Link to presentation video: youtu.be/uNt4rg_cPRc
Link to github repository with interactive components

Delay-Aware Itinerary Planning via Predict-Then-Optimize Framework

Annie Chen, Ashlee Liu, Sukhman Sidhu, Kenny Wongchamcharoen, Charmaine Yuen

Abstract

Travelers frequently face disruptions due to delays and cancellations, yet most trip planners rely only on static flight durations and costs. We aim to build a tool that predicts risks associated with different flights and chooses routes that minimize expected delays and cancellations. Our system serves risk-aware travelers, airlines optimizing connections, and logistics planners under time and budget constraints. Inspired by the "predict-then-optimize" framework (Grigas et al., 2018), we develop a delay-aware itinerary optimization model that combines machine learning and operations research. Using historical flight data from an airline delay dataset, we create a pipeline that predicts expected travel delays and incorporates them into an optimization model to construct robust travel itineraries. We applied machine learning techniques to predict delays and cancellations, embedding these predictions into a simple network flow, single-stage stochastic program formulated in Pyomo, to select an itinerary that minimizes expected cost, travel time, and cancellation risks. This project demonstrates how predictive analytics and optimization can be combined to support real-world travel planning under uncertainty.

Dataset and Processing

This tabular dataset offers a comprehensive overview of flight arrival and delay statistics for U.S. airports, segmented by airline carriers. It covers flights from August 2013 to August 2023, with data aggregated monthly for each unique combination of year, month, airport, and carrier. Each entry includes detailed metrics such as the total number of arriving flights, delays exceeding 15 minutes, as well as the number of cancellations and diversions. Additionally, it breaks down delay causes into categories such as carrier-related issues, weather, the National Airspace System (NAS), security, and delays due to late-arriving aircraft.

Key fields include carrier and airport codes and names, along with a range of performance and delay indicators. Its structured nature allows for targeted exploration of airline and airport performance trends, and supports analysis of the underlying factors contributing to flight delays in the aviation sector. The Dataset schema is shown in Appendix.

The dataset initially contained 200,000+ rows and 130+ airlines, many of which are regional carriers, cargo operators, or codeshare duplicates. These carriers often:

- Do not serve major hubs regularly
- Have very small sample sizes (low arr_flights)
- Introduce noise and outliers into machine learning models and optimization

To ensure better reliability and interpretability, we focused on the top 7 commercial U.S. carriers:

1. United Airlines (UA)
2. Delta Air Lines (DL)
3. American Airlines (AA)
4. Southwest Airlines (WN)
5. Alaska Airlines (AS)
6. JetBlue Airways (B6)
7. Frontier Airlines (F9)

To further reduce our dataset for sharper analysis and more accurate models, we also constrained our geographic scope to realistic travel corridors — specifically, flights that originate from major cities on the West Coast - Seattle, The Bay Area and Los Angeles and end in the greater NYC area. This reflects a common travel use-case where travelers are evaluating options among major carriers flying between two dense metro regions. Instead of modeling all U.S. airports, which would introduce unnecessary complexity and sparse data, we only retained airports that met one of the following criteria:

1. Major layover hubs (e.g., DEN, DFW, ORD)
2. Airports along the likely path of cross-country flights
3. Starting and ending cities as described in the above geographic constraints

This focused subset allowed us to build models that are both accurate and practically useful, while still capturing a wide range of operational behaviors and delay types. The resulting list of selected airports with our rationales is shown in table 1.1. Using this scoping methodology, we significantly reduced the dataset to a more manageable and relevant subset of 12,742 rows, each representing a valid airport-carrier combination with sufficient flight volume for meaningful analysis.

Supervised Learning - Predicting Average Flight Delay time

This section outlines how we used supervised learning to predict average delay time based on airport and carrier-level flight performance data. Each row in our processed dataset includes:

- arr_flights: Number of arriving flights
- arr_delay: Total delay time
- arr_cancelled: Number of cancellations
- delay breakdowns: carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay

These values are aggregated, so each row represents summary statistics rather than individual flights. Since the dataset is aggregated, we derived **per-flight metrics** as modeling targets:

- **Average Delay**: avg_delay = arr_delay / arr_flights
- **Cancellation Rate**: cancel_rate = arr_cancelled / arr_flights

These were computed after removing rows where arr_flights = 0 to avoid divide-by-zero errors. We also imputed missing values (e.g., from airports with low monthly traffic) using median imputation for robustness. We trained models for the main prediction task of expected delay time in minutes; our target avg_delay. We also conducted exploratory data analysis to analyze contribution to delay by type with results shown in Figure 1.2 .

From these results, we decided to exclude irrelevant columns such as year, month, and the target (arr_del15) from the feature matrix, and select delay-related features based on their contributions. Late Aircraft Delay, Carrier Delay, and NAS Delay together accounted for ~95% of total delay time. Whereas security_ct and security_delay features have negligible contribution (<1%), so we dropped them from our analysis. We experimented with multiple ML models:

1. Random Forest

Random Forest (RF) was chosen as flight delay data is inherently non-linear, noisy, and influenced by many factors such as weather, airport congestion, and airline operations. RF is well-suited to this because it can model complex interactions and handle both low- and high-variance features without requiring feature scaling. RF is relatively robust to outliers and multicollinearity, both of which are common in aggregated airline performance data. For hyperparameter tuning, we use RandomizedSearchCV with 30 random combinations. Results are shown in Figure 1.3. The result was impressive, with R^2 ~0.9064 and MSE ~6.05. From Figure 1.3, RF generalized reasonably well for typical delays but slightly underpredicted rare large delays.

2. Gradient Boosted Trees (XGBoost)

To better capture complex patterns and rare large delays, we applied Gradient Boosted Trees using XGBoost as our second model. Additionally, due to the presence of large delays, we applied a log-transform to the target avg_delay to compress extreme delay values and stabilize model training in a hope of improving prediction of rare events. For hyperparameter tuning, we use RandomizedSearchCV with 30 random combinations as well. Results are shown in Figure 1.4. The test set significantly improved from RF, with R^2 ~0.9746 and MSE only ~1.64. From Figure 1.4, we can see tighter clustering around the ideal 1:1 line compared to previous models. From Figure 1.5, we decided to choose XGBoost Regressor with a log-transformed target as our final model. We ran this on our data, preprocessed them for our optimization model in Section 7.

From Figure 1.6 & 1.7, we found that Carriers like JetBlue (B6) and Frontier (F9) exhibit the highest average predicted delays. Alaska (AS) and Delta (DL) tend to have lower and more consistent delays. In terms of Airport-Level Delay Patterns, certain airports such as EWR and DFW have higher

predicted delays, while SJC and OAK are among the lowest. This suggests that both carrier and route trajectory (via airport hubs) significantly affect delays.

Supervised Learning - Predicting Airline Cancellation Risk

To predict whether a group of flights exhibits a high cancellation rate (>5%), we trained and evaluated a RF classifier using historical delay and cancellation data on major U.S. airlines and airports.

Data and Preprocessing

- Target variable: Binary flag where
  - 0 = Normal operation (less than or equal to 5% cancellations)
  - 1 = Problematic (greater than 5% cancellations)
- Features:
  - Categorical: carrier, airport (One-Hot Encoded)
  - Numerical: flight counts and delay types

We further applied the mean to missing numerical values in order to keep the data balanced. The performance of our model was evaluated using a confusion matrix, classification report, and ROC curve.

1. Default Threshold Results

- Accuracy: 93%
- Precision (Class 1 - Canceled): 0.30
- Recall (Class 1 - Canceled): 0.02
- F1-Score (Class 1 - Canceled): 0.03

While the overall accuracy is high (93%), this is misleading due to the severe class imbalance. The model performs very poorly on the minority class (flights with high cancellation rates), achieving only 2% recall for class 1. It essentially classifies most samples as non-canceled.

2. Optimized Threshold Results

- Accuracy: 57%
- Precision (Class 1 - Canceled): 0.12
- Recall (Class 1 - Canceled): 0.82
- F1-Score (Class 1 - Canceled): 0.21

After applying a more aggressive threshold by maximizing the difference between true positive rate and false positive rate, the model significantly improves its ability to identify high-risk groups (recall increases to 82%). However, this comes at the cost of much lower precision and overall accuracy, indicating more false positives (i.e., flagging low-risk groups as high-risk).

The ROC curve we generated in Figure 2.1 further shows a reasonable separation between the two classes with an AUC of 0.74, suggesting that our model has decent ability in discriminating between positive and negative cases. We then computed cancellation probabilities for all flight groups and aggregated these predictions by (carrier, airport) pairs. The result is a ranked list of flight groups by their average predicted cancellation risk, highlighting the combinations most likely to experience frequent disruptions.

Unsupervised Learning - Clustering Analysis (Identifying Delay Risk Profiles)

To uncover patterns in flight delays, we applied unsupervised learning to group similar flights and airport-carrier pairs based on their delay behavior. Importantly, we withheld the known dominant delay cause from the dataset during clustering and later used it for post-hoc validation to assess the quality and interpretability of the discovered clusters.

1. Clustering by Delay Composition

We focused on five specific causes for delays, including carrier, weather, NAS, security, and late aircraft delay. For each flight, we calculated the total contribution of each cause to the total delay and used those proportions to cluster together similar flights.

Using the Elbow Method and Silhouette Analysis, we determined that 4 clusters best captured the variation in delay patterns. Each cluster reflects a distinct type of delay behavior:

- Cluster 0: Late aircraft–dominated delays
- Cluster 1: Mixed cause delays
- Cluster 2: NAS-focused disruptions
- Cluster 3: Weather-heavy profiles
- Cluster 4: Carrier-dominated delays

We validated the clusters by looking at the dominant cause of delay for each flight, and the clusters aligned very well. We summarize the distribution of delay shares across clusters with a heatmap in Figure 3.1 and 3.2.

To quantify this alignment, we mapped each KMeans cluster to its most common true cause and computed the classification accuracy. The resulting clustering accuracy was 79% indicating that the unsupervised model effectively recovered dominant delay types without access to labels during training.

2. Principal Component Analysis (PCA) Visualization

We used PCA to visualize the clusters better by reducing the data to 2D and 3D. The top two components explained ~59% of the variance.

- The 2D (Figure 3.3) and 3D (Figure 3.4) scatter plots show clear separation between clusters.
- Flights from routes like ATL–DL and DFW–AA appeared as outliers with consistently high delays, making them easy to flag for risk analysis.

3. Clustering by Airport-Carrier Delay Statistics

In addition, we categorized the airport-carrier pairs based on three features: average arrival delay, standard deviation of delay, and number of flights.

Applying K-Means with 4 clusters, we assigned each pair a risk level:

- Low Risk
- Moderate Risk
- High Risk
- Very High Risk

This provides an interpretable way to group routes based on performance and variability. PCA plots(Figure 3.5) and boxplots (Figure 3.6) both show that higher risk groups had larger delays and variances in delays, crucial to flag in the optimization phase.

Optimization

From the above, we obtained a prediction of average flight delay time and airline cancellation risk to use for the following optimization modelling which helps travellers avoid unwanted waiting and changes during travelling. The result is consolidated into a new dataset called legs with [airline, origin, delay, cancellation risk, destination].

Assumptions:

1. Average flight delay time and airline cancellation risk predictions are assumed to be accurate and static for optimization purposes. The prediction follows the assumption from the above data analysis, neglecting destination-specific weather, congestions etc.
2. Region-to-region cost is assumed based on distance. Therefore, the price of plane tickets is independent of the airline.
3. All combinations of origin-destination operated by the same airline are considered valid, excluding self-loops (i.e., flights that start and end at the same airport).
4. All transfers are assumed logistically feasible regardless of time between flights for simplicity.

Decision Variable: $x_{ij} = \{1\ if\ leg\ i \in Leg\ is\ selected,\ \ 0\ otherwise\}$ (binary variable)

Parameters:

$delay_i\ (i \in Leg)$: $Predicted\ delay\ in\ minutes\ associated\ with\ leg\ i$

$cancel_i\ (i \in Leg)$: $Predicted\ cancellation\ risk\ as\ a\ probability\ associated\ with\ leg\ i$

- Delay and cancel are based on our machine learning model.

$cost_i\ (i \in Leg)$: $Sampled\ cost\ in\ USD\ for\ leg\ i$

- Assumed based on distance. Sample stochastic cost and truncate below $50 is applied for randomness, assuming the cost follows normal distribution with the specified mean and standard deviation.
- Specific airport is inputted by the user. Cost is determined by the region that the airport is located in.

$origin_i\ (i \in Leg)$: $Origin\ Airport\ for\ leg\ i$

$destination_i\ (i \in Leg)$: $Destination\ Airport\ for\ leg\ i$

- Chosen i needs to have the origin and destination match with the user's input.

Objective Function:

$$min \sum_{i \in Leg} x_i\ (delay_i + \lambda \cdot cancel_i + \mu \cdot \frac{delay_i}{cost_i}),\ \text{where}\ \lambda = 100\ \text{and}\ \mu = 1$$

The weight between delay in minutes and flight cancellation is 1:100. It represents that flight cancellation is considered as extremely undesirable. Delay-to-cost ratio is considered to encourage cost-efficient reliability. It maintains a balance between time and cost.

Constraints:

1. Flow Rule (ensure that the chosen flight is matching the desirable origin and destination)
   For all airports $a \in A$ and $i \in Leg$:

$$\sum_{dest(i) = a} x_i - \sum_{origin(i) = a} x_i = \{-1\ if\ a = origin,\ 1\ if\ a = destination,\ 0\ otherwise\}$$

2. At least one flight must be chosen (ensure an answer is given to the user): $\sum_{i \in Leg} x_i \geq 1$

3. Budget Rule (ensure user can pay for the plane ticket): $\sum_{i \in Leg} cost_i \cdot x_i \leq budget$

Interactive User Input:

We designed an interactive user interface to provide the most desirable (minimum delay time and cancellation risk) based on user requirements. The user is guided to input origin, destination, and budget. The optimization model outputs a summary of their best route option, including the airline, estimated delay time and cancellation risk, and sample cost. It also compares the chosen route and the average options and provides a sample of applying our project into practical usage, illustrated in Figure 4.1.

Takeaways

Our itinerary optimization model successfully demonstrates the integration of predictive analytics and optimization for travel planning under uncertainty. The application of XGBoost for delay prediction and Logistic regression for cancellation risk estimation resulted in robust and accurate predictions, while the clustering analysis identified distinct delay profiles. The model's practical application to real-world travel scenarios highlights its potential to assist risk-aware travelers and logistics planners in making data-driven itinerary decisions.

Appendix

Dataset schema

- year: The year of the data.
- month: The month of the data.
- carrier: Carrier code.
- carrier_name: Carrier name.
- airport: Airport code.
- airport_name: Airport name.
- arr_flights: Number of arriving flights.
- arr_del15: Number of flights delayed by 15 minutes or more.
- carrier_ct: Carrier count (delay due to the carrier).
- weather_ct: Weather count (delay due to weather).
- nas_ct: NAS (National Airspace System) count (delay due to the NAS).
- security_ct: Security count (delay due to security).
- late_aircraft_ct: Late aircraft count (delay due to late aircraft arrival).
- arr_cancelled: Number of flights canceled.
- arr_diverted: Number of flights diverted.
- arr_delay: Total arrival delay.
- carrier_delay: Delay attributed to the carrier.
- weather_delay: Delay attributed to weather.
- nas_delay: Delay attributed to the NAS.
- security_delay: Delay attributed to security.
- late_aircraft_delay: Delay attributed to late aircraft arrival.

Supervised Learning - Predicting Average Flight Delay time

Region	Airports to Include	Why Include
West Coast	SFO, OAK, SJC, LAX, SEA	Starting points + Bay Area/West Coast hubs, SEA hub for Alaska
Rockies	DEN, PHX, SLC	Common layovers from SFO; Mountain West connectors
Midwest	ORD, DFW, MSP, STL	United, American hubs; central U.S. connection points
East Coast	JFK, LGA, EWR, CLT	Final destination cities + New York alternatives
Southeast	ATL	Major Delta hub; strong East–West linkage

Table 1.1: Selected airports and rationales

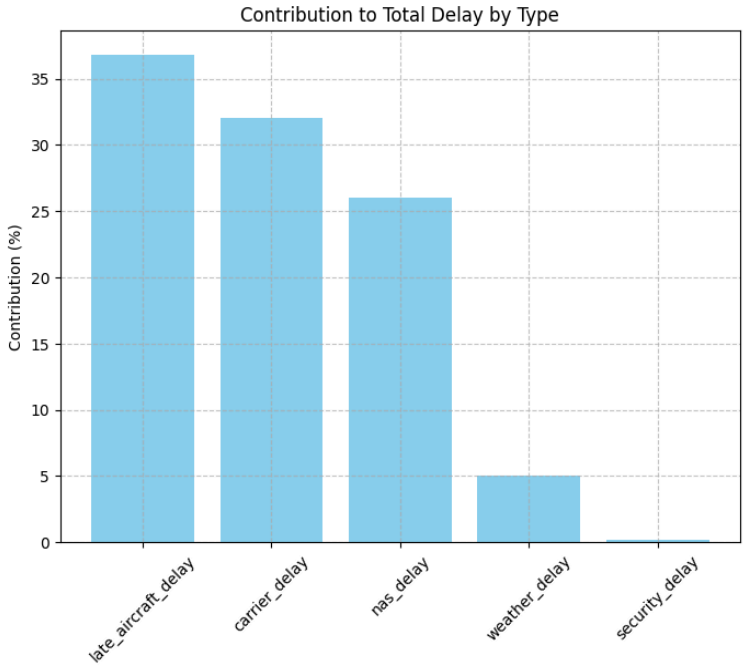


Figure 1.2: Contribution to Total Delay by Type

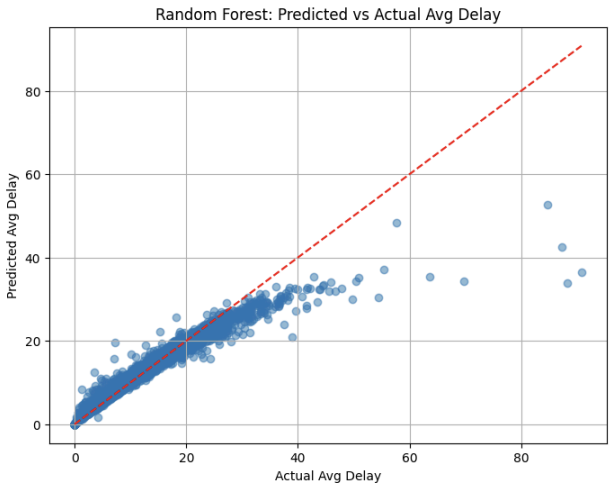


Figure 1.3: Predicted vs Actual average delay for Random Forest

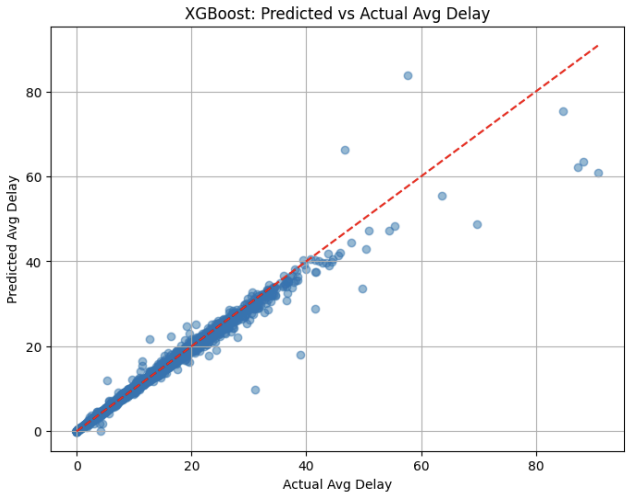


Figure 1.4: Predicted vs Actual average delay for XGBoost

Model	Test R^2	Test MSE	Notes
Linear Regression	~0.20	~47.55	Weak baseline, no complex patterns captured
Random Forest	~0.92	~4.83	Good for common delays but struggles with rare events
XGBoost (log-transformed target)	~0.98	~1.01	Best tradeoff between bias and variance

Figure 1.5: Summary of results from all models

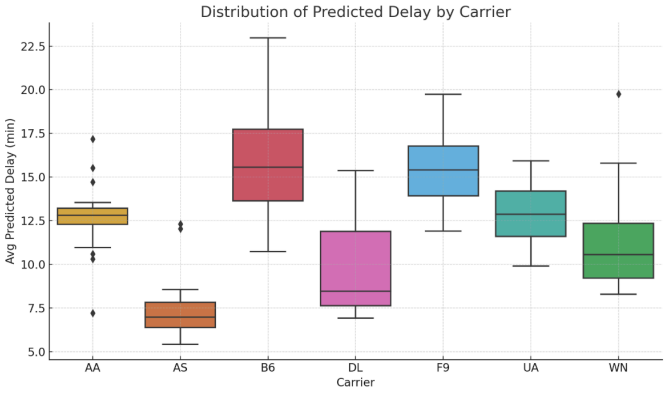


Figure 1.6: Boxplots of Predicted Delay by Carrier

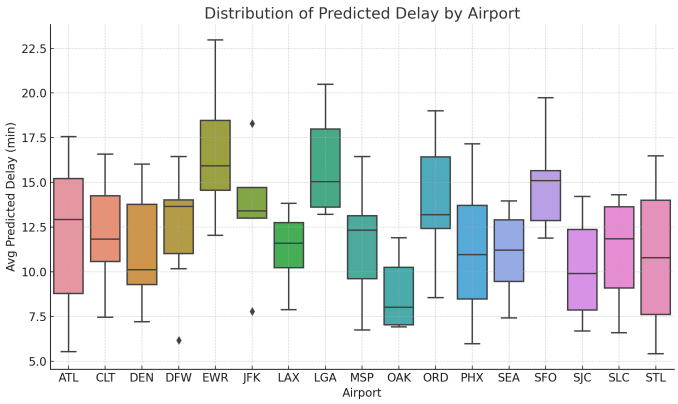


Figure 1.7: Boxplots of Predicted Delay by Airport

Supervised Learning - Predicting Airline Cancellation Risk

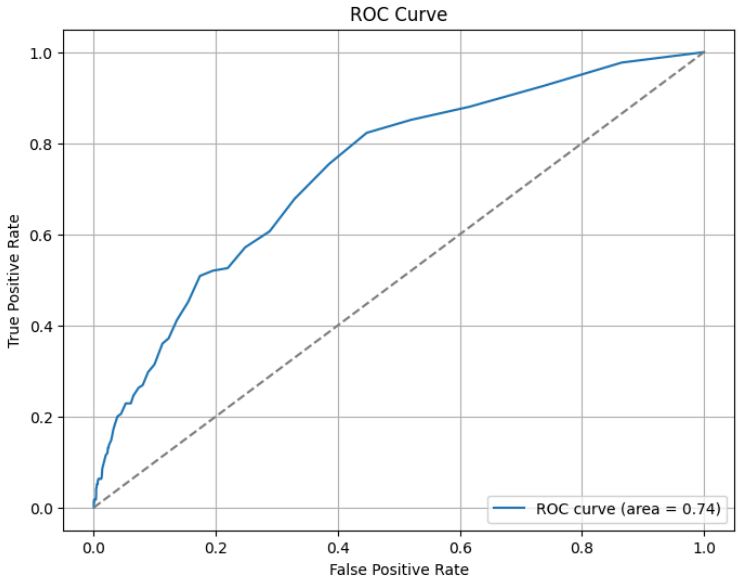


Figure 2.1: ROC Curve for Cancellation Risk

Unsupervised Learning

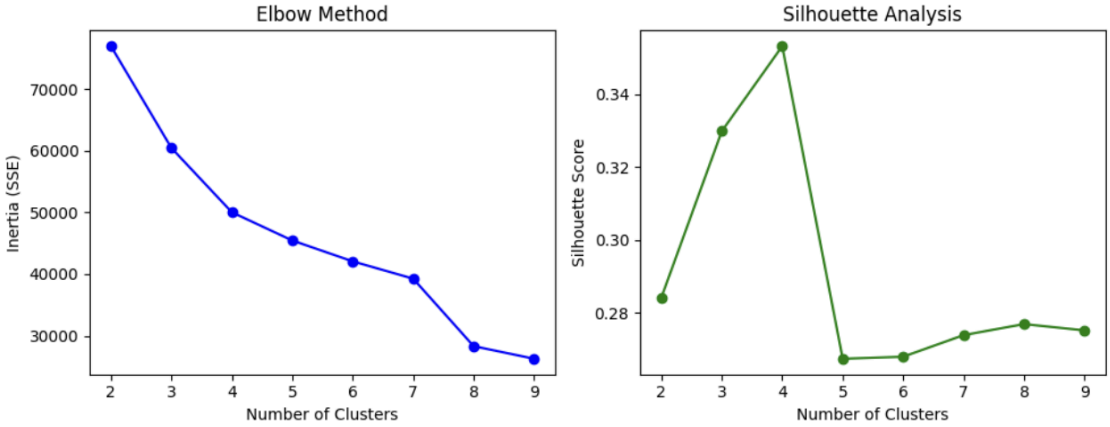


Figure 3.1: Cluster Selection Using Elbow Method and Silhouette Analysis

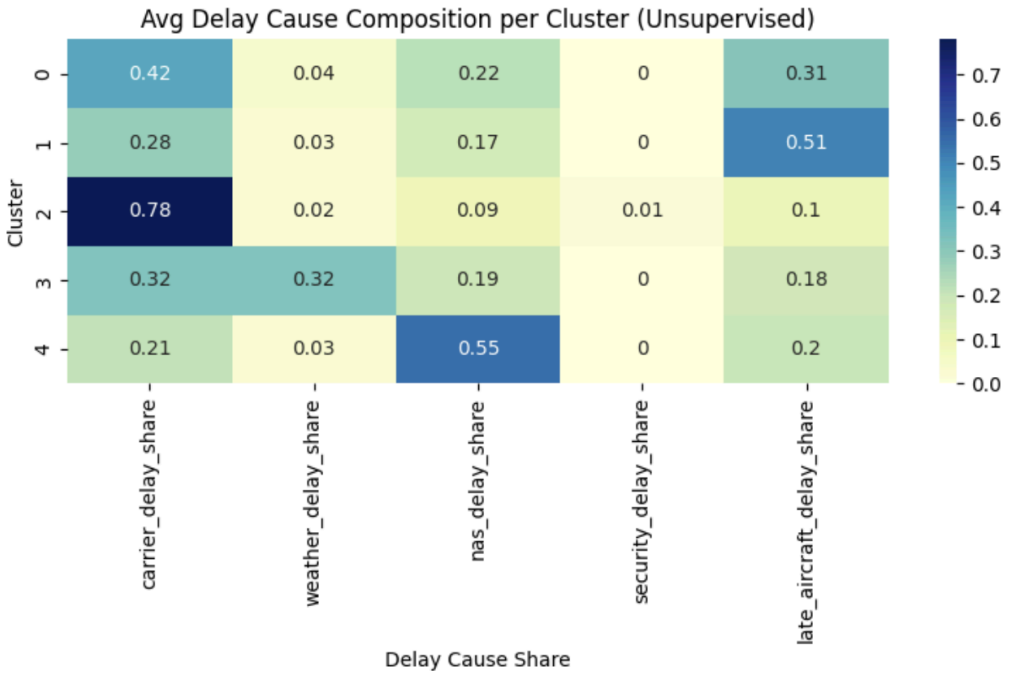
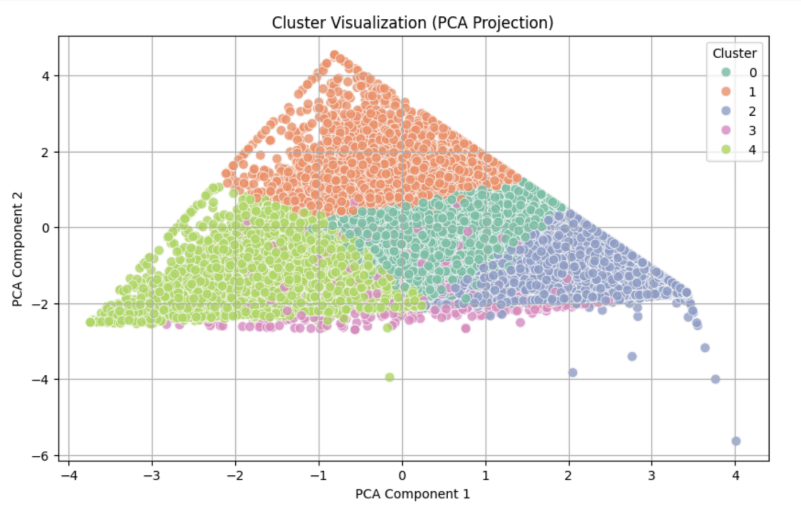


Figure 3.2: Average Delay Cause Share per Cluster




```
print(f"PCA explained variance ratio: {pca.explained_variance_ratio_}")
```

```
PCA explained variance ratio: [0.32041334 0.27191312]
```

Figure 3.3: 2D Scatter Plot for Principal Component Analysis

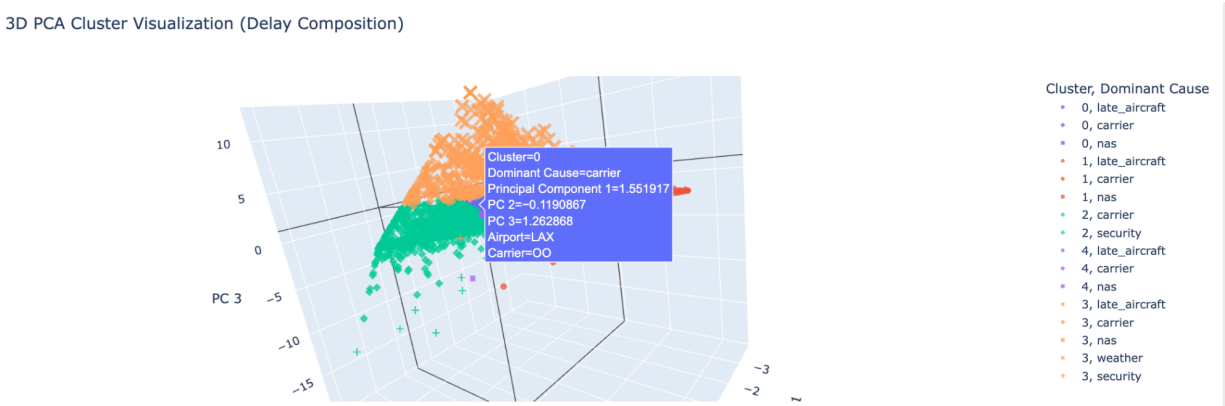


Figure 3.4: 3D Scatter Plot for Principal Component Analysis

Figure 3.5: 2D PCA Visualization of Clusters by Risk Level with Outlier Airports Highlighted

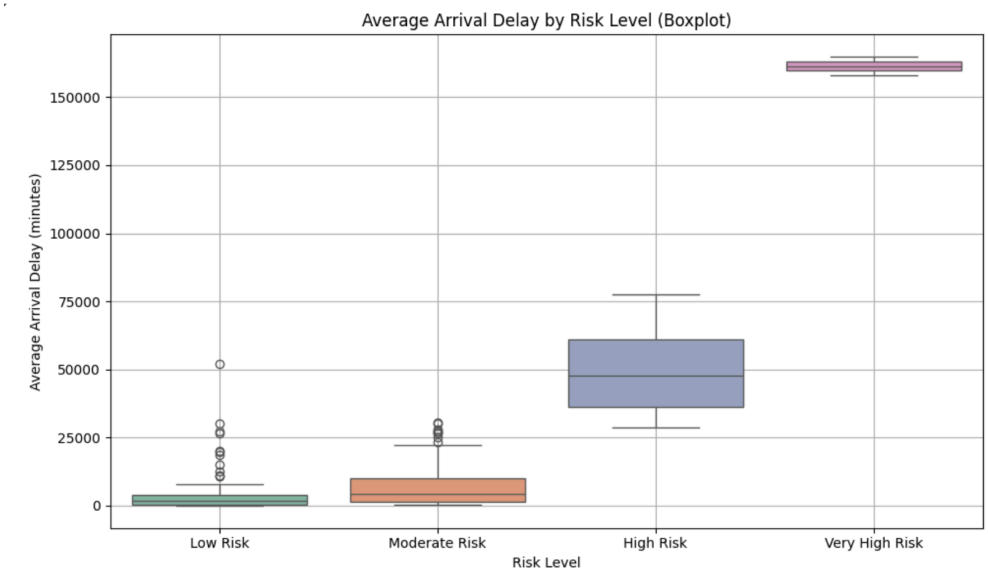


Figure 3.6: Boxplot for Distribution of Average Arrival Delays Across Risk Levels

Optimization

```
🌍 Welcome to the Flight Itinerary Optimizer!
📍 Enter your origin airport code (e.g., SFO): SFO
🧭 Enter your destination airport code (e.g., LAX): JFK
💰 What's your maximum total budget in USD? 800

🧠 Objective: Minimize delay + λ(100) × cancellation risk + μ(1) × delay/cost
💸 Budget constraint: Max $800.00

🧮 Calculating the best itinerary for you... Please wait ✈️💨

✅ ✈️ Optimal Route Found!
    airline origin destination      delay  cancel_risk  cost_sample
935      DL    SFO         JFK  11.880899      0.01719   584.938373

📊 Summary:
💸 Total cost (sampled): $584.94
⏱️ Total predicted delay: 11.9 minutes
⚠️ Total cancellation risk: 0.02
📉 Total delay-to-cost ratio (∑ delayᵢ / costᵢ): 0.020

🎉 You saved approximately:
   ⏳ 0.4 fewer minutes of delay
   🚫 0.06 lower cancellation risk
```

Figure 4.1: Sample Terminal for Interactive User Input