By Amun Intelligence Team

2015 Flight Delays and Cancellation

Team Members

The outstanding results of this project were only possible thanks to the collective effort of our talented team.

01. Taha Ahmed Taha
02. Ahmed Hussein Hassan
03. George Gamil Sedky
04. Mostafa Saeed Mostafa
05. Mariam Mohamed

Agenda

01. Introduction
02. Objectives
03. Data Cleaning
04. Data Analysis
05. Conclusions

01

Introduction

Flight disruptions, including cancellations and diversions, significantly affect airline operations and passenger experiences. This analysis examines these disruptions by exploring cancellation reasons and patterns across airlines and airports. Using data cleaned with Python, the goal is to provide insights that can help improve operational efficiency and reduce the impact of disruptions on travelers.

02

Objective

The goal of this data analysis is to provide a comprehensive understanding of flight disruptions by exploring cancellation reasons and diverted flights across different airlines and airports. The data was cleaned using Python to ensure accurate and reliable results, offering valuable insights into the patterns of flight disruptions. The analysis focuses on identifying the key factors affecting airline and airport performance, providing actionable insights that can help decision-makers improve operational efficiency and reduce the impact of disruptions on passengers

03

Data Cleaning

data cleaning Introduction

The dataset is composed of three CSV files: airlines.csv, airports.csv, and flights.csv.

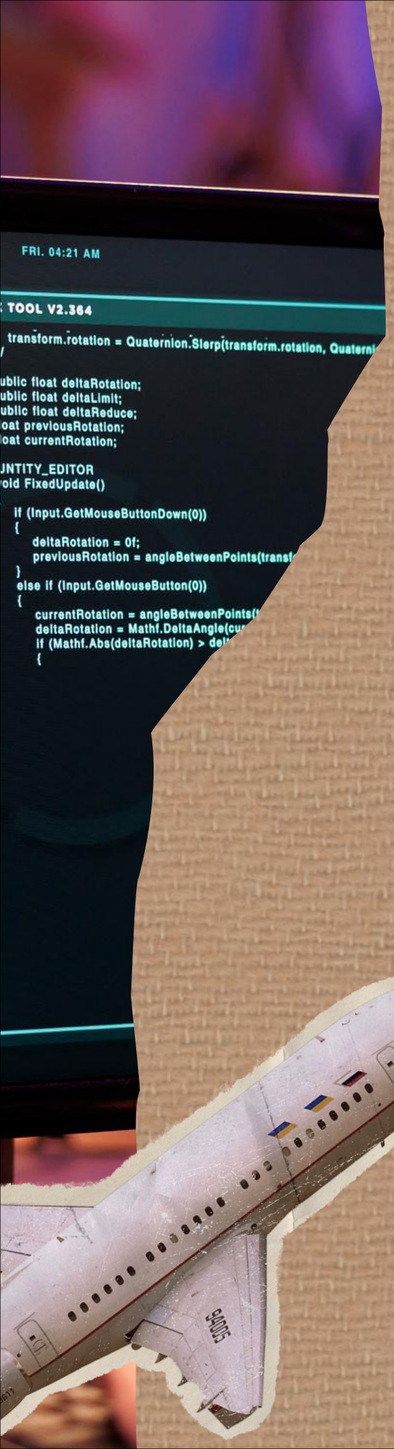
- The primary file, flights.csv, contains key columns, including: Date
- Information: YEAR, MONTH, DAY, DAY_OF_WEEK.
- Flight Identifiers: AIRLINE, FLIGHT_NUMBER, TAIL_NUMBER.
- Flight Locations: ORIGIN_AIRPORT, DESTINATION_AIRPORT.
- Timing Details: SCHEDULED_DEPARTURE, DEPARTURE_TIME, SCHEDULED_ARRIVAL, ARRIVAL_TIME.
- Delays & Cancellations: DEPARTURE_DELAY, ARRIVAL_DELAY, TOTAL_DELAY, CANCELLED, CANCELLATION_REASON.
- Flight Operations: TAXI_OUT, WHEELS_OFF, ELAPSED_TIME, AIR_TIME, DISTANCE, WHEELS_ON, TAXI_IN.

Initial Data Observations

- Data Volume: Around 8 M of flight records spanning a specified time period.
- Common Issues Identified:

1. Missing Values: The CANCELLATION_REASON column has approximately 98.46% missing values, while columns such as AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY, LATE_AIRCRAFT_DELAY, and WEATHER_DELAY contain around 81.72% missing values.
2. Outliers: Extreme or inconsistent values in DEPARTURE_DELAY and ARRIVAL_DELAY.

- Data Type Mismatches: Certain columns (e.g., time data) may have incorrect formats, such as being stored as text.



Data Cleaning Process

1. Handling Missing Values:

- Replacing Null Values:
- Since the percentage of missing values in these columns is small and won't significantly affect our analysis, the following columns were filled with appropriate replacements:
- TAIL_NUMBER: 0.25% missing .dropna()
- DEPARTURE_TIME: 1.5% missing
- DEPARTURE_DELAY: 1.5% missing
- TAXI_OUT: 1.5% missing
- WHEELS_OFF: 1.5% missing
- SCHEDULED_TIME: 0.0001% missing
- ELAPSED_TIME: 1.8% missing
- AIR_TIME: 1.8% missing
- WHEELS_ON: 1.6% missing
- TAXI_IN: 1.6% missing
- ARRIVAL_TIME: 1.6% missing
- ARRIVAL_DELAY: 1.8% missing

```
columnsToFill = ['DEPARTURE_DELAY', 'DEPARTURE_TIME',
                 'DEPARTURE_DELAY', 'TAXI_OUT', 'WHEELS_OFF',
                 'SCHEDULED_TIME', 'ELAPSED_TIME', 'AIR_TIME',
                 'WHEELS_ON', 'TAXI_IN', 'ARRIVAL_TIME',
                 'ARRIVAL_DELAY']
nulls = df.isnull().sum() / df.shape[0]*100
for i in columnsToFill:
    if nulls[i]<81 :
        df[i] = df[i].fillna(0)
```

Data Cleaning Process

- Handling Missing Values:

1. Replacing Values with:

- A: Mechanical issues with the aircraft.
- B: Crew-related problems (e.g., staffing shortages, pilot illness).
- C: Significant weather events (e.g., storms, blizzards, low visibility).
- D: Air traffic control issues, security concerns, or national aviation system delays.

```
df["CANCELLATION_REASON"] =
df["CANCELLATION_REASON"].replace
("A" , "Mechanical issues")
df["CANCELLATION_REASON"] =
df["CANCELLATION_REASON"].replace
("B" , "Crews problems")
df["CANCELLATION_REASON"] =
df["CANCELLATION_REASON"].replace
("C" , "Significant weasther")
df["CANCELLATION_REASON"] =
df["CANCELLATION_REASON"].replace
("D" , "Air traffic Control")
```

1. Replacing Null Values in CANCELLATION_REASON:

- Filled missing values in the CANCELLATION_REASON column with "No Cancellation."

```
df["CANCELLATION_REASON"] =
df["CANCELLATION_REASON"].
fillna( "No Cancellation" )
```

Data Cleaning Process

- Time Format Conversion:
- In this step, we convert time columns (e.g., SCHEDULED_DEPARTURE, DEPARTURE_TIME, ARRIVAL_TIME, SCHEDULED_ARRIVAL) from integer format to a readable "HH
- format. This is done by breaking down the time into hours and minutes using a custom function, format_time(), ensuring that the data is standardized and easier to interpret for further analysis.

```
df['SCHEDULED_DEPARTURE'] =
df['SCHEDULED_DEPARTURE'].astype(int)
df['DEPARTURE_TIME'] =
df['DEPARTURE_TIME'].astype(int)
df['ARRIVAL_TIME'] =
df['ARRIVAL_TIME'].astype(int)
df['SCHEDULED_ARRIVAL'] =
df['SCHEDULED_ARRIVAL'].astype(int)

def format_time(departure_time):
    hours = departure_time // 100
    minutes = departure_time % 100
    return f"{hours:02}:{minutes:02}"
df['SCHEDULED_DEPARTURE'] =
df['SCHEDULED_DEPARTURE'].apply(format_time)
df['DEPARTURE_TIME'] =
df['DEPARTURE_TIME'].apply(format_time)
df['ARRIVAL_TIME'] =
df['ARRIVAL_TIME'].apply(format_time)
df['SCHEDULED_ARRIVAL'] =
df['SCHEDULED_ARRIVAL'].apply(format_time)
```

Data Cleaning Process

- Function Description: fixingTime
- This function takes a string representing time in the format "HH:MM" and corrects any instances where the time is listed as "24:00". Since "24:00" is not a valid time format (it should be "00:00" for midnight), the function converts it to "00:00". For all other time values, the function returns the input string unchanged.
- Input: A string representing time (e.g., "24:00", "14:30")
- Output: The corrected time string (e.g., "00:00" if input is "24:00", otherwise returns the original string)

```python
def fixingTime(stringTime):
  if stringTime == '24:00':
    return '00:00'
  else:
    return stringTime
```

Data Cleaning Process

- Handling Time Formatting Issues:
- To ensure consistent and valid time values, we applied the fixingTime() function to correct any instances where time is incorrectly listed as "24:00", converting it to "00:00" for proper midnight representation. This ensures that the time columns (SCHEDULED_DEPARTURE, DEPARTURE_TIME, ARRIVAL_TIME, SCHEDULED_ARRIVAL) have accurate and standardized values.

```
df['SCHEDULED_DEPARTURE'] =
df['SCHEDULED_DEPARTURE'].apply(fixingTime)

df['DEPARTURE_TIME'] =
df['DEPARTURE_TIME'].apply(fixingTime)

df['ARRIVAL_TIME'] =
df['ARRIVAL_TIME'].apply(fixingTime)

df['SCHEDULED_ARRIVAL'] =
df['SCHEDULED_ARRIVAL'].apply(fixingTime)
```

Data Cleaning Process

- Time Grouping by Period of Day:
- We created a Time_Grouping() function to categorize flight times into three periods: Morning, Afternoon, and Evening. This function converts the time values into hour-based groups for both ARRIVAL_TIME and DEPARTURE_TIME, allowing for easier analysis of flight trends based on the time of day.

```
def Time_Grouping(timestring):
    hour = int(timestring.split(':')[0])
    if (1 <= hour <= 11) or (hour == 0):
        return 'Morning'
    elif 12 <= hour < 18:
        return 'Afternoon'
    else:
        return 'Evening'

df['ArrivalTime_Group'] =
df['ARRIVAL_TIME'].apply(Time_Grouping)
df['DepartureTime_Group'] =
df['DEPARTURE_TIME'].apply(Time_Grouping)
```

Data Cleaning Process

- Converting Float Columns to Integer:

To optimize the dataset and ensure consistency, we identified all columns with floating-point values and converted them to integer type where applicable. This process helps in reducing memory usage and ensures the precision required for further analysis.

```
FloatColumns = df.select_dtypes(include=['float64'])
for j in FloatColumns:
        df[j] = df[j].astype(int)
```

Data After Cleaning Process

	0	1	2	3	4
YEAR	2015	2015	2015	2015	2015
MONTH	1	1	1	1	1
DAY	1	1	1	1	1
DAY_OF_WEEK	4	4	4	4	4
AIRLINE	AS	AA	US	AA	AS
FLIGHT_NUMBER	98	2336	840	258	135
TAIL_NUMBER	N407AS	N3KUAA	N171US	N3HYAA	N527AS
ORIGIN_AIRPORT	ANC	LAX	SFO	LAX	SEA
DESTINATION_AIRPORT	SEA	PBI	CLT	MIA	ANC
SCHEDULED_DEPARTURE	00:05	00:10	00:20	00:20	00:25
DEPARTURE_TIME	23:54	00:02	00:18	00:15	00:24
DEPARTURE_DELAY	-11	-8	-2	-5	-1
TAXI_OUT	21	12	16	15	11
WHEELS_OFF	15	14	34	30	35
SCHEDULED_TIME	205	280	286	285	235
ELAPSED_TIME	194	279	293	281	215
AIR_TIME	169	263	266	258	199
DISTANCE	1448	2330	2296	2342	1448
WHEELS_ON	404	737	800	748	254
TAXI_IN	4	4	11	8	5
SCHEDULED_ARRIVAL	04:30	07:50	08:06	08:05	03:20
ARRIVAL_TIME	04:08	07:41	08:11	07:56	02:59
ARRIVAL_DELAY	-22	-9	5	-9	-21
DIVERTED	Not Diverted	Not Diverted	Not Diverted	Not Diverted	Not Diverted
CANCELLED	Not Cancelled	Not Cancelled	Not Cancelled	Not Cancelled	Not Cancelled
CANCELLATION_REASON	No Cancellation	No Cancellation	No Cancellation	No Cancellation	No Cancellation
AIR_SYSTEM_DELAY	0	0	0	0	0
SECURITY_DELAY	0	0	0	0	0
AIRLINE_DELAY	0	0	0	0	0
LATE_AIRCRAFT_DELAY	0	0	0	0	0
WEATHER_DELAY	0	0	0	0	0
ArrivalTime_Group	Morning	Morning	Morning	Morning	Morning
DepartureTime_Group	Evening	Morning	Morning	Morning	Morning

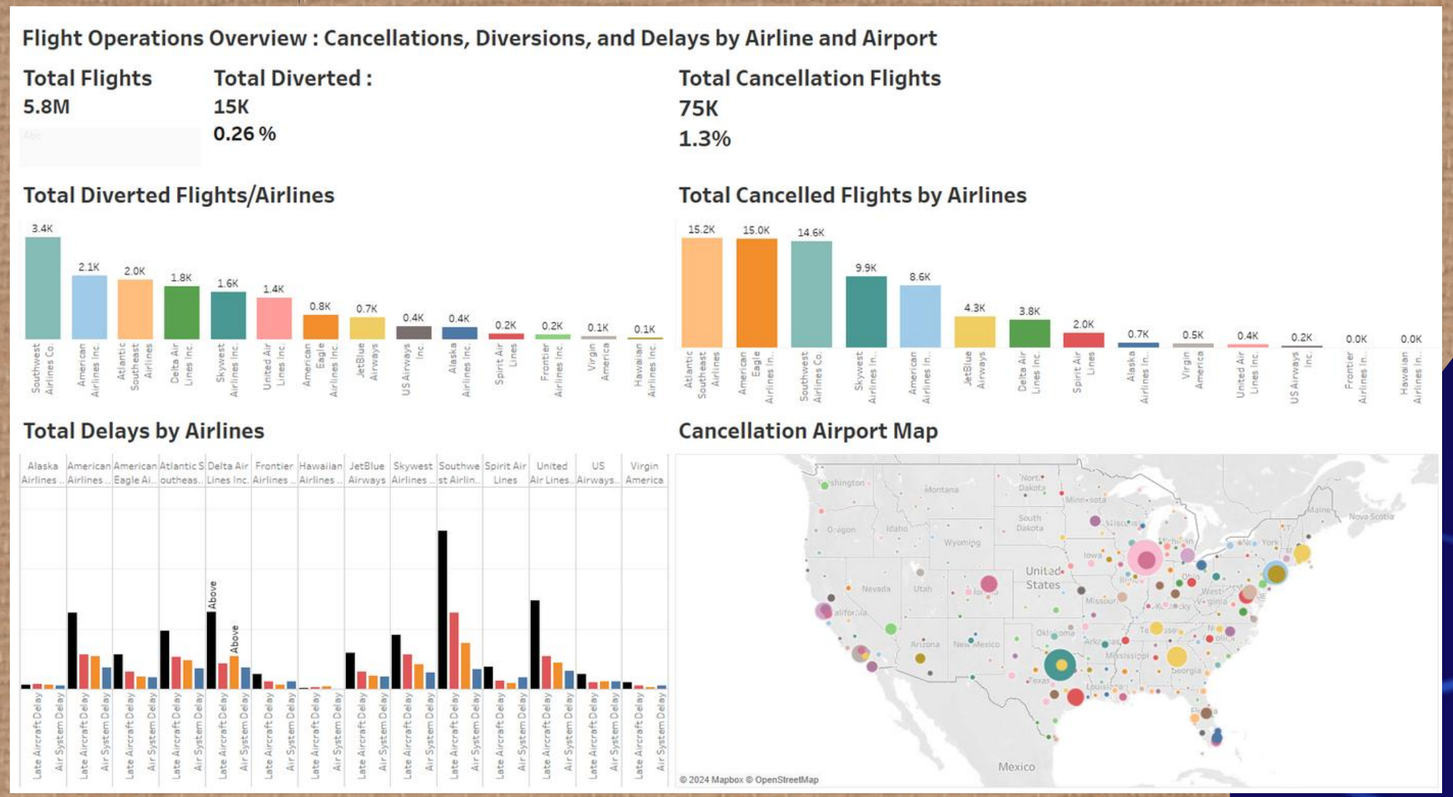
04 data analysis and visualization

data analysis and visualization

• Insight

1. cancelattion and Diversion rate below the acceptable range for civil Aviation Authorities and airlines as below 2-3%

2. Total delays are above the acceptable range 15-20%

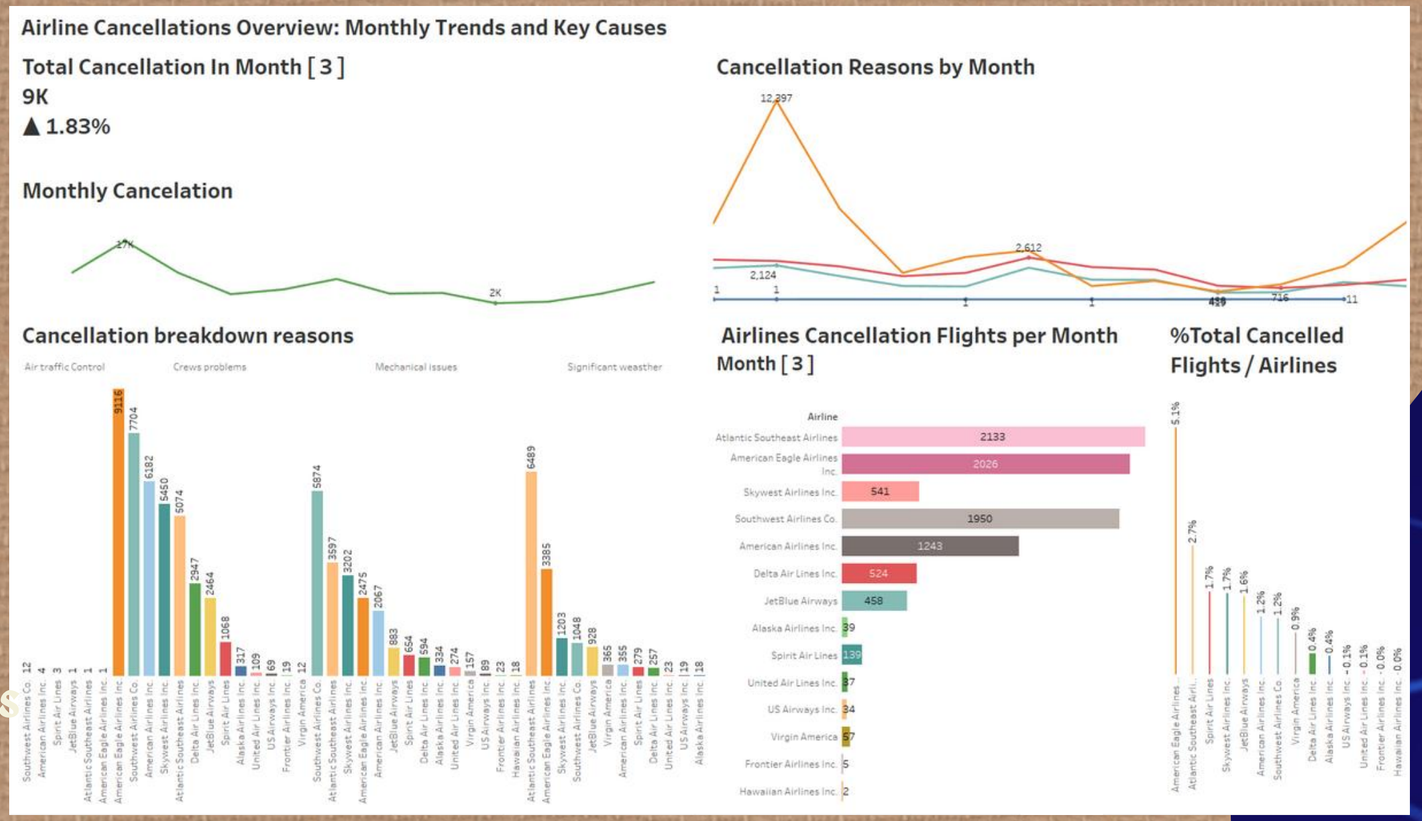


Dashboard 1

data analysis and visualization

Insight

1. American Airlines, Southwest Airlines, and Delta Airlines experience the highest number of cancellations, with crew and mechanical issues being the most common cause.
2. Weather, while a factor, is not the leading cause of cancellations; most issues are related to operational problems (crew and mechanical issues).
3. Spirit Airlines has a notably high percentage of cancellations relative to its flight volume

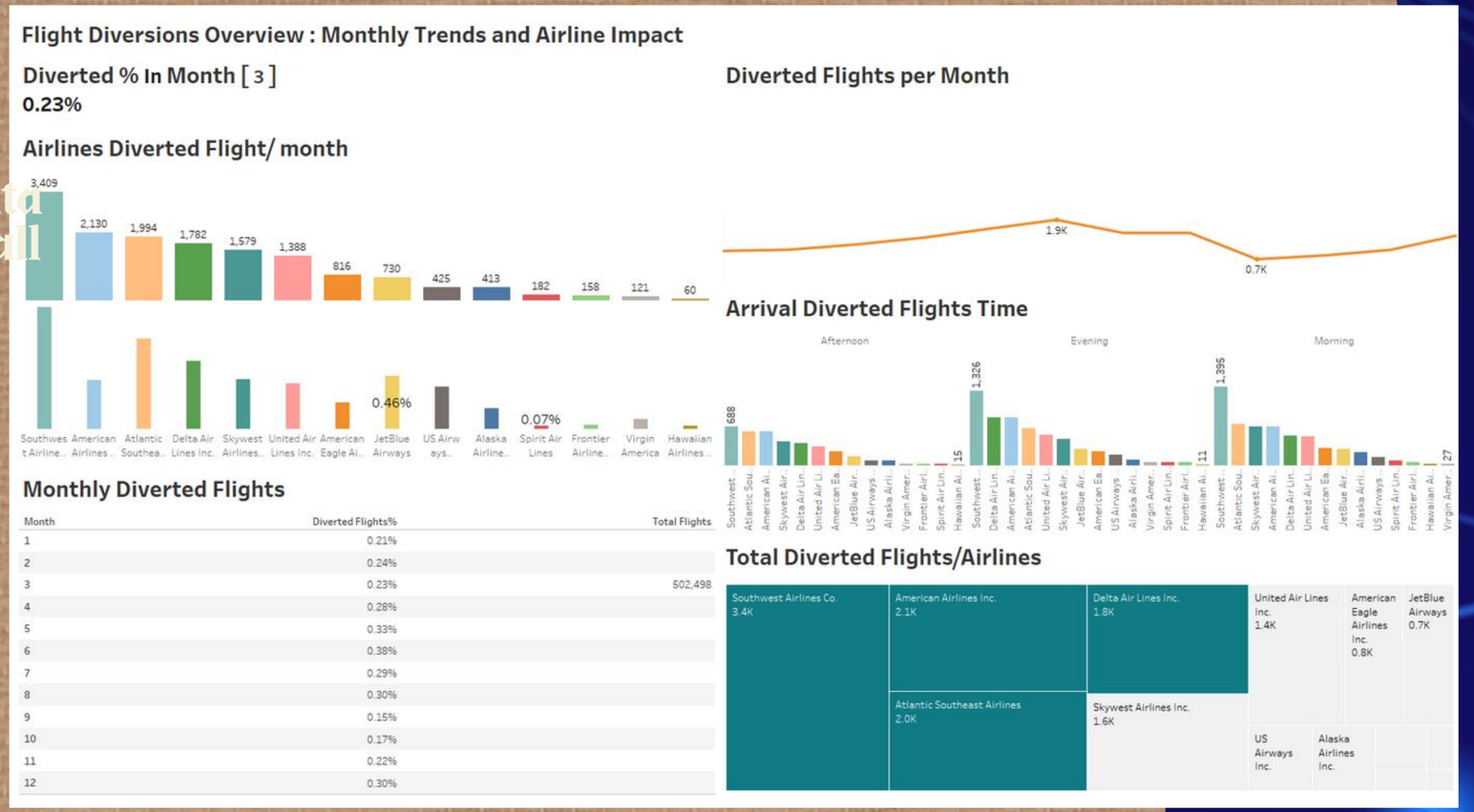


Dashboard 2

data analysis and visualization

- **Insight**

1. June has highest diverted while Sep has the lowest
2. Southwest , Atlantic and Delta has the highest percetange all over thier flights
3. Morning and Evenening has the most Diverted flights

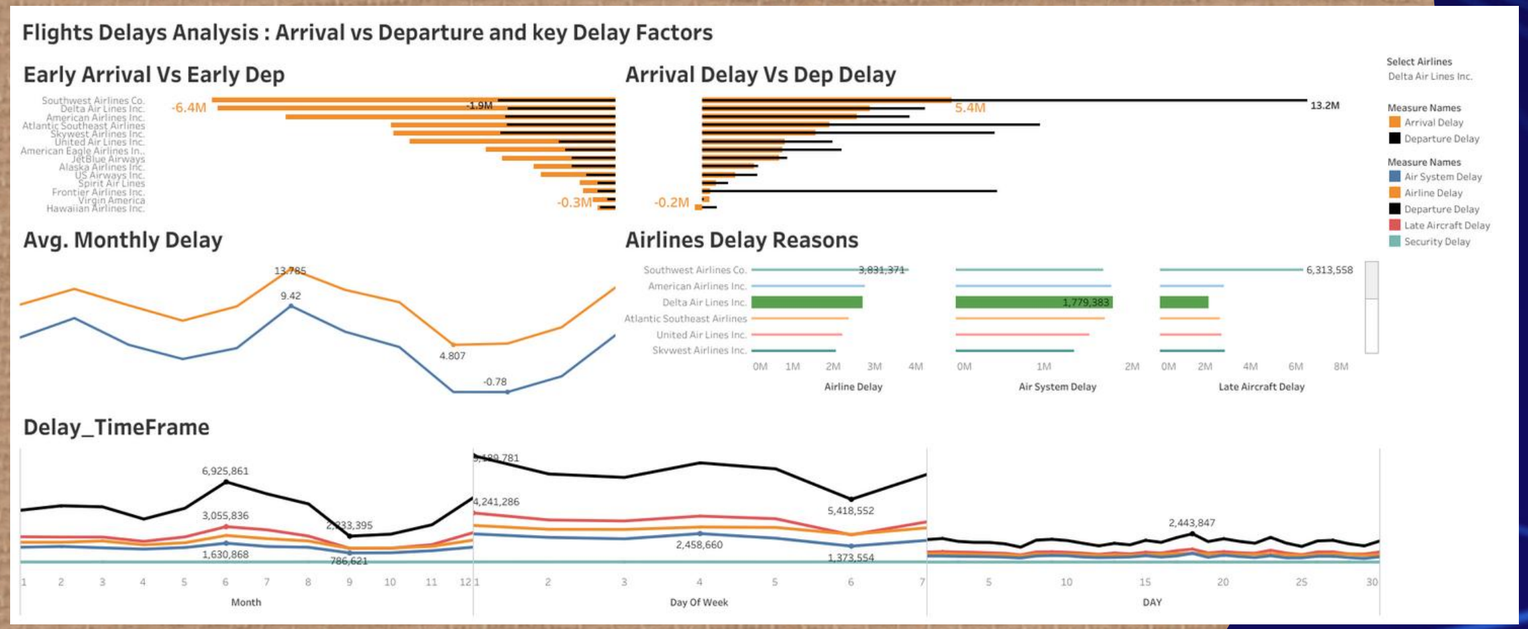


Dashboard 3

data analysis and visualization

• Insight

1. Most flights depart on time or a little early,as same as arriving more early or on time but there are a few flights with significant delays that are impacting the average delay time, same as arriving delay time
2. Delta Airline has high depurture delay but manage to arrive early
3. average monthly delay was more than 13mins as departure delay and 9 min as arrival delay in June
4. AirSystem, Airline , late aircraft , security delay was highest in june
5. Delay was above the average

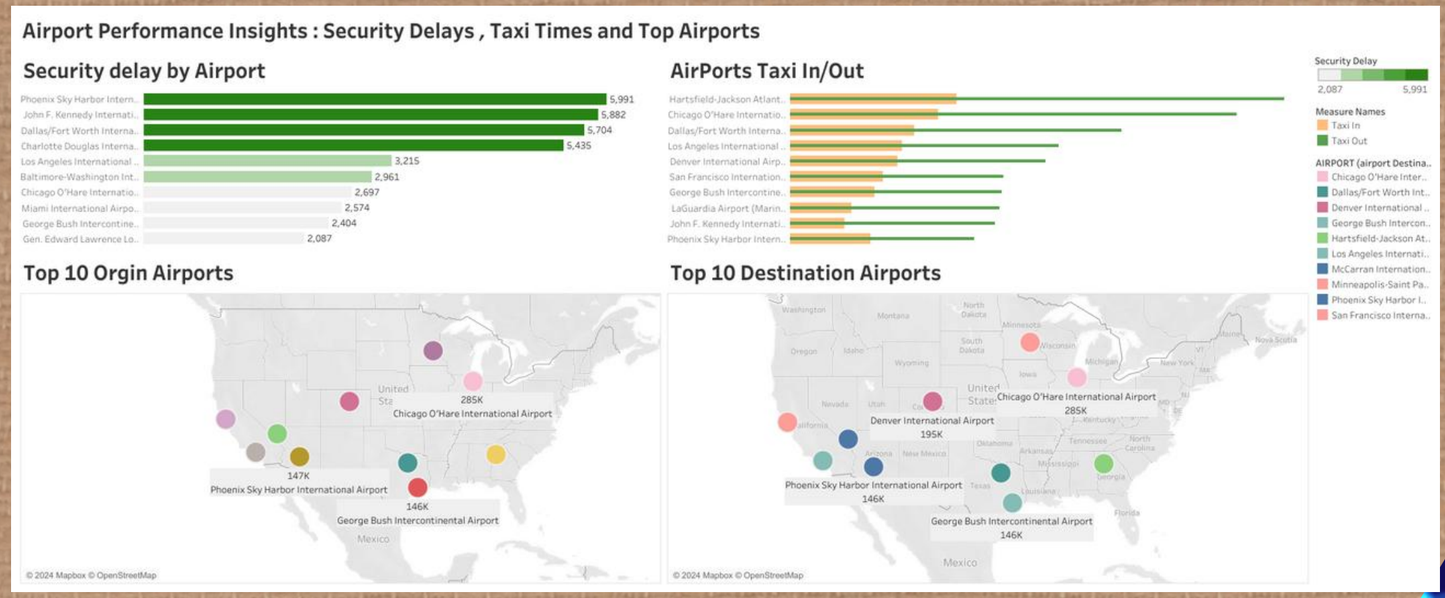


Dashboard 4

data analysis and visualization

- **Insight**

1. Security Delay and Taxi in/out was below the acceptable mins as for taxi in/out 20mins and security delay as 30 mins



Dashboard 5

05

Conclusions

1. Impact: Flight disruptions significantly affect airline operations and passenger experiences.
2. Data Cleaning: Addressed missing values, outliers, and data type mismatches.
3. Rates: Cancellation and diversion rates were below the acceptable range (2-3%).
4. Airlines: American, Southwest, and Delta had the most cancellations due to crew and mechanical issues. Spirit Airlines had a high cancellation rate.
5. Trends: June had the most diverted flights; September had the least. Morning and evening flights were most often diverted.
6. Delays: Most flights were on time, but significant delays impacted averages. Delta had high departure delays but arrived early. June had the highest delays.
7. Efficiency: Security delays and taxi times were within acceptable limits.

Amun Intelligence Team

Thank you