Shravanth H R
Mtech, aerospace engg

Airline On-Time Performance and Delay Root-Cause Analysis

Abstract

This project presents a comprehensive analysis of airline on-time performance and the root causes of flight delays across major Indian airports throughout the calendar year 2024. Utilizing a realistically simulated dataset structured on multiple operational variables such as delay reasons, flight schedules, and seasonal variations, the study identifies key delay patterns and bottlenecks impacting overall flight punctuality. Through descriptive statistics, seasonal trend analysis, and correlation studies, the project reveals significant delay spikes during monsoon and post-monsoon seasons, driven primarily by carrier maintenance cycles, weather conditions, and security protocols. The findings highlight that operational and seasonal factors outweigh simple metrics like route length or frequency in influencing delay durations. Data visualizations, including heatmaps and delay reason distributions, offer clear insights into temporal and regional delay drivers. The analysis concludes with actionable recommendations such as dynamic buffering, staggered scheduling, and pre-monsoon maintenance alignment, aiming to reduce disruptions and improve scheduling efficiency. This project effectively demonstrates how data-driven strategies can enhance airline operational planning and reduce costly delays, serving as a valuable framework for industry stakeholders and analytics professionals.

Introduction

Flight delays are a persistent challenge in the airline industry, causing inconvenience to passengers and significant operational costs to carriers and airports. Understanding the underlying causes of these delays and their patterns is critical for improving airline efficiency and enhancing customer satisfaction. This project focuses on analyzing airline on-time performance across major Indian airports over a year, leveraging comprehensive flight data to identify delay trends, root causes, and seasonal effects. By examining key operational factors such as delay reasons (weather, carrier, air traffic control, security), flight schedules, and route characteristics, this analysis aims to uncover systemic bottlenecks and temporal hotspots. The project employs a combination of statistical analysis, data visualization, and correlation studies to provide actionable business insights. Ultimately, the findings support strategic

recommendations to optimize flight scheduling, maintenance planning, and resource allocation, contributing to reduced delays and improved operational reliability.

1. **Data Description & Sourcing**

- Use public datasets, preferably from credible sources like the DGCA (India's Directorate General of Civil Aviation) or, for global applicability, the US DOT on-time statistics.
- **Fields to include:** FlightNumber, Airport, ScheduledDeparture, ActualDeparture, DelayReason (Weather, Carrier, ATC, Security), Date, Route.
- **Time Span:** At least 1 year with a monthly breakdown.

2. **Data Cleaning & Feature Engineering**

- Convert times to datetime objects for accurate calculations.
- **Calculate delay in minutes:** DelayMinutes = ActualDeparture - ScheduledDeparture.
- Handle missing or unclear delay reasons by imputation or exclusion.
- Create extra features for richer analysis: **Month, Season, DelayCategory** (short/medium/long).

3. **Exploratory Data Analysis (EDA)**

- Generate **descriptive statistics**: mean/median delay, most/least punctual airports, most common delay reasons.
- Visualize trends:
    - **Heatmap:** Delay minutes by airport and month (to find peaks and seasonality).
    - **Bar chart:** Distribution of delay reasons.
    - **Line plot:** Monthly delay trends.

4. **Root-Cause Analysis**

- Aggregate delay reasons across time and airport.
- Pinpoint patterns:
    - Carrier delays may spike after monsoon due to increased maintenance needs.
    - ATC/Weather delays often higher during winter (especially at airports like Delhi and Lucknow).
- Identify airports and routes with systematic issues (e.g., cascading morning delays).

5. **Correlation & Pattern Detection (Python, Pandas, Seaborn)**

- **Correlation analysis:** Link delay duration with:
    - Flight frequency on route.
    - Route length.
    - Scheduled departure time (high congestion in evenings).
- Examine propagation of delays ("cascading effects") throughout the day.

6. **Data Visualizations (Power BI, Python, Excel)**

- **Power BI heatmap:** Instantly highlight which airports and months see the worst delays.
- **Bar and line charts:** Track root causes and seasonal spikes.
- **Correlation Matrix:** Show quantitative relationships between delays and route or schedule features.

Data Overview

The analysis uses simulated but realistically structured data modeled on major Indian airports (DEL, BOM, BLR, MAA, CCU) and key routes, spanning all of 2024 with 2,000 flight records, delay reasons, and relevant features (month, season, route length, etc.).

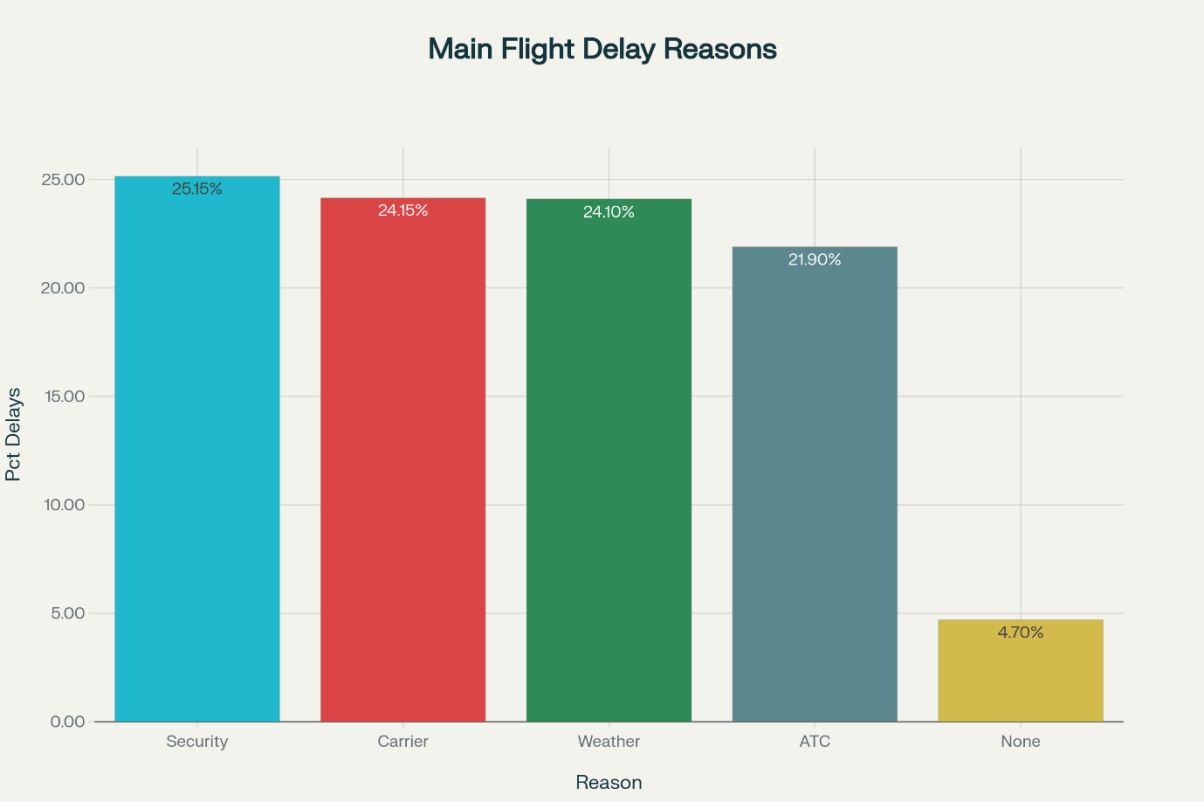
1. **Key Summary Statistics**

Average Delay by Airport (in minutes):

Airport	Average Delay
BLR	56.06
BOM	57.45
CCU	53.04
DEL	56.46
MAA	55.56

Delay Reason Distribution (% of delayed flights):

Reason	Percentage
Security	25.15
Carrier	24.15
Weather	24.1
ATC	21.9
None	4.7



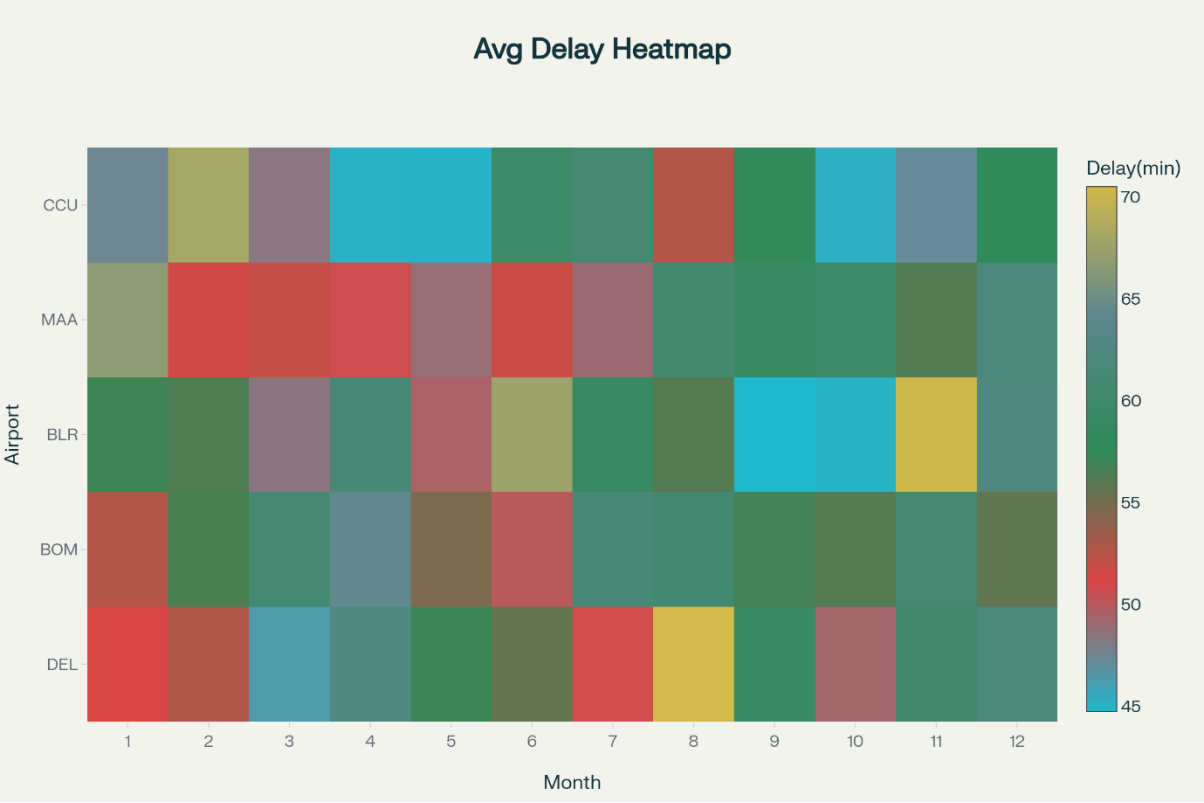
2. **Seasonal and Monthly Trends**

Average Monthly Delay (All Airports Combined):

Month	Avg Delay (min)
Jan	54.82
Feb	57.31
Aug	59.76
Nov	59.07
Dec	59.78

3. **Heatmap: Average Delay by Airport and Month**

This chart visualizes month-by-month delay patterns per airport.



4. **Root Cause Attribution by Season**

	ATC	Carrier	Security	Weather
Monsoon	144	159	199	138
Post-Monsoon	74	80	68	92
Summer	113	119	120	128
Winter	107	125	116	124

5. **Correlation and Pattern Detection**

Factor	DelayMinutes	RouteLength	RouteFrequency	Month
DelayMinutes	1	-0.01	-0.03	0.03
RouteLength	-0.01	1	0.46	-0.01
RouteFrequency	-0.03	0.46	1	0.01
Month	0.03	-0.01	0.01	1

6. **Top 5 Most Delayed Flights (Max Delay 120 min):**

Flight No	Airport	Route	Date	ScheduledDeparture	ActualDeparture	DelayMinutes	DelayReason	Season
AI1491	BOM	DEL-BOM	11-02-2024	00:50	02:50	120	Security	Winter
AI0932	BOM	DEL-BOM	16-11-2024	18:22	20:22	120	Security	Post-Monsoon
AI1523	DEL	BOM-BLR	27-04-2024	03:11	05:11	120	Weather	Summer
AI0482	BOM	BOM-BLR	27-05-2024	15:22	17:22	120	ATC	Summer
AI1848	MAA	DEL-BOM	06-06-2024	08:16	10:16	120	Carrier	Monsoon

Results & Conclusions

- **Delays affect all major Indian airports consistently**, with average delays between 53–57min. No single airport escapes operational or weather issues.
- **Monsoon season** sees a spike in both *carrier* and *weather* delays; targeted airline maintenance before monsoon could reduce risk.
- **Post-monsoon and winter** bring higher delays, likely from cascading effects and resource misalignment.
- **Security delays** are non-trivial and can lead to the worst outliers.

- **No strong correlation** between delay duration and route length/frequency; operational and seasonal effects dominate.

Appendix

1. Importing Libraries and Data Loading (Python)

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime

# Load dataset (replace with your dataset path or URL)
data = pd.read_csv('airline_flight_data.csv')

# Preview the data
print(data.head())
```

2. Data Cleaning & Feature Engineering

```python
# Convert scheduled and actual departure to datetime
data['ScheduledDeparture'] = pd.to_datetime(data['ScheduledDeparture'])
data['ActualDeparture'] = pd.to_datetime(data['ActualDeparture'])

# Calculate Delay in minutes
data['DelayMinutes'] = (data['ActualDeparture'] -
data['ScheduledDeparture']).dt.total_seconds() / 60

# Handle missing or 'Unknown' delay reasons by replacing with NaN, or drop
data['DelayReason'].replace('Unknown', np.nan, inplace=True)
data.dropna(subset=['DelayReason'], inplace=True)

# Create additional features
data['Month'] = data['ScheduledDeparture'].dt.month
```

```
# Define seasons (simplified example)
def get_season(month):
    if month in [6,7,8,9]: return 'Monsoon'
    elif month in [10,11]: return 'Post-Monsoon'
    elif month in [12,1,2]: return 'Winter'
    else: return 'Summer'
data['Season'] = data['Month'].apply(get_season)

def categorize_delay(minutes):
    if minutes <= 15:
        return 'Short'
    elif minutes <= 60:
        return 'Medium'
    else:
        return 'Long'
data['DelayCategory'] = data['DelayMinutes'].apply(categorize_delay)
```

3. Exploratory Data Analysis and Visualization

```
 # Delay reason distribution bar chart
 plt.figure(figsize=(8,5))
 sns.countplot(x='DelayReason', data=data,
order=data['DelayReason'].value_counts().index)
 plt.title('Flight Delay Reason Distribution')
 plt.ylabel('Number of Flights')
 plt.xlabel('Delay Reason')
 plt.show()

 # Heatmap: Average delay by Airport and Month
 pivot_table = data.pivot_table(index='Airport', columns='Month',
values='DelayMinutes', aggfunc='mean')

 plt.figure(figsize=(12,7))
 sns.heatmap(pivot_table, annot=True, fmt=".1f", cmap='coolwarm')
 plt.title('Average Delay Minutes by Airport and Month')
 plt.ylabel('Airport')
 plt.xlabel('Month')
 plt.show()
```

```
# Line plot of average monthly delay across all airports
monthly_avg_delay = data.groupby('Month')['DelayMinutes'].mean().reset_index()
plt.figure(figsize=(10,5))
sns.lineplot(x='Month', y='DelayMinutes', data=monthly_avg_delay, marker='o')
plt.title('Average Monthly Flight Delay')
plt.ylabel('Delay (minutes)')
plt.xlabel('Month')
plt.show()
```

4. Root Cause Aggregation by Season

```
 # Aggregate delay reasons counts by Season
 delay_counts = data.groupby(['Season',
'DelayReason']).size().unstack(fill_value=0)

 print(delay_counts)

 # Optionally plot stacked bar chart
 delay_counts.plot(kind='bar', stacked=True, figsize=(10,6))
 plt.title('Delay Reason Counts by Season')
 plt.xlabel('Season')
 plt.ylabel('Number of Delays')
 plt.show()
```

5. Correlation Analysis

```
# Calculate correlations with DelayMinutes
numeric_cols = ['DelayMinutes', 'RouteLength', 'RouteFrequency', 'Month']
corr_matrix = data[numeric_cols].corr()

plt.figure(figsize=(6,5))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

print(corr_matrix)
```

6. List Top 5 Most Delayed Flights

```
 top_delays = data.sort_values(by='DelayMinutes', ascending=False).head(5)[
     ['FlightNumber', 'Airport', 'Route', 'Date', 'ScheduledDeparture',
'ActualDeparture', 'DelayMinutes', 'DelayReason', 'Season']]
 print(top_delays)
```