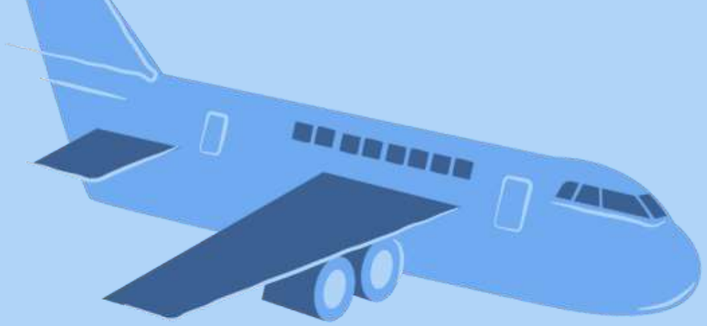
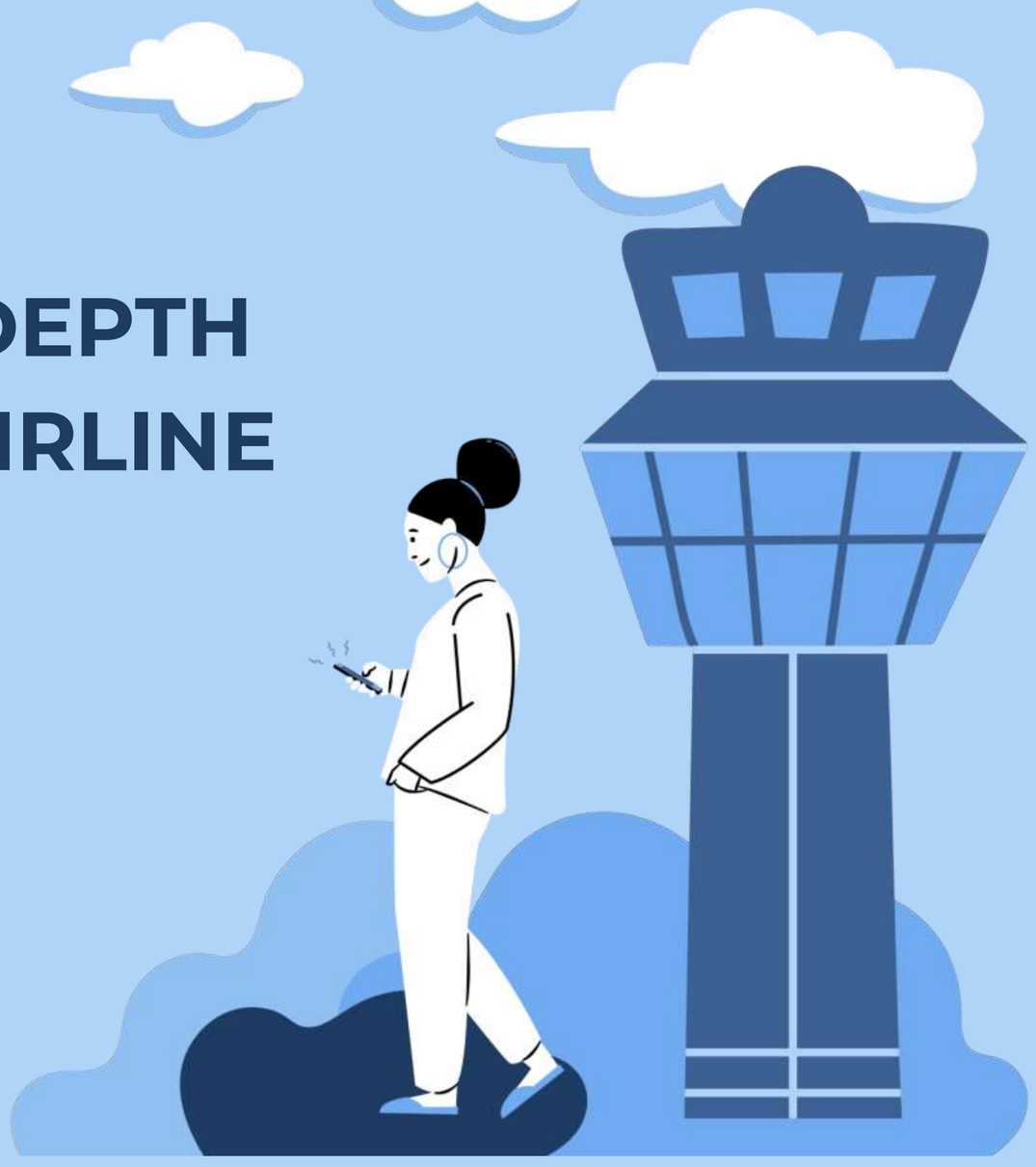


CLEARING THE SKIES: AN IN-DEPTH LOOK AT FACTORS BEHIND AIRLINE DELAYS



GROUP - 2

Ankith Gundeboina

Avinash Appineni

Pavan Kumar Pula

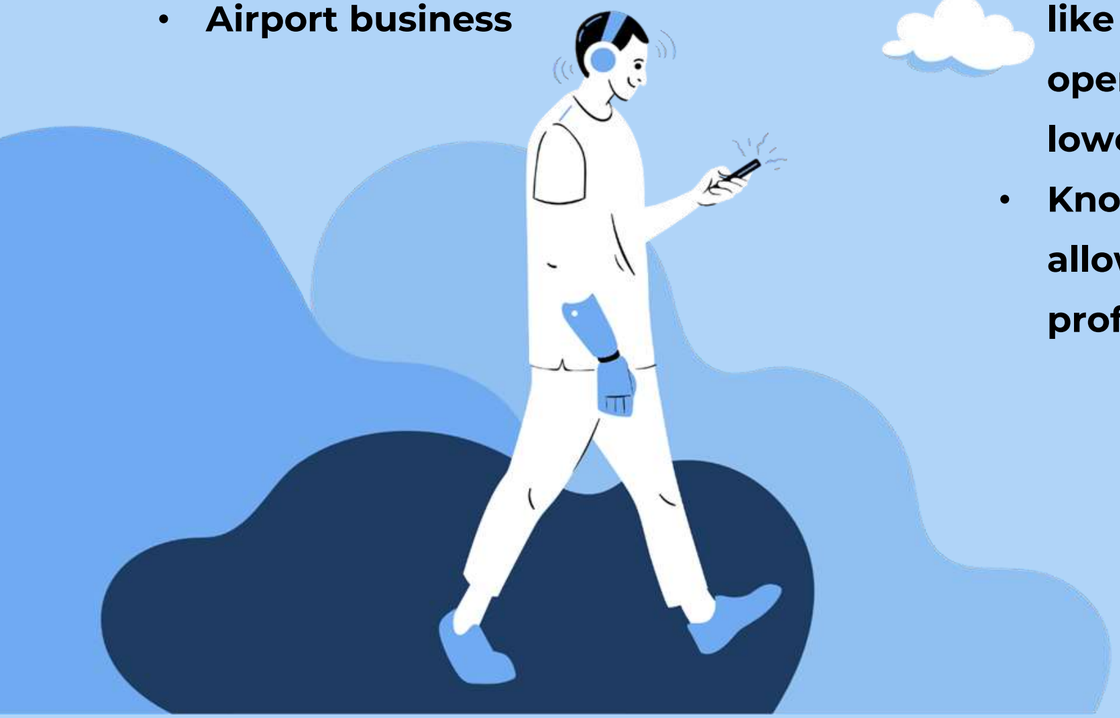
Pavan Sai Teja Kokkura

TABLE OF CONTENTS

INTRODUCTION

IMPACT

- Travelers
- Airline Company
- Airport business



IMPORTANCE

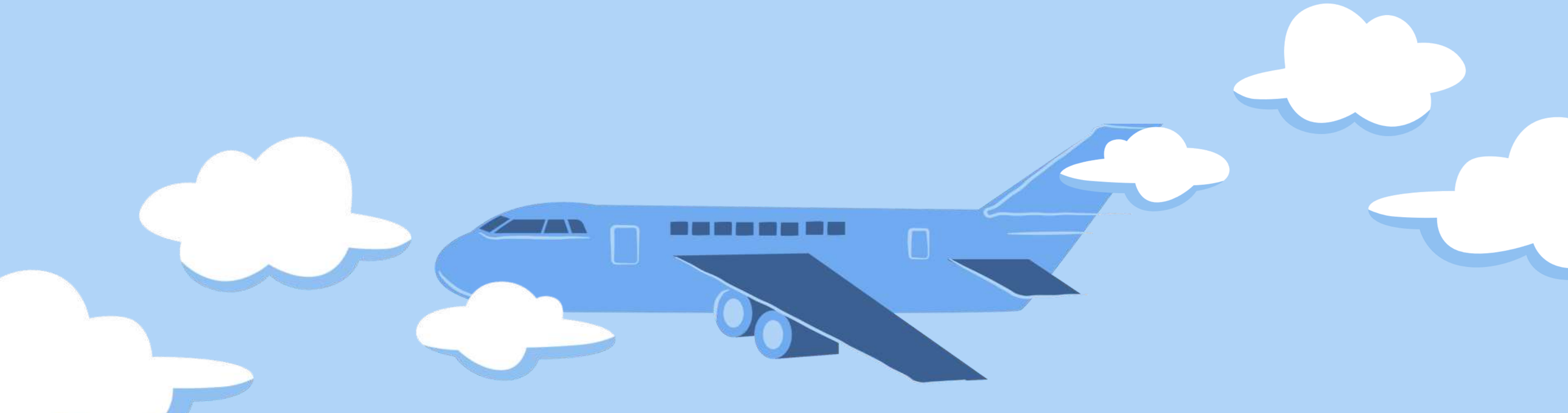
- Considering globalization and Industry growth managing a factor like delay is a crucial to enhancing operational performance and lowering economic losses.
- Knowing the trends in the delay allows airlines to enhance resource profitability while minimizing traffic.

FOCUS

- Analyzing delay factors, including weather, security, and mechanical issues, using a machine learning approach

PROBLEM STATEMENT

Flight delays and cancellations disrupt airline operations, leading to economic losses, reduced productivity, and customer dissatisfaction. These issues stem from various factors, including weather, carrier operations, and systemic inefficiencies. This project aims to leverage machine learning models to accurately predict delays, identify root causes, and classify patterns, providing actionable insights to minimize downtime and improve operational efficiency.



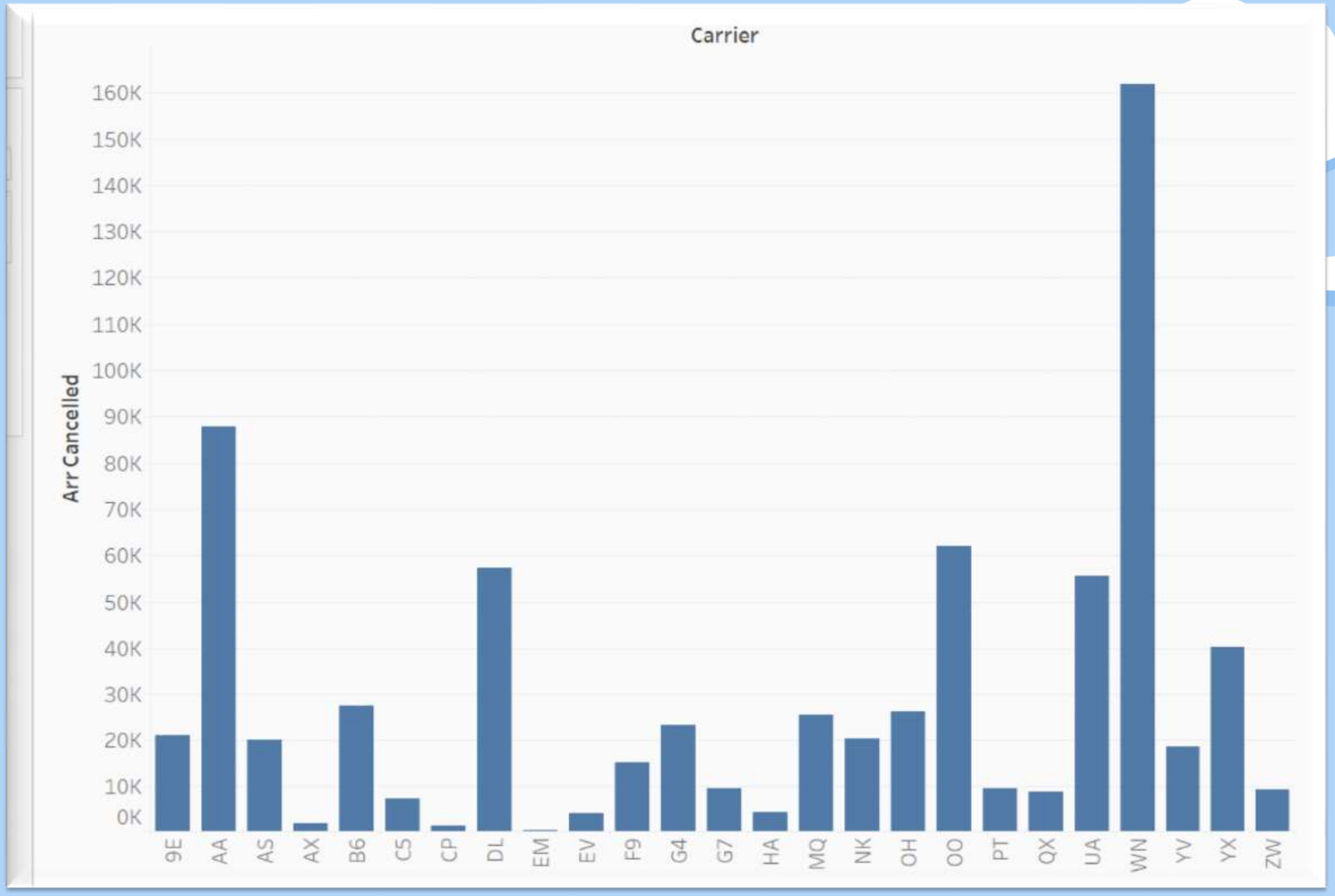
DATA OVERVIEW

- The dataset is titled as “Airline_Delay_Cause”
- Data is collected by **Bureau of Transportation Statistics**, USA
- Data is airlines traffic based
- There are 21 variables and 92k observations in the dataset
- The dataset is recorded yearly (2020 - 2024)
- The attributes in the dataset are about flight operations and levels of delay
- Key variables such as
  - arr_delay (arrival delay),
  - carrier_delay,
  - weather_delay,
  - nas_delay,
  - late_aircraft_delay are influenced by factors like the year, month, carrier, airport, and corresponding counts (carrier_ct, weather_ct, etc.)

DATA ANALYSIS

TREND OF FLIGHT CANCELLATIONS AMONGST CARRIERS

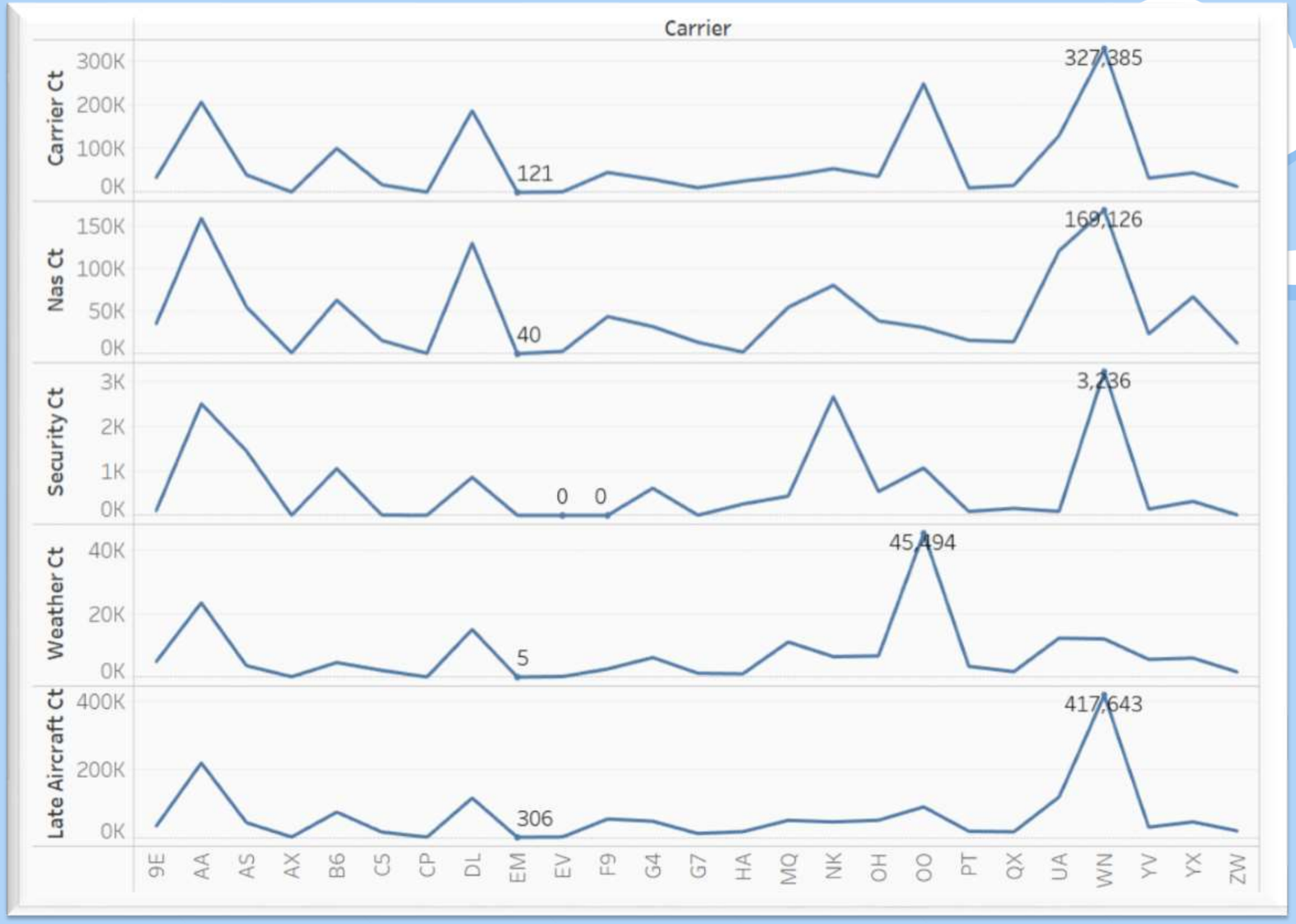
- Shows a cancellation rate per airline.
- Bar Chart: Compare the all cancellations for each carrier.
- It shows which airlines cancel their flights more than others.
- **Southwest , American, Delta** and **Skywest Airlines** are experiencing more cancellations



DATA ANALYSIS

Delays influenced by operational factors that contribute to differences in delay performance.

- **Carrier Delays**: High for WN and OO, indicating internal issues like scheduling or staffing.
- **NAS Delays**: Significant for WN, OO, and UA due to congestion at major hubs.
- **Security Delays**: Moderate across carriers with minimal variability.
- **Weather Delays**: Peaks for WN and OO linked to geographically sensitive areas.
- **Late Aircraft Delays**: High for WN, suggesting cascading effects from earlier flights.

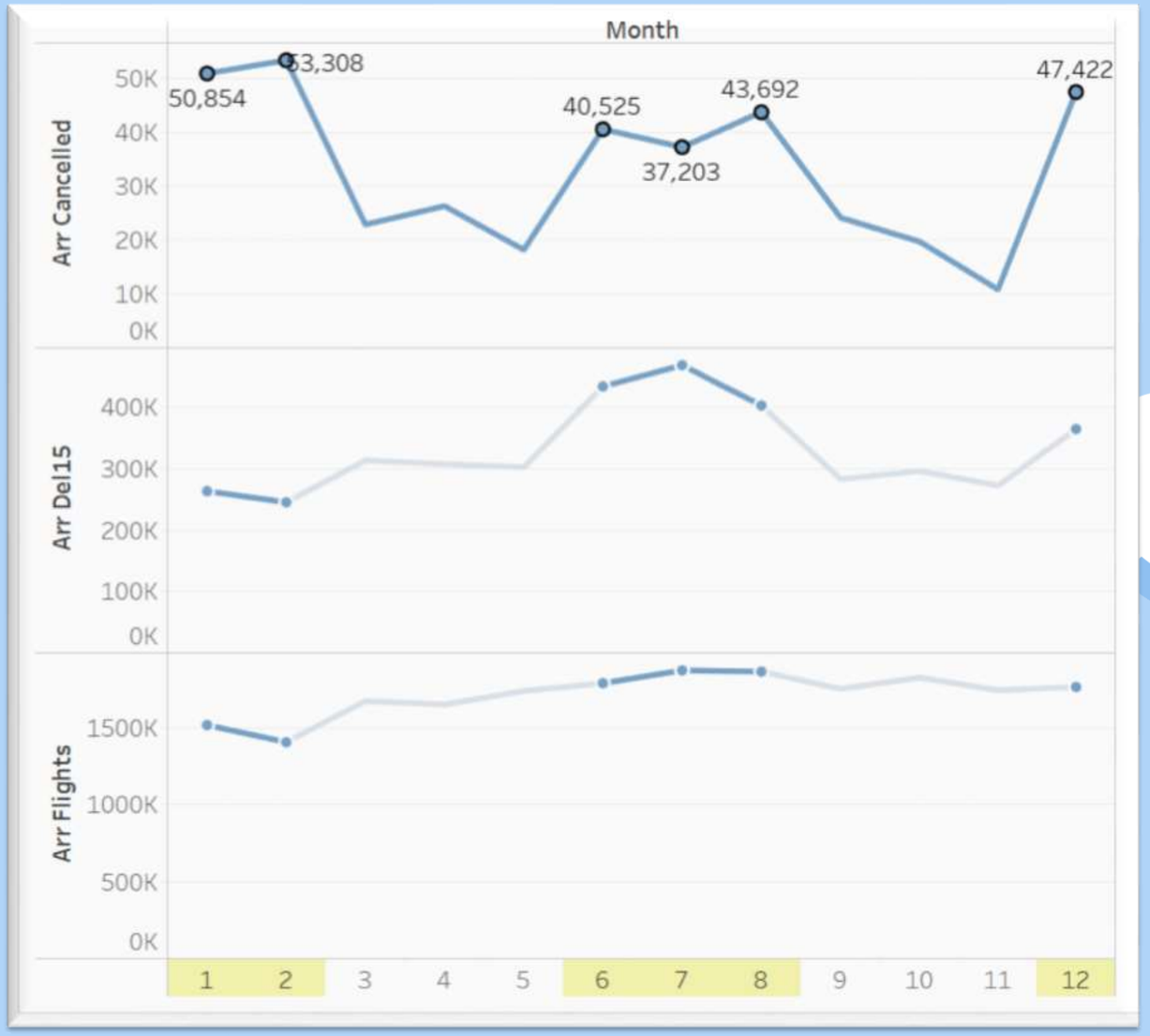


DATA ANALYSIS

A seasonal trend in the number of airline cancellations and delays

- This chart illustrates the relationship between arriving flights, aircraft delays, and cancellations across the months of the year.

- The data highlights that the highest occurrences of delays and cancellations are observed during December, January, February, as well as June, July, and August.



MODELING

Objective:

Create predictive models needed to categorize and predict different flight delays.

Models Used:

1. **Random Forest**: Its features: decision trees with high accuracy and stability against multifactorial delays.
2. **Decision Tree**: Easy and clear to identify main causes of delays.
3. **Linear Regression**: Determines correlation between delay and its causes.

Evaluation Metrics:

Classification accuracy and cross entropy for classifiers data, MSE and MAE for regression.

RESEARCH QUESTIONS

1. Can we predict whether a flight will be canceled based on factors like carrier, weather conditions, NAS issues, and airport location?

Model Used: Random Forest Classifier

- Target variable: arr_cancelled (Flight Cancellation)
- Included variables:
- *Time-related:* Total Time, Arrival Time Lasted
- *Frequency-related:* Arrival Carrier Frequency, Number of Arriving Flights
- *Delay counts:* carrier_ct, weather_ct, etc.

Data Split:

- 80% Training, 20% Testing
- *Model Accuracy:* 97.52%

Key Insights:

- Carrier-specific delays (carrier_ct) have the highest importance.
- Strong correlation between carrier delays and cancellation rates across airlines.

```
Random Forest Accuracy: 0.9752425911355888
Confusion Matrix (Random Forest):
[[204223    262     87 ...      0      0      0]
 [  1085    187     64 ...      0      0      0]
 [   532    118     59 ...      0      0      0]
 ...
 [     0      0      0 ...      0      0      0]
 [     0      0      0 ...      0      0      0]
 [     0      0      0 ...      0      0      0]]
Classification Report (Random Forest):
```

```
              accuracy                           0.98    209715
             macro avg       0.01      0.01      0.01    209715
          weighted avg       0.97      0.98      0.97    209715
```

Visualization: Bar Chart

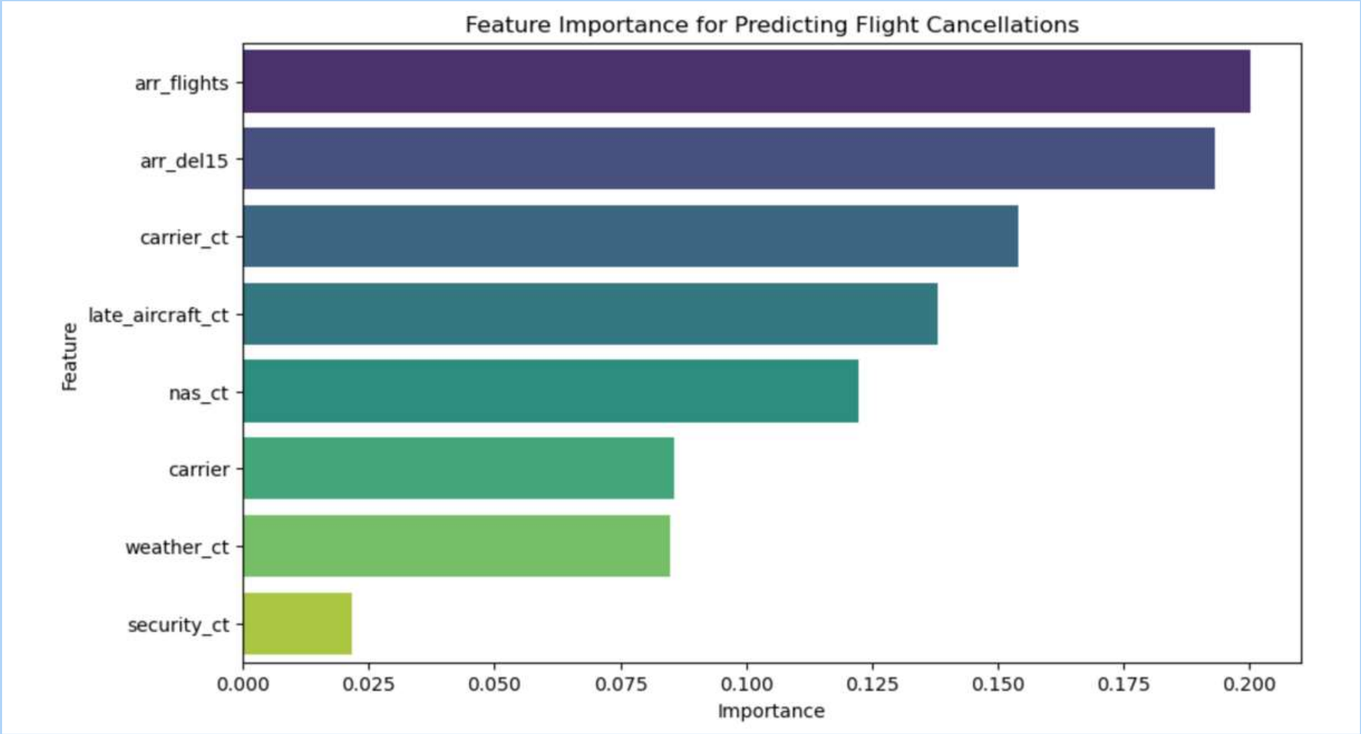
•Represents *Feature Importance* in predicting flight cancellations

Axes:

•*Y-axis:* Features

•*X-axis:* Importance Scores (assigned by Random Forest Classifier)

Purpose: Highlights the *contribution of each feature* to the model's predictions.



2. Can we predict the likelihood of a flight being delayed by more than 15 minutes using variables like the airline, month, weather conditions, and NAS issues?

Model Used: Gradient Boosting Classifier

- *Target:* Predicting flight delays over 15 minutes
- *Key Features:* Month, carrier, airport, and delay counts
- *Accuracy:* 91% with LightGBM Classifier (100 estimators)

Main Findings:

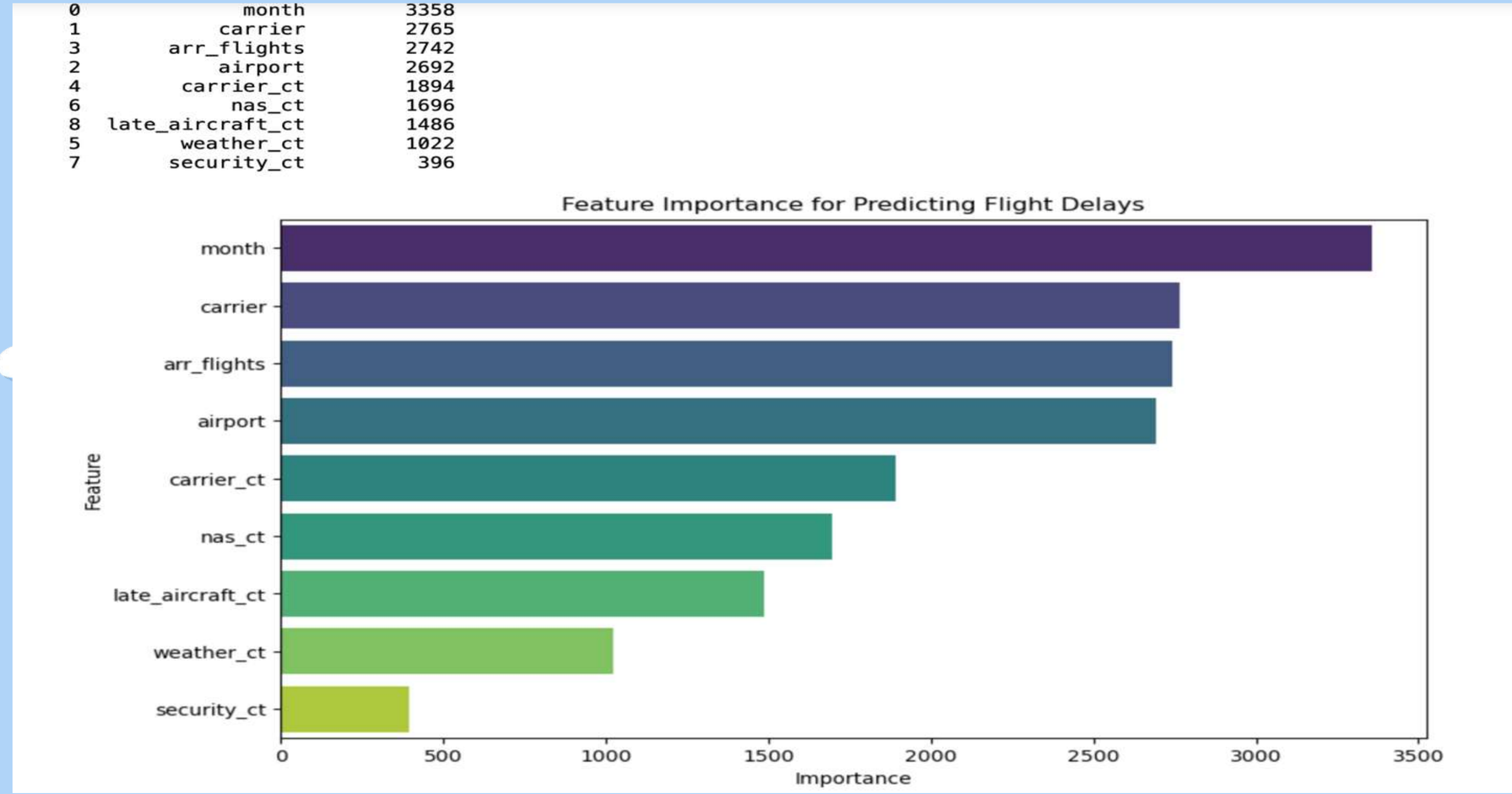
- Month, carrier, and airport are most important for predictions.
- Lower recall for delayed flights indicates room for improvement.

Insights: Identified key factors related to operational and seasonal delays.

```
LightGBM Accuracy: 0.906047842326949

Confusion Matrix:
[[56927    19     0 ...     0     0     0]
 [    0    28     0 ...     0     0     0]
 [    0    49     0 ...     0     0     0]
 ...
 [    0     0     0 ...     0     0     0]
 [    0     0     0 ...     0     0     0]
 [    0     0     0 ...     0     0     0]]
```

- **Top Predictors**: Month, carrier, number of arriving flights
- **Impact**: Most likely to influence arrival delay
- **Application**: Helps estimate optimal flight connections and minimize delays.

0	month	3358
1	carrier	2765
3	arr_flights	2742
2	airport	2692
4	carrier_ct	1894
6	nas_ct	1696
8	late_aircraft_ct	1486
5	weather_ct	1022
7	security_ct	396



3. Can we classify the root cause of flight delays (carrier, weather, NAS, security, or late aircraft) based on flight data?

- **Model Used**: Decision Tree Classifier
- **High Performance**: Good for carrier, NAS, late aircraft, weather delays
- **Challenges**: Lower precision/recall for weather delay; low performance for security delays (few incidents)
- **Strengths**: High overall accuracy (98.02%); easy to interpret and identify delay causes
- **Outcome**: Successfully predicts flight delay reasons.

```
Decision Tree Accuracy: 0.9801974826388888
Classification Report:
                    precision    recall  f1-score   support

       carrier_ct       0.99      0.99      0.99      9792
late_aircraft_ct        0.97      0.97      0.97      4478
           nas_ct       0.98      0.98      0.98      3938
      security_ct       0.25      1.00      0.40         1
       weather_ct       0.94      0.86      0.90       223

         accuracy                           0.98     18432
        macro avg       0.83      0.96      0.85     18432
     weighted avg       0.98      0.98      0.98     18432
```

·**Accuracy**: 98.02% for Decision Tree Classifier

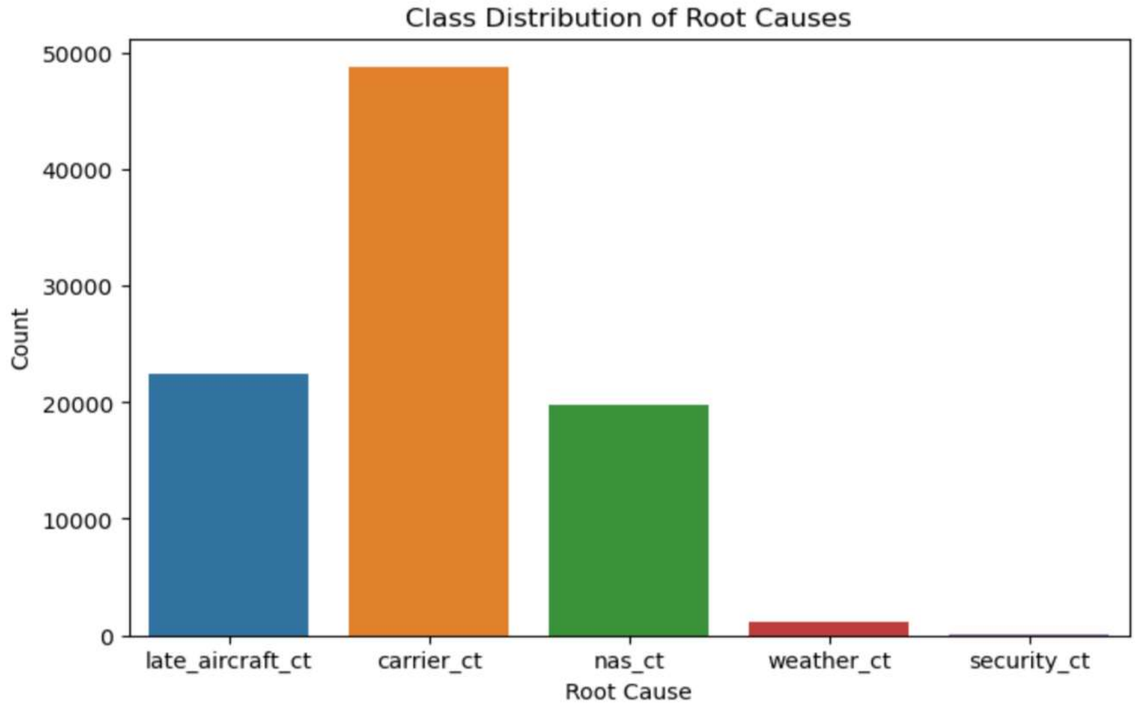
·**High Probabilities**: For carrier, NAS, late aircraft delays

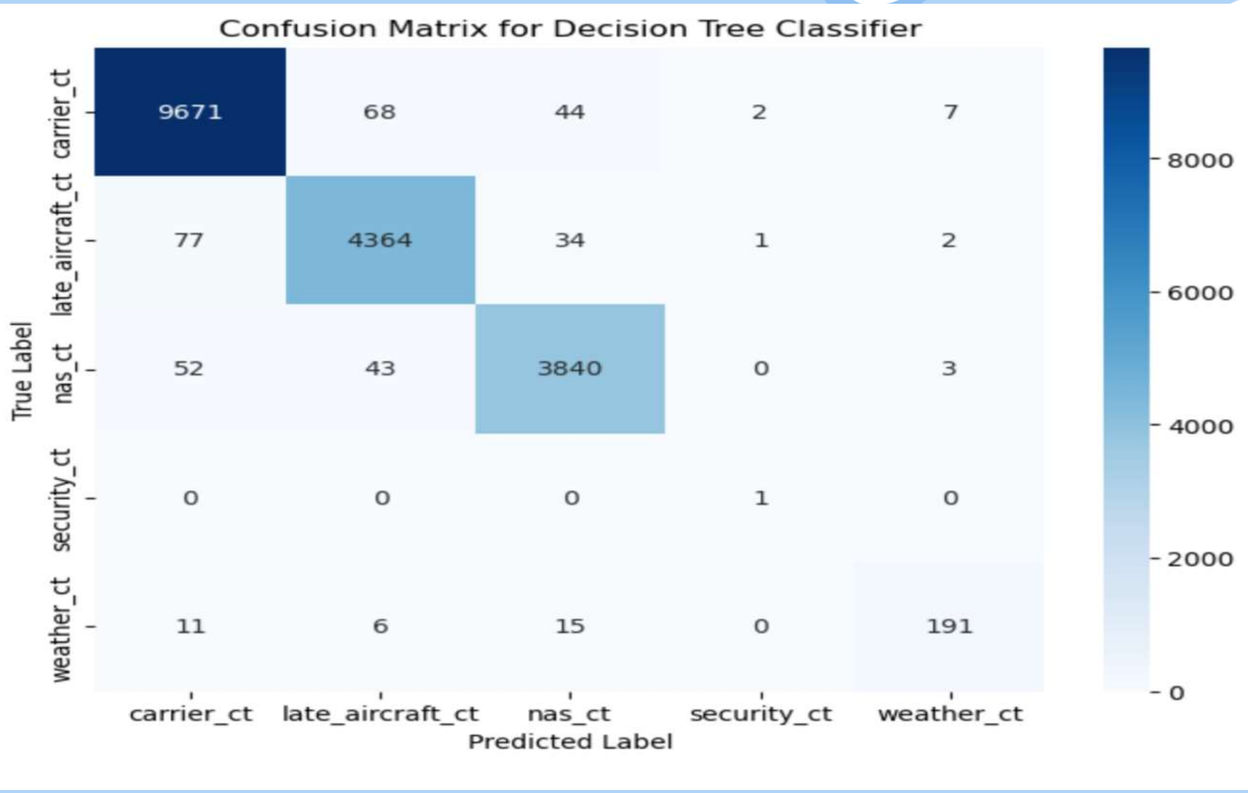
·**Weather Delays**: Good performance, slightly lower precision/recall

·**Model Strength**: Effective at assigning probability scores for delay causes

·**Areas for Improvement**: Security class imbalance

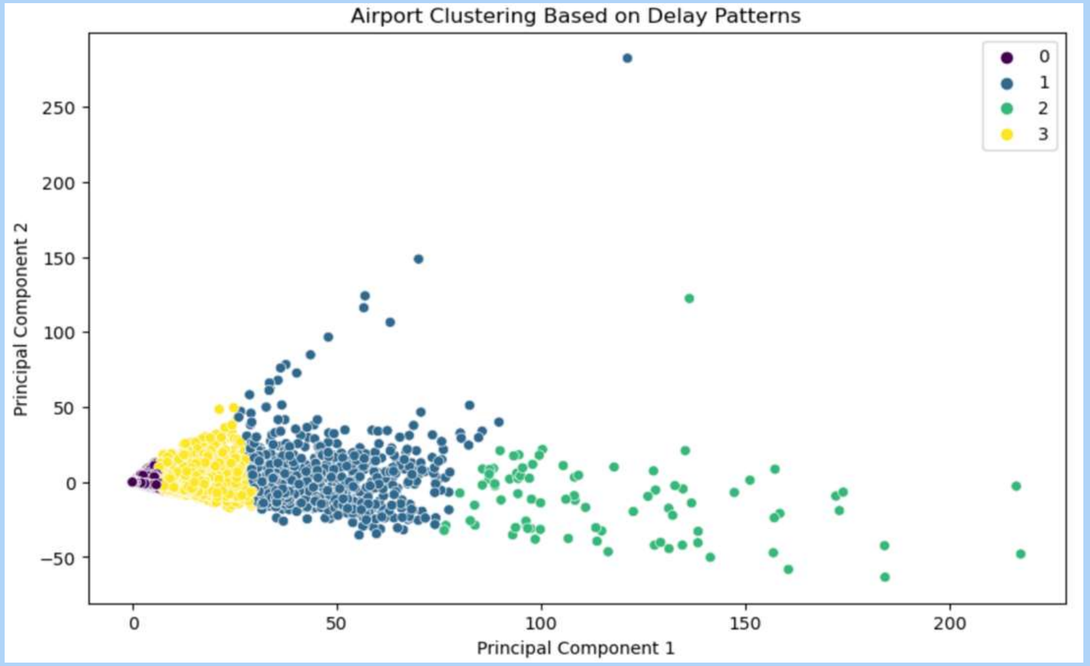
·**Confusion Matrix**: Clear overview of model performance by delay type.



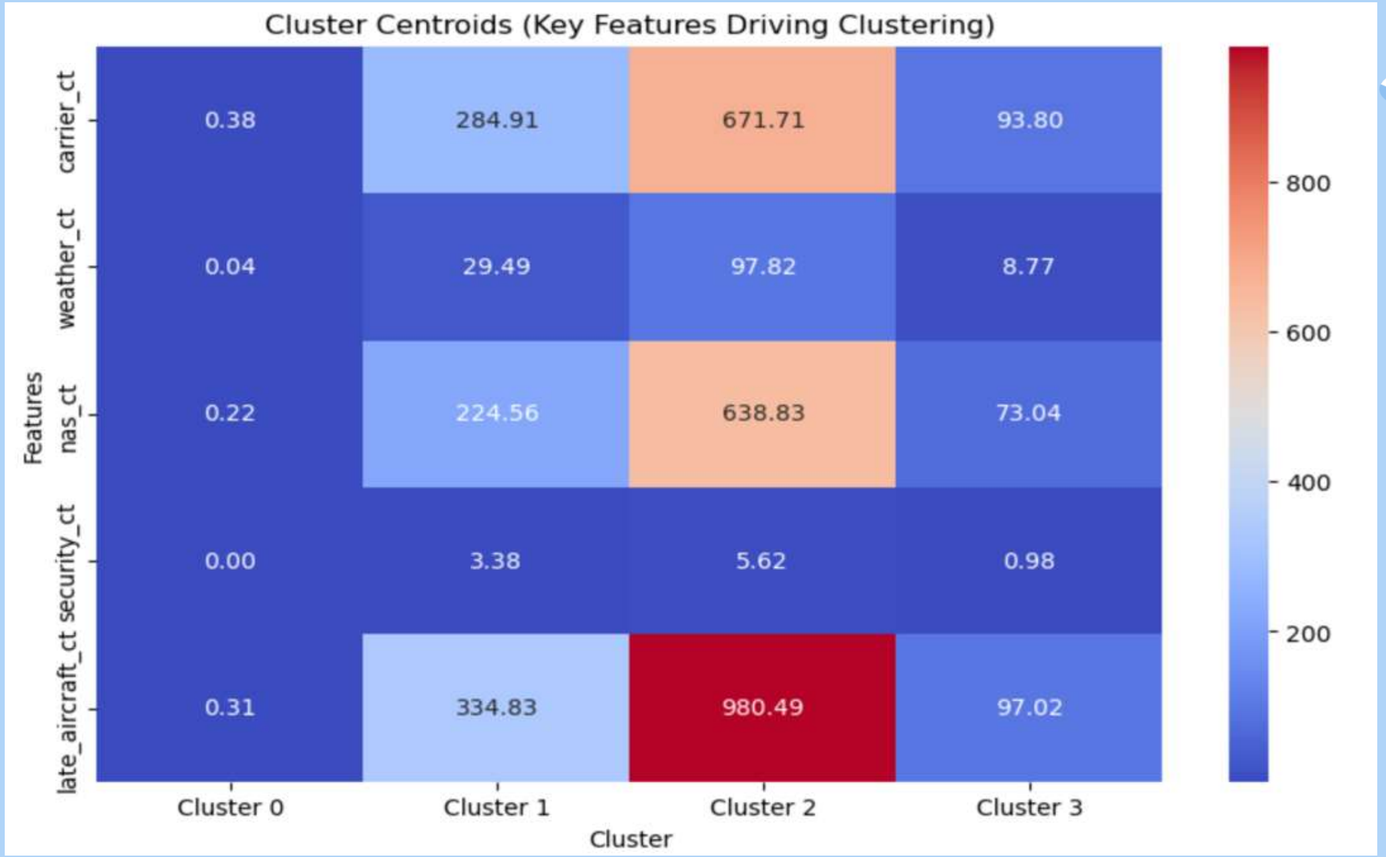


4. Can we cluster airports based on their delay patterns to identify those with the highest likelihood of delays due to specific factors (weather, congestion, etc.)?

- **Cluster 0**: Low delays, small/well-dispersed airports
- **Cluster 1**: Moderate delays, contract-related, NAS, carrier delays, medium-sized airports
- **Cluster 2**: High delays, NAS, carrier, late aircraft delays, busy hub airports
- **Cluster 3**: Small delays, slightly higher carrier_ct, small airports with occasional disturbances
- **Scatter Plot**: 2D visualization using PCA, points color-coded by cluster.



- **Clustering Insight**: Identifies airports with distinct delay patterns
- **Cluster 2**: Represents operational issues
- **Clusters 0 & 3**: Indicate effective management strategies
- **Visuals**: Heatmap shows delays clearly by cluster



5. Can we predict the peak months for flight cancellations using historical data on cancellations, weather patterns, and flight volumes?

High Accuracy: 97.17% accurate (class 0)

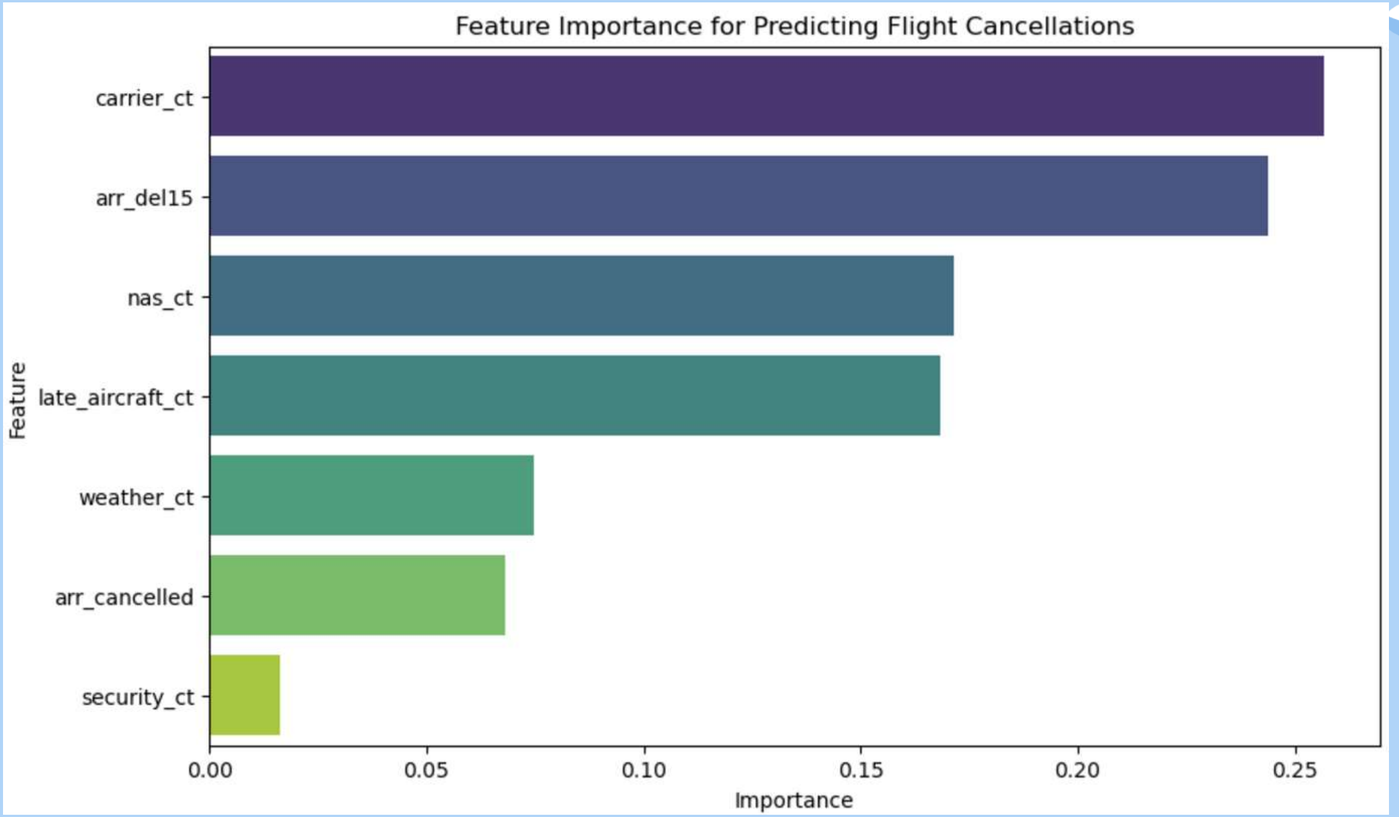
Imbalance Problem: Small classes (1, 2, 3) have low precision and recall

```
Random Forest Accuracy: 0.971685382542975
Confusion Matrix (Random Forest):
[[201475    508    365    391]
 [   654    641    502    504]
 [   557    501    840    472]
 [   550    479    455    821]]
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    202739
           1       0.30      0.28      0.29      2301
           2       0.39      0.35      0.37      2370
           3       0.38      0.36      0.37      2305

    accuracy                           0.97    209715
   macro avg       0.51      0.50      0.50    209715
weighted avg       0.97      0.97      0.97    209715
```

Most Influential Factor: Carrier-related delays (carrier_ct)

Other Key Factors: Arrival delays (arr_del15), National Airspace System issues (nas_ct), late aircraft arrivals (late_aircraft_ct)

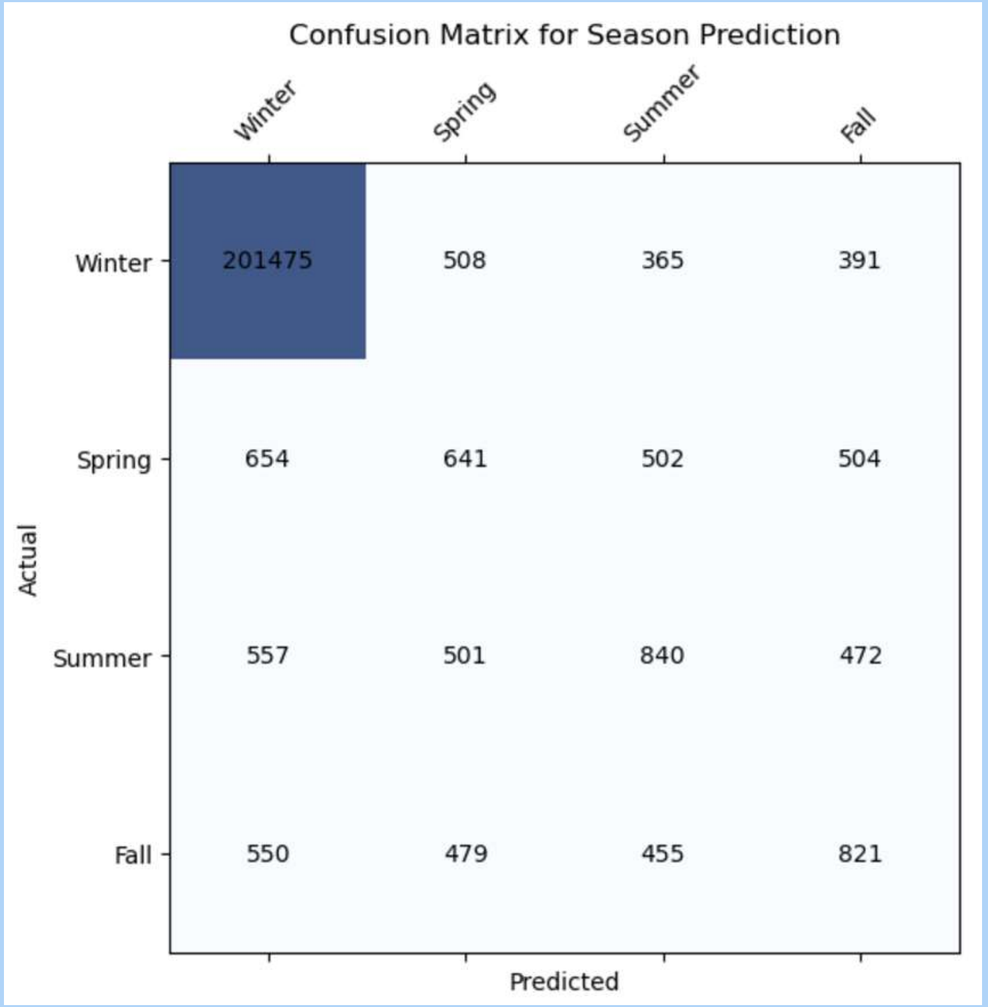
Least Influential Factors: Weather-related delays (weather_ct), prior cancellations (arr_cancelled), security delays (security_ct)

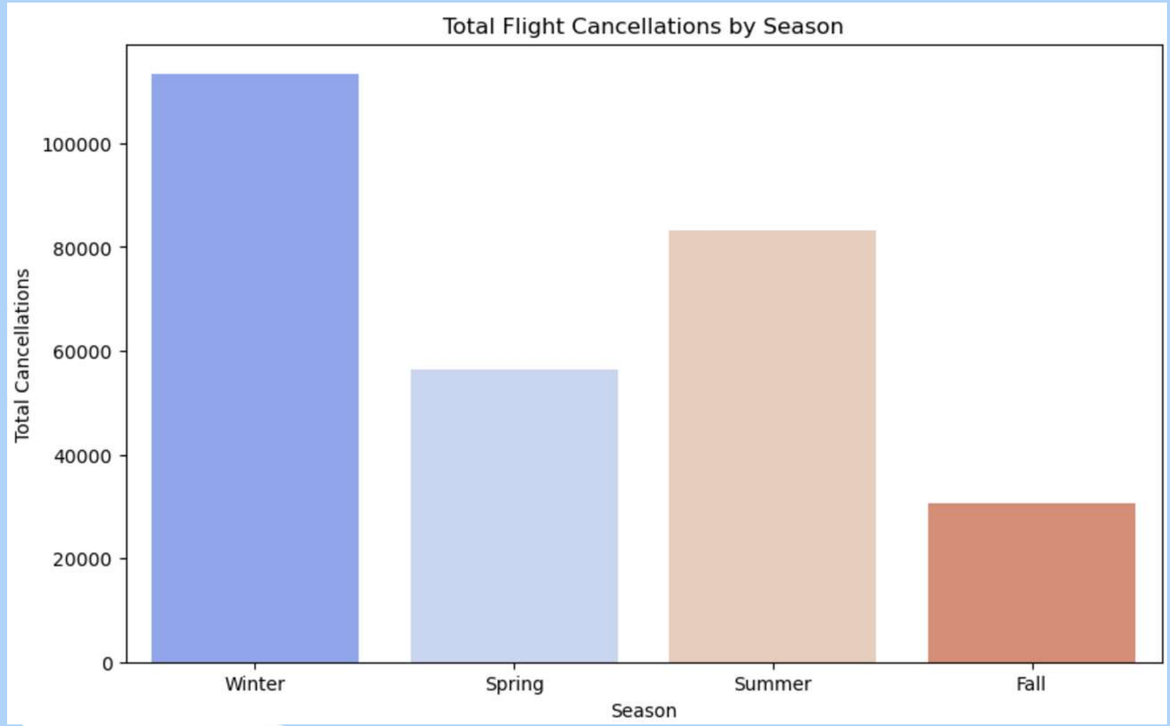


High Accuracy for Winter: 201,475 correct predictions

Misclassifications Exist: Spring often confused with Winter (654 times)

Key Insight: Model performs well in Winter but needs improvement for other seasons

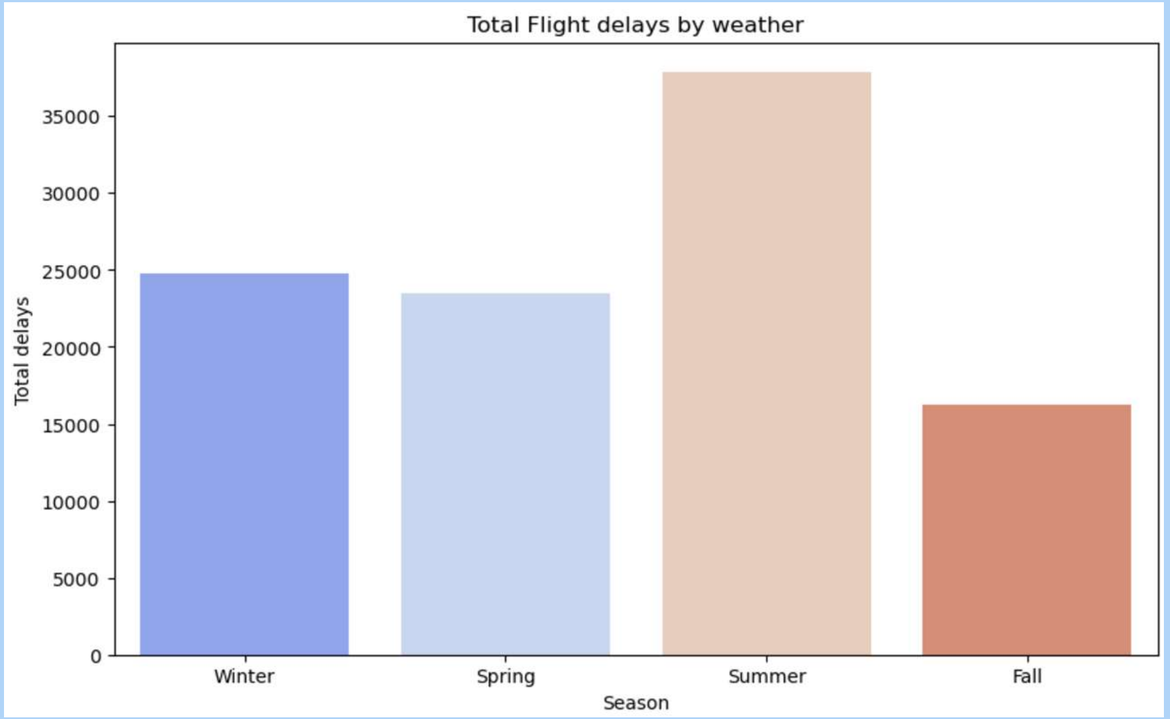




Winter: Highest cancellations (over 100,000)

Summer: Second highest due to storms and demand

Fall: Fewest cancellations, stable weather



Summer: Most delays due to weather, over 35,000

Winter and Spring: Moderate delays, around 25,000 each

Fall: Fewest delays, indicating more stable weather conditions

LIMITATIONS

I

CLASS IMBALANCE

Rare delay types (like security delays) are not well-represented, making models biased.

II

FEATURE LIMITATION

The model's accuracy depends on the chosen features; missing important factors can lower accuracy.

III

MODEL COMPLEXITY

Gradient Boosting needs careful tuning and may overfit, affecting performance on complex data.

IV

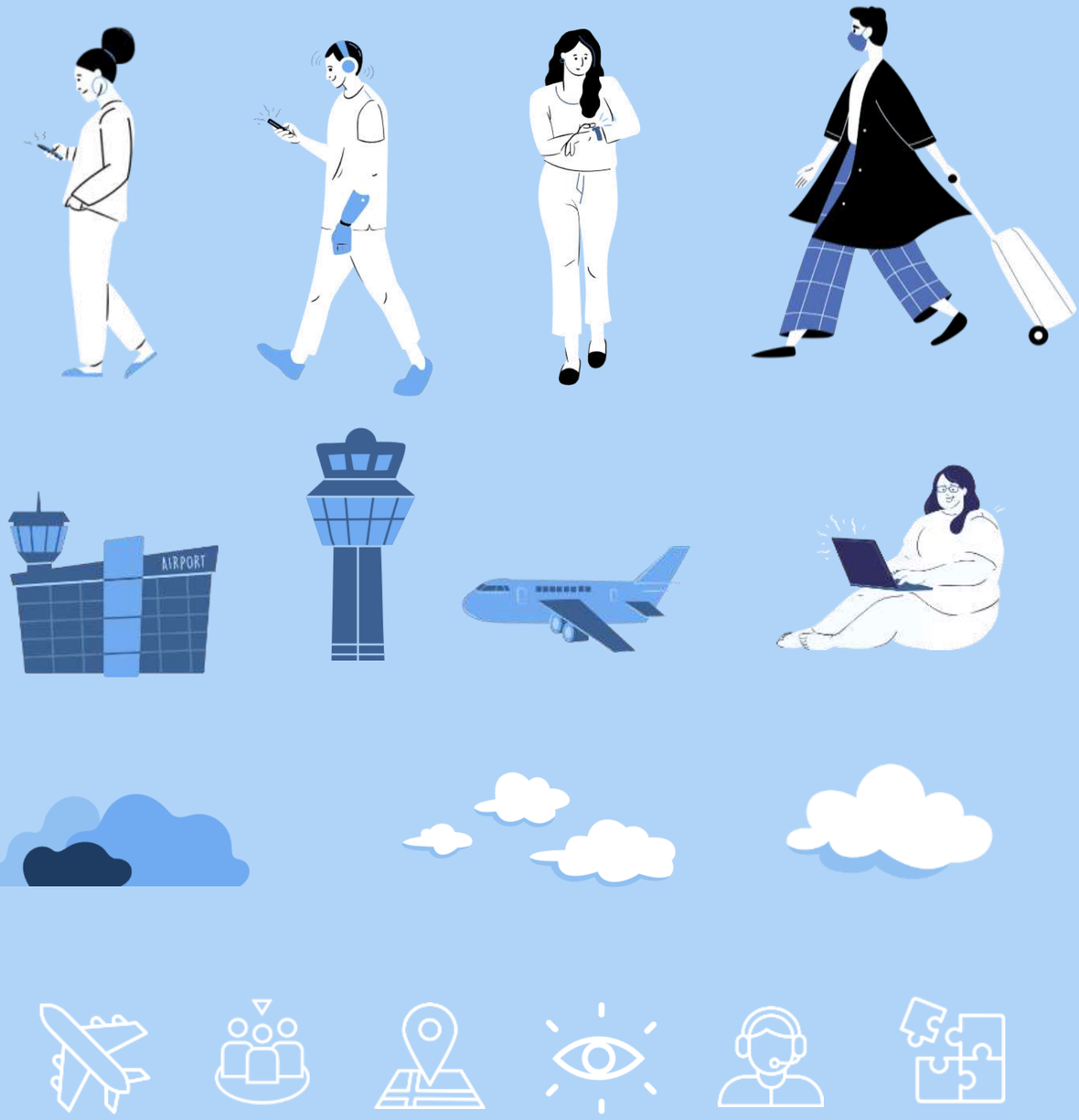
CLUSTER UNDERSTANDING

It can be hard to make sense of clusters without specific knowledge about the industry.

CONCLUSION

In conclusion, this project analyzed airline delays and cancellations to uncover key trends and causes. Predictive models helped forecast delays, estimate durations, identify cascading delays, and cluster airports by delay patterns, offering valuable insights for passengers, airlines, and airports.

While the models performed well, challenges like class imbalance and limited features remain. Future efforts can focus on tailored models for specific airlines or regions to enhance accuracy and efficiency in flight operations.



THANK YOU!