Balanced flight cancellation and delay : 2019-2023

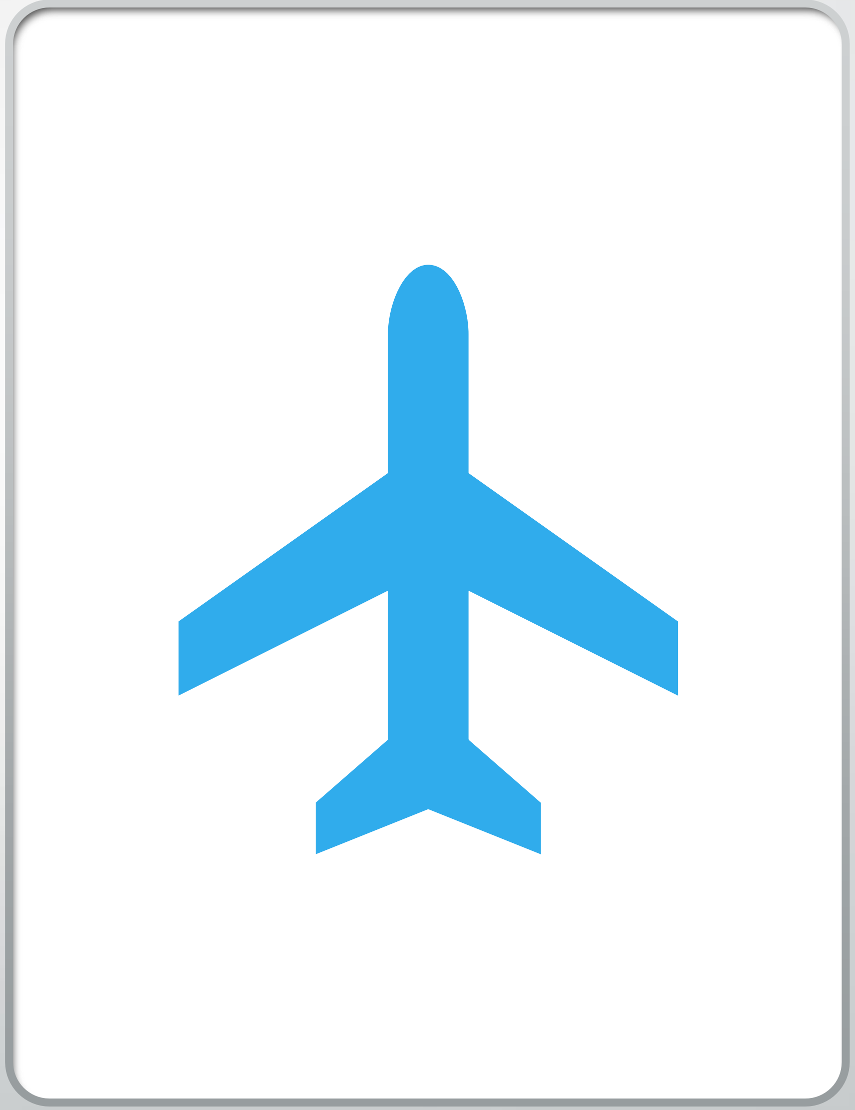
Objective

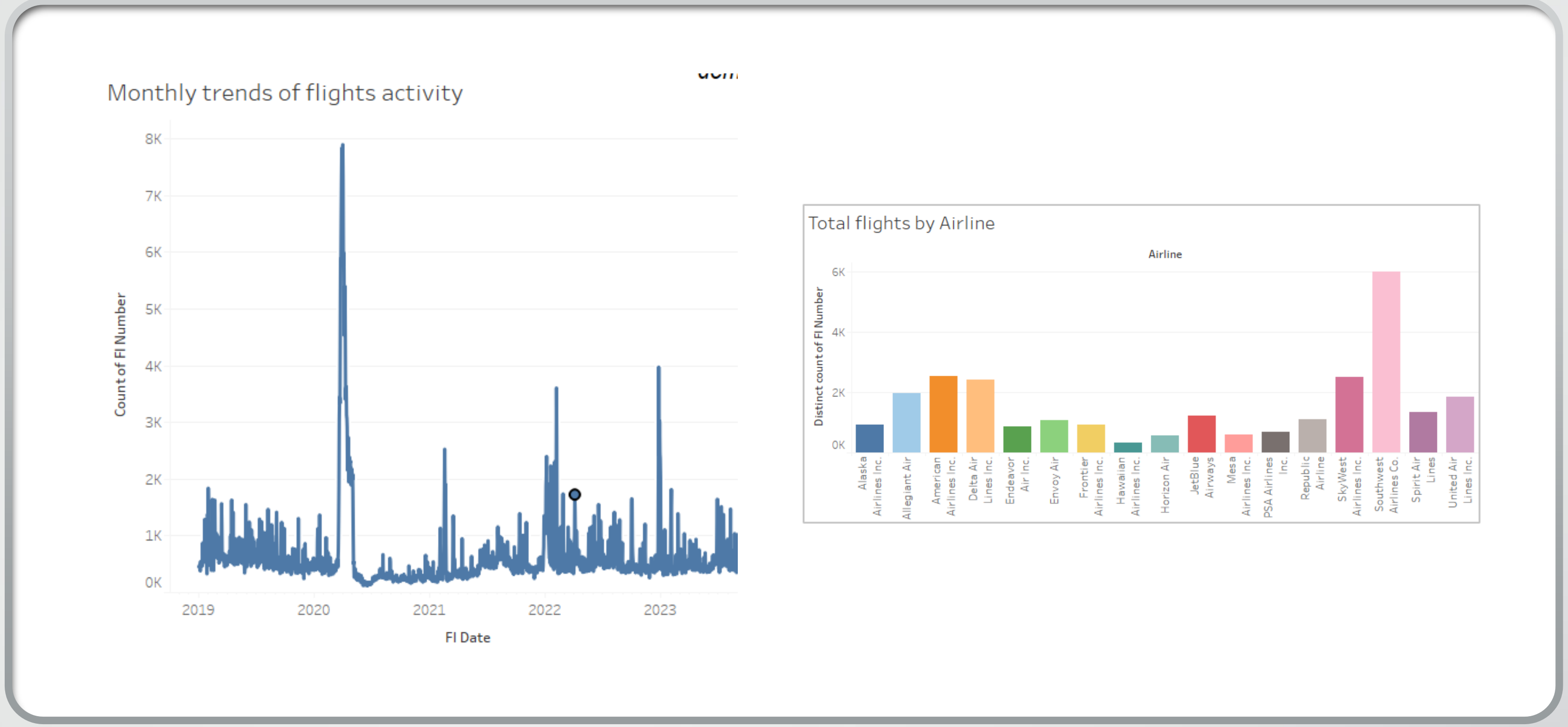
The primary goal of selecting this dataset is to conduct a comprehensive analysis of Flight Delay and Cancellation Metrics. Specific objectives include:

1. Departure Delays, Arrival Delays, and Cancellation Rates:
   Analyze patterns and trends in departure and arrival delays.
   Investigate the frequency and reasons behind flight cancellations.
2. Airline Performance:
   Evaluate the on-time performance of different airlines.
   Examine average delay times and cancellation rates for each airline.
3. Distance and Route Analysis:
   Explore the relationship between the distance of flights and delays.
   Analyze delays on specific routes or between particular cities.
4. Time Analysis:
   Study the impact of departure and arrival times on delays.
   Identify peak periods of delay occurrences.
5. Airport Analysis:
   Utilize the 'airport_location.csv' file to analyze delays and cancellations at various airports.
6. Reasons for Delay:
   Investigate the causes of delays, including aircraft issues, weather conditions, and other contributing factors.

Data:
The dataset is sourced from the U.S. Department of Transportation's Bureau of Transportation Statistics, covering Airline Flight Delay and Cancellation Data from August 2019 to August 2023. The dataset is accessible on Kaggle.

Skills & Tools:
Big Data
Understanding Data Ethics
Tableau
Python
Data Mining
Predictive Analysis
Time Series Analysis and Forecasting
Regression and Cluster Analysis





Monthly trends of flights activity

Total flights by Airline

Exploration and Analysis

Key Questions:

Are there dominant airlines in terms of total flight activity?

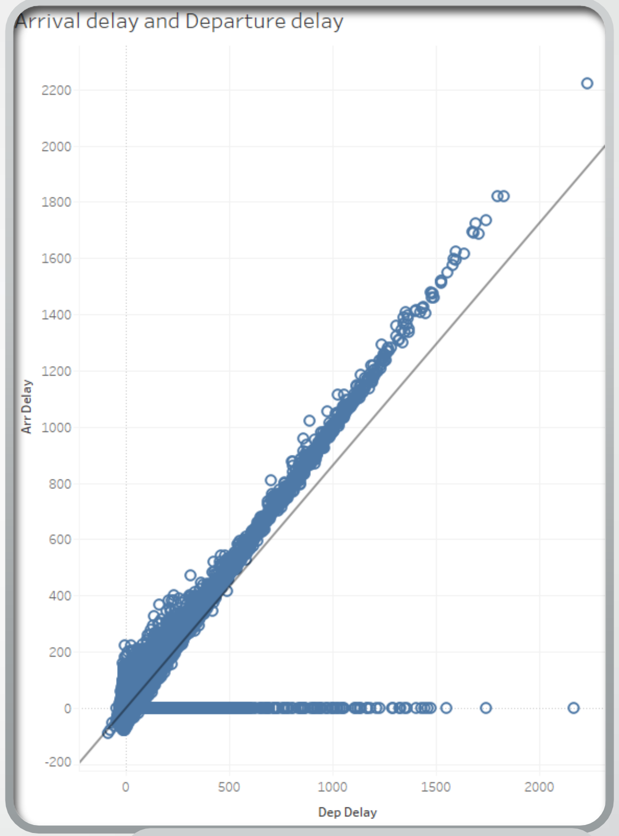
Top 5 busiest airlines based on the distinct count of flight numbers?

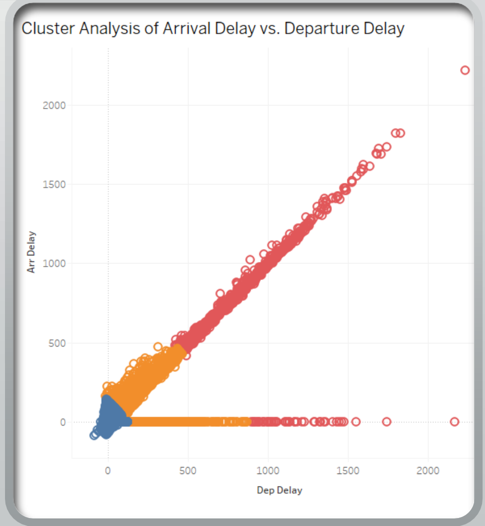
Are there specific months where airlines can anticipate increased demand or challenges?

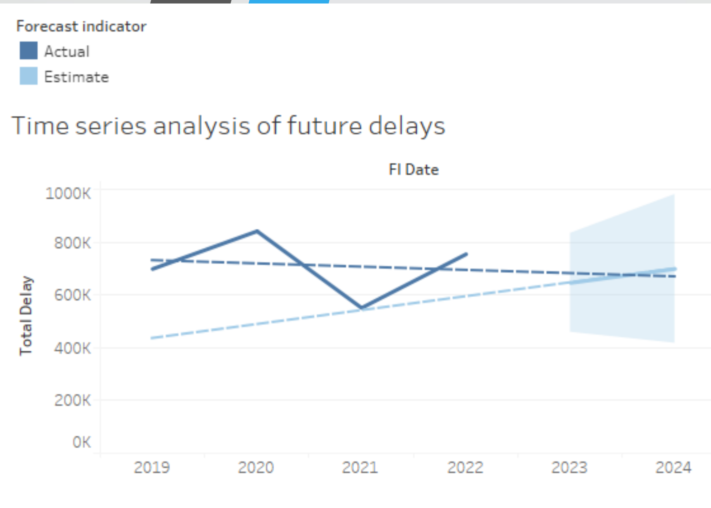
Exploration and Analysis

- **Hypothesis:**
- There is a significant positive relationship between arrival delay and departure delay.
- Based on the following information it is clear that there is a significant positive relationship between arrival delay and departure delay.
- **ARR_DELAY and Other Variables:**
- ARR_DELAY has a strong positive correlation with DEP_DELAY (0.965265), confirming the strong relationship between arrival delays and departure delays.
- ARR_DELAY has a moderate positive correlation with DELAY_DUE_WEATHER (0.293201), indicating a moderate association between arrival delays and delays due to weather.

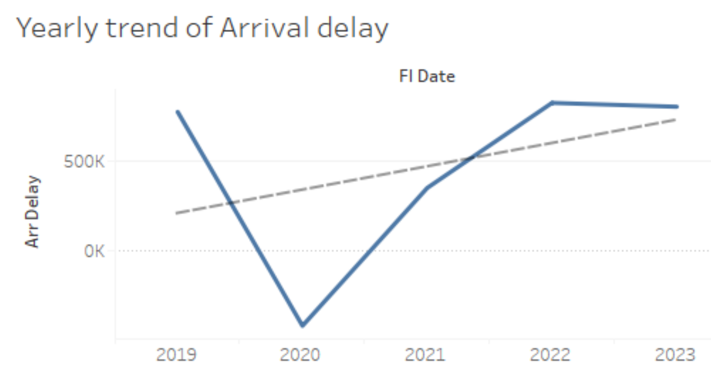
The clustering provides additional granularity by revealing different groups of flights with distinct delay characteristics.


Arrival delay and Departure delay

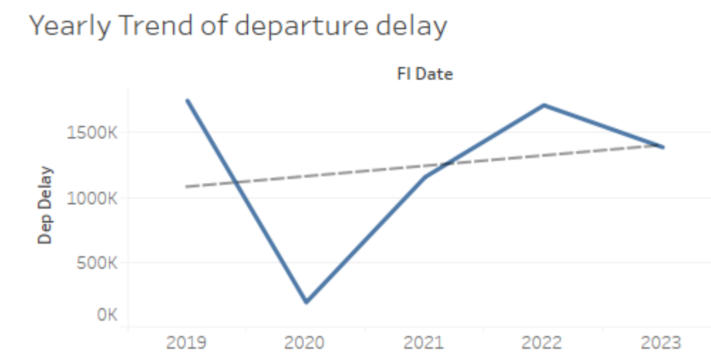

Cluster Analysis of Arrival Delay vs. Departure Delay



Time series analysis of future delays



Yearly trend of Arrival delay



Yearly Trend of departure delay

Regression Analysis

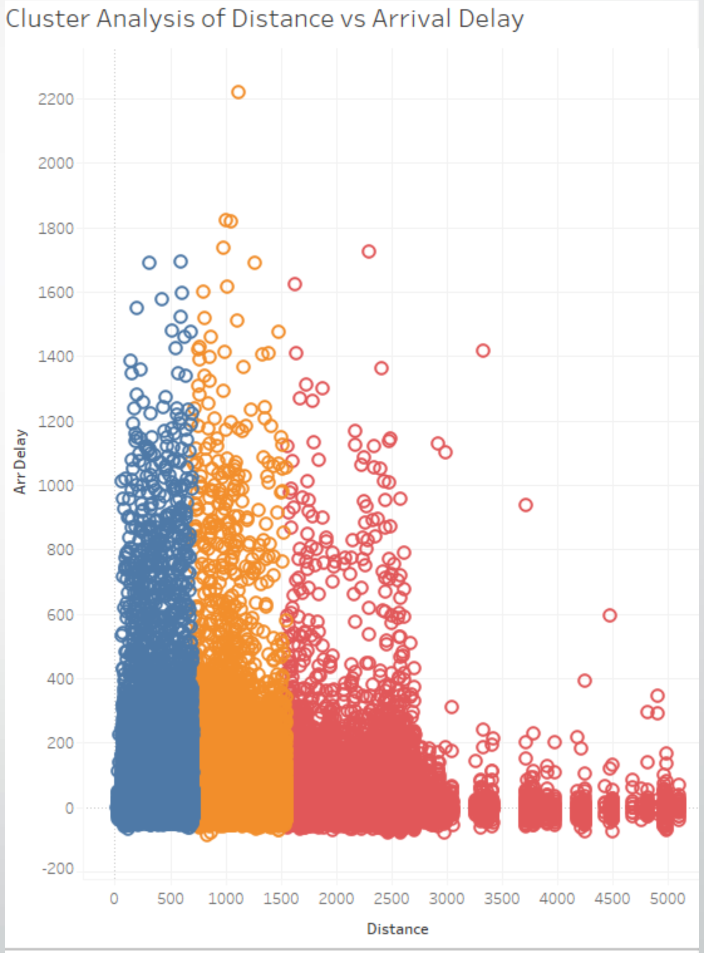
Arrival Delay and Departure Delay:
Equation: Arrival Delay = 0.863512 * Departure Delay - 2.7914
P-value: < 0.0001
Interpretation: Strong positive linear relationship between departure delay and arrival delay, statistically significant.
Total Delay and Years (Estimated Trend Line):
Equation: Total Delay = -12304.5 * Year of Fl Date + 2.55726e+07
P-value: 0.869988
Interpretation: Estimated decreasing trend in total delays over the years, but the trend is not statistically significant.
Delay Due Weather and Years:
Equation: Delay Due Weather = -2453.9 * Year of Fl Date + 5.03856e+06
P-value: 0.829682
Interpretation: Estimated decreasing trend in delays due to weather over the years, but the trend is not statistically significant.
Arrival Delay and Years:
Equation: Arrival Delay = 129245 * Year of Fl Date - 2.60741e+08
P-value: 0.520419
Interpretation: Estimated increasing trend in arrival delays over the years, but the trend is not statistically significant.
Dep Delay and Year:
Equation: Dep Delay = 80416.5 * Year of Fl Date + -1.61282e+08
P-value: 0.746779
Interpretation: Estimated increasing trend in departure delays over the years, but the trend is not statistically significant.
Total Delay Trend with Intersection:
Equation: Actual and Estimated Trend Lines intersect, with the estimated trend going higher than the actual trend after the intersection.
Interpretation: Divergence between the observed total delays and the model's predicted trend, indicating potential limitations or areas for improvement in the current modeling approach.

Cluster Analysis

General Relationship
There is no straightforward linear relationship between distance and arrival delay, departure delay, or delay due to weather. The concentration of data points suggests that the relationship may be influenced by various factors.

Relationship between Distance, Arr_Delay
The distinct clusters suggest that there are different phases or conditions associated with arrival delays at various distance ranges. Each cluster may represent a specific set of influencing factors.

Arrival delays show distinct clusters, suggesting that specific distance ranges or conditions are associated with different patterns of arrival delays.



Conclusion

Impact of COVID-19:
March 2020's spike in flight counts aligns with the onset of the COVID-19 pandemic, indicating notable shifts in travel behavior.
Airline Dominance:
Southwest Airlines leads with the highest flight count, demonstrating operational resilience during dynamic conditions.
Operational Resilience:
Southwest Airlines consistently maintains high flight counts, reflecting adaptability and efficiency in operations.
Hypothesis Confirmation:
The linear relationship between arrival delay (arr_delay) and departure delay (dep_delay) is confirmed through statistical analysis, affirming their correlation.
Visual Insights:
Graphs, including time series charts, bar charts for airline comparison, and bubble charts, visually represent trends, market share, and operational aspects of each airline.
Other Summary Statistics:
Summary statistics, such as correlation coefficients, p-values, and coefficients from regression analyses, provide quantitative support for observed relationships.
Recommendations:
Continuous monitoring of external factors, data-driven decision-making, and proactive customer communication are recommended for operational excellence.
Future Considerations:
Ongoing adaptation to external factors, collaboration within the industry, and prioritizing customer experience remain critical for sustained success.