FINAL PROJECT

FLIGHT OPERATIONS ANALYSIS AND DELAY RISK PREDICTION

Author: Lam Thao My

BUSINESS CONTEXT

- Flight delays and cancellations impact cost, customer experience and operational efficiency
- Airlines need data-driven insight to move from reactive to proactive operations
- Understanding delay patterns is critical for risk management

BUSINESS PROBLEM

- Operational disruptions are not random
- High-level KPIs hide localized delay risks
- Lack of early identification limits proactive intervention

PROJECT OBJECTIVES

- Analyze global flight delay patterns
- Identify temporal and geographic risk factors
- Evaluate predictive potential of machine learning models
- Translate findings into actionable business insight

DATA SOURCE

- Dataset collected from Kaggle - Airline Dataset
- Data generated as synthetic airline operational data
- Designed to simulate real-world airline operations
- Covers:
  - Passenger information
  - Airport and geographic details
  - Flight schedules and statuses

Data Overview

- Total observations: 98,619 flight records
- Total features: 15 variables
- Data scope:
  - Global coverage across multiple continents
  - Various airports, routes, and passengers
- Key target variable:
  - Flight Status (On Time, Delayed, Cancelled)
- Data structure:
  - Mix of numerical, categorical, and date fields

SQL PROCESSING: DATA VALIDATION & EXPLORATION

Validated data quality

```
--Kiểm tra NULL
SELECT
    SUM(CASE WHEN Passenger_ID IS NULL THEN 1 ELSE 0 END) AS null_passenger,
    SUM(CASE WHEN Age IS NULL THEN 1 ELSE 0 END) AS null_age,
    SUM(CASE WHEN Flight_Status IS NULL THEN 1 ELSE 0 END) AS null_status
FROM THCK.dbo.Airline;
```

	null_passenger	null_age	null_status
1	0	0	0

Analyzed flight status distribution

```
--Trạng thái chuyến bay
SELECT Flight_Status, COUNT(*) AS total
FROM THCK.dbo.Airline
GROUP BY Flight_Status;
```

	Flight_Status	total
1	Delayed	32831
2	Cancelled	32942
3	On Time	32846

DATA TRANSFORMATION & FEATURE ENGINEERING

Create processed table with metadata & time features

```
CAST(GETDATE() AS DATE) AS load_date,

-- ======================
-- TIME FEATURES
-- ======================
YEAR(Departure_Date) AS dep_year,
MONTH(Departure_Date) AS dep_month,
DATENAME(MONTH, Departure_Date) AS dep_month_name,
DATENAME(WEEKDAY, Departure_Date) AS dep_weekday,
```

Create business flags for operational issues

```
CASE WHEN Flight_Status = 'Delayed' THEN 1 ELSE 0 END AS is_delayed,
CASE WHEN Flight_Status = 'Cancelled' THEN 1 ELSE 0 END AS is_cancelled,

CASE
    WHEN Flight_Status IN ('Delayed','Cancelled') THEN 'Issue'
    ELSE 'OnTime'
END AS flight_status_group,
```

Segment passengers and regions

```
CASE
    WHEN Age < 18 THEN 'Under 18'
    WHEN Age BETWEEN 18 AND 35 THEN '18-35'
    WHEN Age BETWEEN 36 AND 60 THEN '36-60'
    ELSE '60+'
END AS age_group,

CASE
    WHEN Continents IN (
        'Europe','North America','Asia',
        'South America','Africa','Oceania'
    )
    THEN Continents
    ELSE 'Other'
END AS continent_group
```

Store processed data

```
INTO THCK.dbo.Airline_Processed
FROM THCK.dbo.Airline;
```

DATA QUALITY CHECKS

Purpose

- Validate data completeness, consistency, and reliability before analysis.

Checks Performed

- Missing values assessment
- Duplicate records detection
- Data type validation
- Business logic consistency checks

Result

- Dataset passed all quality checks and is ready for analysis and modeling.

MISSING VALUES CHECK

Method

- Verified null values across all key columns.

Result

- No missing values detected in the dataset.

Conclusion

- Data completeness is 100%, no imputation required.

```
df.isnull().sum()
```

✓ 0.1s

```
Passenger_ID                0
First_Name                  0
Last_Name                   0
Gender                      0
Age                         0
Nationality                 0
Airport_Name                0
Airport_Country_Code        0
Country_Name                0
Continents                  0
Departure_Date              0
Arrival_Airport             0
Pilot_Name                  0
Flight_Status               0
load_date                   0
dep_year                    0
dep_month                   0
dep_month_name              0
dep_weekday                 0
is_delayed                  0
is_cancelled                0
flight_status_group         0
age_group                   0
continent_group             0
dtype: int64
```

DUPLICATE RECORDS CHECK

Method

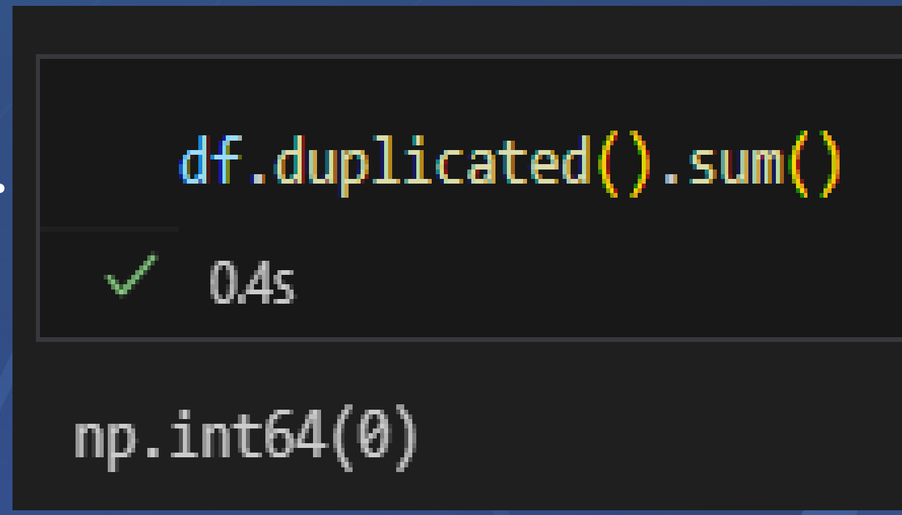
- Checked for duplicated rows at record level.

Result

- No duplicate observations found.

Conclusion

- Each record represents a unique flight-passenger instance.



BUSINESS LOGIC VALIDATION

Flight_Status vs delay and cancellation flags

```
# Validate flight status vs flags
df[['Flight_Status', 'is_delayed', 'is_cancelled']].value_counts()
```

✓ 0.0s

Flight_Status	is_delayed	is_cancelled	
Cancelled	0	1	32942
On Time	0	0	32846
Delayed	1	0	32831

Name: count, dtype: int64

Segmentation consistency

```
# Validate segmentation logic
df['age_group'].value_counts()
df['continent_group'].value_counts()
```

✓ 0.0s

continent_group

North America	32033
Asia	18637
Oceania	13866
Europe	12335
Africa	11030
South America	10718

Name: count, dtype: int64

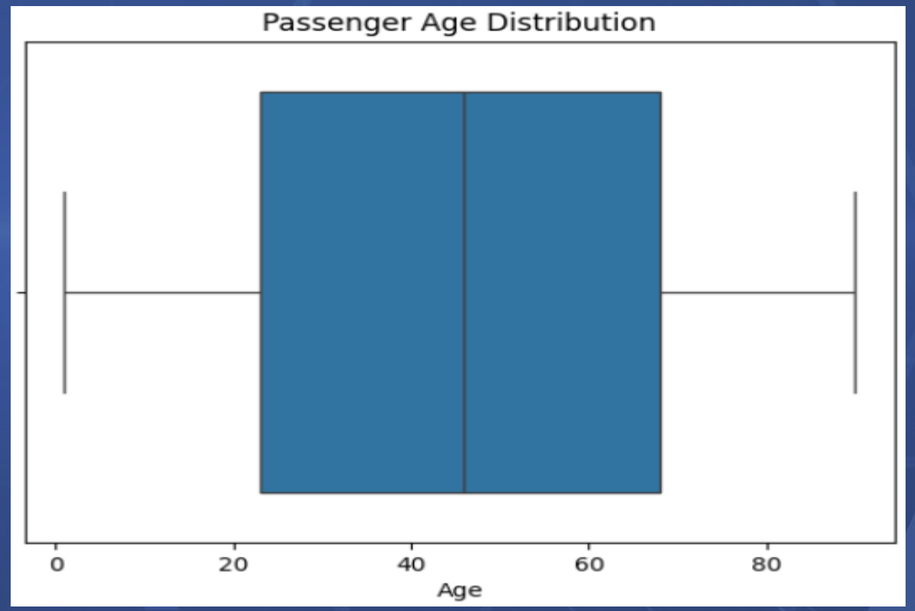
OUTLIER CHECK (AGE)

- Passenger age ranges from 1 to 90.
- No unrealistic or extreme outliers detected.

```
df['Age'].describe()
```

0.0s

```
count    98619.000000
mean        45.504021
std         25.929849
min          1.000000
25%         23.000000
50%         46.000000
75%         68.000000
max         90.000000
Name: Age, dtype: float64
```



ANALYTICS-READY DATASET

Final Dataset Includes

- Passenger demographics
- Temporal features
- Geographic segmentation
- Flight status and delay indicators

```
analysis_cols = [
    'Passenger_ID', 'Gender', 'Age', 'age_group', 'Nationality', 'Airport_Name', 'Country_Name', 'continent_group',
    'Departure_Date', 'dep_month', 'dep_month_name', 'dep_weekday', 'Flight_Status', 'is_delayed','is_cancelled']
df_analysis = df[analysis_cols].copy()
df_analysis.head(10)
```

✓ 1.2s Python

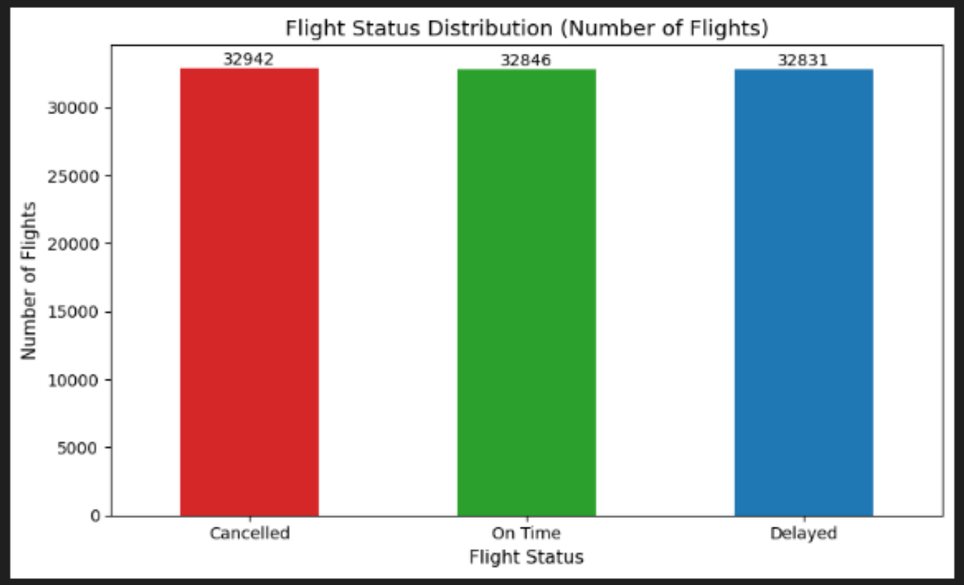
	Passenger_ID	Gender	Age	age_group	Nationality	Airport_Name	Country_Name	continent_group	Departure_Date	dep_month	dep_month_name	dep_weekday	Flight_Status	is_delayed	is_cancelled
0	j3Y5Po	Female	65	60+	China	Bazhong Enyang Airport	China	Asia	2022-01-04	1	January	Tuesday	On Time	0	0
1	zDEwOE	Male	36	36-60	Philippines	Syangboche Airport	Nepal	Asia	2022-09-22	9	September	Thursday	Cancelled	0	1
2	IDQYQJ	Male	3	Under 18	Montenegro	Sim Airport	Papua New Guinea	Oceania	2022-04-08	4	April	Friday	Delayed	1	0
3	nKDtL3	Male	19	18-35	Canada	West Angelas Airport	Australia	Oceania	2022-07-01	7	July	Friday	Delayed	1	0
4	as6oao	Female	30	18-35	China	Mc Cook Ben Nelson Regional Airport	United States	North America	2022-05-12	5	May	Thursday	Delayed	1	0
5	yS0oR7	Male	75	60+	Portugal	Penzance Heliport	United Kingdom	Europe	2022-03-26	3	March	Saturday	On Time	0	0
6	qbfwY5	Male	26	18-35	China	Cochise County Airport	United States	North America	2022-01-15	1	January	Saturday	Cancelled	0	1
7	hgZ1qd	Male	41	36-60	Albania	Nukutavake Airport	French Polynesia	Oceania	2022-04-03	4	April	Sunday	Cancelled	0	1
8	AmtY4N	Male	11	Under 18	Indonesia	Golmud Airport	China	Asia	2022-06-02	6	June	Thursday	Delayed	1	0
9	wASeJa	Female	73	60+	China	Alagoinhas Airport	Brazil	South America	2022-09-10	9	September	Saturday	Cancelled	0	1

OVERALL FLIGHT PERFORMANCE

- Flight statuses (On Time, Delayed, Cancelled) are nearly evenly distributed.
- No unrealistic or extreme outliers detected.

```python
# Count số lượng chuyến bay theo trạng thái
status_count = df_analysis['Flight_Status'].value_counts()
# Định nghĩa màu theo ý nghĩa business
color_map = {
    'On Time': '#2ca02c',
    'Delayed': '#1f77b4',
    'Cancelled': '#d62728'    }
# Ánh xạ màu theo thứ tự index
colors = [color_map.get(status, '#7f7f7f') for status in status_count.index]
# Vẽ biểu đồ
ax = status_count.plot(
    kind='bar',
    color=colors)
# Trang trí biểu đồ
plt.title("Flight Status Distribution (Number of Flights)")
plt.xlabel("Flight Status")
plt.ylabel("Number of Flights")
plt.xticks(rotation=0)
# Hiển thị label số lượng trên mỗi cột
for i, v in enumerate(status_count.values):
    ax.text(i, v, f"{v}", ha='center', va='bottom')
plt.tight_layout()
plt.show()
```

✓ 0.1s



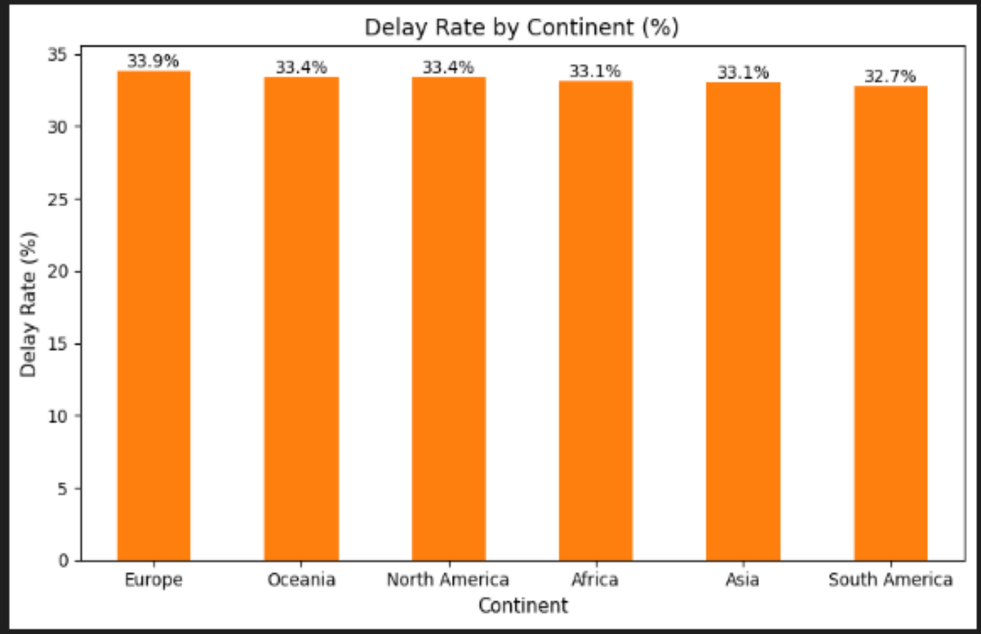
DELAY RATE BY CONTINENT

- Delay rates are highly consistent across continents (~32.7%–33.9%).
- No continent shows extreme deviation from the global average.

```
delay_by_continent = (
    df_analysis
    .groupby('continent_group')['is_delayed']
    .mean()
    .sort_values(ascending=False) * 100)
ax = delay_by_continent.plot(
    kind='bar',
    color=COLOR_WARN)
plt.title("Delay Rate by Continent (%)")
plt.xlabel("Continent")
plt.ylabel("Delay Rate (%)")
plt.xticks(rotation=0)

# Label %
for i, v in enumerate(delay_by_continent.values):
    ax.text(i, v, f"{v:.1f}%", ha='center', va='bottom')
plt.tight_layout()
plt.show()
```

✓ 0.1s



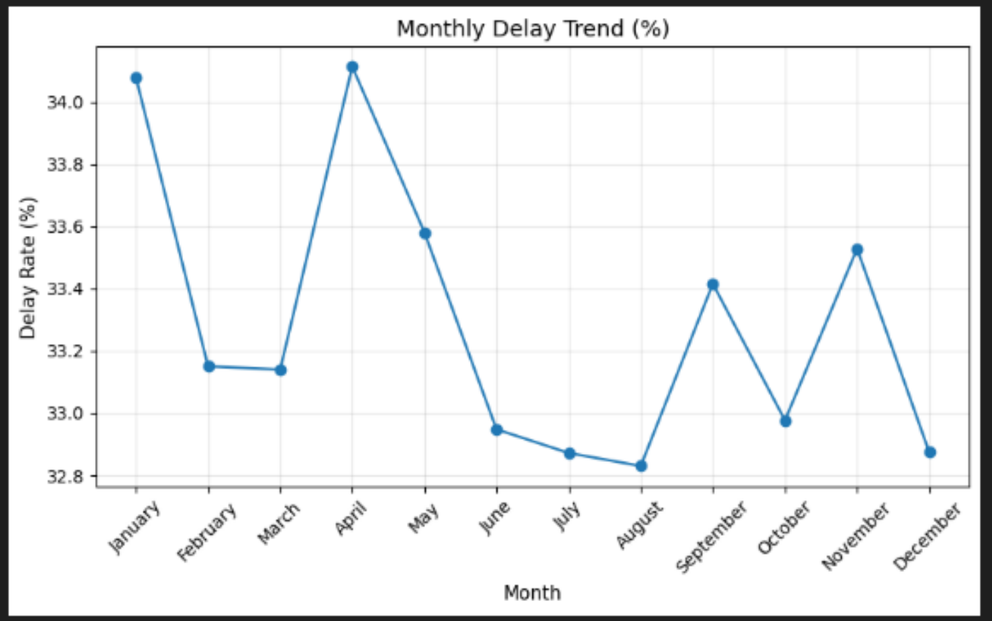
MONTHLY DELAY TREND

- Delay rates peak in January and April.
- Lowest delay levels occur around July-August.
- Overall variation remains moderate throughout the year.

```
delay_by_month = (
    df_analysis
    .groupby(['dep_month', 'dep_month_name'])['is_delayed']
    .mean()
    .reset_index()
    .sort_values('dep_month'))
plt.plot(
    delay_by_month['dep_month_name'],
    delay_by_month['is_delayed'] * 100,
    marker='o',
    color=COLOR_NEUTRAL)
plt.title("Monthly Delay Trend (%)")
plt.xlabel("Month")
plt.ylabel("Delay Rate (%)")
plt.xticks(rotation=45)
plt.grid(alpha=0.3)

plt.tight_layout()
plt.show()
```

✓ 0.2s

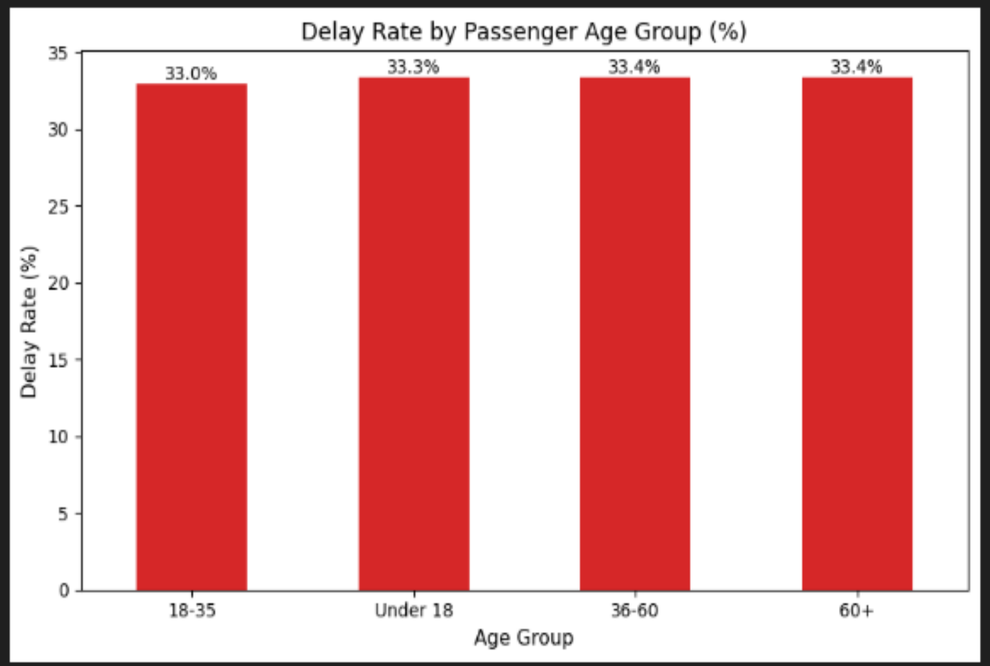


DELAY RATE BY PASSENGER AGE GROUP

- Delay rates are nearly identical across all age groups (~33%).
- No age group exhibits significantly higher or lower delay risk.

```
delay_by_age = (
    df_analysis
    .groupby('age_group')['is_delayed']
    .mean()
    .sort_values() * 100)
ax = delay_by_age.plot(
    kind='bar',
    color=COLOR_BAD)
plt.title("Delay Rate by Passenger Age Group (%)")
plt.xlabel("Age Group")
plt.ylabel("Delay Rate (%)")
plt.xticks(rotation=0)
for i, v in enumerate(delay_by_age.values):
    ax.text(i, v, f"{v:.1f}%", ha='center', va='bottom')
plt.tight_layout()
plt.show()
```

✓ 2.4s

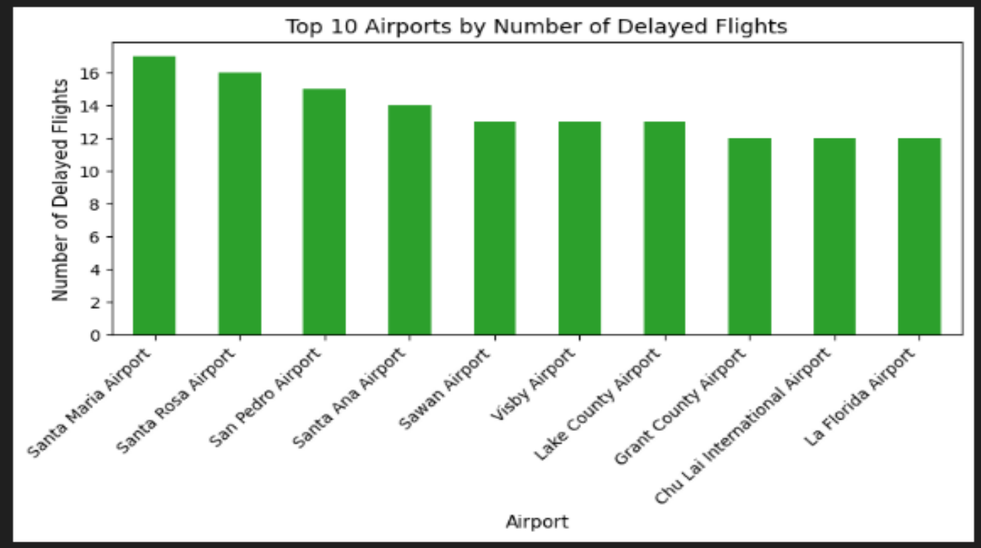


TOP AIRPORTS CONTRIBUTING TO DELAYS

- A small number of airports contribute disproportionately to total delays.
- Santa Maria Airport and Santa Rosa Airport show the highest delayed flight counts.

```
airport_delay = (
    df_analysis
    .groupby('Airport_Name')['is_delayed']
    .agg(
        total_flights='count',
        delayed_flights='sum')
    .reset_index())
airport_delay['delay_rate'] = (
    airport_delay['delayed_flights'] /
    airport_delay['total_flights'])
top_airports = airport_delay.sort_values(
    'delayed_flights',
    ascending=False
).head(10)
ax = top_airports.plot(
    x='Airport_Name',
    y='delayed_flights',
    kind='bar',
    color=COLOR_GOOD,
    legend=False)
plt.title("Top 10 Airports by Number of Delayed Flights")
plt.xlabel("Airport")
plt.ylabel("Number of Delayed Flights")
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

✓ 1.7s



VISUALIZATION OVERVIEW: OPERATIONAL PERFORMANCE

- Global coverage: 235 countries and 9,061 airports provide a broad geographic footprint suitable for cross-region comparisons.
- Regional concentration: North America contributes the highest volume at roughly 32K flights, followed by Asia at about 19K and Oceania at about 14K, so overall traffic is strongly driven by North America.
- Passenger mix: Gender distribution is nearly even, with around 50K male and 49K female passengers, suggesting minimal demographic skew at the top-line level.
- Monthly stability: Flight volume remains within a narrow band of about 7.7K to 8.5K flights per month, showing mild seasonality rather than large fluctuations.
- Outcome distribution: On-time, Delayed, and Cancelled are close to evenly split at roughly 33K each, which supports fair comparison across outcomes and reduces class imbalance for analysis.

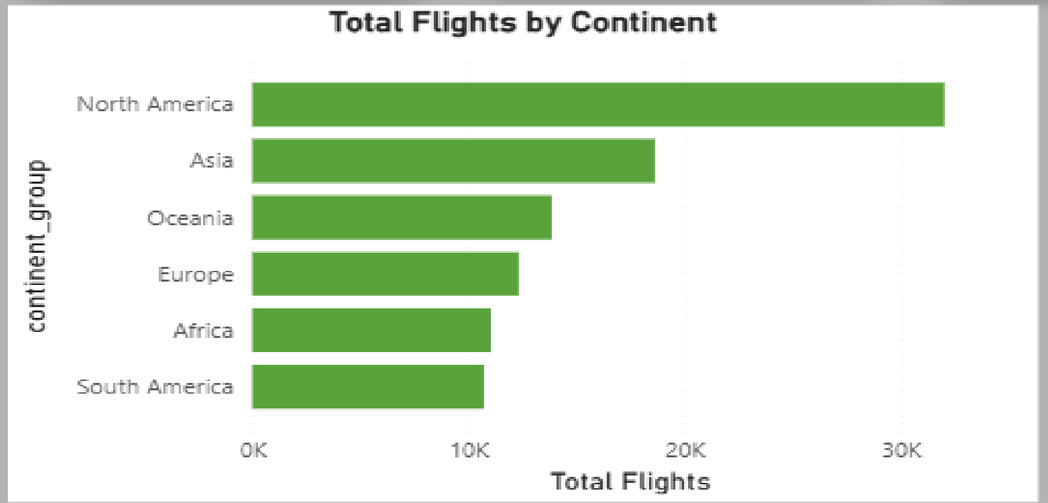
EXECUTIVE OVERVIEW

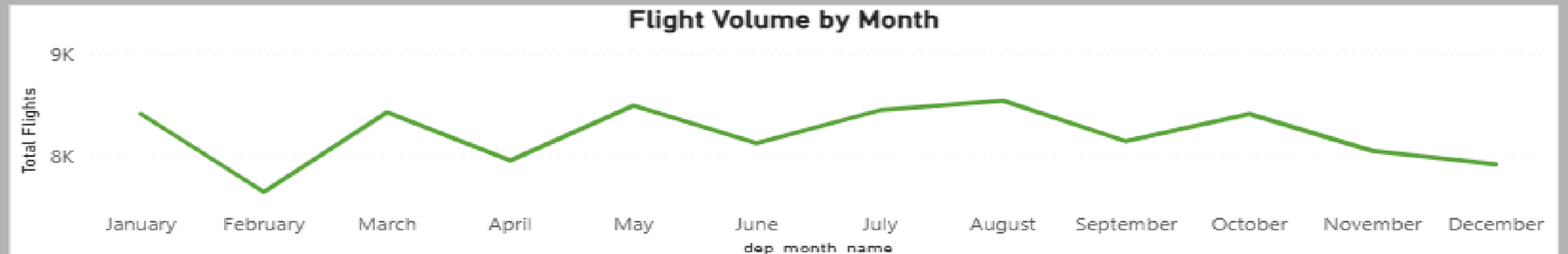
98619
Total Flights

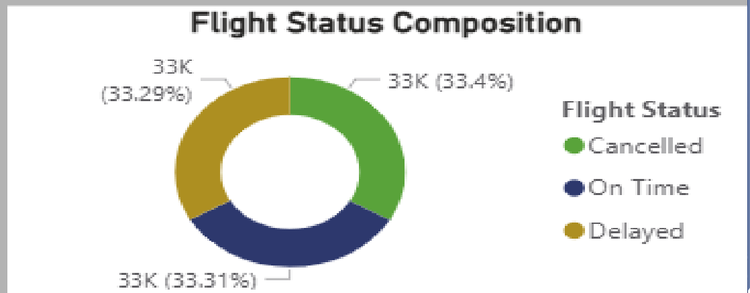
235
Total Countries

9061
Total Airports

dep_month
1
12







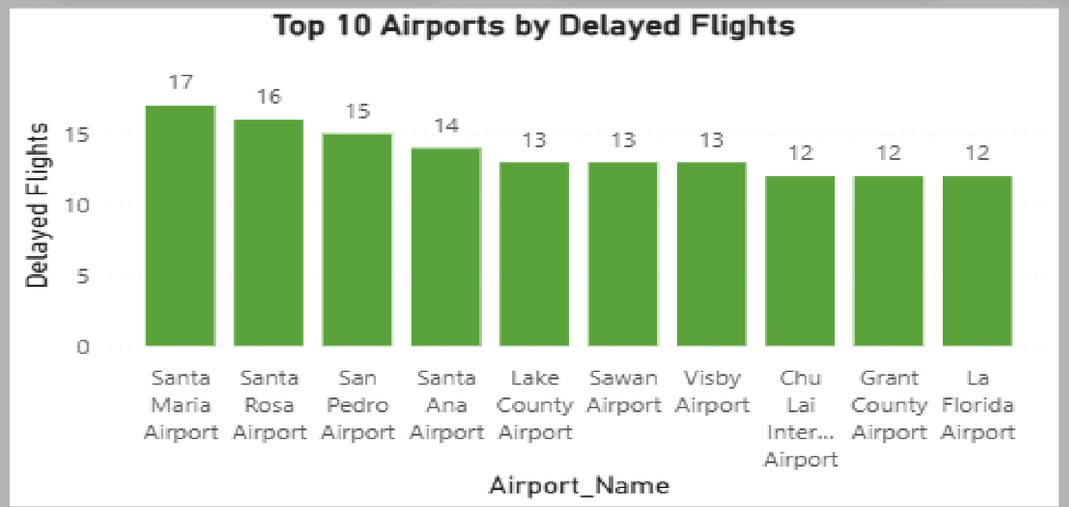
VISUALIZATION INSIGHTS: DELAYS & OPERATIONAL HOTSPOTS

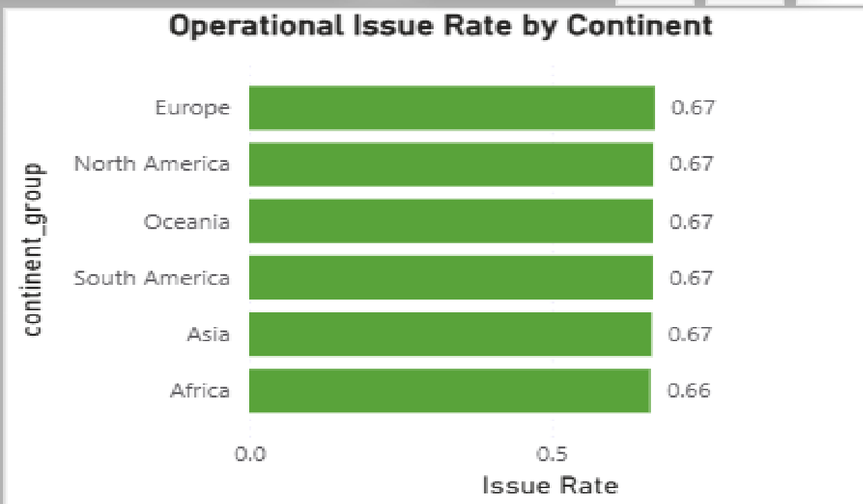
PERFORMANCE & HOTSPOTS

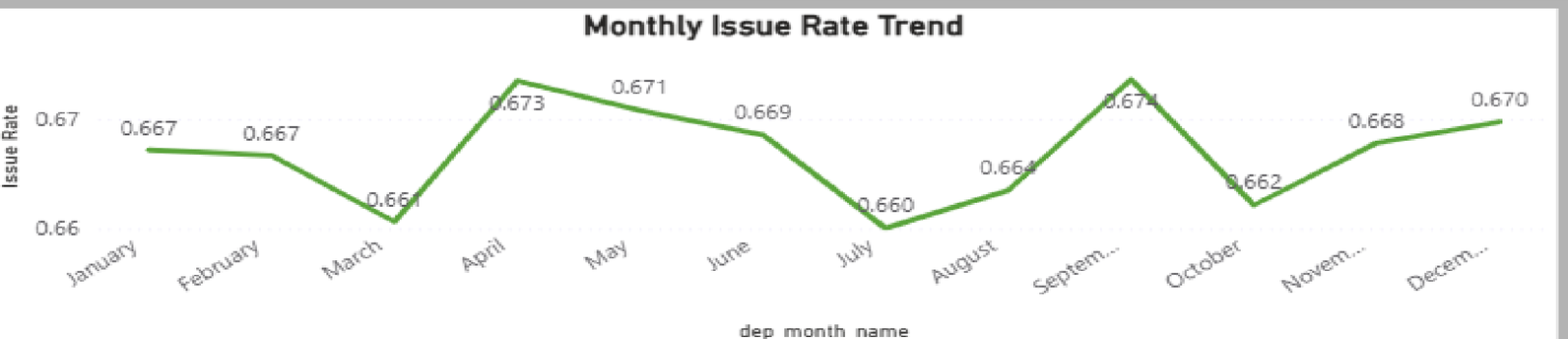
- Service reliability is weak: On-time Rate 0.33 vs Issue Rate 0.67, with a Reliability Score of 33.31.
- Rates are similar across continents: Issue Rate is 0.67 for most regions; Africa is slightly better at 0.66 (best Reliability Score 33.72).
- North America is the main exposure driver: Highest volume (32,033 flights), so it contributes the most to total issues even with similar rates.
- Performance is stable by month: Issue Rate stays within 0.660–0.673; highest in April (0.673), lowest in July (0.660).
- Delay hotspots are concentrated: Top delayed airports range 12–17 delays; highest is Santa Maria Airport (17) and Santa Rosa Airport (16).

0.33	0.67	33.31
On-time Rate	Issue Rate	Reliability Score

dep_month: 1 – 12







Operational Performance Summary by Continent

continent_group	Total Flights	On-time Rate	Issue Rate	Reliability Score
Africa	11030	0.34	0.66	33.72
Asia	18637	0.33	0.67	33.49
Europe	12335	0.33	0.67	32.93
North America	32033	0.33	0.67	33.23
Oceania	13866	0.33	0.67	33.27
South America	10718	0.33	0.67	33.27
Total	**98619**	**0.33**	**0.67**	**33.31**

MODEL TRAINING OVERVIEW

Objective: predict flight delays using classification models
Target variable: is_delayed (1 = delayed, 0 = others)
Approach:

- Train multiple models for comparison
- Handle class imbalance with class weighting

Models:

- Logistic regression
- Random forest
- Gradient boosting

MODEL COMPARISON

- Accuracy: Overall prediction correctness.
- Precision: Reliability of predicted delays.
- Recall: Ability to correctly identify delayed flights (key focus).
- F1-score: Balance between precision and recall.
- ROC-AUC: Ability to separate delayed vs non-delayed classes.

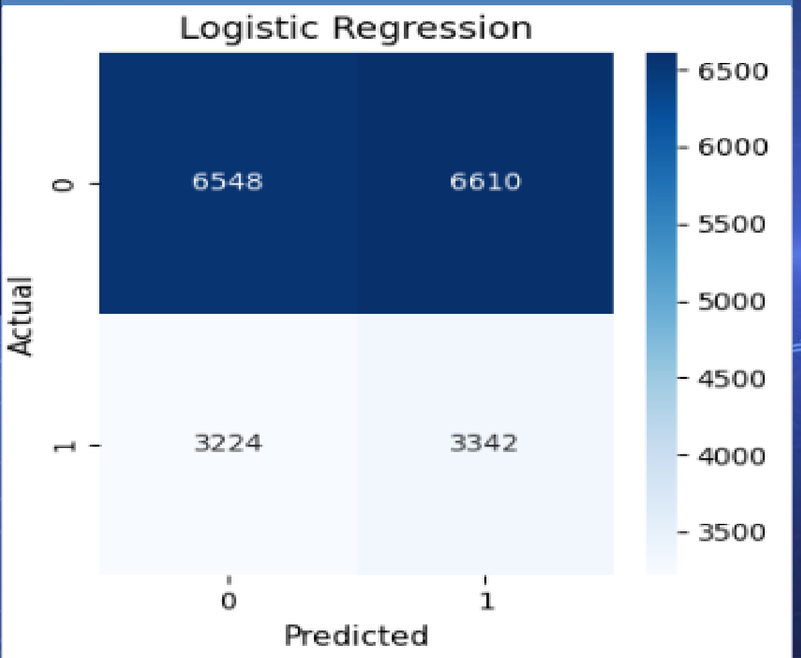
EVALUATION METRICS

	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.501420	0.335812	0.508986	0.404649	0.501444
Random Forest	0.535591	0.336978	0.408316	0.369233	0.503197
Gradient Boosting	0.667055	0.333333	0.000152	0.000304	0.503202

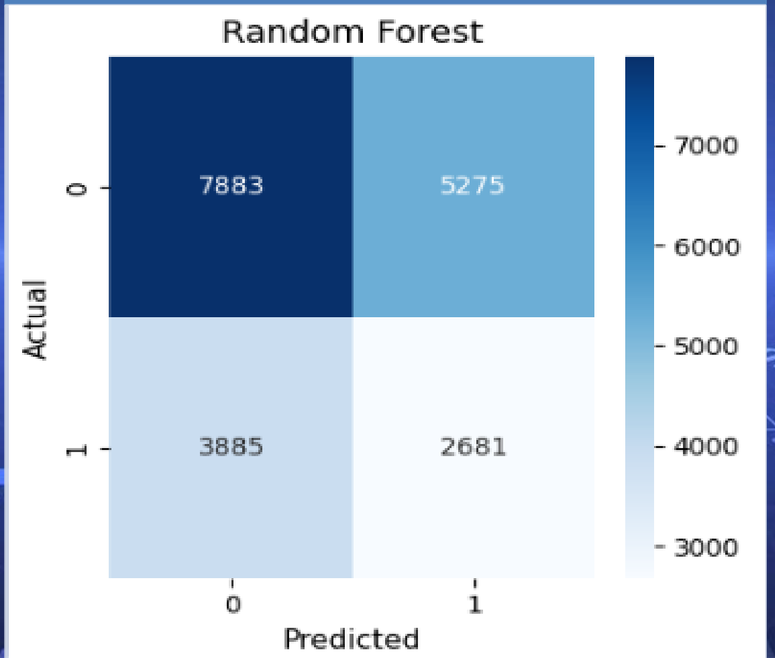
- All models show ROC-AUC ≈ 0.50.
- Limited separation between delayed vs non-delayed flights.
- Logistic Regression achieves the highest recall.
- Tree-based models trade recall for accuracy.

CONFUSION MATRIX

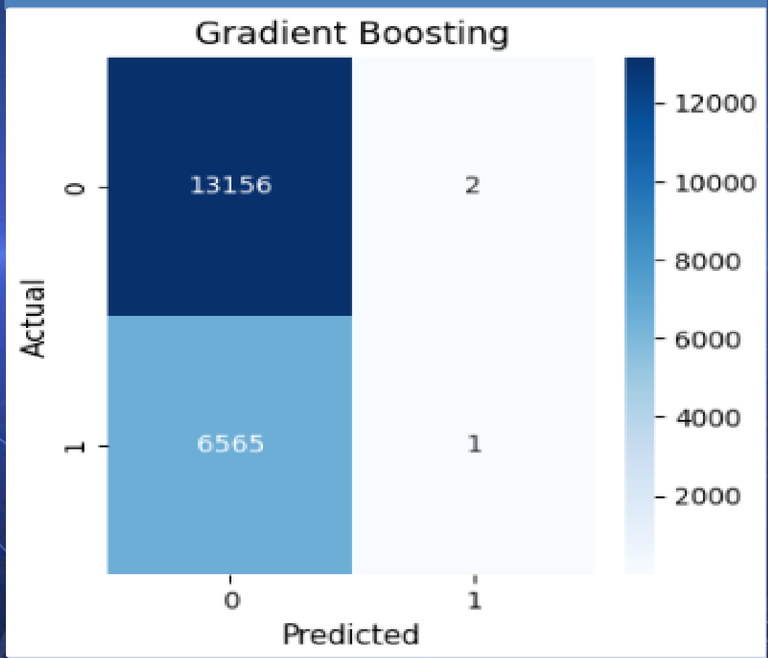
- Detects the most delayed flights
- Accepts higher false positives



Misses more delayed flights

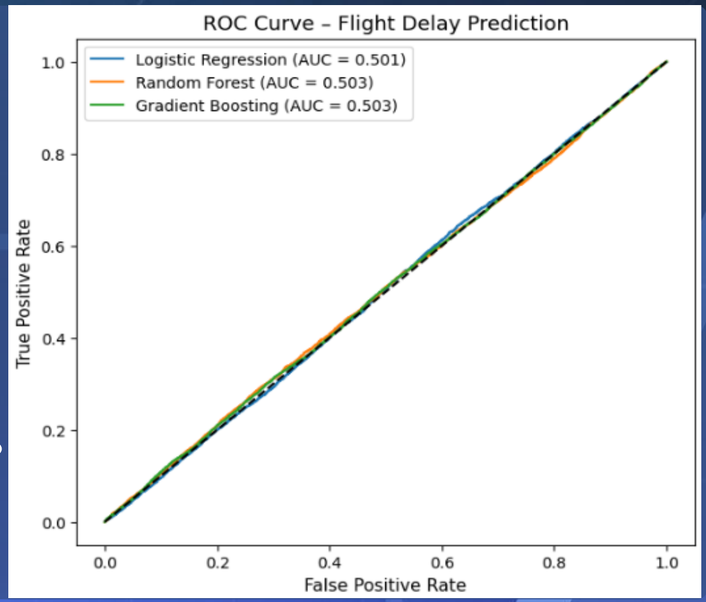


Almost fails to detect delays



ROC CURVE SUMMARY

- ROC curves are close to the diagonal.
- Indicates near-random classification performance .
- Performance limited by feature quality, not model choice.



FINAL MODEL SELECTION

- ***Selected Model:*** Logistic Regression.
- ***Reason:***
  - Best recall for delay detection.
  - More suitable for operational risk screening.
  - Preferred when missing delays is costly.

MODELING LIMITATIONS

- No weather or operational constraint data.
- Delay patterns are globally consistent.
- Weak predictive signal in available features.

RECOMMENDATIONS

- Use Logistic Regression as early warning tool.
- Focus improvements on:
  - Weather data
  - Airport congestion
  - Aircraft turnaround time
- Rebuild models with enriched features

THANK YOU FOR LISTENING!