Appendix

Key Variable Summary:

Top Predictive Features:

- DEP_HOUR_SIN: Sinusoidal encoding of departure hour (highest importance)
- SEASON: Categorical representation of flight season
- IS_REDEYE: Binary flag for overnight flights (11pm-5am)
- SOURCE_FILE: Indication of which month the flight data came from
- ORIGIN_WEATHER_ICON: Weather condition icon at origin airport

Weather Variables:

- ORIGIN_CONDITIONS/DEST_CONDITIONS: Weather conditions at origin/destination
- MAX_WEATHER_SEVERITY: Maximum weather severity level (0-3 scale)
- ORIGIN_PRECIPITATION/DEST_PRECIPITATION: Precipitation amounts
- ORIGIN_SNOW/DEST_SNOW: Snow measurements
- ORIGIN_WIND_SPEED/DEST_WIND_SPEED: Wind speed measurements

Temporal Variables:

- WEEK_OF_YEAR: Calendar week number
- DEP_HOUR_COS: Cosine encoding of departure hour
- IS_MORNING_RUSH/IS_EVENING_RUSH: Peak travel time indicators
- IS_WEEKEND: Weekend travel indicator
- IS_HOLIDAY_SEASON: Holiday period indicator

Flight Characteristics:

- OP_UNIQUE_CARRIER: Operating airline
- DISTANCE: Flight distance
- ORIGIN/DEST: Airport codes for origin and destination

Data Analysis Strategy:

1. Flight Data Pre-Processing

- Sets missing DEP_DELAY and ARR_DELAY to 0 only for non-cancelled flights.
- Parsed dates and standardized time formats
- Filtered records to retain relevant fields: Flight Date, Origin, Destination, Scheduled and Actual Arrival, Arrival Delay
- Created a binary target label: DELAYED = 1 if DEP_Delay or ARR_DELAY >= 15 min, else 0
- Reduced data types for memory efficiency (e.g., Int64 → Int8/Int16)

- Fills missing DEP_TIME and ARR_TIME using group-wise median imputation by OP_UNIQUE_CARRIER, ORIGIN, and DEST.
- Based on departure hour, created DEP_HOUR, DEP_HOUR_SIN/COS, IS_MORNING_RUSH (6-9am), IS_EVENING_RUSH (4-7pm), IS_REDEYE (11pm-5am), IS_MIDDAY (10am-3pm), and TIME_CATEGORY (5 time-of-day segments).

2. Weather Data Enrichment

- **Geolocation Matching:** Airport coordinates were assigned using the GitHub Airport Dataset, and flights were mapped to their nearest weather datapoint using the Haversine function.
- **Clustering Strategy:** Airports were grouped into 100 km regional clusters to reduce redundant API calls and improve spatial accuracy for weather alignment.
- **Weather Integration:** Pulled data from the Visual Crossing API (temperature, wind speed, visibility, precipitation, cloud cover).
- **Derived Features:** Added 12 new features — 6 binary flags (e.g., ORIGIN_EXTREME_WEATHER, HAS_WEATHER_DELAY) and 6 severity scores (e.g., DEST_WEATHER_SEVERITY, CARRIER_DELAY_SEVERITY) — to quantify weather impact on delays.

3. Data Integration
Airports were clustered using the Haversine formula (≤100 km) to reduce API calls by fetching weather per region. Weather data from Visual Crossing API was merged to each flight at origin and destination, adding severity scores and binary flags for extreme conditions. U.S. holidays were labeled to capture temporal effects, resulting in a final enriched dataset of 3.6M+ rows × 110 columns.

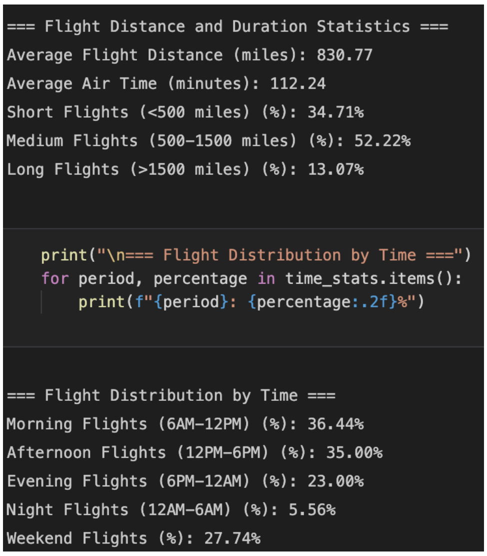
4. Feature Engineering The pipeline performs advanced spatio-temporal feature engineering by integrating regional weather conditions and holiday context with flight records. Key transformations included categorical encoding, numerical normalization, and creation of derived features like weather severity buckets, adding 54 new features to enrich model input for delay prediction.

5. Missing Value Handling:

- The enhanced dataset shows strong integrity 100% complete geographic clustering, nearly all weather metrics present (only 39 visibility values missing out of 3.6M), full temporal features (dates, weekdays, holiday flags), and structured missingness only where expected: delay causes (79.8% null), cancellation codes (98.7% null), and holiday names (82.4% null).
- All key delay, time, and weather columns were cleaned using conditional imputation (e.g., delay = 0 if not cancelled/diverted), with missingness flags added. Categorical fields like CANCELLATION_CODE were one-hot encoded, delay causes were converted to severity scores and binary flags, visibility was median-imputed, and new features like

WEATHER_IMPACT_SCORE, SEVERITY_DISTANCE_EFFECT, and SEASON were engineered to enrich predictive value.

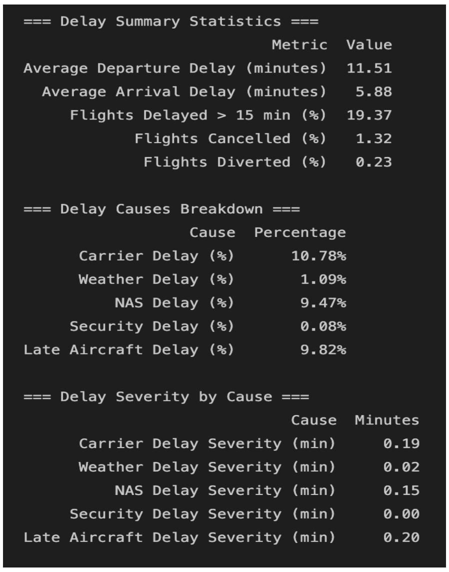
<u>Statistical analysis Output:</u>



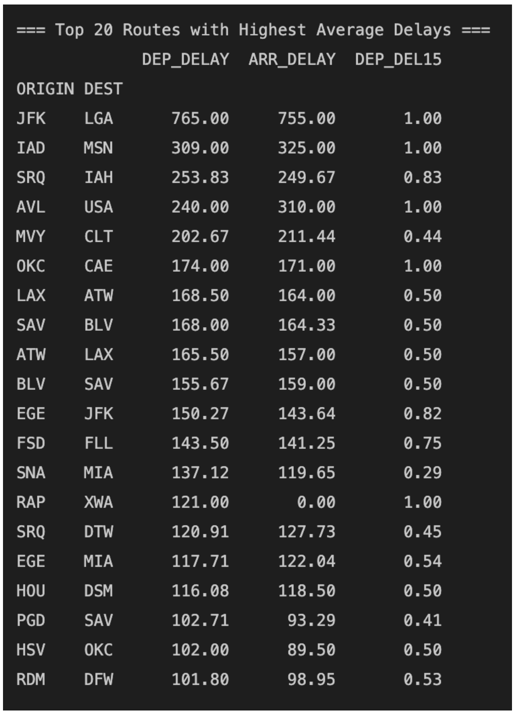
```
=== Flight Distance and Duration Statistics ===
Average Flight Distance (miles): 830.77
Average Air Time (minutes): 112.24
Short Flights (<500 miles) (%): 34.71%
Medium Flights (500-1500 miles) (%): 52.22%
Long Flights (>1500 miles) (%): 13.07%
```

```
print("\n=== Flight Distribution by Time ===")
for period, percentage in time_stats.items():
    print(f"{period}: {percentage:.2f}%")
```

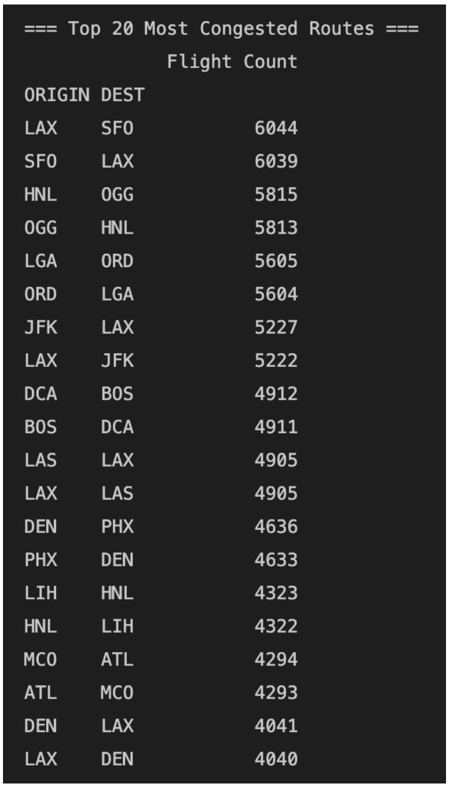
```
=== Flight Distribution by Time ===
Morning Flights (6AM-12PM) (%): 36.44%
Afternoon Flights (12PM-6PM) (%): 35.00%
Evening Flights (6PM-12AM) (%): 23.00%
Night Flights (12AM-6AM) (%): 5.56%
Weekend Flights (%): 27.74%
```



```
=== Delay Summary Statistics ===
                           Metric  Value
Average Departure Delay (minutes)  11.51
  Average Arrival Delay (minutes)   5.88
     Flights Delayed > 15 min (%)  19.37
            Flights Cancelled (%)   1.32
             Flights Diverted (%)   0.23

=== Delay Causes Breakdown ===
                  Cause  Percentage
      Carrier Delay (%)      10.78%
      Weather Delay (%)       1.09%
          NAS Delay (%)       9.47%
     Security Delay (%)       0.08%
Late Aircraft Delay (%)       9.82%

=== Delay Severity by Cause ===
                              Cause  Minutes
      Carrier Delay Severity (min)     0.19
      Weather Delay Severity (min)     0.02
          NAS Delay Severity (min)     0.15
     Security Delay Severity (min)     0.00
Late Aircraft Delay Severity (min)     0.20
```

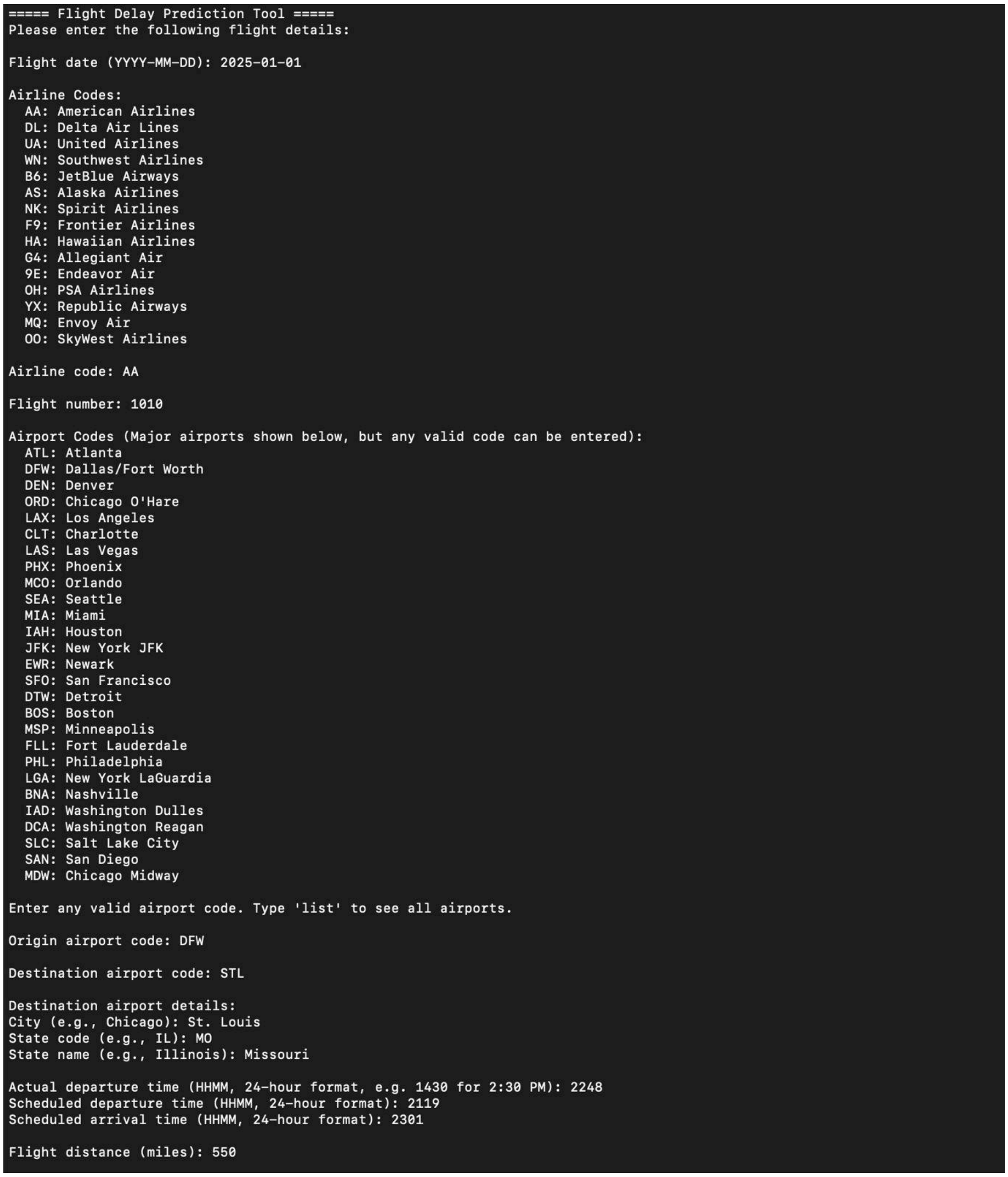


```
=== Top 20 Routes with Highest Average Delays ===
             DEP_DELAY  ARR_DELAY  DEP_DEL15
ORIGIN DEST
JFK    LGA      765.00     755.00       1.00
IAD    MSN      309.00     325.00       1.00
SRQ    IAH      253.83     249.67       0.83
AVL    USA      240.00     310.00       1.00
MVY    CLT      202.67     211.44       0.44
OKC    CAE      174.00     171.00       1.00
LAX    ATW      168.50     164.00       0.50
SAV    BLV      168.00     164.33       0.50
ATW    LAX      165.50     157.00       0.50
BLV    SAV      155.67     159.00       0.50
EGE    JFK      150.27     143.64       0.82
FSD    FLL      143.50     141.25       0.75
SNA    MIA      137.12     119.65       0.29
RAP    XWA      121.00       0.00       1.00
SRQ    DTW      120.91     127.73       0.45
EGE    MIA      117.71     122.04       0.54
HOU    DSM      116.08     118.50       0.50
PGD    SAV      102.71      93.29       0.41
HSV    OKC      102.00      89.50       0.50
RDM    DFW      101.80      98.95       0.53
```

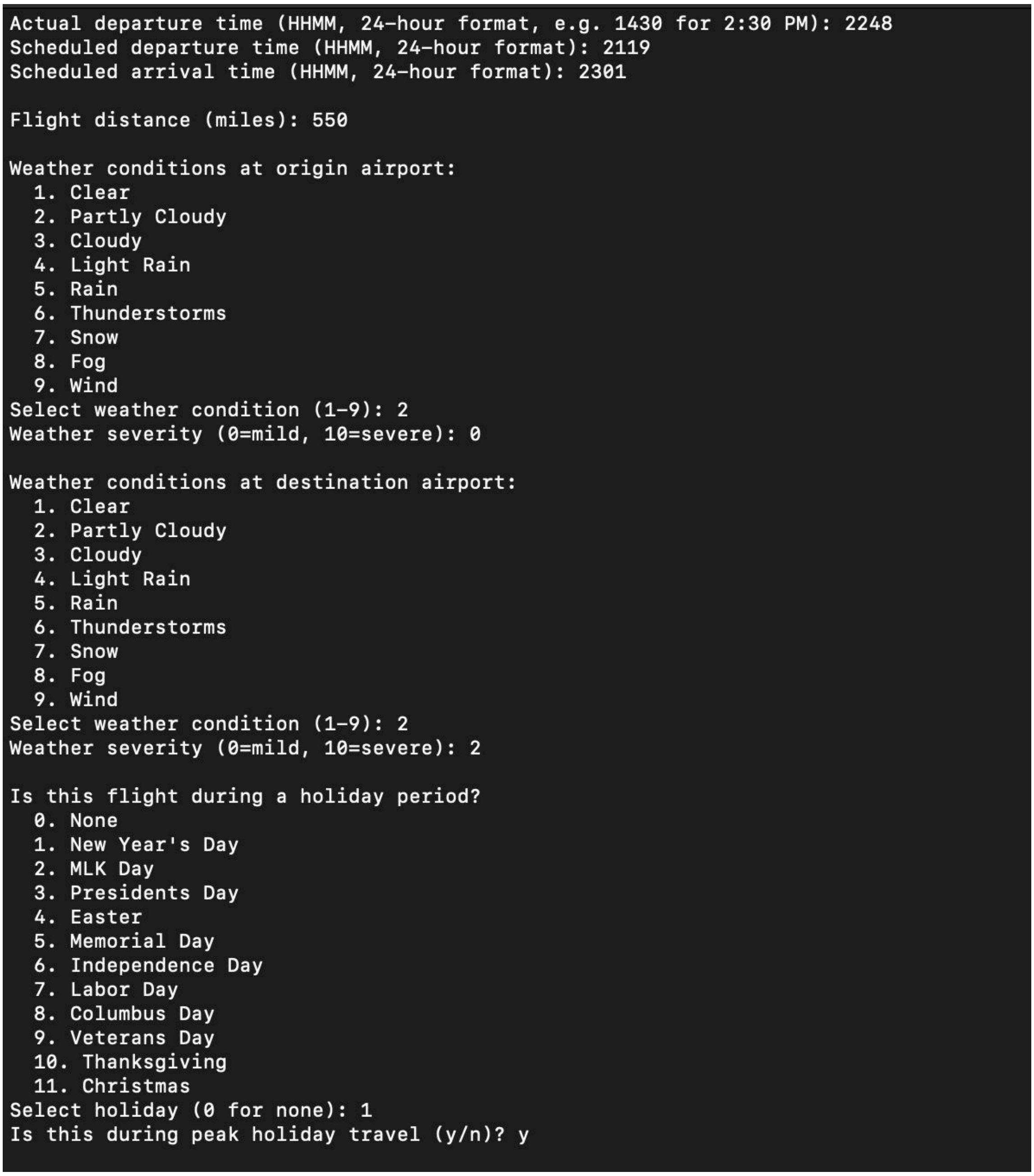


```
=== Top 20 Most Congested Routes ===
             Flight Count
ORIGIN DEST
LAX    SFO           6044
SFO    LAX           6039
HNL    OGG           5815
OGG    HNL           5813
LGA    ORD           5605
ORD    LGA           5604
JFK    LAX           5227
LAX    JFK           5222
DCA    BOS           4912
BOS    DCA           4911
LAS    LAX           4905
LAX    LAS           4905
DEN    PHX           4636
PHX    DEN           4633
LIH    HNL           4323
HNL    LIH           4322
MCO    ATL           4294
ATL    MCO           4293
DEN    LAX           4041
LAX    DEN           4040
```

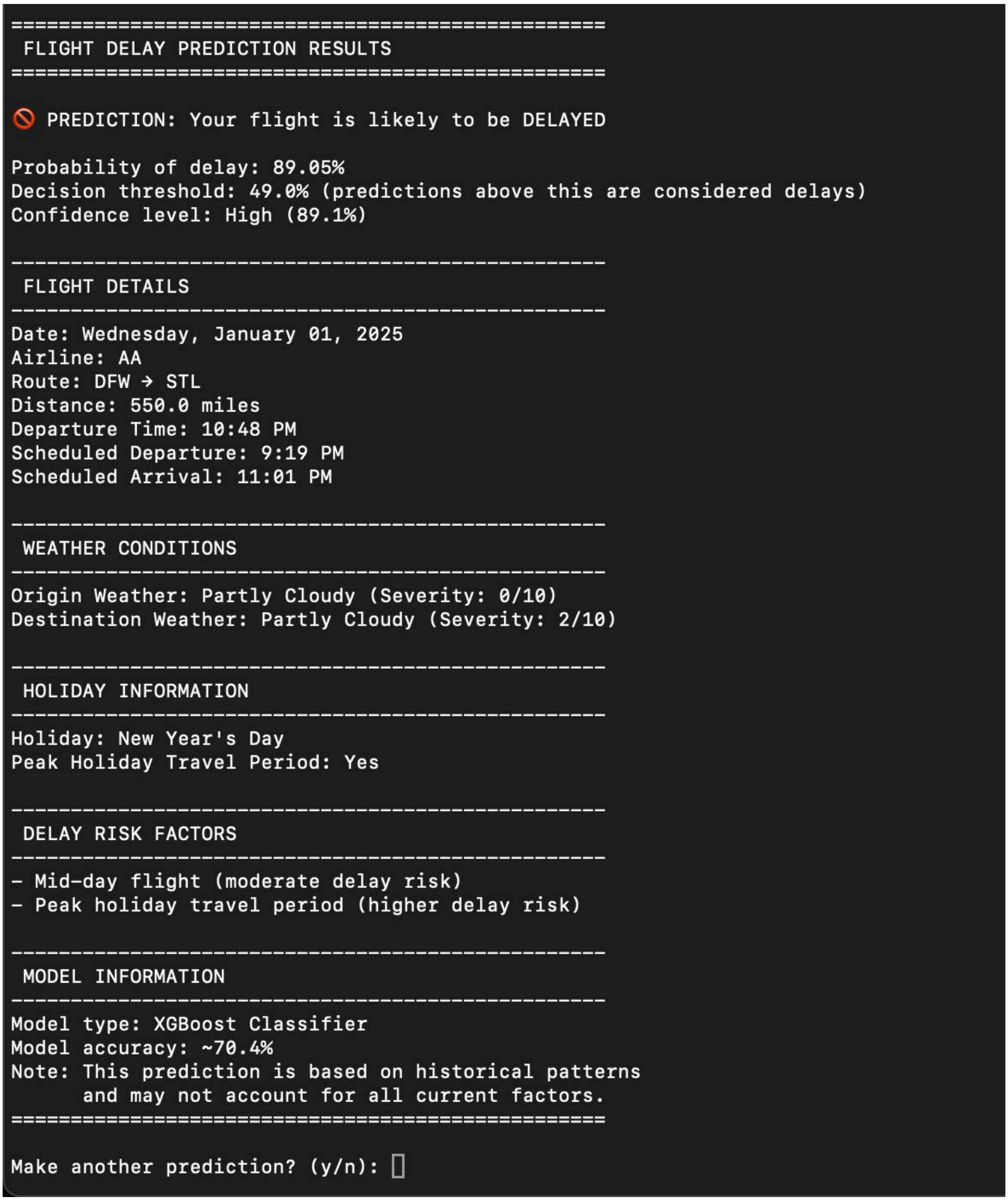
Screenshots of Interactive tool: Case Study: Flight-Level Prediction



```
===== Flight Delay Prediction Tool =====
Please enter the following flight details:

Flight date (YYYY-MM-DD): 2025-01-01

Airline Codes:
  AA: American Airlines
  DL: Delta Air Lines
  UA: United Airlines
  WN: Southwest Airlines
  B6: JetBlue Airways
  AS: Alaska Airlines
  NK: Spirit Airlines
  F9: Frontier Airlines
  HA: Hawaiian Airlines
  G4: Allegiant Air
  9E: Endeavor Air
  OH: PSA Airlines
  YX: Republic Airways
  MQ: Envoy Air
  OO: SkyWest Airlines

Airline code: AA

Flight number: 1010

Airport Codes (Major airports shown below, but any valid code can be entered):
  ATL: Atlanta
  DFW: Dallas/Fort Worth
  DEN: Denver
  ORD: Chicago O'Hare
  LAX: Los Angeles
  CLT: Charlotte
  LAS: Las Vegas
  PHX: Phoenix
  MCO: Orlando
  SEA: Seattle
  MIA: Miami
  IAH: Houston
  JFK: New York JFK
  EWR: Newark
  SFO: San Francisco
  DTW: Detroit
  BOS: Boston
  MSP: Minneapolis
  FLL: Fort Lauderdale
  PHL: Philadelphia
  LGA: New York LaGuardia
  BNA: Nashville
  IAD: Washington Dulles
  DCA: Washington Reagan
  SLC: Salt Lake City
  SAN: San Diego
  MDW: Chicago Midway

Enter any valid airport code. Type 'list' to see all airports.

Origin airport code: DFW

Destination airport code: STL

Destination airport details:
City (e.g., Chicago): St. Louis
State code (e.g., IL): MO
State name (e.g., Illinois): Missouri

Actual departure time (HHMM, 24-hour format, e.g. 1430 for 2:30 PM): 2248
Scheduled departure time (HHMM, 24-hour format): 2119
Scheduled arrival time (HHMM, 24-hour format): 2301

Flight distance (miles): 550
```



```
Actual departure time (HHMM, 24-hour format, e.g. 1430 for 2:30 PM): 2248
Scheduled departure time (HHMM, 24-hour format): 2119
Scheduled arrival time (HHMM, 24-hour format): 2301

Flight distance (miles): 550

Weather conditions at origin airport:
  1. Clear
  2. Partly Cloudy
  3. Cloudy
  4. Light Rain
  5. Rain
  6. Thunderstorms
  7. Snow
  8. Fog
  9. Wind
Select weather condition (1-9): 2
Weather severity (0=mild, 10=severe): 0

Weather conditions at destination airport:
  1. Clear
  2. Partly Cloudy
  3. Cloudy
  4. Light Rain
  5. Rain
  6. Thunderstorms
  7. Snow
  8. Fog
  9. Wind
Select weather condition (1-9): 2
Weather severity (0=mild, 10=severe): 2

Is this flight during a holiday period?
  0. None
  1. New Year's Day
  2. MLK Day
  3. Presidents Day
  4. Easter
  5. Memorial Day
  6. Independence Day
  7. Labor Day
  8. Columbus Day
  9. Veterans Day
  10. Thanksgiving
  11. Christmas
Select holiday (0 for none): 1
Is this during peak holiday travel (y/n)? y
```



```
==================================================
 FLIGHT DELAY PREDICTION RESULTS
==================================================

🚫 PREDICTION: Your flight is likely to be DELAYED

Probability of delay: 89.05%
Decision threshold: 49.0% (predictions above this are considered delays)
Confidence level: High (89.1%)

--------------------------------------------------
 FLIGHT DETAILS
--------------------------------------------------
Date: Wednesday, January 01, 2025
Airline: AA
Route: DFW → STL
Distance: 550.0 miles
Departure Time: 10:48 PM
Scheduled Departure: 9:19 PM
Scheduled Arrival: 11:01 PM

--------------------------------------------------
 WEATHER CONDITIONS
--------------------------------------------------
Origin Weather: Partly Cloudy (Severity: 0/10)
Destination Weather: Partly Cloudy (Severity: 2/10)

--------------------------------------------------
 HOLIDAY INFORMATION
--------------------------------------------------
Holiday: New Year's Day
Peak Holiday Travel Period: Yes

--------------------------------------------------
 DELAY RISK FACTORS
--------------------------------------------------
- Mid-day flight (moderate delay risk)
- Peak holiday travel period (higher delay risk)

--------------------------------------------------
 MODEL INFORMATION
--------------------------------------------------
Model type: XGBoost Classifier
Model accuracy: ~70.4%
Note: This prediction is based on historical patterns
      and may not account for all current factors.
==================================================

Make another prediction? (y/n): 
```