Navigating the Skies: Analyzing Flight Delay Trends in the US

A Data-Driven Approach to Analyzing and predicting Flight Delays Across US Airports

Index

- Business Questions
- Dataset
- Data Scrubbing
- EDA
- Delay Prediction Model
- Conlusion
- Recommendation
- Next Steps
- One Pager

Business Questions

Business Questions

- Which airports have the highest percentage of delayed flights?
- What is the average duration of flight delays?
- Are there certain times of day, days of the week, or months of the year when delays are more common?
- What is the impact of weather, air traffic control, and other external factors on flight delays?
- How do different types of delays (e.g. late aircraft, airline-related delays, weather-related delays) affect the overall on-time performance of airlines?
- Are there opportunities for airlines to improve their operations and reduce delays, and if so, what specific strategies should they pursue?

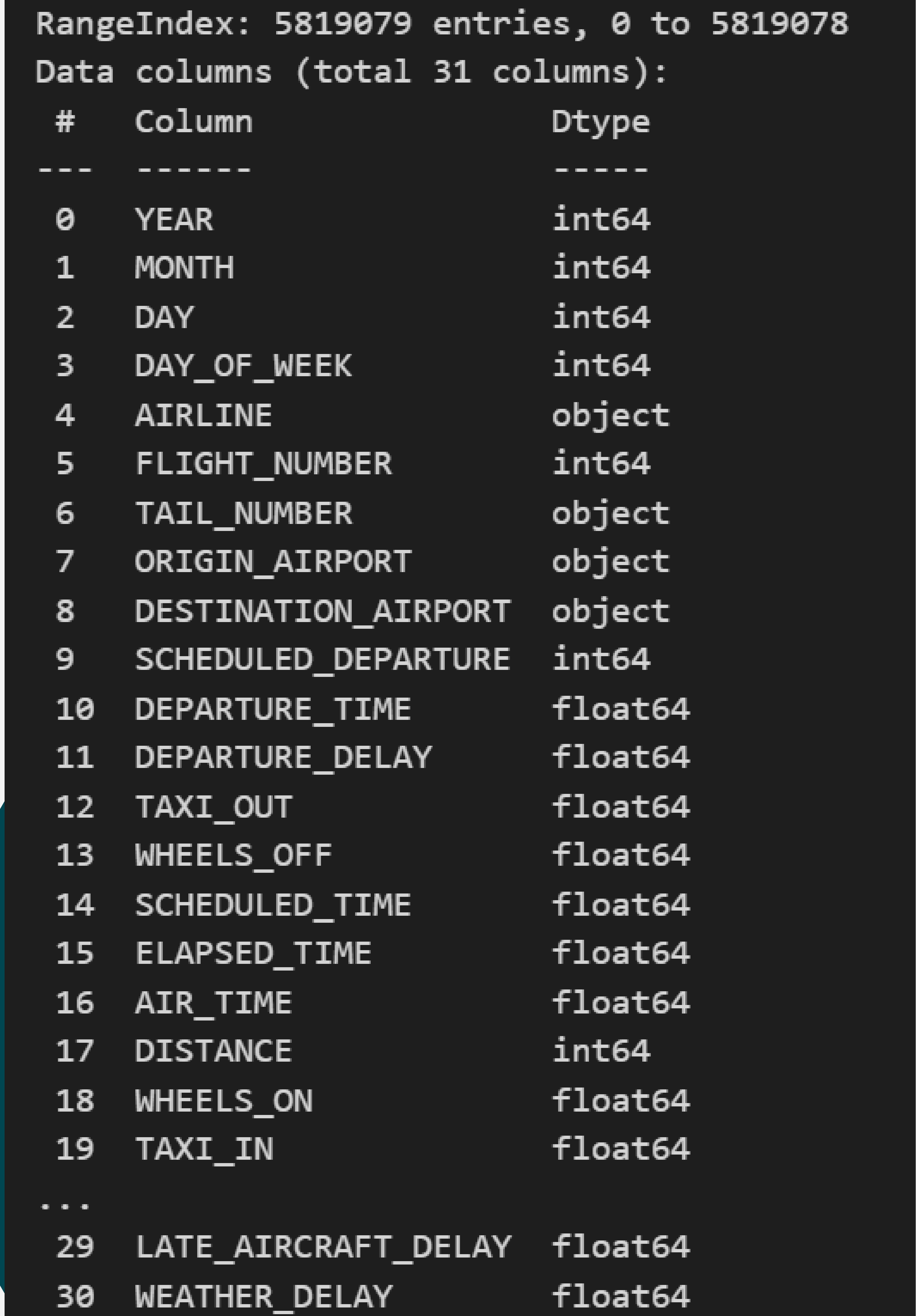
Dataset

https://www.kaggle.com/datasets/usdot/flight-delays

The Data

- Columns:

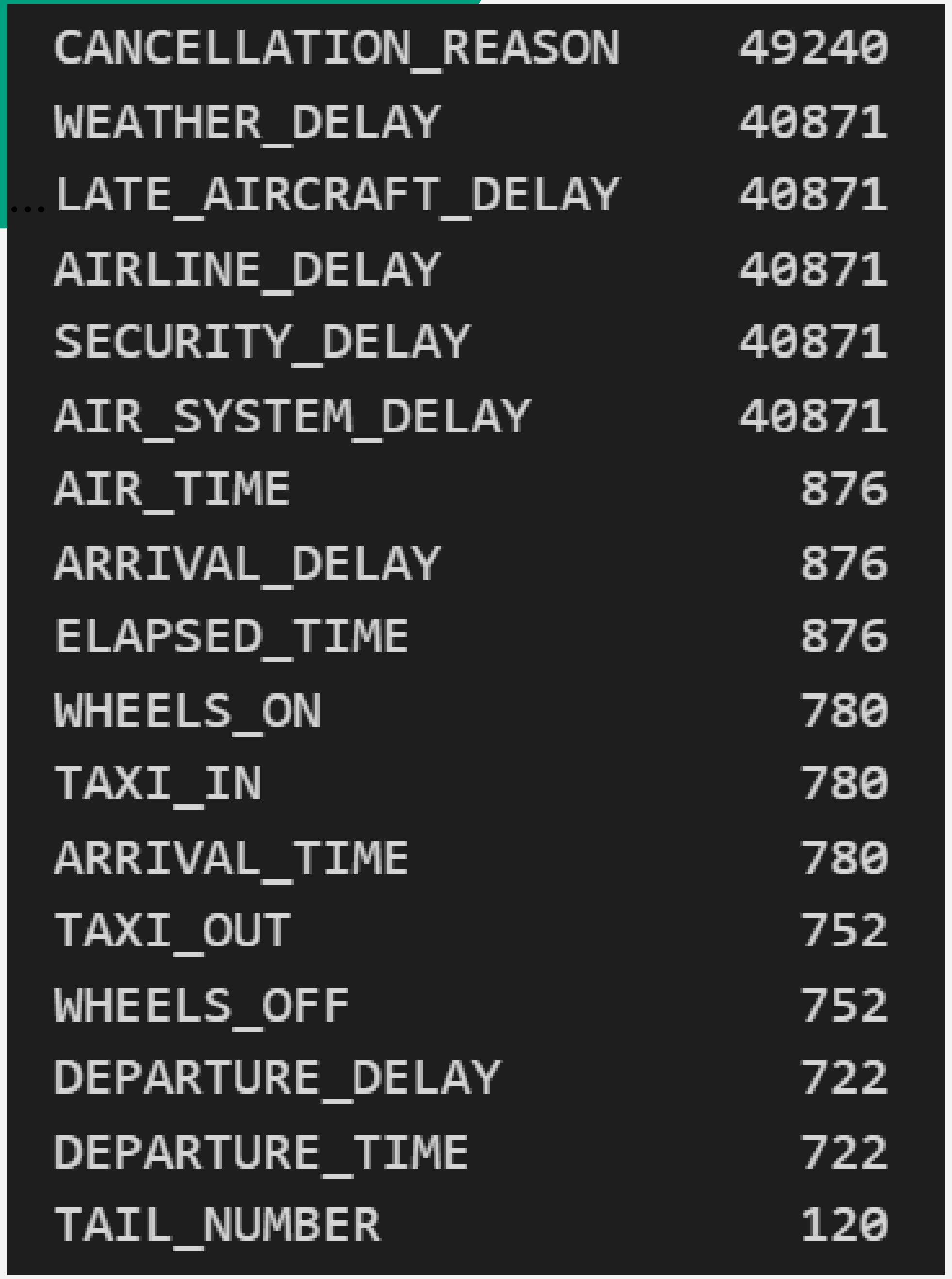
['YEAR', 'MONTH', 'DAY', 'DAY_OF_WEEK', 'AIRLINE', 'FLIGHT_NUMBER', 'TAIL_NUMBER', 'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT', 'SCHEDULED_DEPARTURE', 'DEPARTURE_TIME', 'DEPARTURE_DELAY', 'TAXI_OUT', 'WHEELS_OFF', 'SCHEDULED_TIME', 'ELAPSED_TIME', 'AIR_TIME', 'DISTANCE', 'WHEELS_ON', 'TAXI_IN', 'SCHEDULED_ARRIVAL', 'ARRIVAL_TIME', 'ARRIVAL_DELAY', 'DIVERTED', 'CANCELLED', 'CANCELLATION_REASON', 'AIR_SYSTEM_DELAY', 'SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY']



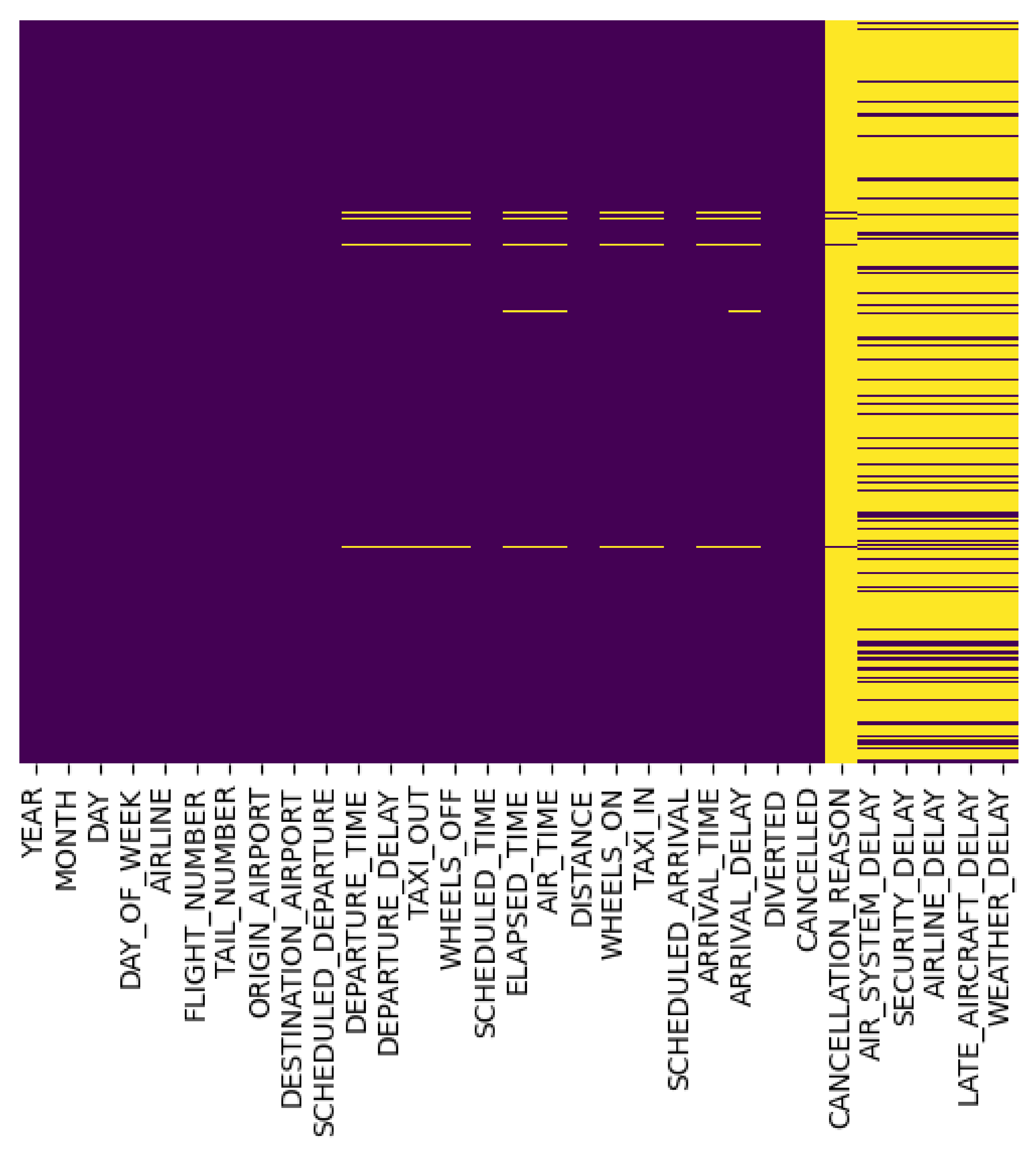
```
RangeIndex: 5819079 entries, 0 to 5819078
Data columns (total 31 columns):
 #   Column               Dtype
---  ------               -----
 0   YEAR                 int64
 1   MONTH                int64
 2   DAY                  int64
 3   DAY_OF_WEEK          int64
 4   AIRLINE              object
 5   FLIGHT_NUMBER        int64
 6   TAIL_NUMBER          object
 7   ORIGIN_AIRPORT       object
 8   DESTINATION_AIRPORT  object
 9   SCHEDULED_DEPARTURE  int64
 10  DEPARTURE_TIME       float64
 11  DEPARTURE_DELAY      float64
 12  TAXI_OUT             float64
 13  WHEELS_OFF           float64
 14  SCHEDULED_TIME       float64
 15  ELAPSED_TIME         float64
 16  AIR_TIME             float64
 17  DISTANCE             int64
 18  WHEELS_ON            float64
 19  TAXI_IN              float64
...
 29  LATE_AIRCRAFT_DELAY  float64
 30  WEATHER_DELAY        float64
```

Sample Data

Sample of 50,000



```
CANCELLATION_REASON    49240
WEATHER_DELAY          40871
LATE_AIRCRAFT_DELAY    40871
AIRLINE_DELAY          40871
SECURITY_DELAY         40871
AIR_SYSTEM_DELAY       40871
AIR_TIME                 876
ARRIVAL_DELAY            876
ELAPSED_TIME             876
WHEELS_ON                780
TAXI_IN                  780
ARRIVAL_TIME             780
TAXI_OUT                 752
WHEELS_OFF               752
DEPARTURE_DELAY          722
DEPARTURE_TIME           722
TAIL_NUMBER              120
```



Data Scrubbing

Dealing with Missing Values

CANCELLATION_REASON

Delete column.

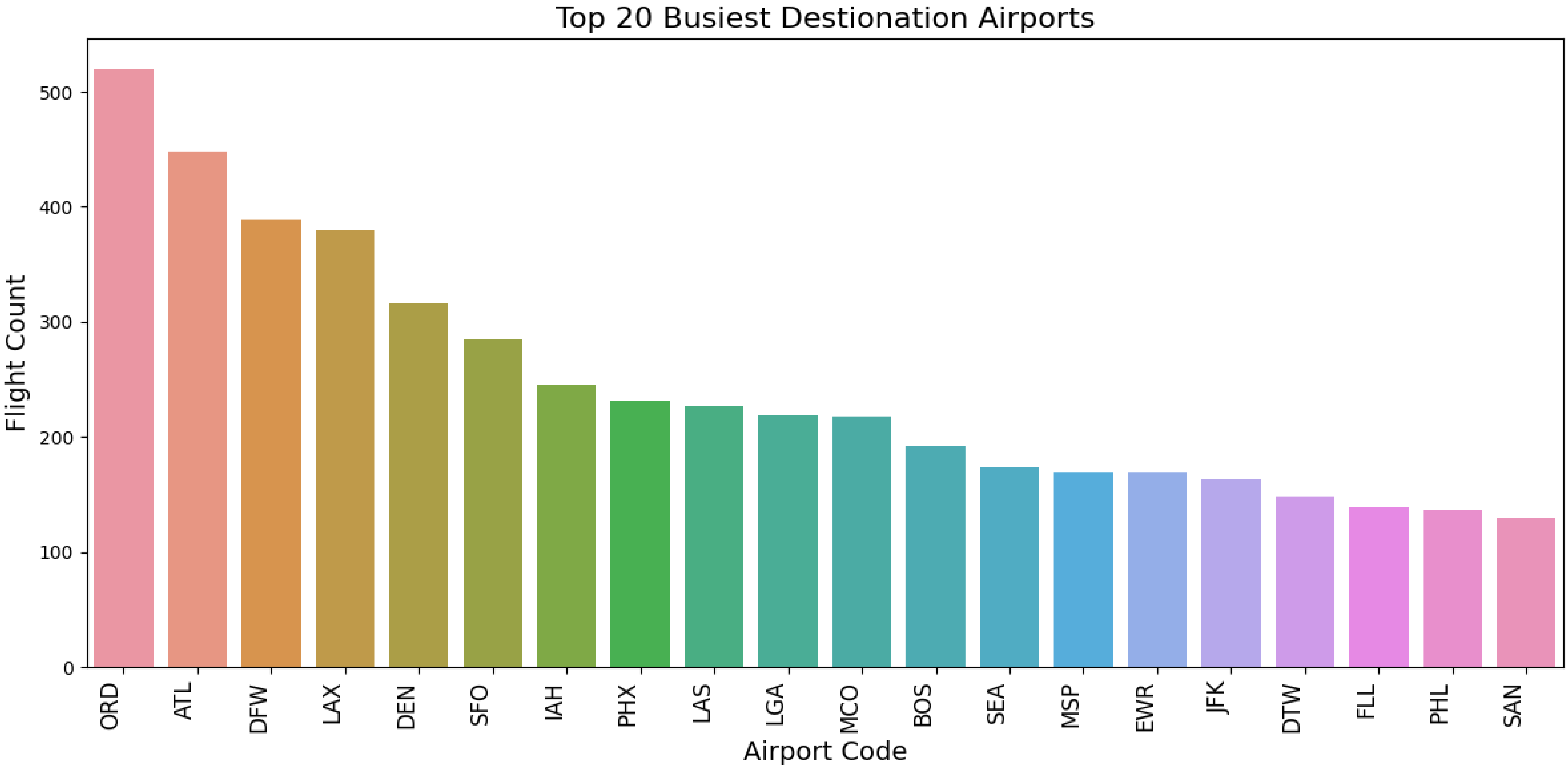
WEATHER_DELAY, LATE_AIRCRAFT_DELAY, AIRLINE_DELAY, SECURITY_DELAY, AIR_SYSTEM_DELAY

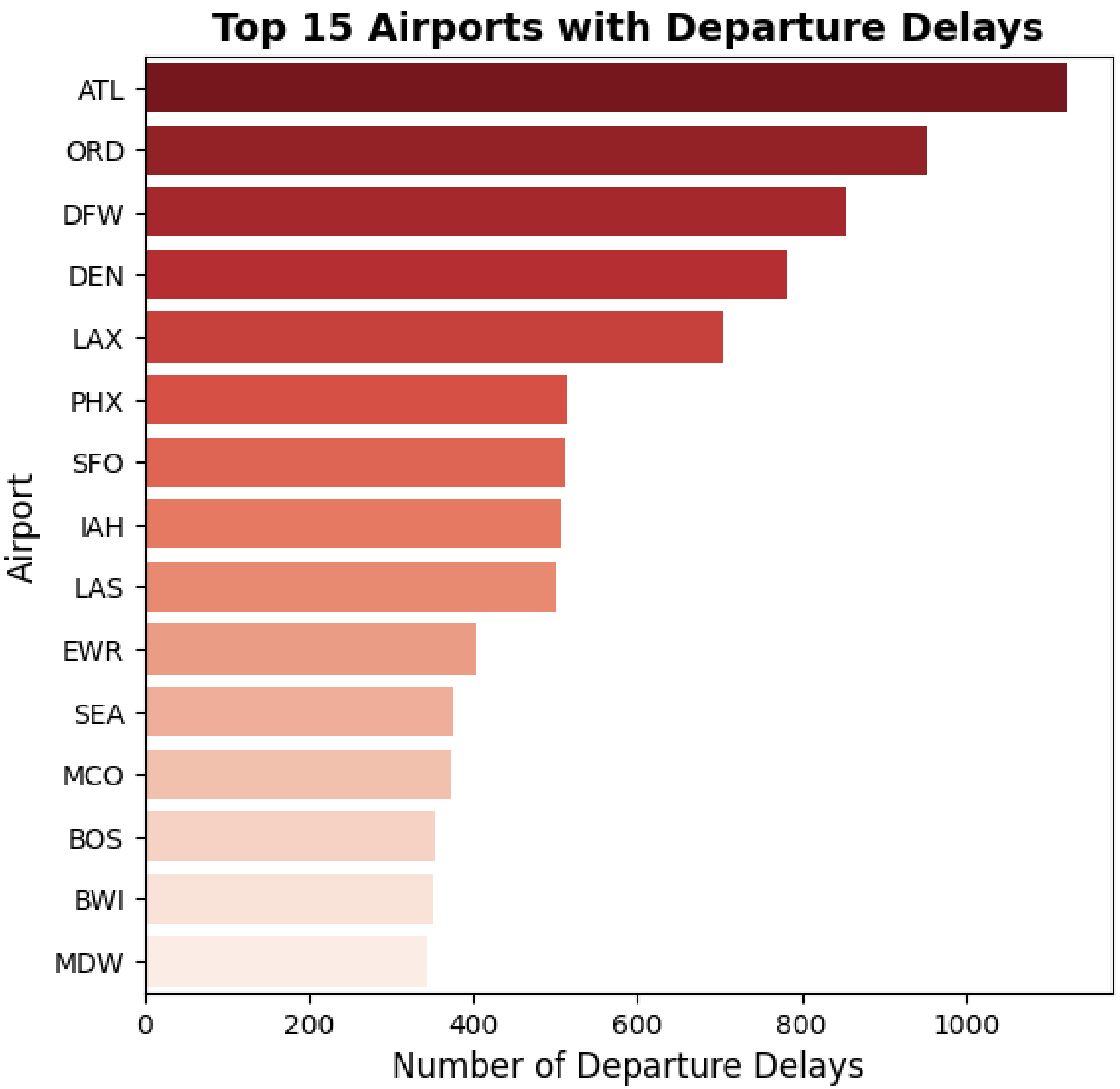
Fill with 0

Rows with Missing Values

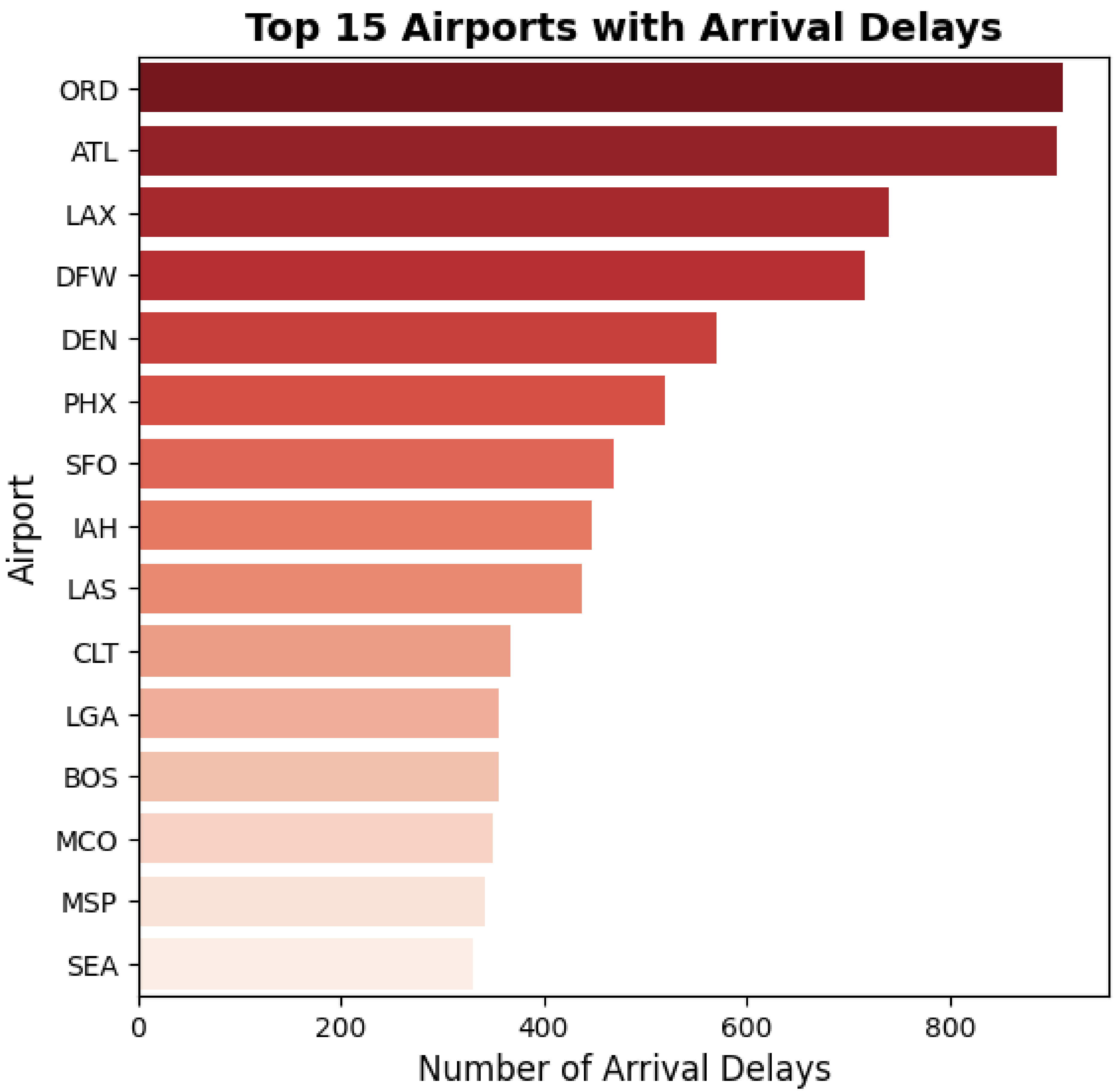
Drop Row

EDA

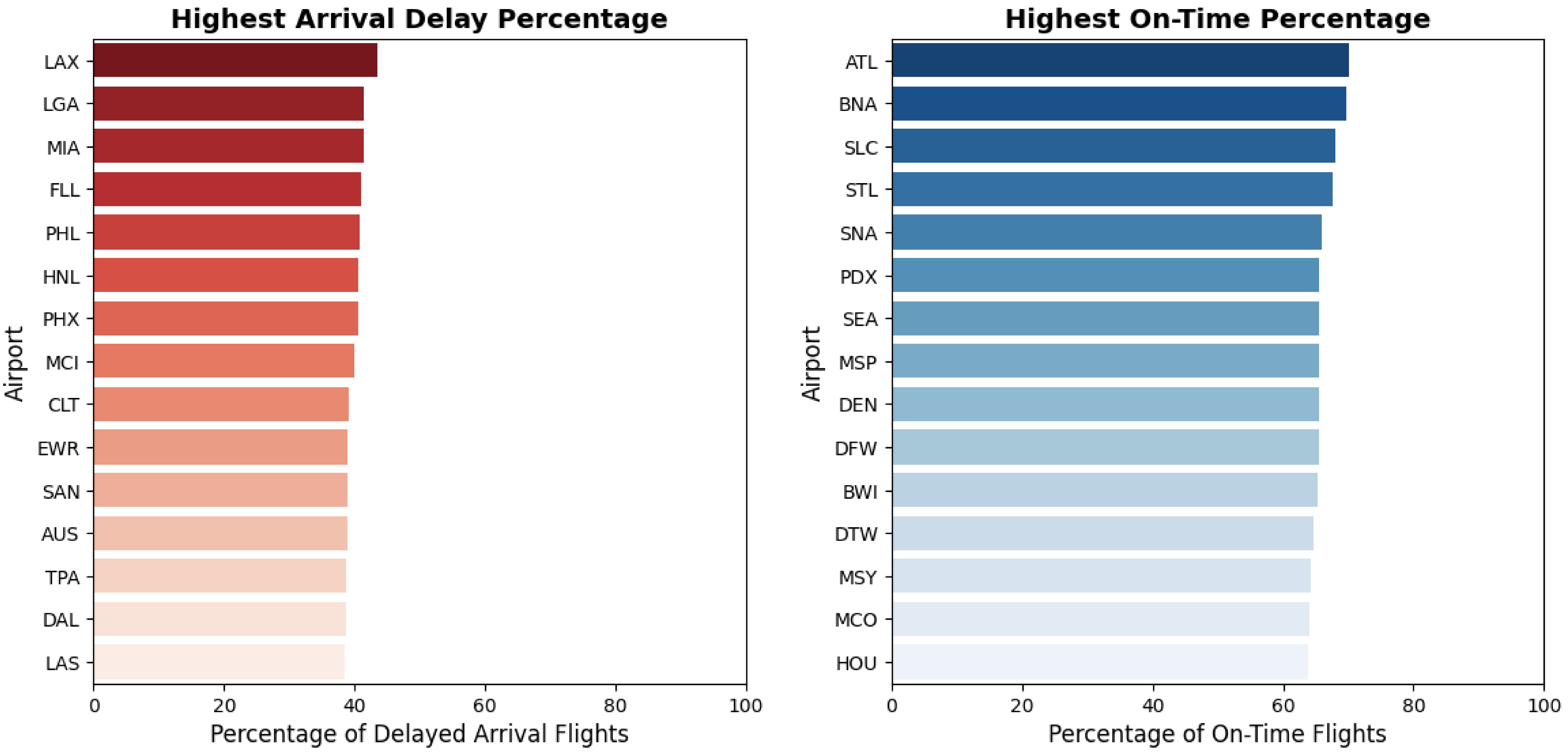

Top 20 Busiest Destionation Airports
Flight Count
0
100
200
300
400
500
ORD
ATL
DFW
LAX
DEN
SFO
IAH
PHX
LAS
LGA
MCO
BOS
SEA
MSP
EWR
JFK
DTW
FLL
PHL
SAN
Airport Code

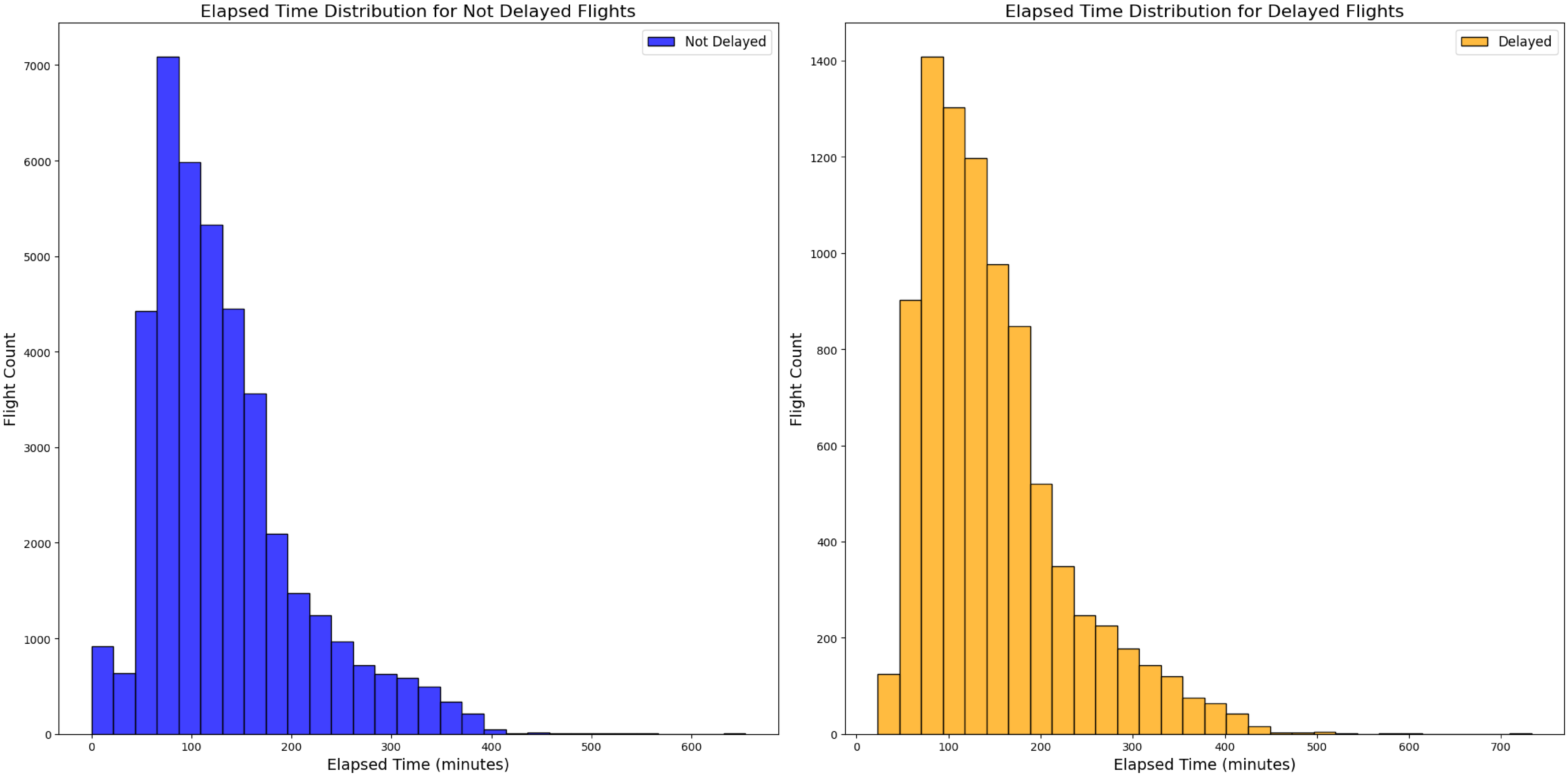


Top 15 Airports with Departure Delays



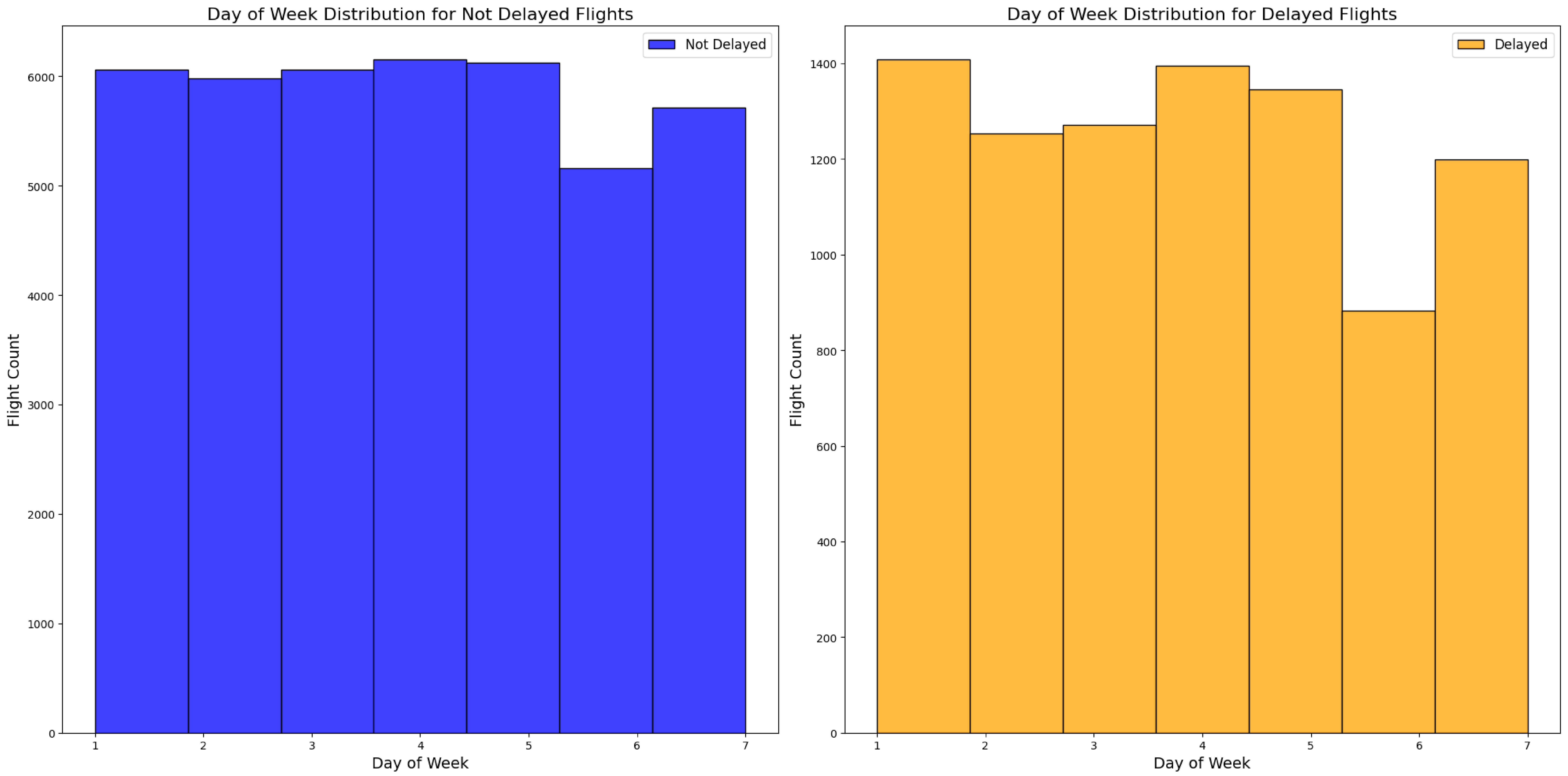
Top 15 Airports with Arrival Delays

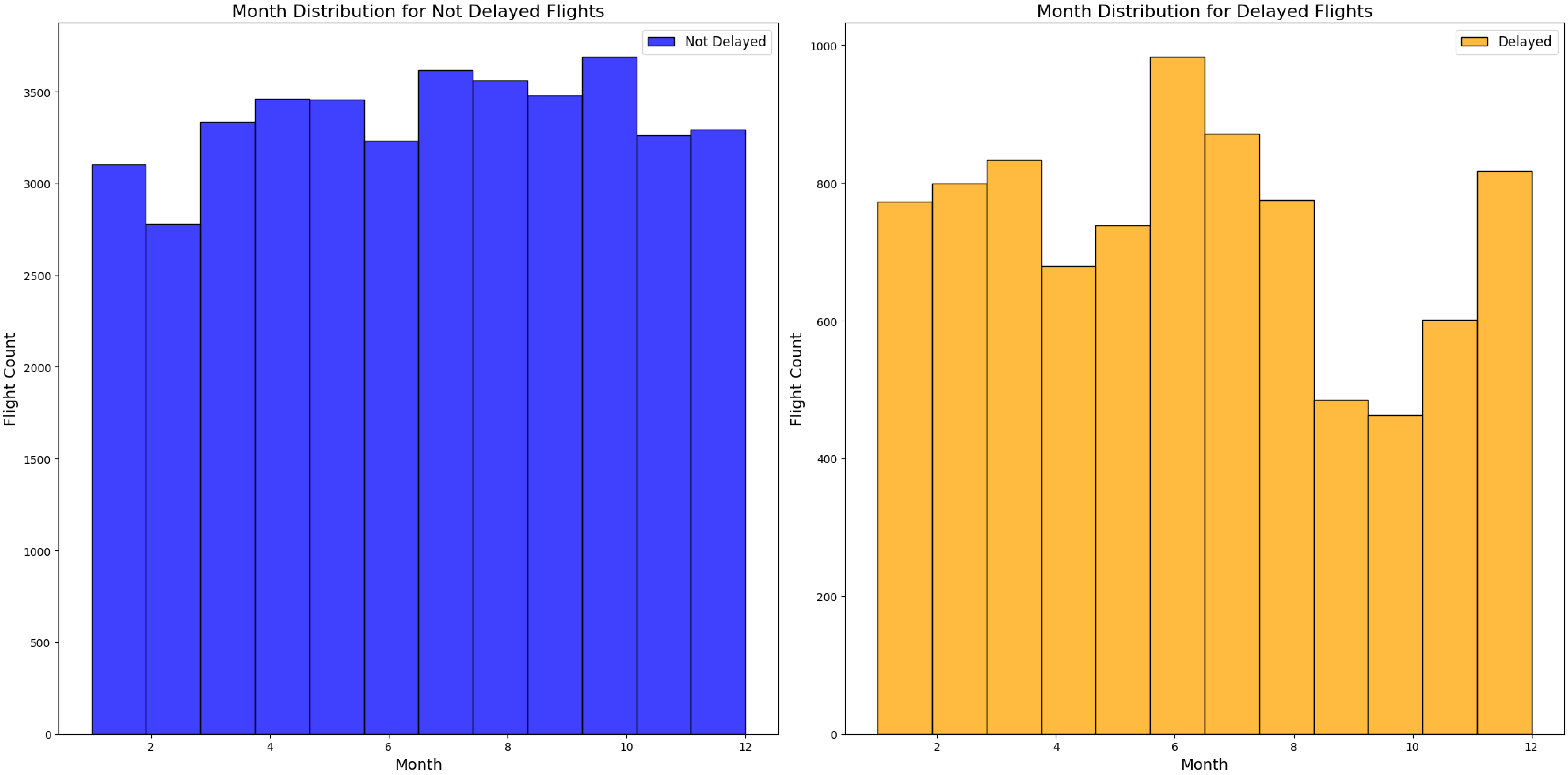

Highest Arrival Delay Percentage
Airport
LAX
LGA
MIA
FLL
PHL
HNL
PHX
MCI
CLT
EWR
SAN
AUS
TPA
DAL
LAS
0
20
40
60
80
100
Percentage of Delayed Arrival Flights
Highest On-Time Percentage
Airport
ATL
BNA
SLC
STL
SNA
PDX
SEA
MSP
DEN
DFW
BWI
DTW
MSY
MCO
HOU
0
20
40
60
80
100
Percentage of On-Time Flights

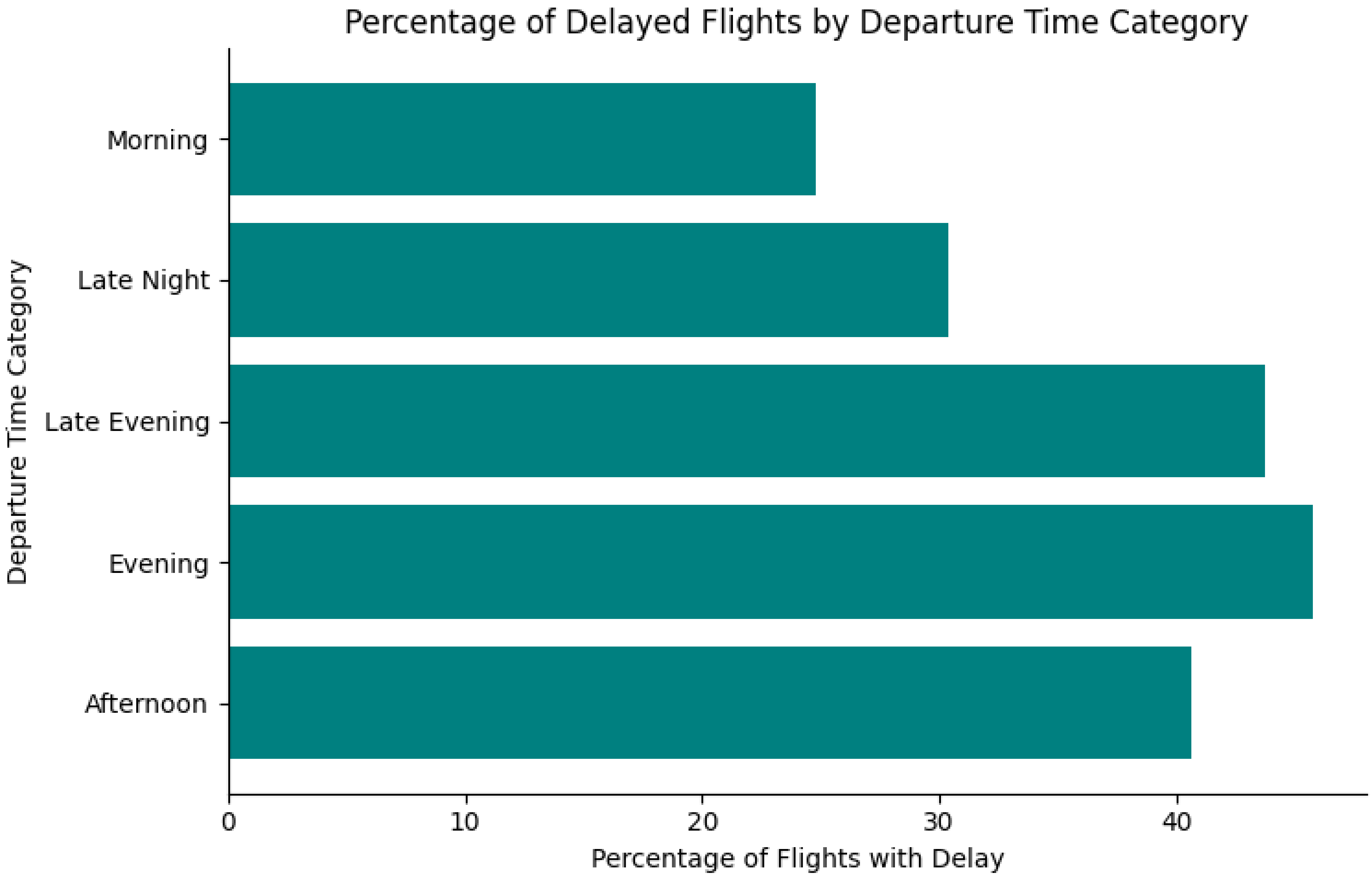


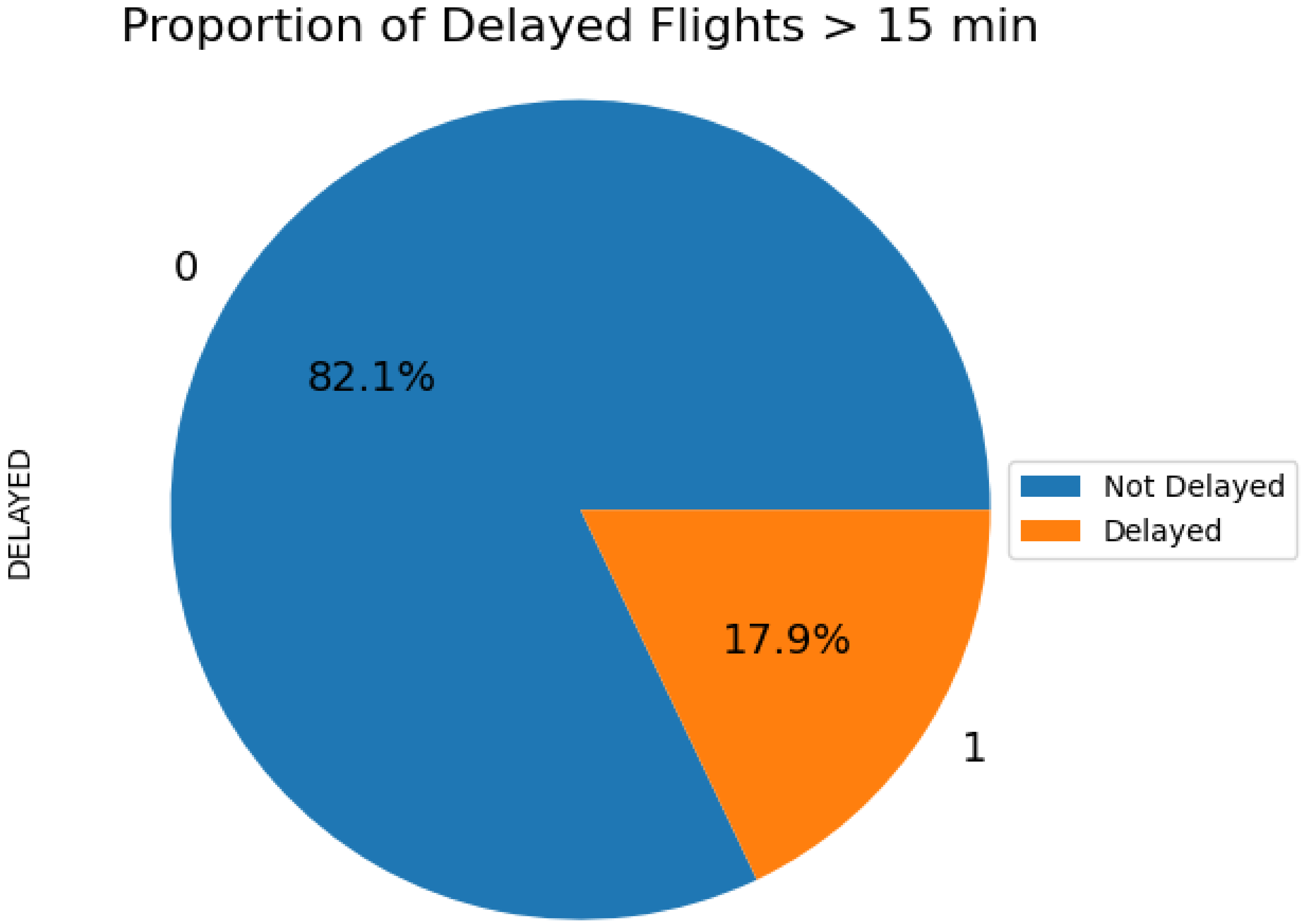
Elapsed Time Distribution for Not Delayed Flights

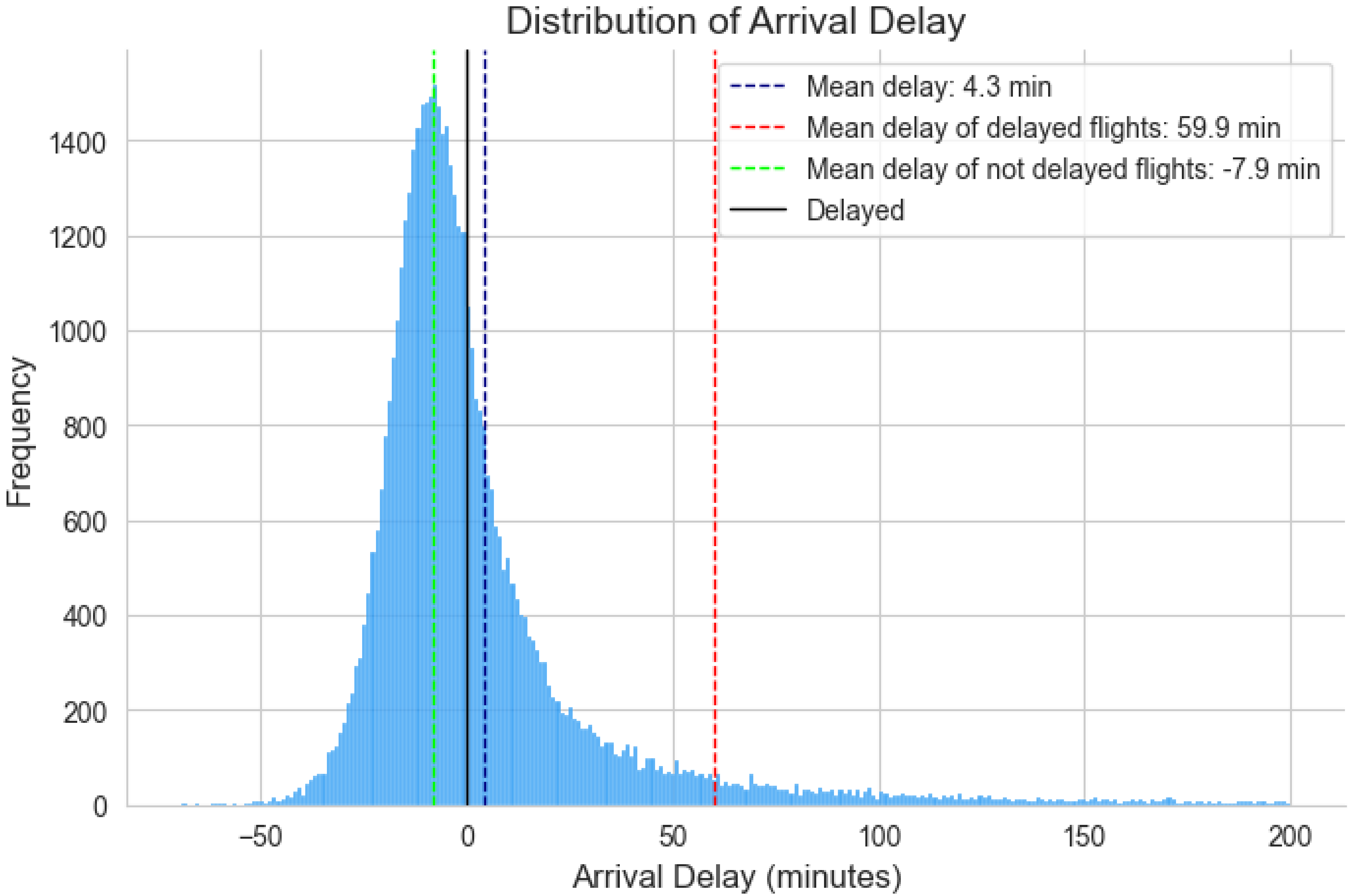
Elapsed Time Distribution for Delayed Flights

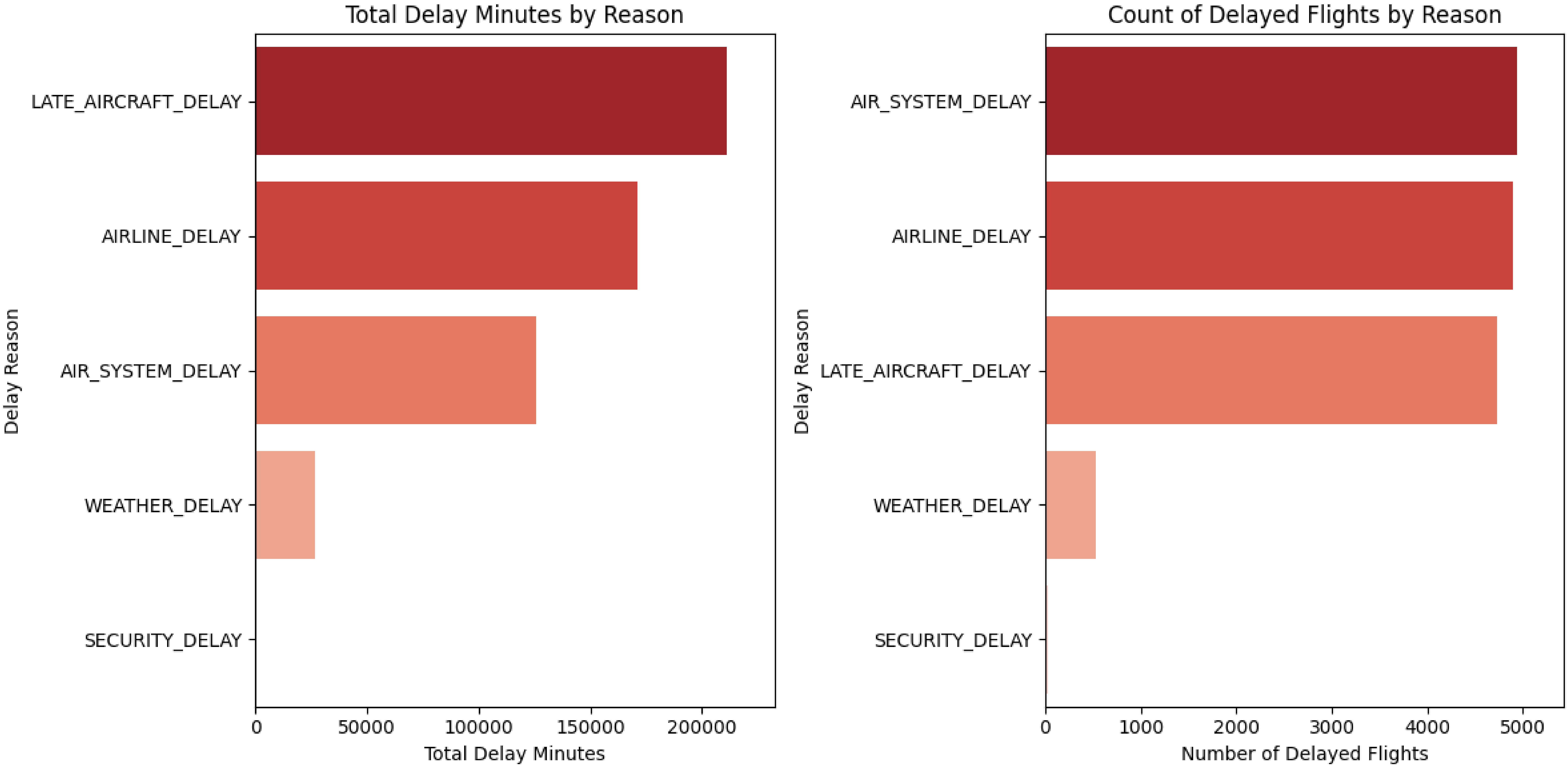

Day of Week Distribution for Not Delayed Flights
Not Delayed
Flight Count
0
1000
2000
3000
4000
5000
6000
Day of Week
1
2
3
4
5
6
7
Day of Week Distribution for Delayed Flights
Delayed
Flight Count
0
200
400
600
800
1000
1200
1400
Day of Week
1
2
3
4
5
6
7

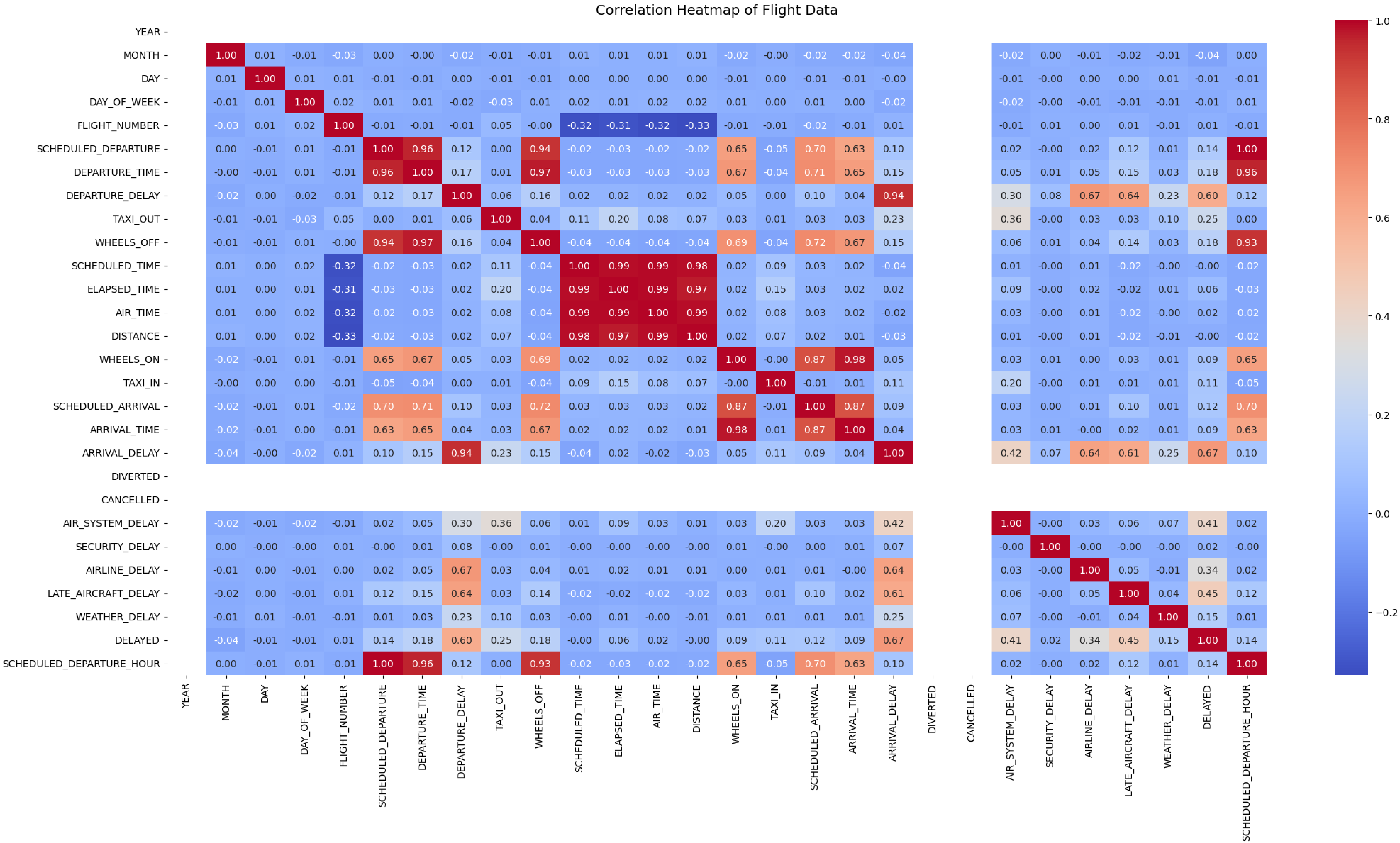

Month Distribution for Not Delayed Flights
Not Delayed
Flight Count
0
500
1000
1500
2000
2500
3000
3500
Month
2
4
6
8
10
12
Month Distribution for Delayed Flights
Delayed
Flight Count
0
200
400
600
800
1000
Month
2
4
6
8
10
12


Percentage of Delayed Flights by Departure Time Category
Departure Time Category
Morning
Late Night
Late Evening
Evening
Afternoon
0
10
20
30
40
Percentage of Flights with Delay


Proportion of Delayed Flights > 15 min
DELAYED
0
82.1%
17.9%
1
Not Delayed
Delayed


Distribution of Arrival Delay
Mean delay: 4.3 min
Mean delay of delayed flights: 59.9 min
Mean delay of not delayed flights: -7.9 min
Delayed
Frequency
0
200
400
600
800
1000
1200
1400
−50
0
50
100
150
200
Arrival Delay (minutes)


Total Delay Minutes by Reason
Delay Reason
LATE_AIRCRAFT_DELAY
AIRLINE_DELAY
AIR_SYSTEM_DELAY
WEATHER_DELAY
SECURITY_DELAY
0
50000
100000
150000
200000
Total Delay Minutes
Count of Delayed Flights by Reason
Delay Reason
AIR_SYSTEM_DELAY
AIRLINE_DELAY
LATE_AIRCRAFT_DELAY
WEATHER_DELAY
SECURITY_DELAY
0
1000
2000
3000
4000
5000
Number of Delayed Flights



Correlation Heatmap of Flight Data

	YEAR	MONTH	DAY	DAY_OF_WEEK	FLIGHT_NUMBER	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS_OFF	SCHEDULED_TIME	ELAPSED_TIME	AIR_TIME	DISTANCE	WHEELS_ON	TAXI_IN	SCHEDULED_ARRIVAL	ARRIVAL_TIME	ARRIVAL_DELAY	DIVERTED	CANCELLED	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY	DELAYED	SCHEDULED_DEPARTURE_HOUR
YEAR																												
MONTH		1.00	0.01	-0.01	-0.03	0.00	-0.00	-0.02	-0.01	-0.01	0.01	0.01	0.01	0.01	-0.02	-0.00	-0.02	-0.02	-0.04			-0.02	0.00	-0.01	-0.02	-0.01	-0.04	0.00
DAY		0.01	1.00	0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	-0.01	-0.00			-0.01	-0.00	0.00	0.00	0.01	-0.01	-0.01
DAY_OF_WEEK		-0.01	0.01	1.00	0.02	0.01	0.01	-0.02	-0.03	0.01	0.02	0.01	0.02	0.02	0.01	0.00	0.01	0.00	-0.02			-0.02	-0.00	-0.01	-0.01	-0.01	-0.01	0.01
FLIGHT_NUMBER		-0.03	0.01	0.02	1.00	-0.01	-0.01	-0.01	0.05	-0.00	-0.32	-0.31	-0.32	-0.33	-0.01	-0.01	-0.02	-0.01	0.01			-0.01	0.01	0.00	0.01	-0.01	0.01	-0.01
SCHEDULED_DEPARTURE		0.00	-0.01	0.01	-0.01	1.00	0.96	0.12	0.00	0.94	-0.02	-0.03	-0.02	-0.02	0.65	-0.05	0.70	0.63	0.10			0.02	-0.00	0.02	0.12	0.01	0.14	1.00
DEPARTURE_TIME		-0.00	-0.01	0.01	-0.01	0.96	1.00	0.17	0.01	0.97	-0.03	-0.03	-0.03	-0.03	0.67	-0.04	0.71	0.65	0.15			0.05	0.01	0.05	0.15	0.03	0.18	0.96
DEPARTURE_DELAY		-0.02	0.00	-0.02	-0.01	0.12	0.17	1.00	0.06	0.16	0.02	0.02	0.02	0.02	0.05	0.00	0.10	0.04	0.94			0.30	0.08	0.67	0.64	0.23	0.60	0.12
TAXI_OUT		-0.01	-0.01	-0.03	0.05	0.00	0.01	0.06	1.00	0.04	0.11	0.20	0.08	0.07	0.03	0.01	0.03	0.03	0.23			0.36	-0.00	0.03	0.03	0.10	0.25	0.00
WHEELS_OFF		-0.01	-0.01	0.01	-0.00	0.94	0.97	0.16	0.04	1.00	-0.04	-0.04	-0.04	-0.04	0.69	-0.04	0.72	0.67	0.15			0.06	0.01	0.04	0.14	0.03	0.18	0.93
SCHEDULED_TIME		0.01	0.00	0.02	-0.32	-0.02	-0.03	0.02	0.11	-0.04	1.00	0.99	0.99	0.98	0.02	0.09	0.03	0.02	-0.04			0.01	-0.00	0.01	-0.02	-0.00	-0.00	-0.02
ELAPSED_TIME		0.01	0.00	0.01	-0.31	-0.03	-0.03	0.02	0.20	-0.04	0.99	1.00	0.99	0.97	0.02	0.15	0.03	0.02	0.02			0.09	-0.00	0.02	-0.02	0.01	0.06	-0.03
AIR_TIME		0.01	0.00	0.02	-0.32	-0.02	-0.03	0.02	0.08	-0.04	0.99	0.99	1.00	0.99	0.02	0.08	0.03	0.02	-0.02			0.03	-0.00	0.01	-0.02	-0.00	0.02	-0.02
DISTANCE		0.01	0.00	0.02	-0.33	-0.02	-0.03	0.02	0.07	-0.04	0.98	0.97	0.99	1.00	0.02	0.07	0.02	0.01	-0.03			0.01	-0.00	0.01	-0.02	-0.01	-0.00	-0.02
WHEELS_ON		-0.02	-0.01	0.01	-0.01	0.65	0.67	0.05	0.03	0.69	0.02	0.02	0.02	0.02	1.00	-0.00	0.87	0.98	0.05			0.03	0.01	0.00	0.03	0.01	0.09	0.65
TAXI_IN		-0.00	0.00	0.00	-0.01	-0.05	-0.04	0.00	0.01	-0.04	0.09	0.15	0.08	0.07	-0.00	1.00	-0.01	0.01	0.11			0.20	-0.00	0.01	0.01	0.01	0.11	-0.05
SCHEDULED_ARRIVAL		-0.02	-0.01	0.01	-0.02	0.70	0.71	0.10	0.03	0.72	0.03	0.03	0.03	0.02	0.87	-0.01	1.00	0.87	0.09			0.03	0.00	0.01	0.10	0.01	0.12	0.70
ARRIVAL_TIME		-0.02	-0.01	0.00	-0.01	0.63	0.65	0.04	0.03	0.67	0.02	0.02	0.02	0.01	0.98	0.01	0.87	1.00	0.04			0.03	0.01	-0.00	0.02	0.01	0.09	0.63
ARRIVAL_DELAY		-0.04	-0.00	-0.02	0.01	0.10	0.15	0.94	0.23	0.15	-0.04	0.02	-0.02	-0.03	0.05	0.11	0.09	0.04	1.00			0.42	0.07	0.64	0.61	0.25	0.67	0.10
DIVERTED																												
CANCELLED																												
AIR_SYSTEM_DELAY		-0.02	-0.01	-0.02	-0.01	0.02	0.05	0.30	0.36	0.06	0.01	0.09	0.03	0.01	0.03	0.20	0.03	0.03	0.42			1.00	-0.00	0.03	0.06	0.07	0.41	0.02
SECURITY_DELAY		0.00	-0.00	-0.00	0.01	-0.00	0.01	0.08	-0.00	0.01	-0.00	-0.00	-0.00	-0.00	0.01	-0.00	0.00	0.01	0.07			-0.00	1.00	-0.00	-0.00	-0.00	0.02	-0.00
AIRLINE_DELAY		-0.01	0.00	-0.01	0.00	0.02	0.05	0.67	0.03	0.04	0.01	0.02	0.01	0.01	0.00	0.01	0.01	-0.00	0.64			0.03	-0.00	1.00	0.05	-0.01	0.34	0.02
LATE_AIRCRAFT_DELAY		-0.02	0.00	-0.01	0.01	0.12	0.15	0.64	0.03	0.14	-0.02	-0.02	-0.02	-0.02	0.03	0.01	0.10	0.02	0.61			0.06	-0.00	0.05	1.00	0.04	0.45	0.12
WEATHER_DELAY		-0.01	0.01	-0.01	-0.01	0.01	0.03	0.23	0.10	0.03	-0.00	0.01	-0.00	-0.01	0.01	0.01	0.01	0.01	0.25			0.07	-0.00	-0.01	0.04	1.00	0.15	0.01
DELAYED		-0.04	-0.01	-0.01	0.01	0.14	0.18	0.60	0.25	0.18	-0.00	0.06	0.02	-0.00	0.09	0.11	0.12	0.09	0.67			0.41	0.02	0.34	0.45	0.15	1.00	0.14
SCHEDULED_DEPARTURE_HOUR		0.00	-0.01	0.01	-0.01	1.00	0.96	0.12	0.00	0.93	-0.02	-0.03	-0.02	-0.02	0.65	-0.05	0.70	0.63	0.10			0.02	-0.00	0.02	0.12	0.01	0.14	1.00

Delay Prediction Model

Variables

X = ['MONTH', 'DAY',
'SCHEDULED_DEPARTURE',
'SCHEDULED_ARRIVAL',
'DIVERTED', 'CANCELLED',
'AIR_SYSTEM_DELAY',
'SECURITY_DELAY',
'AIRLINE_DELAY',
'LATE_AIRCRAFT_DELAY',
'WEATHER_DELAY']

y = DELAYED

1 = DELAYED (arrival delay > 15 minutes)
0 = Not DELAYED

Model Creation

Data Split

test_size = .20

Data Scaling

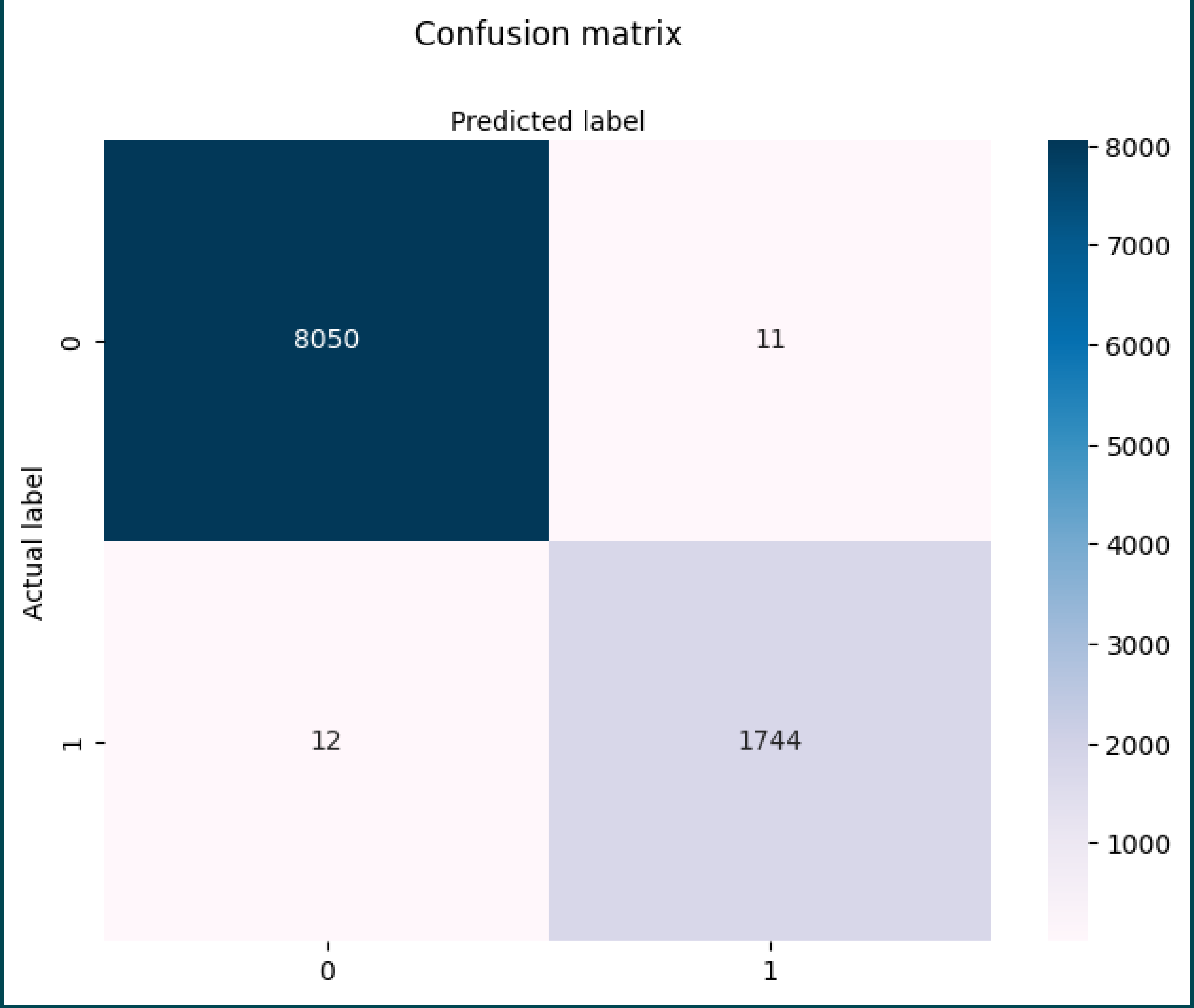
Standard Scaler

Model Fit

Decision Tree Classifier

Testing

CLF prediction on X_test


Confusion matrix
Predicted label
Actual label
0
1
8050
11
12
1744
0
1
8000
7000
6000
5000
4000
3000
2000
1000

Model Results

AUC Score
99.59%

Accuracy
99.8%

Precision
99.4%

Recall
99.4%

Conlusions

Conclusions

- Two primary causes of delays: late aircrafts and airline-related delays.
- Late aircrafts are the most common cause of delays in the US.
- Airline-related delays, which include factors such as crew scheduling, maintenance, and other operational issues, are the second most common cause of delays.
- Other factors that contribute to flight delays include: weather, air traffic control, security, and other issues.
- The data also shows that some airlines and routes are more prone to delays than others, and that delays tend to be more common during certain times of day and certain months of the year.

Recommendations

Recommendations

Based on our findings, we recommend that airlines focus on:

- Strategies to improve aircraft turnaround times
- Optimize crew and maintenance schedules
- Address other operational issues that contribute to delays.

Next Steps

Next steps

- Preemptive rebooking
- Resource allocation
- Operational optimization
- Customer communication
- Continuous improvement

One Pager

Navigating the Skies: Analyzing Flight Delay Trends in the US

Flight delays are a significant problem in the US aviation industry, causing inconvenience and frustration for passengers and incurring costs for airlines. Despite efforts to improve on-time performance, delays continue to be a persistent issue, with many flights experiencing delays due to a variety of factors.

Buisiness Challenges

Preemptive rebooking:

- Airlines can use a delay prediction model to anticipate delays before they occur and proactively rebook passengers on alternative flights to minimize the impact of delays.

Resource allocation:

- By predicting delays in advance, airlines can adjust crew schedules, gate assignments, and other resources to minimize the impact of delays on passengers and maintain a smooth operation.

Operational optimization:

- Delay prediction models can be used to identify patterns and root causes of delays, which can help airlines optimize their operations and improve on-time performance.

Continuous improvement:

- By analyzing delay data and continuously improving the delay prediction model, airlines can achieve better accuracy and reduce the number of delays over time.

Resources & Governance

Resources needed:

- A delay prediction model.
- IT systems.
- Trained personnel.
- Data analytics tools.

Possible costs:

- Development and maintenance of the delay prediction model.
- IT infrastructure costs
- Personnel costs.

Impact and Key KPIs

On-time performance (OTP) - measures the percentage of flights that depart and arrive on time.

- Benefit: OTP is a critical measure of an airline's overall performance and is closely tied to customer satisfaction. High OTP can lead to repeat business, positive reviews, and enhanced reputation.

Average delay time - measures the average amount of time that flights are delayed.

- Benefit: By monitoring average delay time, airlines can identify areas for improvement in their operations and focus on reducing delays to enhance the customer experience

Resource utilization - measures the percentage of available resources (such as gates, crew, and aircraft) that are utilized effectively.

- Benefit: By optimizing resource utilization, airlines can reduce costs associated with idle resources and improve operational efficiency.

Barriers / Constraints

Data quality and availability:

- Airlines may struggle to obtain timely and reliable data from various sources, which can hinder the accuracy of delay prediction models.

Technology infrastructure:

- Implementing delay prediction models and other IT systems can be costly and resource-intensive.

Operational complexity:

- Coordinating resources and communications across multiple stakeholders can be difficult, particularly during unexpected events.

Employee resistance:

- Changes to operational processes can meet with resistance from employees, requiring proper buy-in, training, and support.