NEW YORK FLIGHT OPERATION ANALYSIS

OVERVIEW OF SCENARIO

- **Context:** New York's airports serve as critical hubs for air traffic within the United States. With millions of passengers passing through each year, efficient airport operations are essential for the state's economy, tourism industry, and business activities (Reference: New York State Department of Transportation, "New York State Airport System Plan" - 2021).

- **Challenges:** Managing the large volume of air traffic poses significant challenges, compounded by the region's unpredictable weather patterns. Delays and cancellations are common occurrences, impacting both airlines and passengers alike.

- **Analysis Goals:** Our analysis aims to gain insights into various aspects of airport operations and air travel dynamics over an 11-year period, spanning from 2013 to 2023. Key objectives include understanding patterns of delays and cancellations, evaluating the performance of airports and airlines, identifying peak travel times, and assessing the impact of weather conditions on flight operations.

- **Target Audience:** The primary audience for this analysis includes senior management teams within airports and airlines operating in the New York region. By providing actionable insights derived from comprehensive data analysis, we aim to empower decision-makers to enhance operational efficiency, improve customer satisfaction, and mitigate the impact of adverse weather conditions on air travel.

BUSINESS QUESTIONS

1. Provide a month-by-month breakdown of departures by destination, Airline and Airport departing New York from 2013 to 2023.

Justification: A month-by-month breakdown of departures by destination, airline, and airport from New York over the past decade will help identify trends and seasonal fluctuations in travel, aiding in strategic planning and resource allocation. This data can enhance operational efficiency and marketing strategies by pinpointing high-demand periods and popular destinations. It can also be used to tailor promotions and capacity adjustments.

2. Provide a month-by-month breakdown of departure delays by Airline, Airport, and Air temperature leaving New York from 2013 to 2023.

Justification: Analysing departure delays by airlines and airports alongside air temperature can improve punctuality and operational efficiency. Higher temperatures reduce aircraft lift, necessitating weight restrictions on hotter days. Lower temperatures bring challenges such as ice and snow, which can cause significant delays and hazards. Understanding these effects helps prepare for adverse conditions, ensure safety, and minimise disruptions. (Coffel & Horton, 2015)

3. What is the total number of flights arriving at New York Airports by morning, daytime, evening and night by year by Airline, by airport from 2013 to 2023?

Justification: Examining the arrivals of flights at New York airports, categorised by time of day, airline, and airport between 2013 and 2023, can effectively enhance staff allocation, security measures, and terminal services. This comprehensive analysis offers valuable insights into air traffic patterns, enabling informed decision-making to meet the ever-changing air travel demands.

4. What is the number of flights cancelled by airline, airport, destination and cancellation reasons from New York airports by year from 2013 to 2023?

Justification: By understanding the specific reasons behind flight cancellations, such as technical faults or adverse weather conditions, airlines and airports can implement targeted strategies for crisis management and advanced planning. This proactive approach enables stakeholders to reduce the occurrence of cancellations, minimize disruptions to travel plans, and ultimately elevate the overall experience for passengers.

5. How many flights were diverted from New York airports LGA and JFK by humidity, wind speed and visibility from 2013 to 2023?

Justification: Collecting data on the number of flights diverted from the two busiest New York airports, LGA and JFK, humidity, wind speed, and visibility from 2013 to 2023 enables stakeholders to grasp how specific atmospheric conditions influence flight paths and airport operations, critical for safety and efficiency. High humidity impacts aircraft performance and engine efficiency, strong winds pose challenges during take-off and landing, and poor visibility affects pilots' ability to navigate safely. Analysing these factors aids in enhancing operational planning, safety protocols, and overall airport efficiency. (Met Office, n.d.)

PRIMARY & SECONDARY DATA SET

Primary Data Sources : Airline Reporting Carrier On-Time Performance Dataset and airline names from Bureau of Transportation Statistics (Transtats, n.d.) and list of airports and their (Wikipedia, 2023).

from Wikipedia.

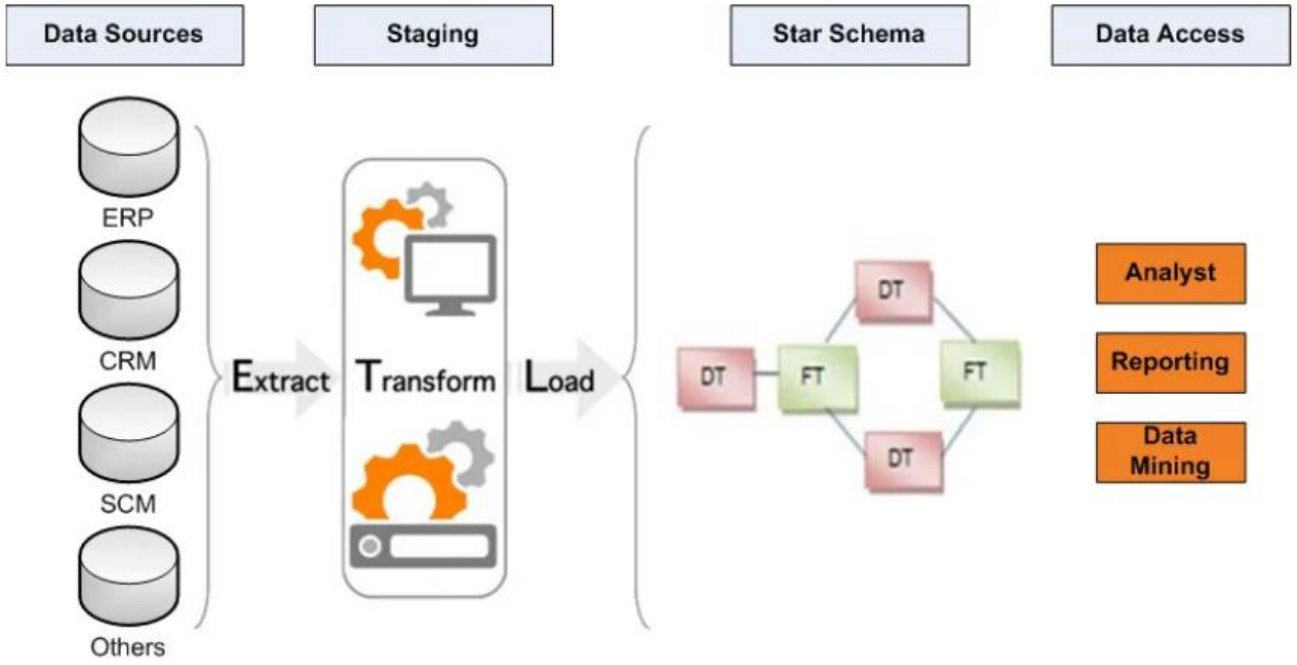
Justification : The Airline Reporting Carrier On-Time Performance Dataset from the Bureau of Transportation Statistics (BTS) was chosen due to its comprehensive coverage and reliability in providing detailed airline performance data. Complemented by additional datasets containing airline names from BTS and airport codes from Wikipedia, this selection ensures the data's accuracy and completeness. These datasets were chosen as they are essential for effectively analyzing airline operations and performance trends, directly addressing the business questions at hand.

Key Features :

Variables	Description
FlightDate	The date when the flight occurred.
Reporting_Airline	The airline that reported the flight information.
Origin	The code representing the departure airport.
OriginCityName	The name of the city where the departure airport is located.
OriginState	The state where the departure airport is situated.
Dest	The code representing the destination airport.
DestCityName	The name of the city where the destination airport is located.
DestState	The state where the destination airport is situated.
DepTime	The time (in local time) when the flight departed.
DepDel15	Indicates whether the flight departure was delayed by 15 minutes or more.
ArrTime	The time (in local time) when the flight arrived.
Cancelled	Indicates whether the flight was cancelled.
Diverted	Indicates whether the flight was diverted.
Airport_Code	The code representing the airport.
Airport_Name	The name of the airport.
CARRIERNAME	The name of the carrier (airline).

DATA WAREHOUSE DEVELOPMENT METHODOLOGY

- The data warehouse development methodology chosen for this project is the Kimball Lifecycle Methodology (KLDM), which emphasizes a bottom-up approach to data integration and dimensional modeling.



- KLDM was chosen for its compatibility with the project's goals of integrating flight and weather data, as well as its focus on iterative development and user involvement throughout the project lifecycle.

Justification:

- Enabling the construction of data marts before their integration into a unified data warehouse, the bottom-up approach offers a streamlined methodology for data consolidation.

- This methodology boasts reduced development timelines, ensuring operational efficiency, while its support for a diverse team ensures inclusivity across business sectors.

- This approach provides enhanced flexibility, enabling adjustments at lower levels without requiring extensive revisions to higher-level plans

- Facilitates targeted data integration, aligning seamlessly with specific business needs and objectives (Kimball et al., 2008).

BDDS AND DI TECHNIQUES UTILISED

- **SparkR within R Studio**

Chosen for its seamless integration with R Studio and its capability for distributed data processing using Apache Spark. It offers parallel processing, enabling efficient analysis of large-scale datasets compared to traditional R environments.

- **Parquet File Format**

Optimises storage and improves query performance by storing data in a columnar format. This allows all transformations to occur within a single file in R Studio, facilitating quick data movement to virtual machines and efficient querying in Hive environments. Parquet files are up to 34x faster than CSV files in query operations and 87% less storage thus reducing costs significantly. (Chopra, 2023)

- **Hadoop Deployment on Azure**

Utilized for its robust distributed storage (HDFS) and data processing capabilities (MapReduce) within the Microsoft Azure cloud environment. Azure's scalability and data management features make it a suitable choice for handling big data workloads.

- **Ambari for Cluster Management**

Chosen for its efficient management of Hadoop clusters, facilitating file uploads and ensuring system reliability. Ambari simplifies Hadoop cluster administration, making it easier to manage and monitor cluster resources.

- **Hive for Data Management**

Selected for its SQL-like interface and ability to provide structured access to data stored in Hadoop. Hive enables the creation of data marts and supports complex querying, enhancing the efficiency of data operations compared to other Hadoop-based storage solutions.

- **Cloudera Hadoop-Tableau Connector**

Chosen for its seamless integration between Hive data marts and Tableau, enabling sophisticated data visualization and dashboard creation. The connector ensures smooth data transfer between Hive and Tableau, simplifying the process of deriving actionable insights from the data.

BDDS AND DI TECHNIQUES UTILISED

- **Data Transformation**

Converts data from CSV to Parquet format, leveraging Parquet's superior storage efficiency and query speed. This transformation is vital for optimizing data filtering based on project-specific requirements.

- **Data Consolidation**

Combining multiple datasets within SparkR enables comprehensive analyses by merging data sources into a unified format, facilitating easier data processing and analysis

- **Data Movement**

Efficiently transferring data across various storage solutions, from local storage to distributed file systems in HDFS via Ambari in the Azure environment, ensures seamless data accessibility and availability.

- **Data Access with Hive**

Utilizing Hive to access and manage data stored in HDFS, creating structured tables for the data mart, streamlines data retrieval and enhances data management capabilities.

- **Data Visualisation with Tableau**

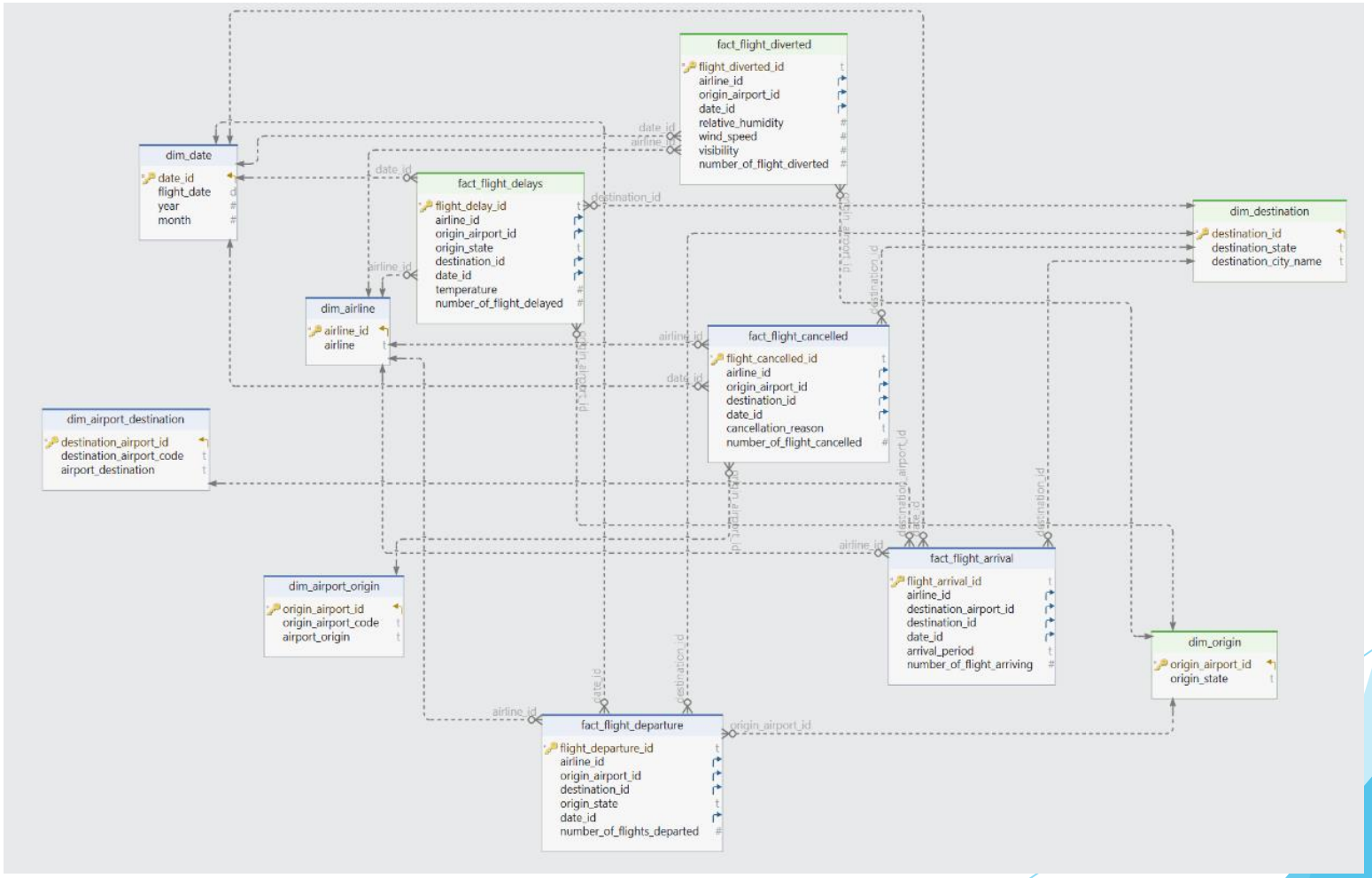
Integrating processed data into Tableau enables the creation of interactive visualizations and dashboards, empowering stakeholders to derive actionable insights and make informed decisions.

ETL

- We opted for ETL (Extract, Transform, Load) due to its ability to refine data before loading, ensuring accuracy and consistency. This process enables thorough data cleaning, transformation, and enhancement, enhancing overall data quality and usability for analysis.
- In contrast, ELT (Extract, Load, Transform) loads raw data directly, potentially compromising data integrity and complicating subsequent processing steps. Therefore, ETL was chosen to streamline data preparation and optimize it for analytical insights.

DATA MART

Data marts, focused subsets of data warehouses, cater to specific departments with relevant data for analysis. Their Constellation schema design, with a central fact table and streamlined dimensions, simplifies complex queries and boosts performance, making it ideal for data analysis tasks within data marts.

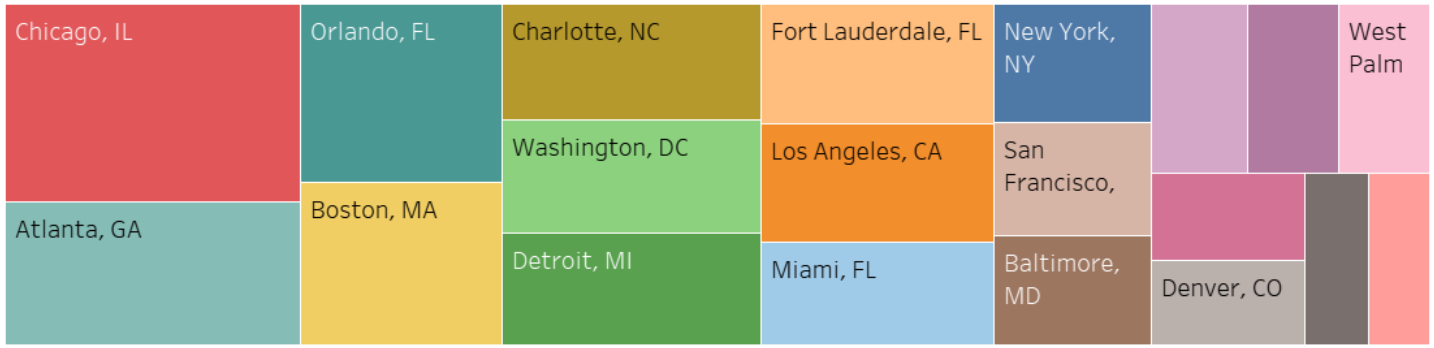


BQ1.Month-by-Month breakdown of departures by destination, Airline and Airport departing New York from 2013 to 2023

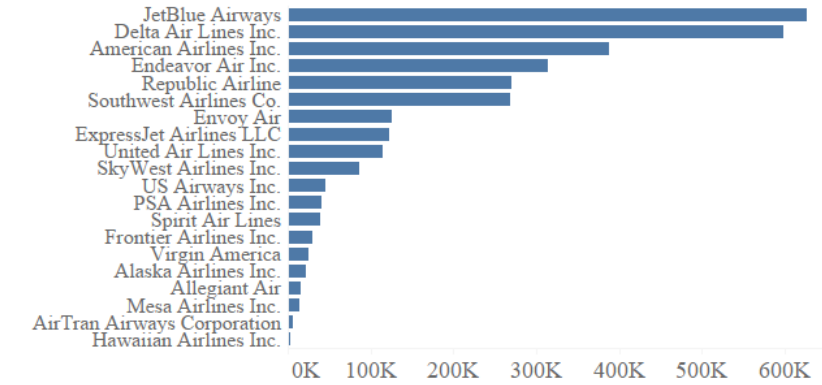
Departure Flight Dashboard

Number of Departure Flight
3,158,609

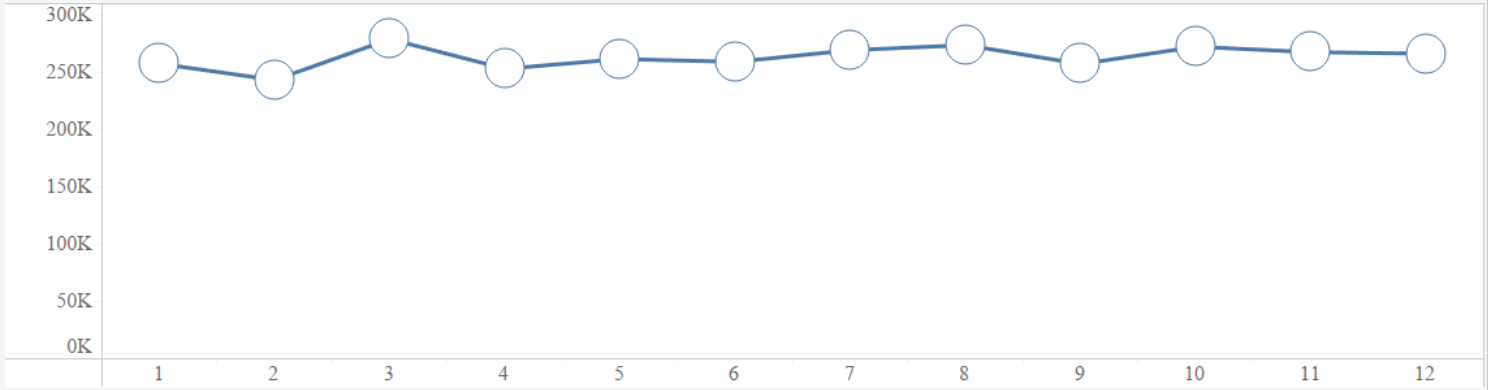
Top Twenty Destination



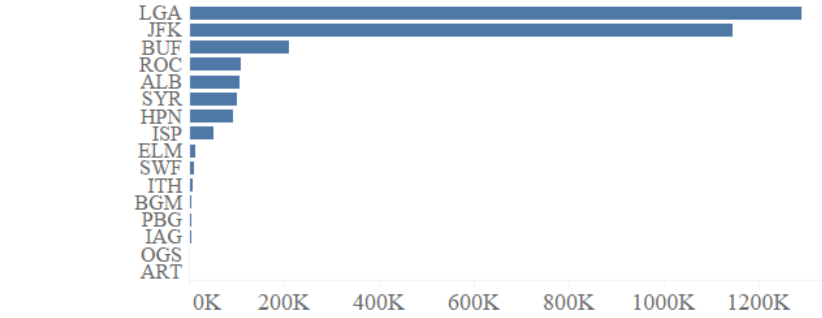
Departure by Airline



Departure Flight by Month Through the Years



Departure Flight by Airport



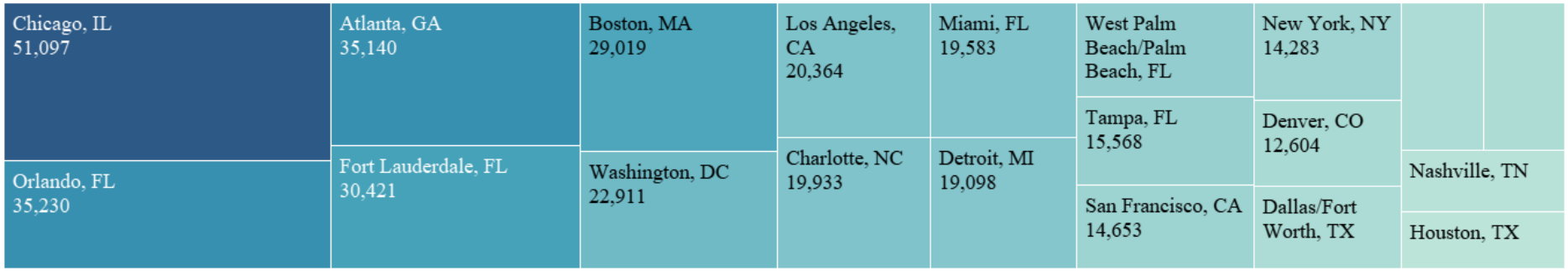
BQ2. Provide a month-by-month breakdown of departure delays by Airline, Airport, and Air temperature leaving New York from 2013 to 2023.

Flight DelayDashboard

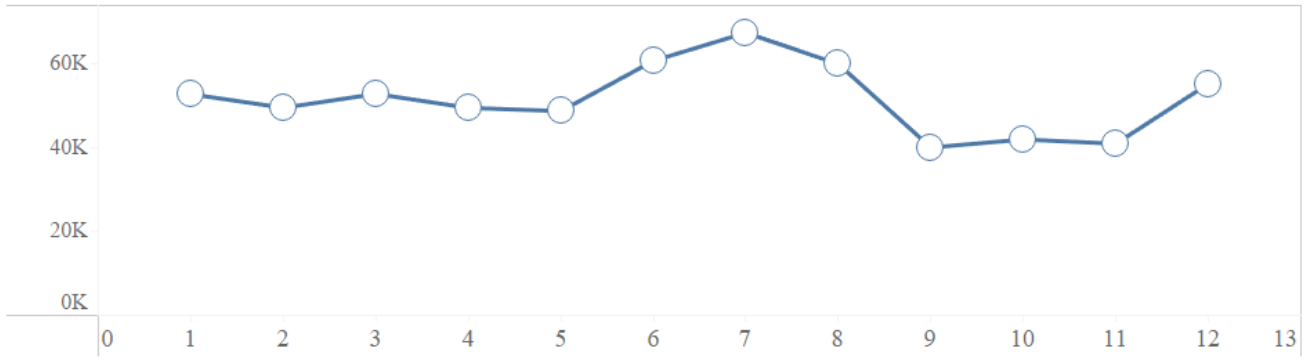
Temperature

1 - 25 (F)	26 - 50 (F)	51 - 75 (F)	-24 - 0 (F)	>= 76 (F)
32,837	206,933	239,288	578	140,466

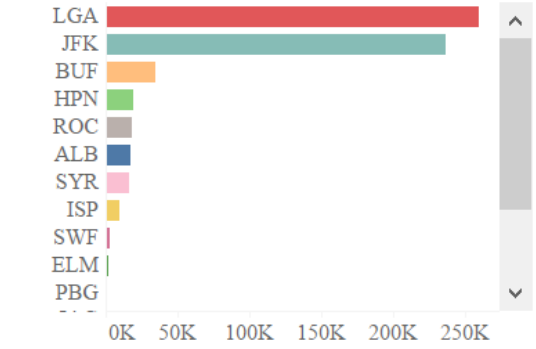
Top 20 Destination by Delay



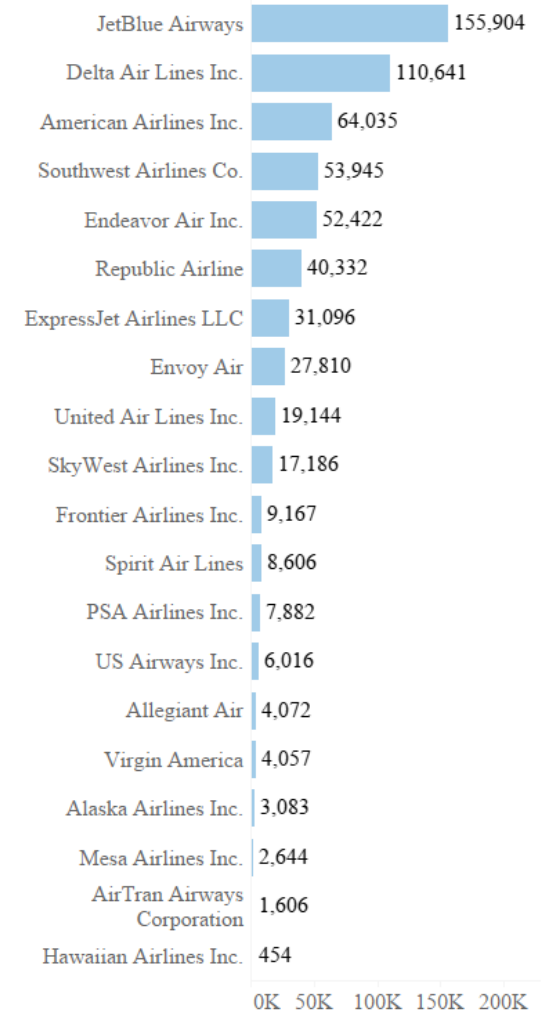
Delay by month



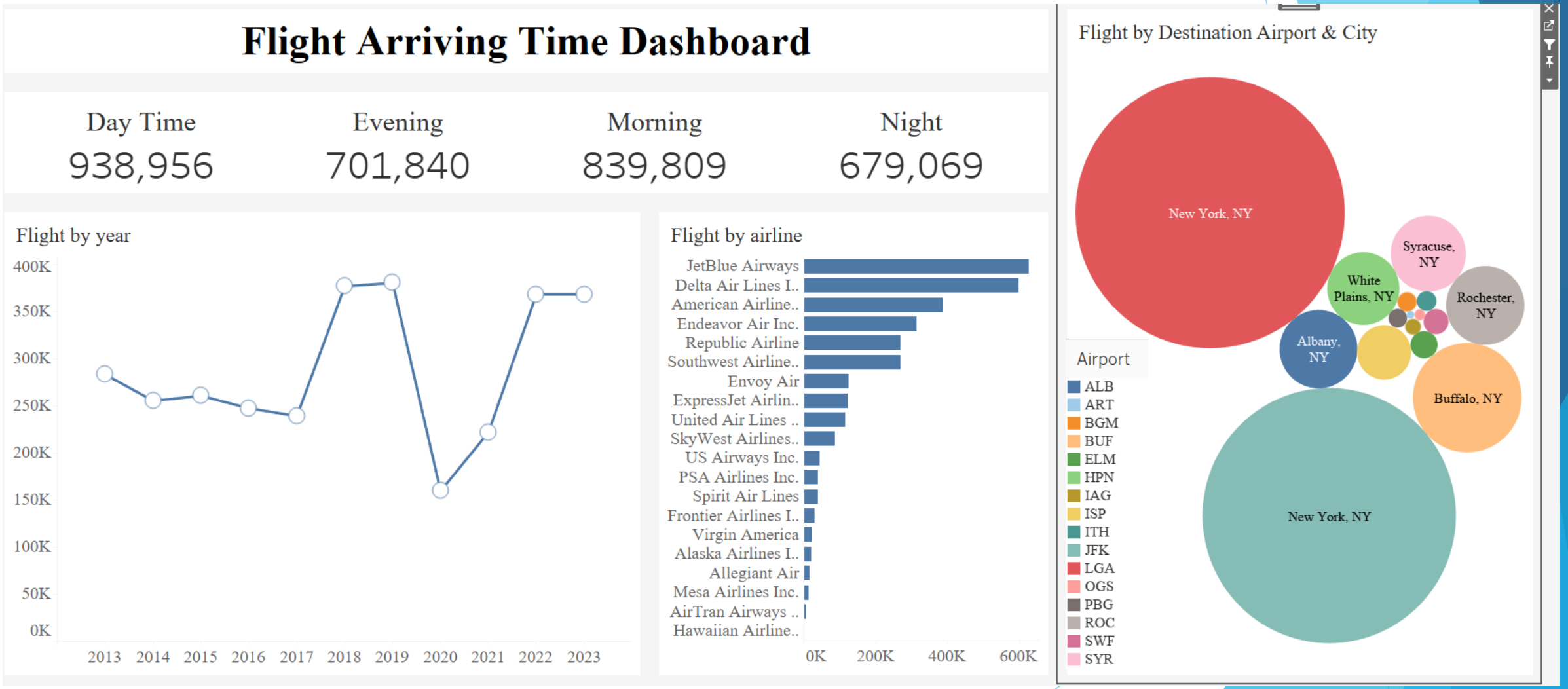
Delay by Airport



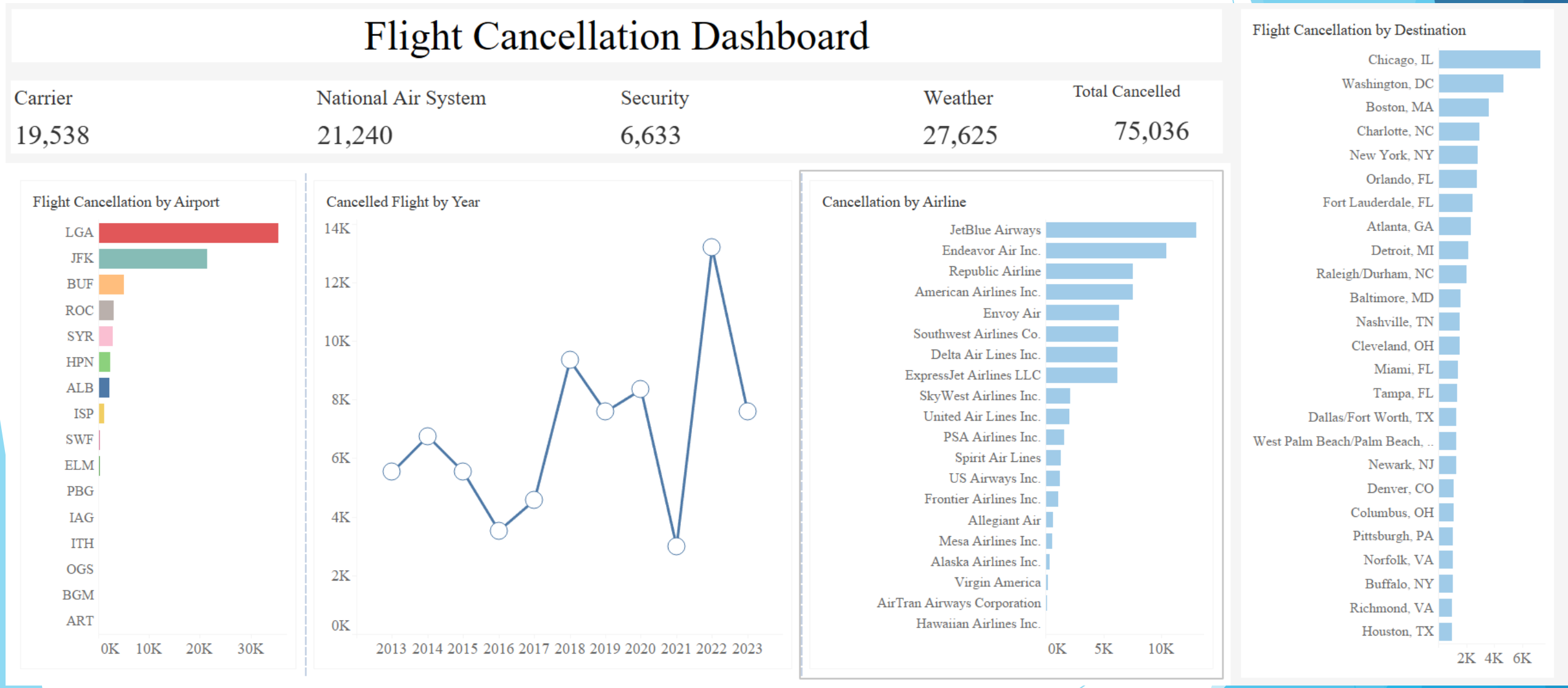
Delay by Airline



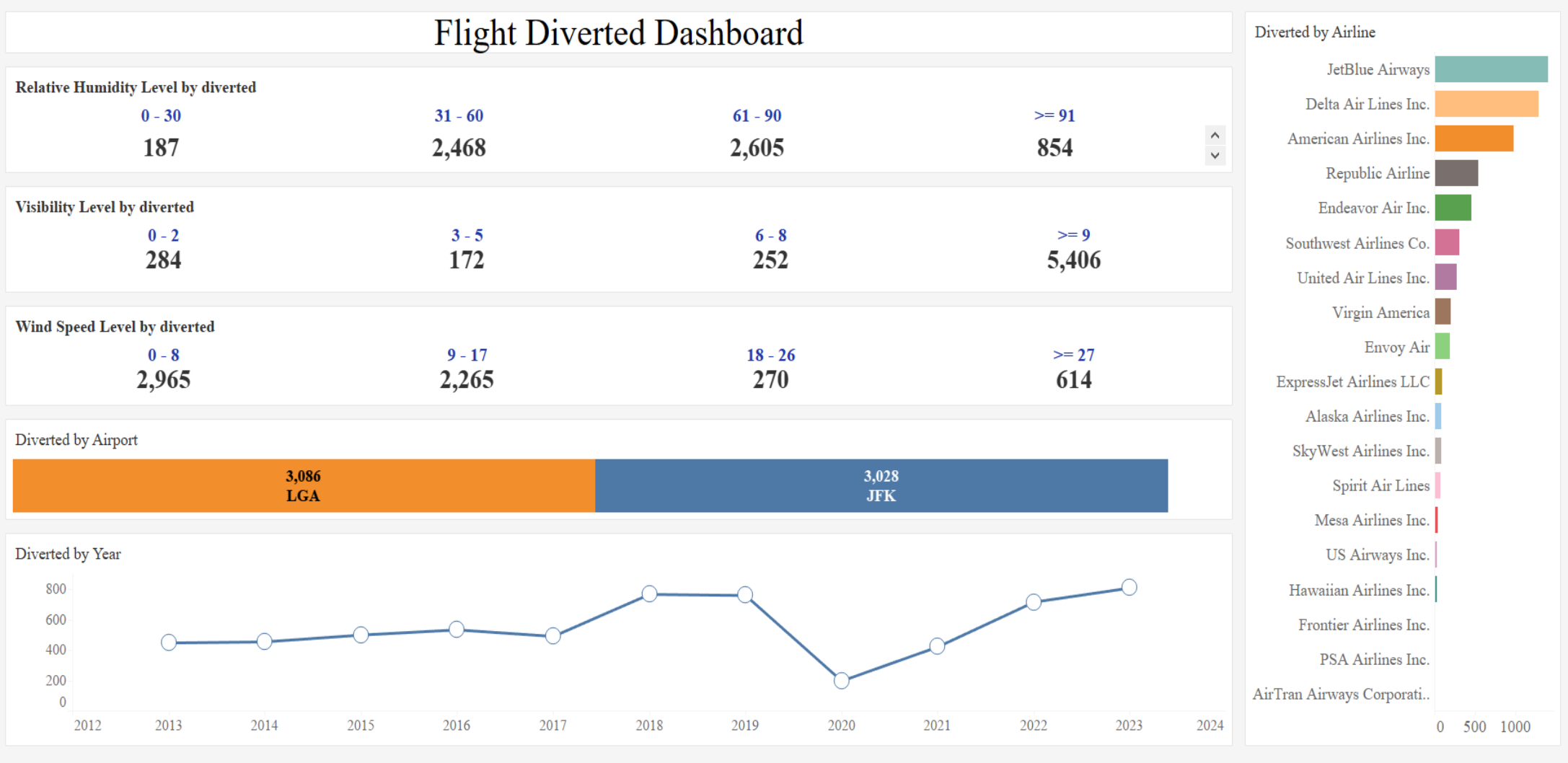
BQ3. What is the total number of flights arriving at New York Airports by morning, daytime, evening and night by year by Airline, by airport from 2013 to 2023?



BQ4. What is the number of flights cancelled by Airline, Airport, destination, and cancellation reason from New York airports by year from 2013 to 2023?



BQ5. **How many flights were diverted from New York airports LGA and JFK by humidity, wind speed and visibility from 2013 to 2023?**



FINAL RECOMMENDATIONS

Optimisation of Airline Schedules:
From the "Departure Flight by Month Through the Year" section, we can see that in March, there were more departures compared to other months, indicating a high demand during this month. Utilising this trend, airlines could increase flights during this peak period to optimise ticket sales and resource utilisation.

Enhanced Crisis Management for Delays:
The "Delay by Month" graph shows an increase in delays toward the middle of the year, peaking in July. This could be used to advise airlines and airports to ramp up their operational and customer service staff during the summer season to handle potential delays more effectively.

Resource Allocation at Airports:
According to the "Flight Arriving Time Dashboard", morning and daytime see higher numbers of arrivals, with 938,956 flights during the daytime and 839,809 flights in the morning throughout the analysed period. This data can be used to argue for increased staffing and better resource distribution during these times to handle the higher passenger load.

Strategic Responses to Cancellations:
The "Flight Cancellation Dashboard" details that weather caused 27,625 out of 75,036 cancellations over the eleven-year period. This statistic underscores the need for better weather forecasting tools and more flexible scheduling practices to mitigate the impact of weather on flight operations.

Safety and Efficiency Improvements:
The data on flight diversions due to visibility shows that high visibility (> 9 miles) caused 5406 diversions. Enhancing weather prediction and communication systems to address high visibility can significantly reduce such diversions and improve safety.

FINAL RECOMMENDATIONS

Dynamic Pricing and Capacity Management:
The "Departure Flight by Airport" shows LaGuardia and JFK as major hubs with significant traffic. Airlines could use this data to adjust ticket prices dynamically. For instance, during March, where there is a spike in departures, prices could be increased slightly to manage demand, whereas in slower months like February, prices could be reduced to attract more passengers.

Improve Passenger Communication:
The "Delay by Airline" data shows that JetBlue Airways leads in volume of departures but also experiences significant delays. Recommendations could include JetBlue developing a dedicated notification system within their app to alert passengers of expected delays, specifically during peak delay periods like July and December, enhancing transparency and improving customer satisfaction.

Weather Adaptation Plan:
From the "Flight Cancellation Dashboard" weather emerges as a major cause of cancellations, particularly at JFK and LaGuardia airports. A recommendation could be the installation of advanced weather prediction systems at these airports to better anticipate adverse weather and adjust flight schedules proactively to minimise cancellations.

Enhanced Staff Training and Flexibility:
The "Flight Arriving Time Dashboard" shows high numbers of arrivals in the morning and daytime. An effective strategy might be to implement shift flexibility and peak time bonuses for staff during these periods to ensure that all positions are covered, and service quality remains high, especially in customer service and baggage handling areas.

Developing Contingency Plans for Diversions:
Considering the data showing a significant number of diversions at JFK and LaGuardia due to visibility issues, as detailed in the "Flight Diversions" dashboard, airports could develop agreements with alternative landing sites and arrange for efficient passenger transfers. This could include shuttle services or partnership deals with local transport companies to look after passenger movement during diversions or significant delays.