Flight Schedule Optimization Project Report

Introduction

Airports are complex systems that face the challenge of managing hundreds or thousands of flights every day while minimizing delays and congestion. John F. Kennedy International Airport (JFK) is one of the busiest and most complex airspaces globally, making it an ideal case study for optimizing flight schedules.

This project addresses the challenge of optimizing flight schedules to reduce congestion delays at JFK. The solution can also be adapted for other busy airports such as Mumbai or Delhi, which face similar operational constraints.

Data Overview and Preprocessing

Data Source

Used publicly available flight data from the Bureau of Transportation Statistics (BTS) TranStat database, focusing on flights involving JFK as either the origin or destination airport. The dataset includes detailed flight information, such as scheduled and actual departure/arrival times, delays, cancellation flags, and delay reasons.

Ref : https://www.transtats.bts.gov/

Data Cleaning and Transformation

The raw BTS dataset needed significant preprocessing to create a reliable, analysis-ready dataset:

- **Filtering for JFK Flights:** I selected only flights where JFK was the origin or destination to focus the problem.
- **Date and Time Parsing:** The raw flight date (`FL_DATE`) was converted to datetime format. Since scheduled and actual times were provided as integers in HHMM format, I converted these into `timedelta` and then combined them with dates to form proper timestamps (`CRS_DEP_TIMESTAMP`, `DEP_TIMESTAMP`, `CRS_ARR_TIMESTAMP`, `ARR_TIMESTAMP`).
- **Handling Overnight Flights:** Some flights arrive after midnight, changing the calendar day. I corrected for this by adjusting arrival timestamps to the next day if arrival time was earlier than departure time.

Flight Schedule Optimization Project Report

- **Categorical Optimization:** Key categorical fields (`Carrier`, `TailNum`, `Origin`, `Dest`, `Day of Week`) were converted to pandas categories for memory and processing efficiency.
- **Feature Engineering:**
  - A weekend binary flag was created for better modeling of weekly patterns.
  - A delay classification was implemented to categorize departure delays as none, minor, or major.
- **Missing Data:** Missing or invalid values in delay and time columns were safely coerced and replaced to maintain data integrity.
- **Storage:** Cleaned data was saved as a Parquet file for efficient loading in subsequent steps.

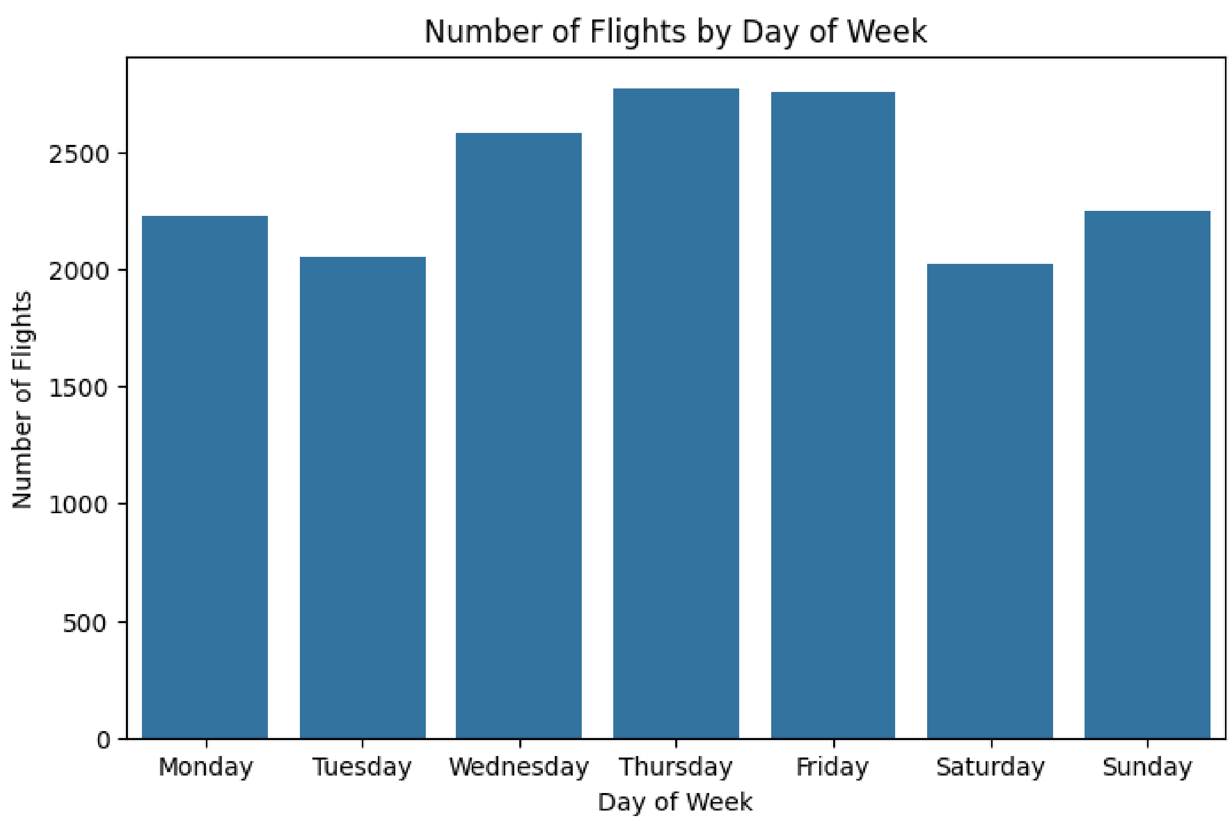
```
                              FL_DATE OP_UNIQUE_CARRIER  \
count                           16652             16652
unique                            NaN                 8
top                               NaN                B6
freq                              NaN              5155
mean    2025-01-15 18:06:18.765313536               NaN
...                               ...               ...
25%               2025-01-08 00:00:00               NaN
50%               2025-01-16 00:00:00               NaN
75%               2025-01-24 00:00:00               NaN
max               2025-01-31 00:00:00               NaN
std                               NaN               NaN

        OP_CARRIER_AIRLINE_ID TAIL_NUM  OP_CARRIER_FL_NUM  ORIGIN_AIRPORT_ID  \
count            16652.000000    16631       16652.000000       16652.000000
unique                    NaN     1504                NaN                NaN
top                       NaN   N3132J                NaN                NaN
freq                      NaN       83                NaN                NaN
mean             20136.375571      NaN        2012.693971       12732.593142
...                       ...      ...                ...                ...
25%              19790.000000      NaN         606.000000       12478.000000
50%              20409.000000      NaN        1438.000000       12478.000000
75%              20409.000000      NaN        2825.000000       13204.000000
max              20452.000000      NaN        8816.000000       15304.000000
std                310.734132      NaN        1792.763569        1124.595634
...
max         23.000000      23.000000          NaN
std          4.877632       5.455280          NaN

[11 rows x 39 columns]
```

Flight Schedule Optimization Project Report

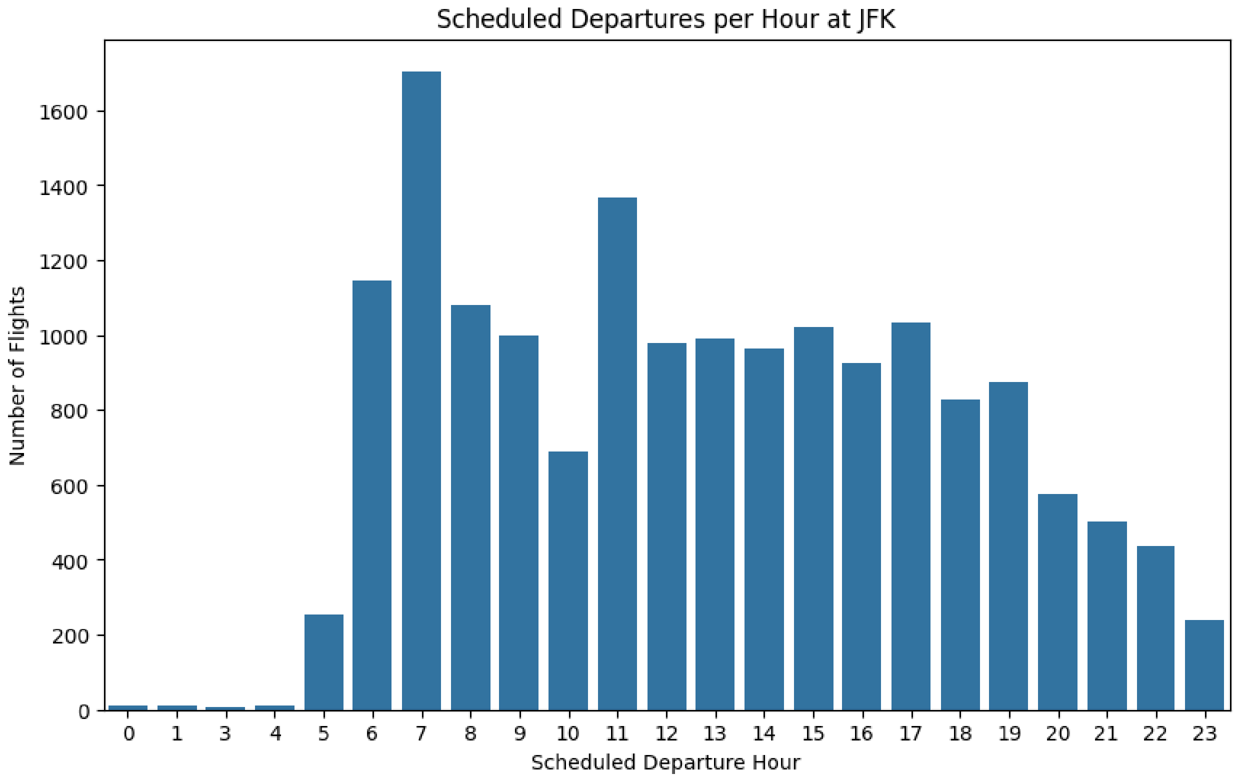
Key Data Statistics

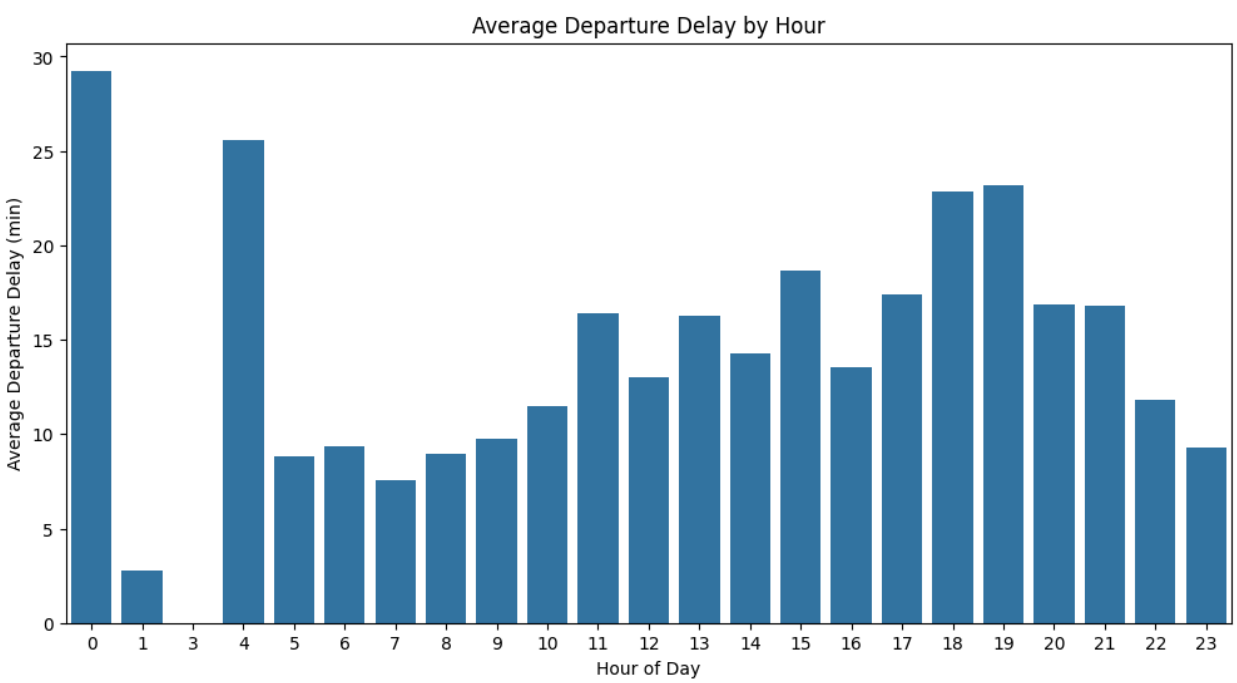
- Total flights involving JFK in January: **16,652**.
- The cancellation rate was about **1.77%**, indicating most flights operate but some are cancelled due to various reasons.
- Departure delays showed a heavily right-skewed distribution, with most flights departing on time or with minor delays, and a few experiencing major delays exceeding 100 minutes.
- Average departure delays varied by hour, with peak delays concentrated in the afternoon and evening hours, consistent with real-world congestion patterns.
- Delay reasons such as carrier issues and late aircraft arrivals contributed the most to total delay minutes.

These analytics guided my approach to focus optimization on smoothing peak congestion times, reducing cascading delays.

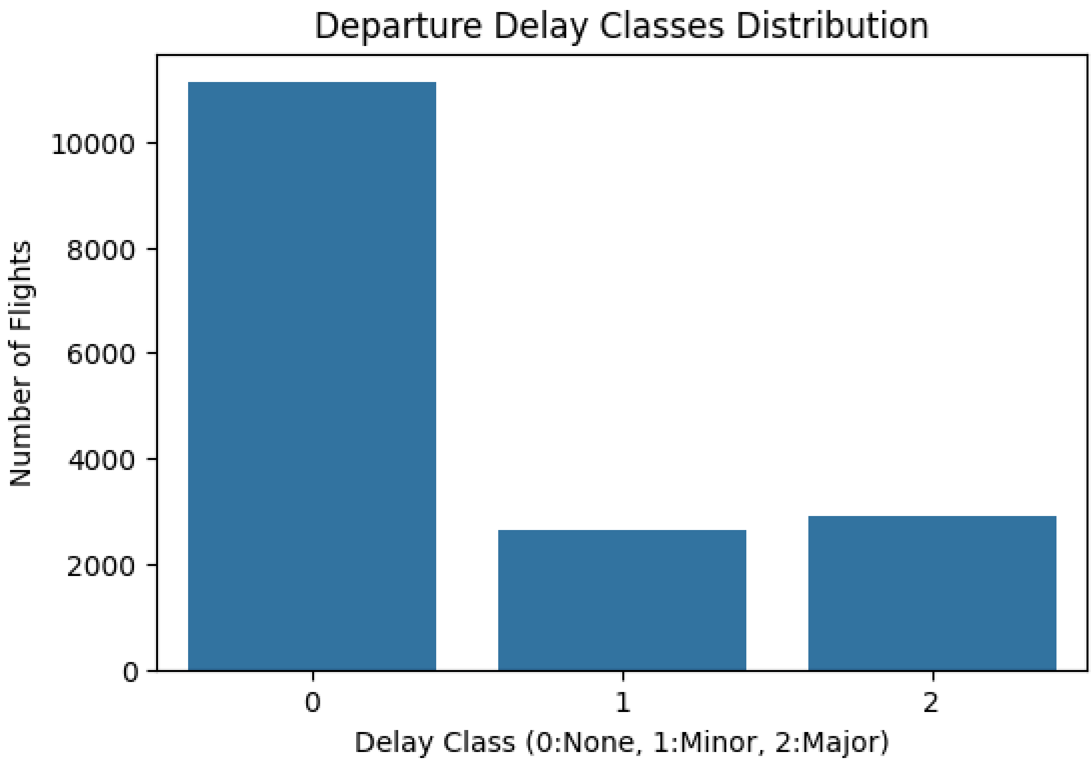


Flight Schedule Optimization Project Report





Flight Schedule Optimization Project Report



Flight Schedule Optimization Project Report

Optimization Problem Definition

Objective

To minimize overall congestion-related delays and distribute flights more evenly within daily operational hours while respecting airport capacity constraints.

Challenges Identified

- **Capacity Limits:** JFK airport operating capacities vary depending on weather conditions and operational modes, with differing arrival and departure throughput per hour.
- **Scheduling Flexibility:** Each flight can be shifted within a +/- 15-minute window to enable better distribution without drastically impacting passenger itineraries.
- **Capacity Modes:** The airport switches between modes (e.g., Visual Arrival Priority, Marginal Weather) affecting capacity. Optimizations need to account for these variable constraints.

Flight Schedule Optimization Project Report

Methodology

Airport Capacity Modeling

integrated the detailed JFK capacity profile from FAA sources, mapping distinct operation modes to their corresponding arrival and departure hourly capacities.

Scheduler Design

- Flight departure times were bucketed in 5-minute intervals.
- Each flight was assigned feasible alternate buckets within a ±15-minute window.
- Capacity constraints enforced maximum allowed flights per bucket based on mode-specific capacity.
- The objective minimized the total deviation of optimized departure times from original schedules.

Optimization Approach

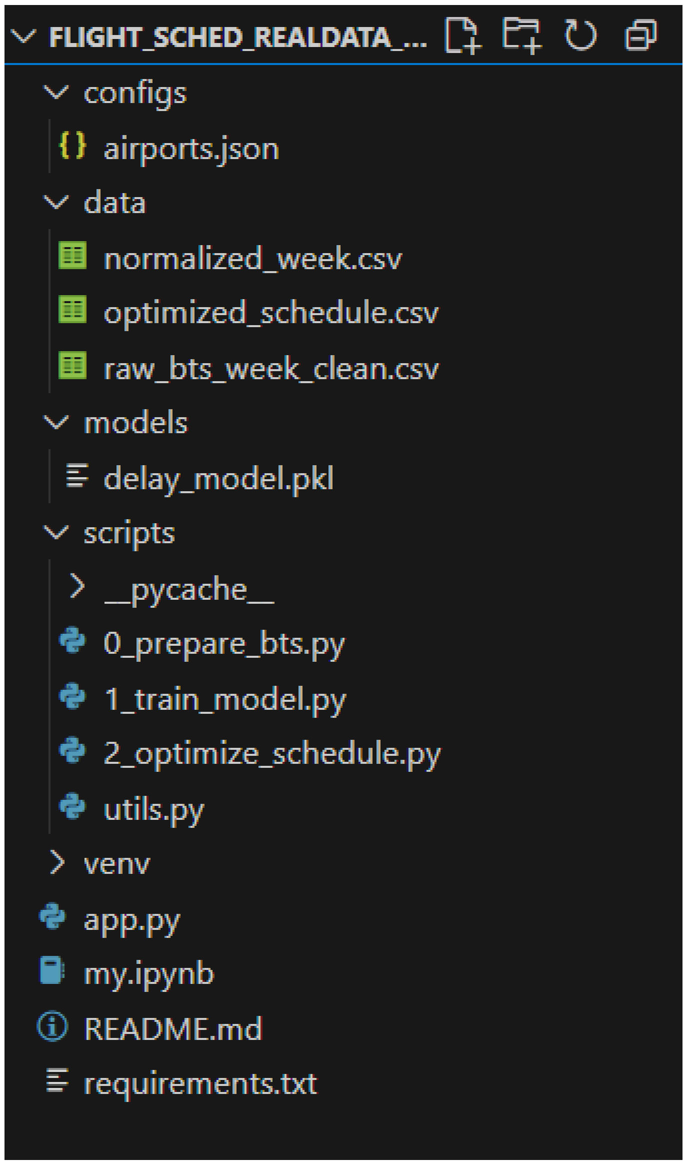
Used Google OR-Tools CP-SAT solver because it efficiently handles combinatorial problems with constraints:

- Decision variables represented flight assignment to specific buckets.
- Constraints enforced one bucket assignment per flight.
- Bucket capacities ensured airport throughput limits were respected.
- Objective function minimized total scheduling deviation.

Fallback greedy heuristics were used when solver was unavailable to ensure robustness.

Flight Schedule Optimization Project Report

Implementation Details



Flight Schedule Optimization Project Report

Key Code Highlights

Github:https://github.com/Hemanth-044/Flight-Schedule-Optimization

- **Data Preprocessing:** (0_prepare_bts.py)

  Converts raw times to timestamps, corrects overnight flights, encodes categorical features, and prepares delay classes.

- **Training Model:** (1_train_model.py)

  Uses gradient boosting regression on extended features including categorized delays.

- **Optimization:** (2_optimize_schedule.py)

  Uses cp_model to assign flights to 5-min buckets within ±15 minutes, respecting capacity constraints dynamically selected by capacity mode.

- **Visualization:** (app.py)

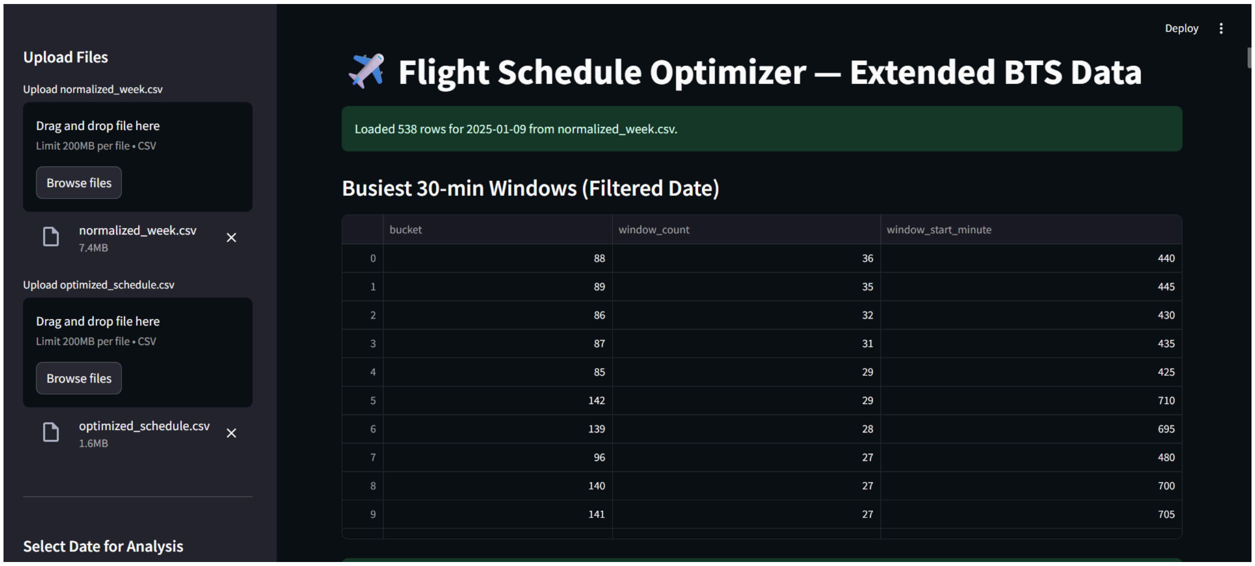
  Interactive choice of capacity modes, date filtering, crash-free loading. Visualizes congestion before/after, delay distributions, delay reasons, cancellations, and optimized schedules.

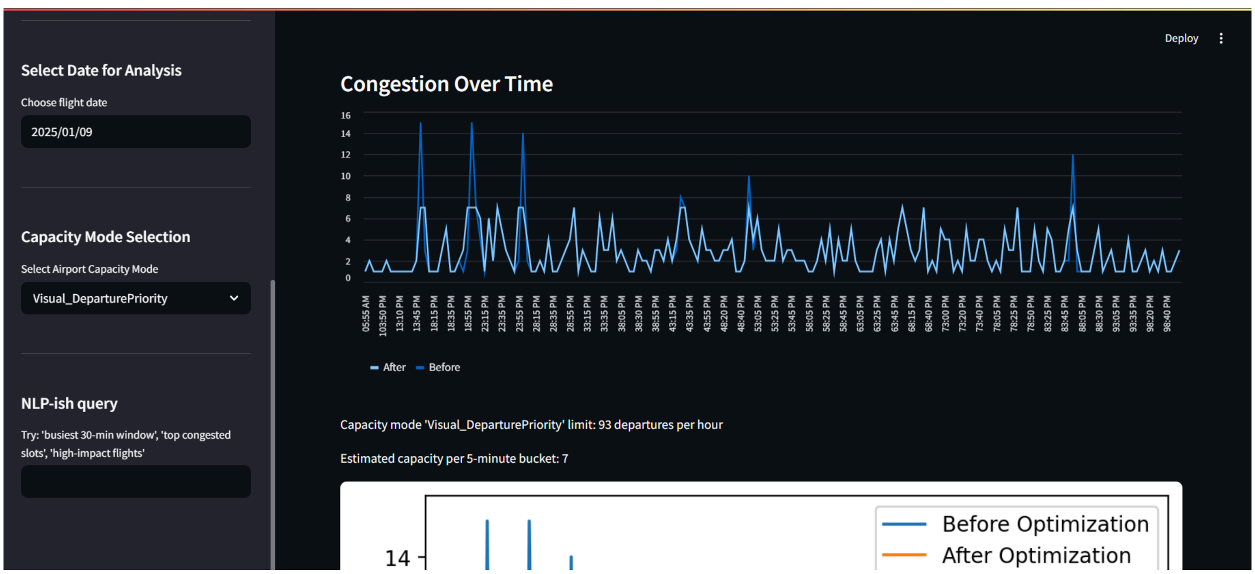
Flight Schedule Optimization Project Report

Results Summary and Insights

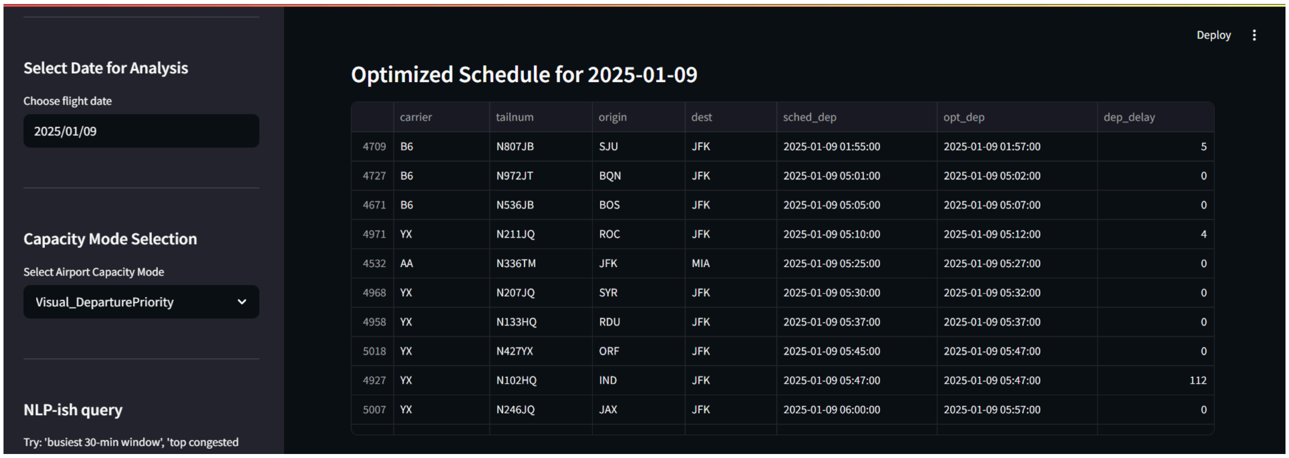
- Trained a machine learning model to predict departure delays using extensive historical flight data.
- The model achieved a test **mean absolute error (MAE) of 6.80 minutes**, demonstrating accurate and reliable delay predictions.
- These predictions support effective scheduling optimization by enabling prioritization and retiming of flights with higher delay risk.
- The optimization significantly smooths congestion **peaks across 5-minute departure buckets**, reducing peak loads.
- Most flights are shifted only moderately in departure time, balancing operational feasibility with passenger convenience.
- Capacity mode selection allows simulation of different real-world weather and operational scenarios, reflecting airport constraints realistically.
- Visualizations of delay reasons and cancellations provide actionable insights to diagnose root causes and guide further improvements.
- The modeling framework and optimized schedules generalize well and can be adapted to other busy airports like Mumbai and Delhi by updating capacity profiles in `configs/airports.json`.

Flight Schedule Optimization Project Report





Flight Schedule Optimization Project Report



References:

1. Bureau of Transportation Statistics (BTS) TranStat Database
   The primary source of flight data used in preprocessing and analysis.
   URL: https://www.transtats.bts.gov/
2. FAA Capacity Profiles for JFK Airport
   Official FAA reports provide detailed capacity constraints and operational mode information used in modeling.
   Reference: FAA John F. Kennedy International Airport Capacity Profile, 2014.
3. Google OR-Tools Constraint Programming Solver
   The optimization engine used for schedule assignment under capacity constraints.
   URL: https://developers.google.com/optimization

4. Machine Learning Techniques for Delay Prediction
   Gradient Boosting Regressor and related regression techniques for flight delay modeling.
   Reference: Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 29(5), 1189-1232.
5. Streamlit for Interactive Data Visualization
   The framework used to develop the dashboard and NLP-powered insights.
   URL: https://streamlit.io/
6. Related Literature on Airport Scheduling Optimization
   Examples of modeling approaches and optimization for airline scheduling and delay mitigation.
   - Kanamori, A., et al. (2005). Airport ground delay program optimization.
   - Bertsimas, D., & Ruan, D. (2019). Predictive and prescriptive methods for delay reduction.
7. Data Science and Delay Analysis Techniques
   Methods for time-series analysis, delay cause categorization, and modeling cascading disruptions in airport operations.