 15+ minutes delayed OR cancelled

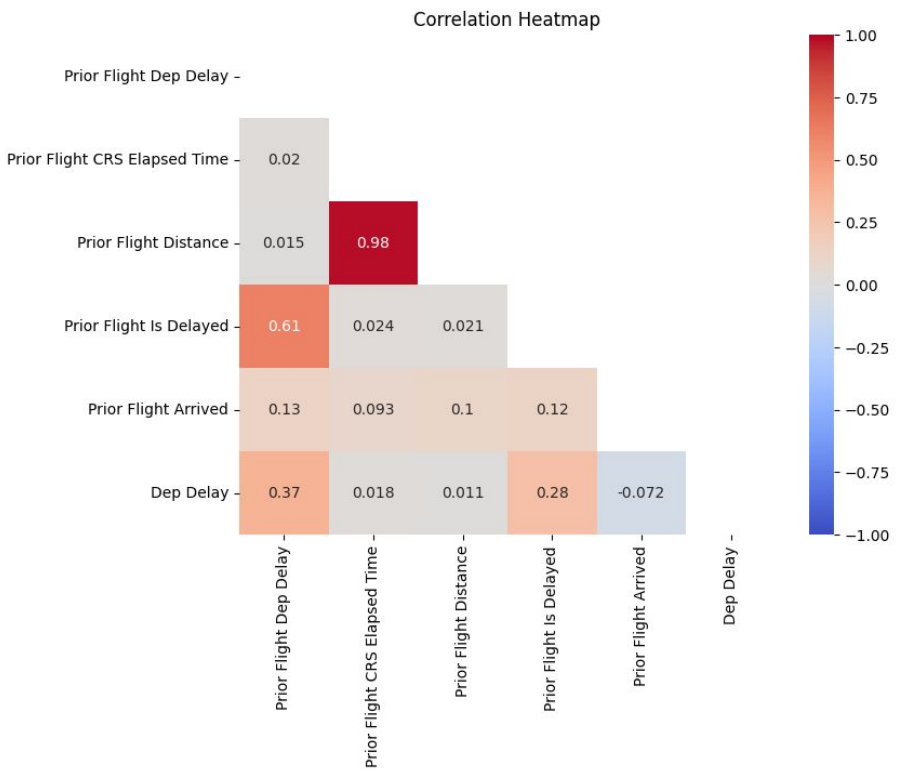




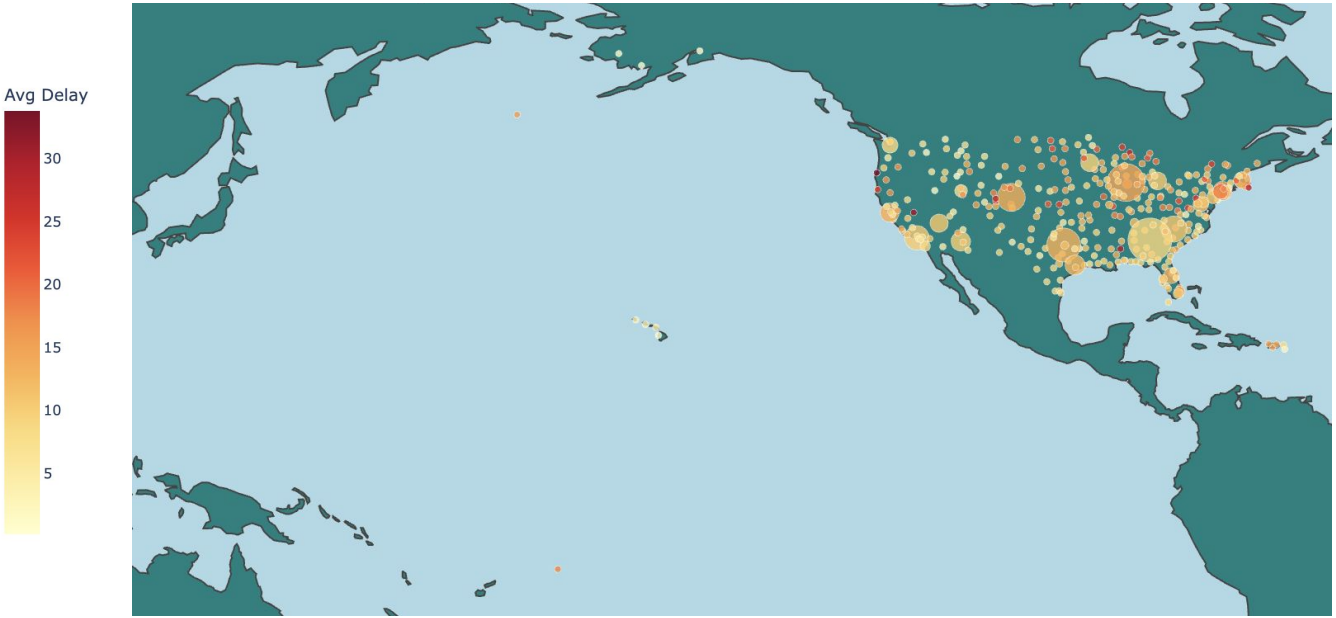
Data Preparation

Delay EDA


Weather Features Correlation with Departure Delays
Dry Bulb Temp
Wet Bulb Temp 0.94
Wind Speed 0.00 -0.03
Precipitation -0.01 0.03 0.03
Visibility 0.23 0.14 0.02 -0.23
Wind Gust 0.00 -0.03 0.44 0.08 0.00
Pressure Δ 0.11 0.08 0.01 0.02 -0.02 0.01
Humidity -0.27 0.03 -0.12 0.14 -0.38 -0.12 -0.09
Departure Delay 0.02 0.02 0.05 0.06 -0.05 0.06 0.01 0.01
Dry Bulb Temp
Wet Bulb Temp
Wind Speed
Precipitation
Visibility
Wind Gust
Pressure Δ
Humidity
Departure Delay


Correlation Heatmap
Prior Flight Dep Delay
Prior Flight CRS Elapsed Time 0.02
Prior Flight Distance 0.015 0.98
Prior Flight Is Delayed 0.61 0.024 0.021
Prior Flight Arrived 0.13 0.093 0.1 0.12
Dep Delay 0.37 0.018 0.011 0.28 -0.072
Prior Flight Dep Delay
Prior Flight CRS Elapsed Time
Prior Flight Distance
Prior Flight Is Delayed
Prior Flight Arrived
Dep Delay
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Airports



Join Pipeline

Time Zones

Created a UDF to get Time zones from Airport's Lat, Lon and ID as a helper table

Flights

Cleaned and deduplicated the flights table, filtering for only the necessary features.

ICAO

Joined the flights data with the airport codes table to obtain the ICAO code

Stations

Identified the closest station for each airport and joined the flights with the stations to retrieve the station ID

UTC

Incorporate time zone information and calculate the timestamps for UTC-2 and UTC-4.

Weather

Prepare weather, join flights with weather on station id and depart utc between depart UTC-2 and UTC-4

Join Pipeline

Time Zones

Created a UDF to get Time zones from Airport's Lat, Lon and ID as a helper table

Flights

Cleaned and deduplicated the flights table, filtering for only the necessary features.

ICAO

Joined the flights data with the airport codes table to obtain the ICAO code

Stations

Identified the closest station for each airport and joined the flights with the stations to retrieve the station ID

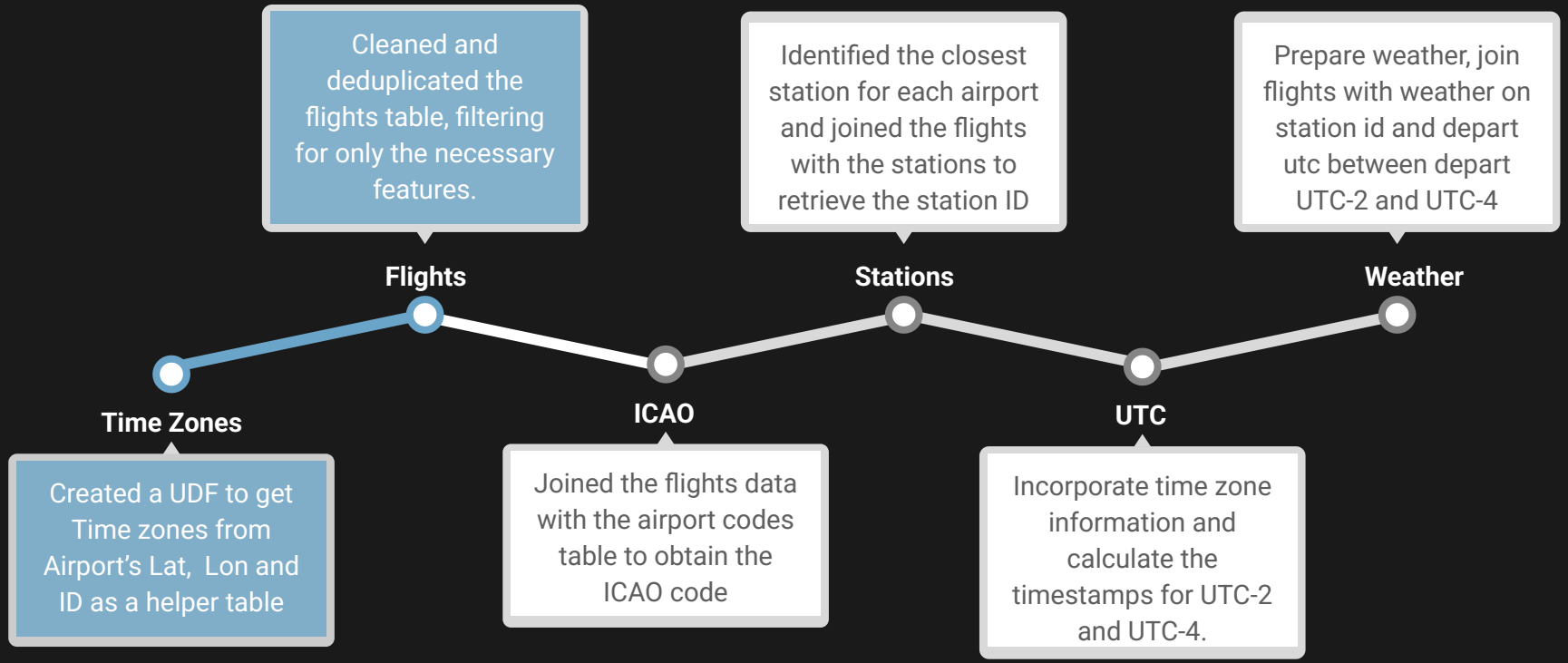
UTC

Incorporate time zone information and calculate the timestamps for UTC-2 and UTC-4.

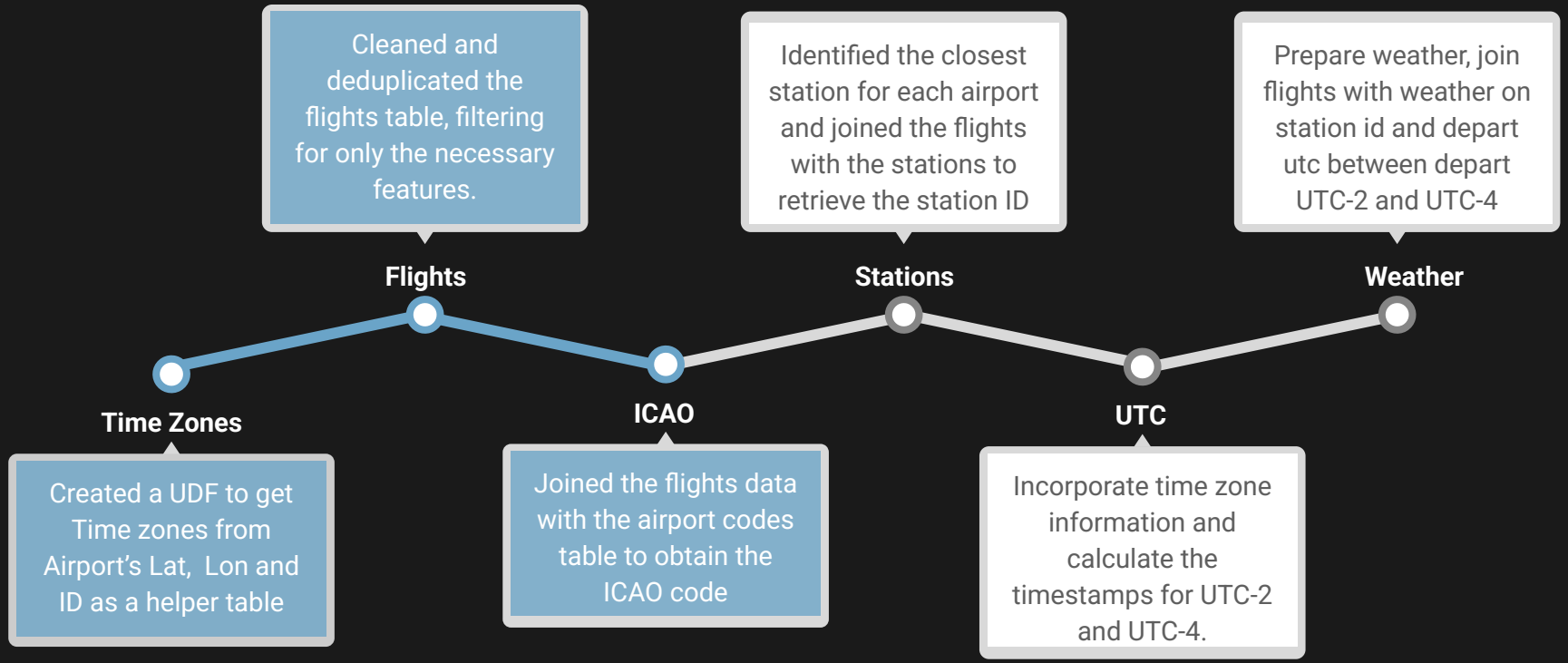
Weather

Prepare weather, join flights with weather on station id and depart utc between depart UTC-2 and UTC-4

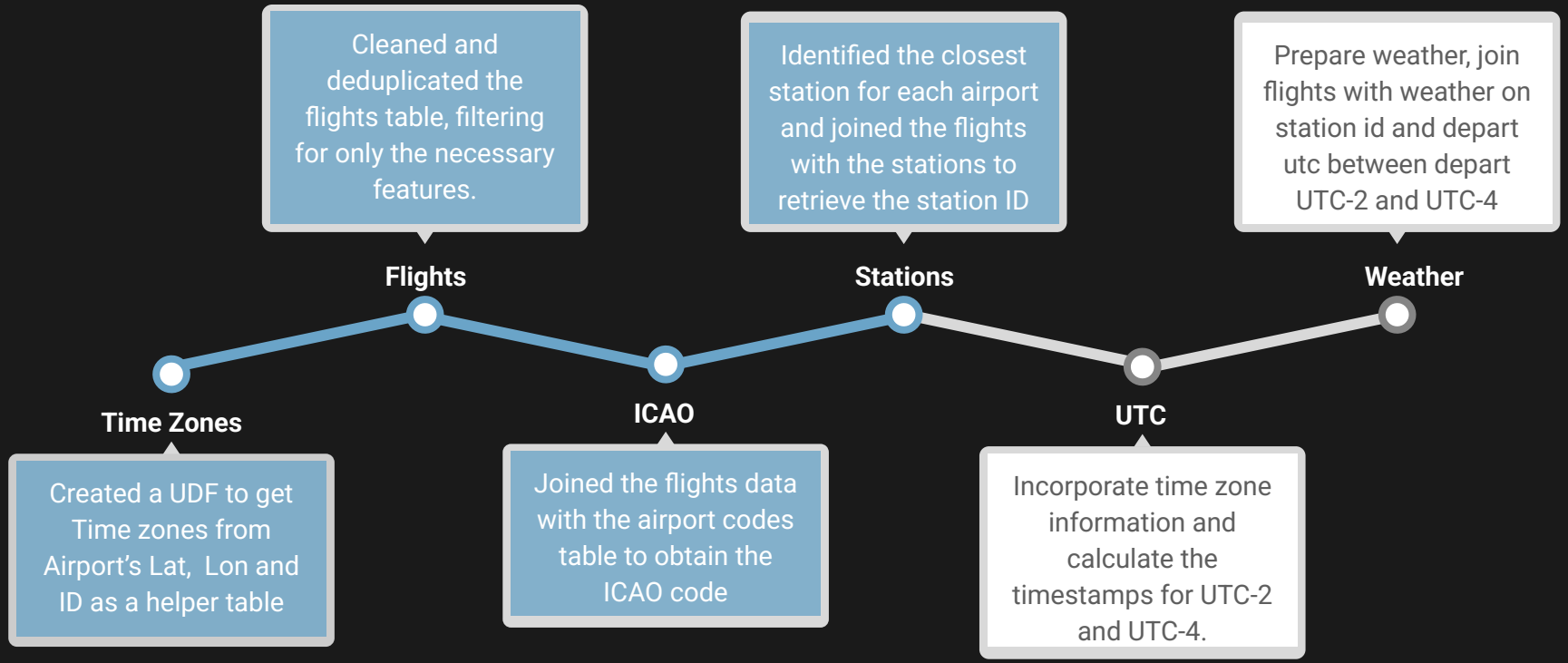
Join Pipeline



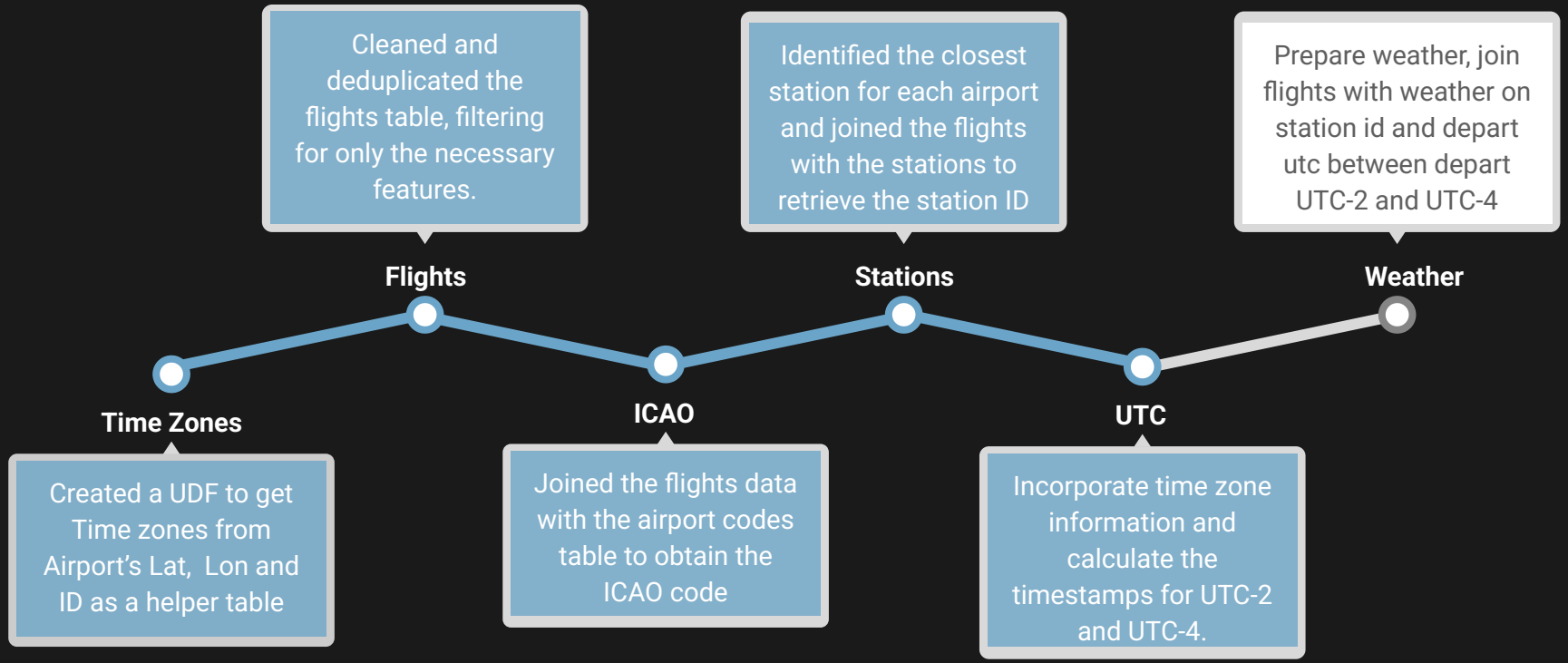
Join Pipeline



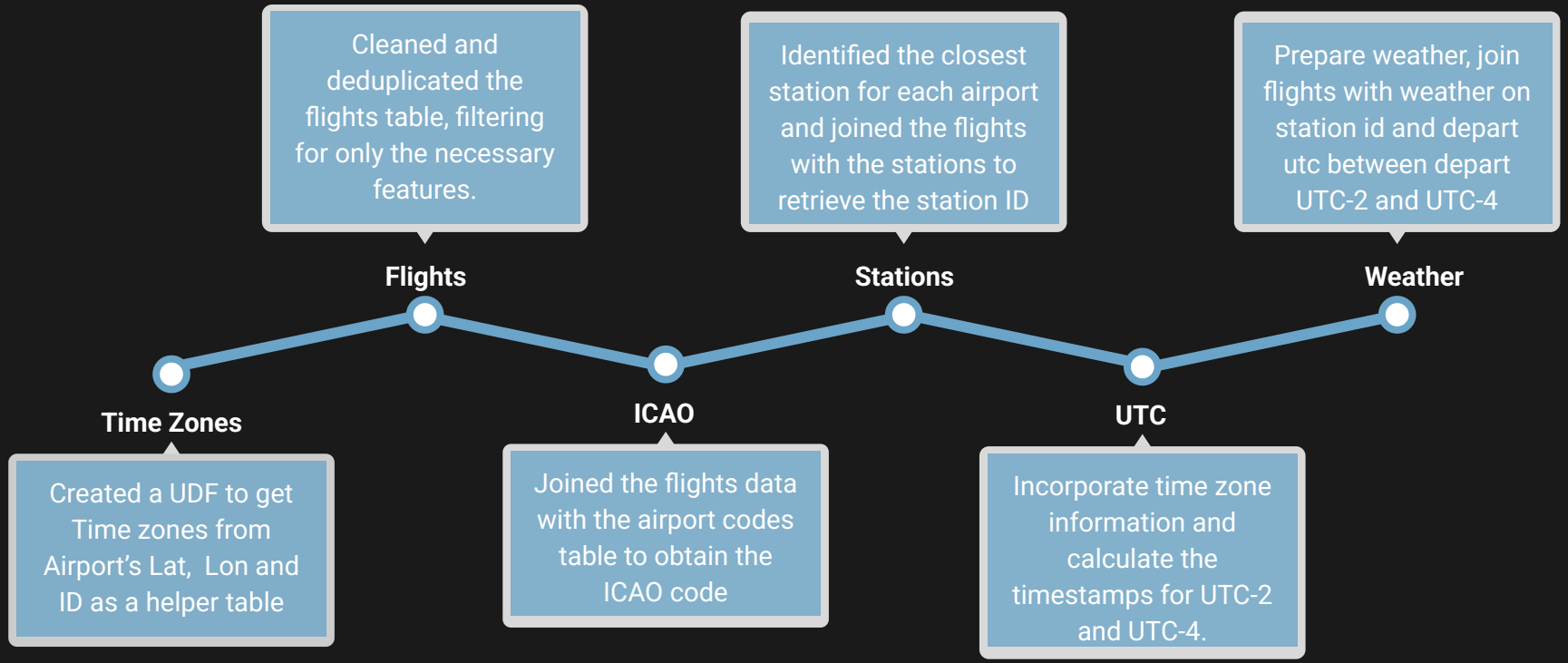
Join Pipeline



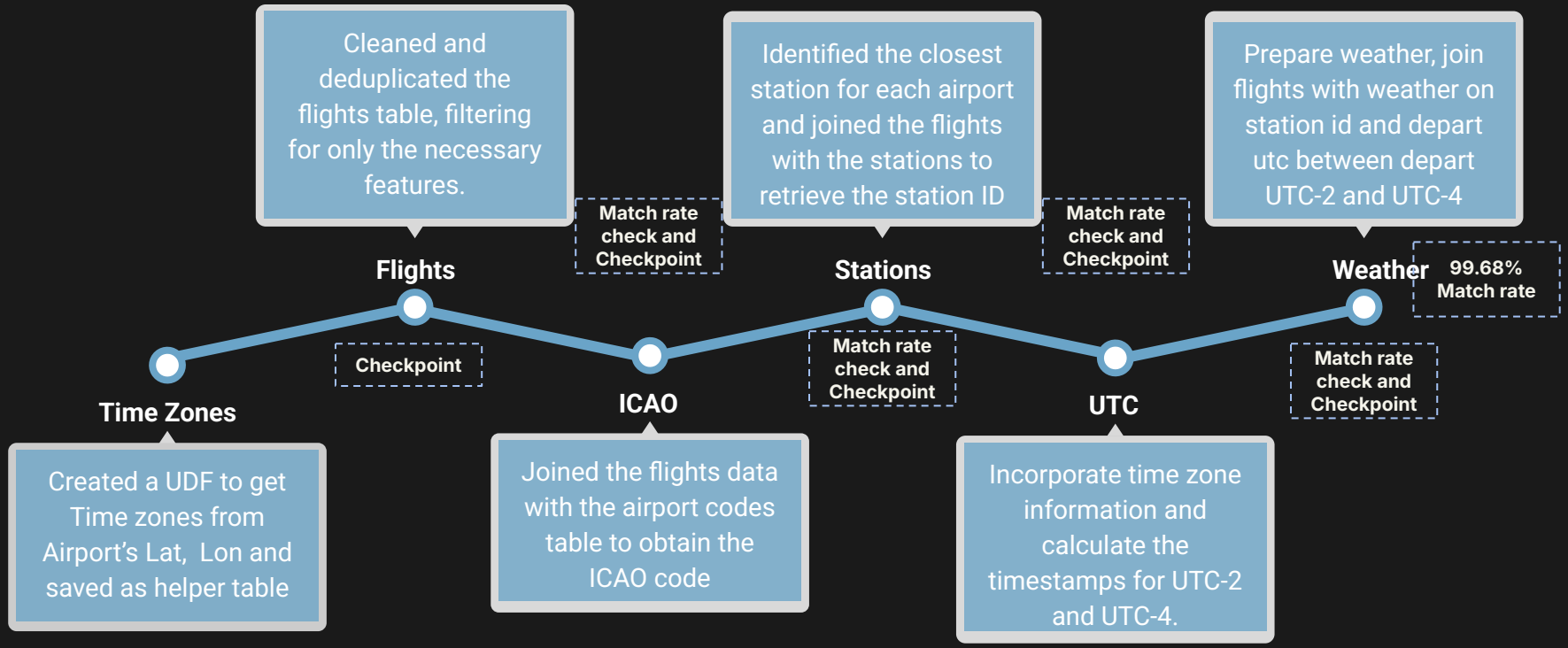
Join Pipeline



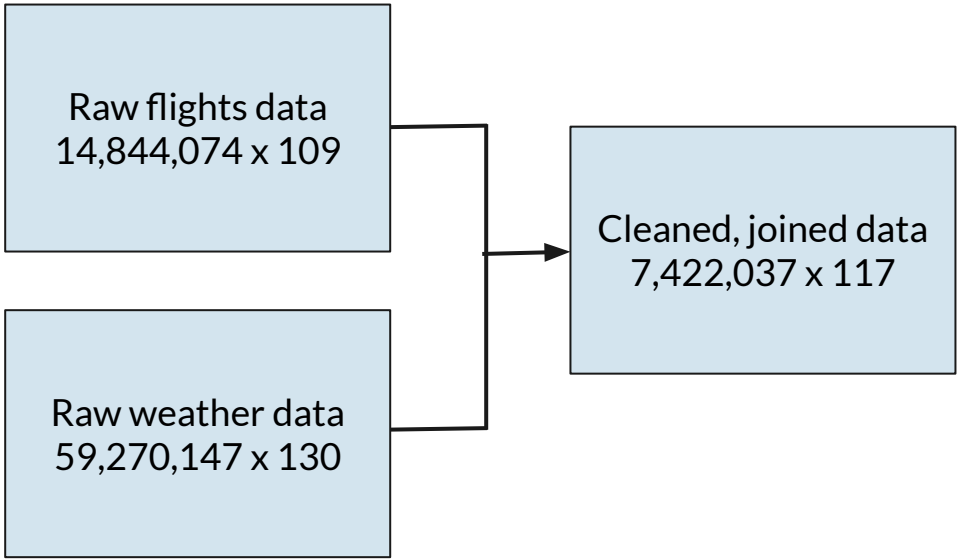
Join Pipeline



Join Pipeline



Data Preparation Result

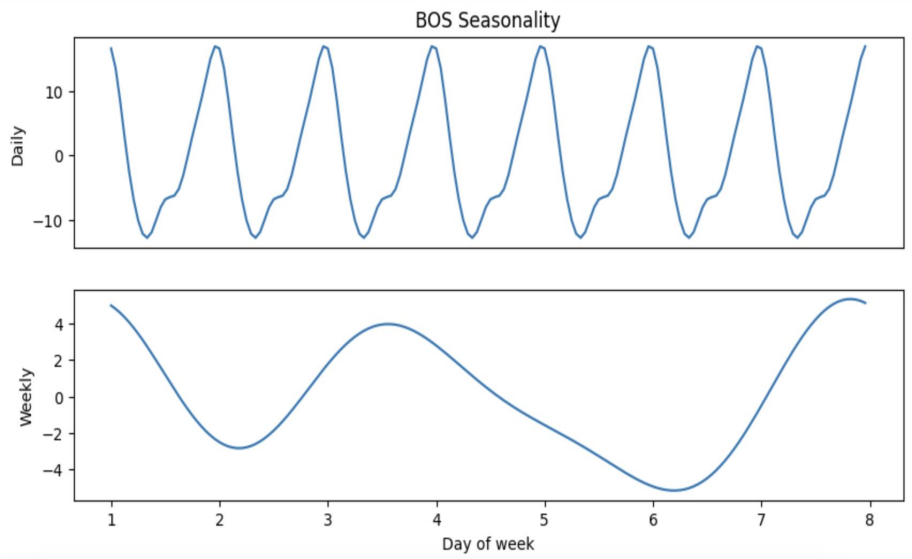


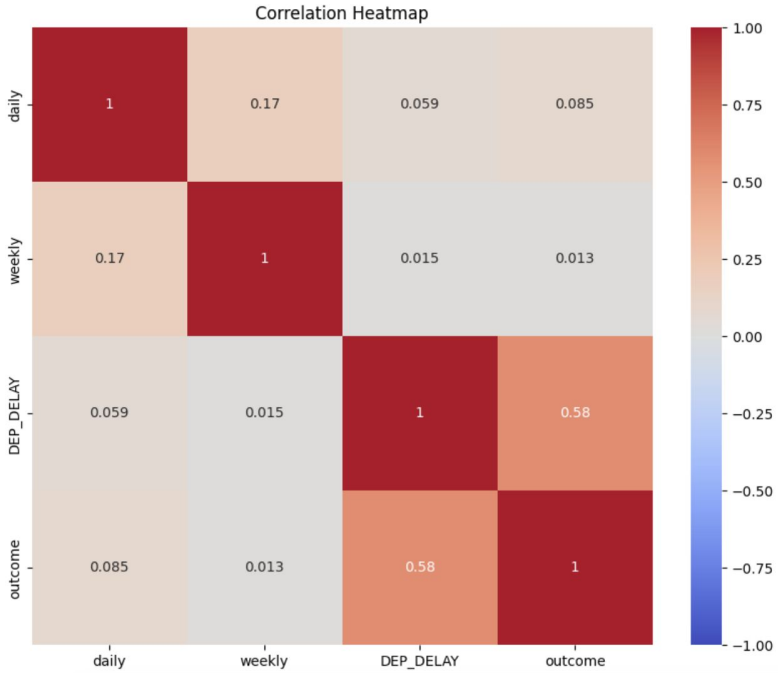
Cleaned data: Outcome variable

Delayed	Not Delayed
1,494,498	5,927,539

Feature Engineering

Delay Seasonality: Prophet Modeling





Recency and Frequency Features

American Airlines

Tail Number N98TW

Prior Flight Definition:

- Prior Destination = Origin
- Within 24 hrs

	ORIGIN	DEST	sched_depart_utc	priorflight_deptime_final	priorflight_isdelayed	priorflight_arrived	priorflght_est_arr_time_final	est_tail_turnaround_window_min
1	PIT	DFW	2019-01-01T12:15:00.000+00:00	null	null	0	null	null
2	DFW	RDU	2019-01-01T16:55:00.000+00:00	2019-01-01T12:15:00.000+00:00	0	0	2019-01-01 15:50:00	65
3	RDU	DFW	2019-01-01T20:33:00.000+00:00	2019-01-01T16:55:00.000+00:00	0	0	2019-01-01 19:34:00	59
4	DFW	PNS	2019-01-02T12:55:00.000+00:00	2019-01-01T20:33:00.000+00:00	0	1	2019-01-01 23:55:00	780

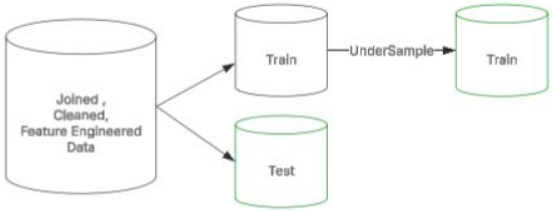
* Using actual arrival if available

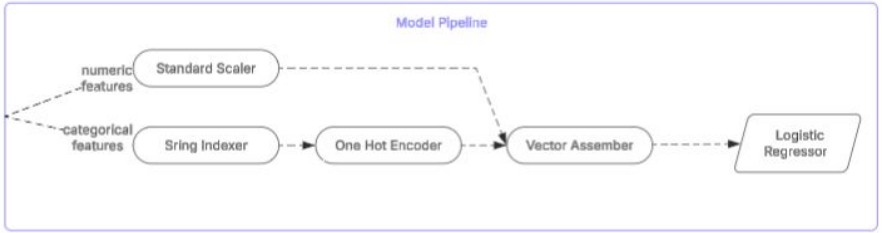
Notes:

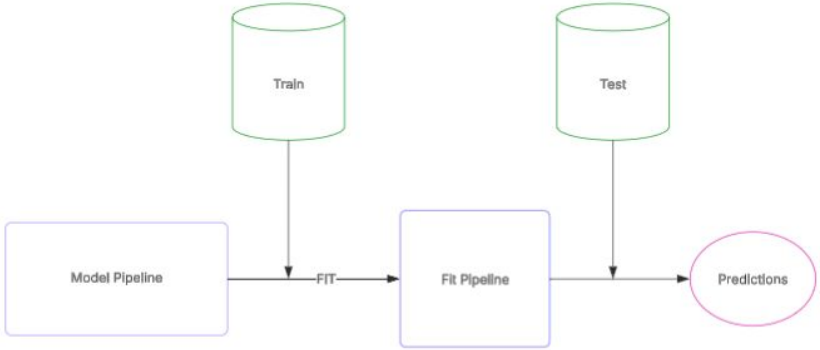
- Assume tail number is known
- Time based prior flight features could be known 2 hrs before expected departure

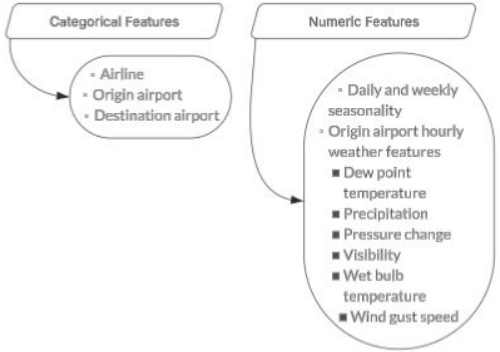
Modeling

Baseline Model Pipeline









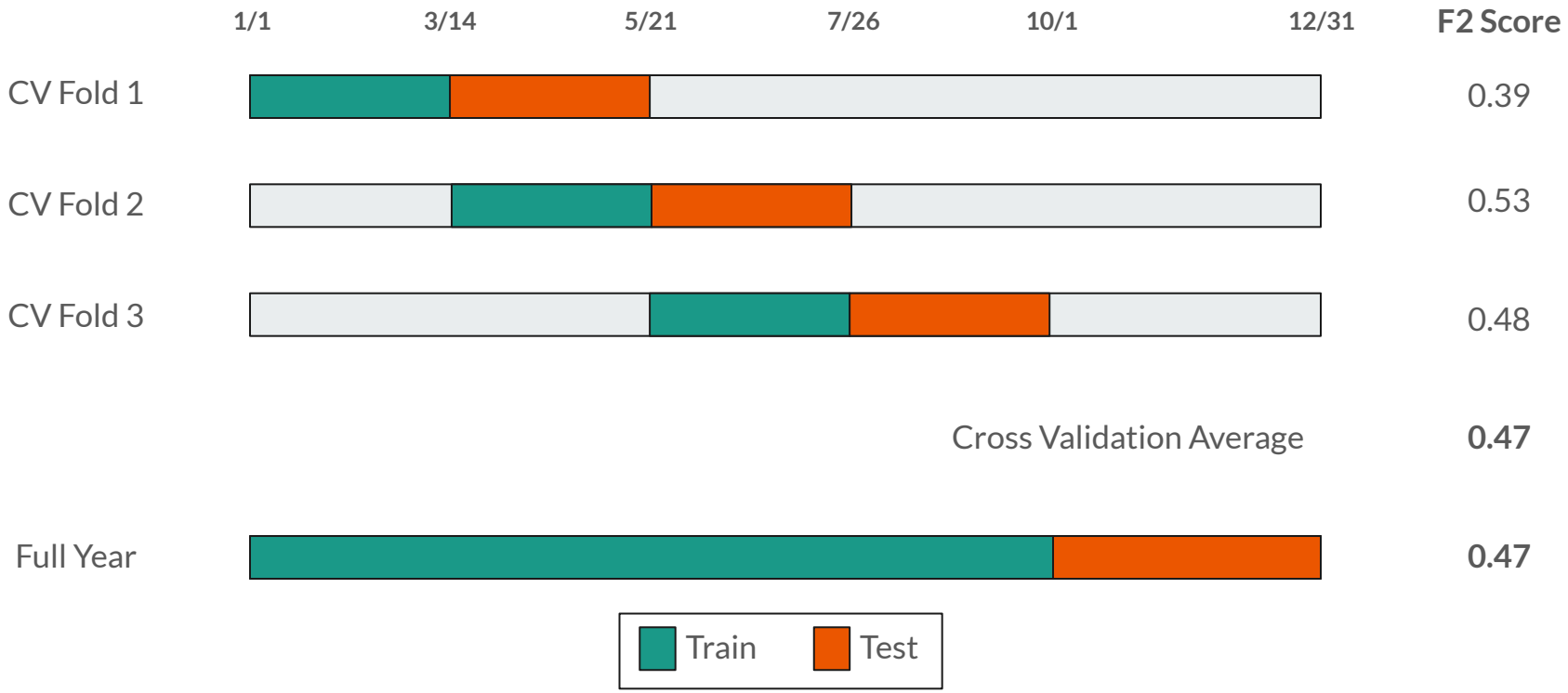
Model Evaluation

Error	Meaning	Consequences	Cost
Type I	Predict delay, depart on time	Confusion, unnecessary changes	Acceptable: we prioritize caution
Type II	Predict on time, depart delayed	Missed connections, poor customer satisfaction, operational disruptions	Costly: Major disruptions can have immense monetary cost

$$F_2 = \frac{5}{\frac{4}{Precision} + \frac{1}{Recall}} = \frac{5}{\frac{4TP+4FN}{TP} + \frac{TP+FP}{TP}}$$

Emphasis on recall!

Preliminary Baseline Results: F2 Score



Next Steps

Phase II

- Feature engineering and processing
  - Delay frequency features
  - Non-numeric weather features
  - Interaction terms
- Feature selection
  - Lasso
  - Explore PCA
- Baseline model hyperparameter tuning

Phase III

- Explore more sophisticated models
  - Tree ensembles
  - Multi-layer perceptron
- Additional feature engineering
  - Graph features
  - Additional seasonality components
- Select and tune a final model

Questions?