UNVEILING PATTERNS: ANALYSIS AND PREDICTION OF AIRLINES DELAYS

Presented to:

FREYA SYSTEMS

Prepared by:

Drexel LeBow Graduate Students

Group - 6

Faculty: Mr. Bennett Foster

Table of Contents:

Introduction

Business Problem

Addressing Flight Delays and the Cost to Airlines.

Understanding and presenting information about arrival and departure delays in the airline industry are key components of this problem. The objective is to propose potential solutions to lessen the financial toll that these delays have on airlines, for smaller carriers like Republic Airlines, Envoy Air, PSA Airlines Inc, Endeavor Air Inc, Mesa Airlines Inc, Horizon Air, Piedmont Airlines, Air Wisconsin Airlines Corp, CommuteAir LLC, GoJet Airlines LLC, Empire Airlines Inc and Peninsula Airways Inc.

Scope of Project

- To predict and analyze the arrival delays specifically for small carriers.
- Focus on understanding the factors contributing to delays in small carriers.
- Provide insights and recommendations to improve the performance of small carriers.

Data

Data Collection: https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=

Data Interpretation:

5 years Monthly Data of Flights in USA: We compiled flight data from 2018 to 2022 into an Excel spreadsheet. Initially, this dataset contained approximately 36 million rows. This data encompassed various attributes such as Year, Month, Day of Month, Day of Week, Marketing Airlines, Operating Airlines, Origin City Name, Origin State Name, Destination City, Destination State, Scheduled Departure Time, Actual Departure Time, Departure Delay 15, Taxi Out, Taxi In, Wheels Off, Wheels On, Scheduled Arrival Time, Actual Arrival time, Arrival Delay 15, Air Time, Distance, Distance Group, Carrier Delay, Weather Delay, NAS Delay, Security Delay, and Late Aircraft Delay. Our team underwent a data cleaning process to eliminate unnecessary and empty rows, streamlining the dataset for easier analysis and utilization.

Carrier information, Flight schedules, Arrival and departure times, and delay information: Each airline possessed comprehensive data regarding the arrival, departure, and delay timings in the dataset. Our efforts were directed towards minimizing the delay duration for every airline. It's worth noting that certain flights experienced delays in their arrival schedules, while others encountered delays in their departure schedules. Each carrier information was helpful to reduce the delay time for small airlines.

Data Preparation:

Missing Values: During our dataset analysis, we identified instances where certain values were missing in both rows and columns. Consequently, our team undertook the task of rectifying these missing values through data cleaning procedures and deleted the missing values from the datasheet.

Redundant variables: As we progressed in our dataset modeling and analysis, we became aware that certain variables within the sheet were unnecessary and were merely contributing to the overall dataset size. To streamline and reduce the dataset's size, we opted to eliminate these variables from the entire dataset.

Joining Huge Monthly Dataset: Upon downloading the dataset from the provided link, we noticed that the data was organized monthly. To enhance the ease of modeling and visualization, we collaboratively merged five years' worth of data into a single Excel file. Each monthly dataset contained approximately 600,000 rows and columns, prompting us to perform data cleaning and consolidation for more efficient utilization on various platforms.

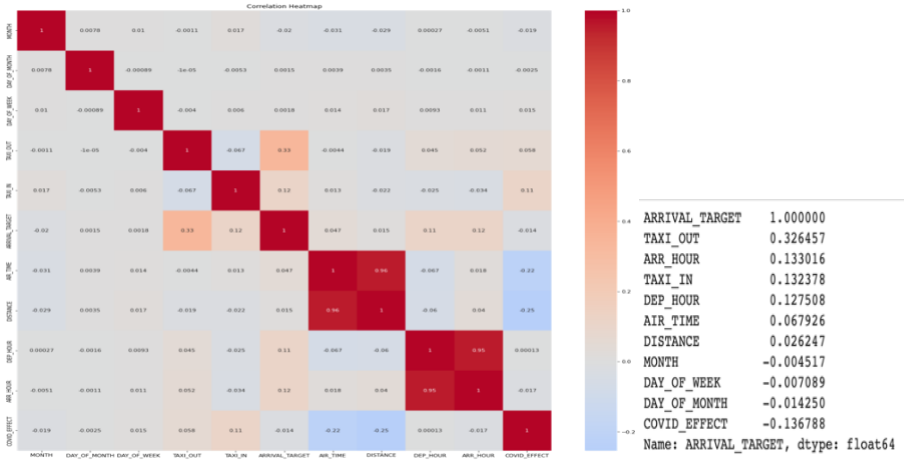
Methodology

Data Cleaning:

We obtained the dataset by downloading it from a provided link. This dataset contained approximately 36 million rows, encompassing data for a five-year period from 2018 to 2022. To streamline the data and make it more manageable, we embarked on a systematic cleaning process at the level of individual months. This approach reduced the number of unnecessary rows and columns, resulting in a dataset of around 250,000 entries per month. During the data cleaning phase, we executed various tasks such as eliminating unnecessary spaces, empty rows, and columns. Additionally, we categorized airlines into "big" and "small" operators. Employing the VLOOKUP function, we established associations between distinct carriers and their corresponding marketing and operating airlines. Subsequently, we used conditional statements (IF conditions) to generate an indicator that facilitated comparisons between marketing and operating airlines. To enhance consistency, we converted numerical values into time formats for fields like Departure Time, Arrival Time, Actual Departure Time, and Actual Arrival Time. Another significant

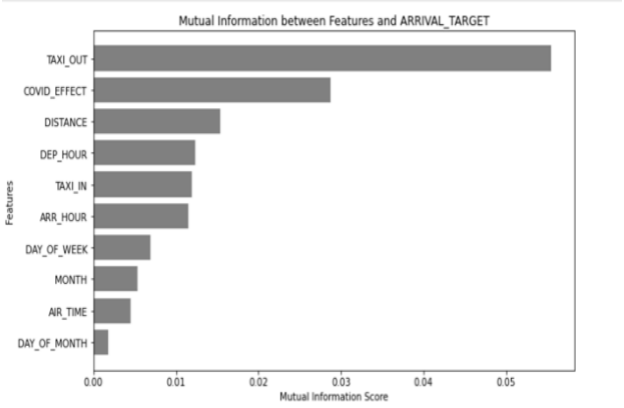
transformation involved the conversion of delay columns (e.g., carrier delay, weather delay, NAS delay, security delay, late aircraft delay) into binary format. Delays with values equal to or greater than 1 were labelled as "1," while others were classified as "0."

Recognizing the importance of relevant features, we excluded irrelevant columns that didn't contribute to our modelling and visualization goals. Furthermore, as our analysis focus was specifically on small airlines, we adjusted the dataset accordingly by eliminating rows associated with large airlines. After these preparations, we exported the cleaned data sheets to Python for further processing. A Python script was then utilized to remove rows where the marketing and operating airlines were identical, ensuring data integrity and quality throughout the dataset.




```
ARRIVAL_TARGET    1.000000
TAXI_OUT          0.326457
ARR_HOUR          0.133016
TAXI_IN           0.132378
DEP_HOUR          0.127508
AIR_TIME          0.067926
DISTANCE          0.026247
MONTH            -0.004517
DAY_OF_WEEK      -0.007089
DAY_OF_MONTH     -0.014250
COVID_EFFECT     -0.136788
Name: ARRIVAL_TARGET, dtype: float64
```

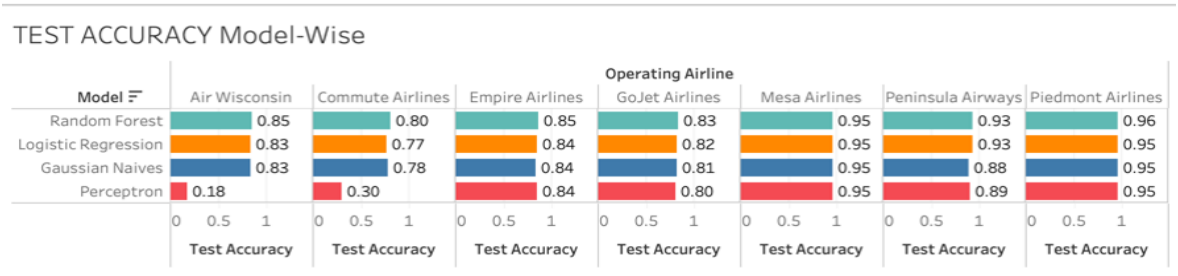
In the correlation analysis, "ARRIVAL TARGET" naturally shows a perfect positive correlation with itself (correlation coefficient 1.000). Notably, "TAXI OUT" exhibits a moderate positive correlation (0.326) with "ARRIVAL TARGET," suggesting longer taxi-out times correspond to extended arrival target times. "ARR HOUR," "TAXI IN," "DEP HOUR," and "AIR TIME" display weaker positive correlations (coefficients between 0.067 and 0.133), implying some influence on arrival target time. In contrast, "DISTANCE," "MONTH," "DAY OF WEEK," "DAY OF MONTH," and "COVID EFFECT" have notably weaker correlations, close to zero or slightly negative, suggesting limited linear impact on arrival target time.

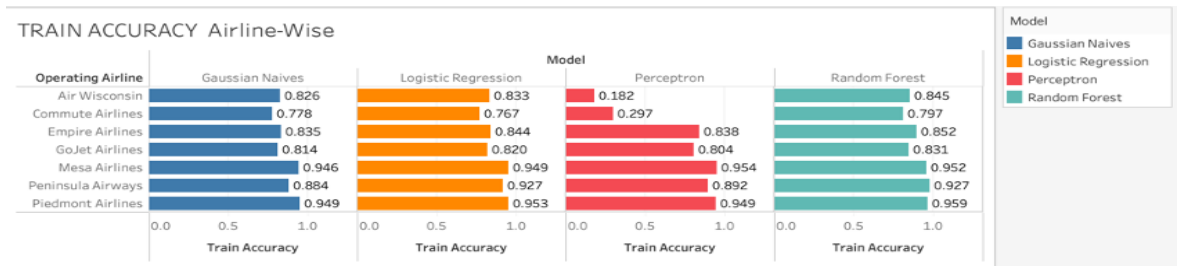


	Feature	MI_Score
3	TAXI_OUT	0.055503
9	COVID_EFFECT	0.028708
6	DISTANCE	0.015342
7	DEP_HOUR	0.012345
4	TAXI_IN	0.011946
8	ARR_HOUR	0.011503
2	DAY_OF_WEEK	0.006929
0	MONTH	0.005314
5	AIR_TIME	0.004550
1	DAY_OF_MONTH	0.001781

Mutual information guides feature selection and model building by helping you decide which features to use, enhancing model performance, interpretability, and efficient resource use. MI scores aid in identifying highly relevant features linked to the target variable, reducing data dimensionality, enhancing model efficiency, and detecting non-linear relationships. They rank features by importance, helping prioritize feature engineering for improved model performance, especially when dealing with complex, non-straightforward interactions among variables. This ultimately enhances model generalization to unseen data.

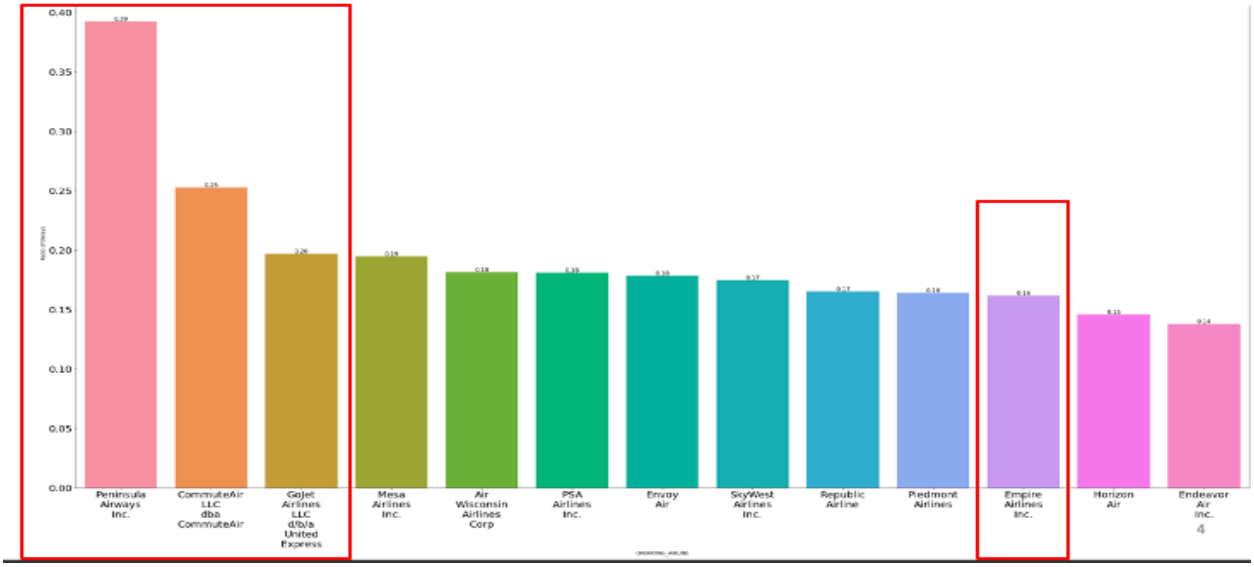
Modeling:





Following the completion of the data cleaning procedure, the dataset was divided into training and testing sets. We used stratification while splitting the data into train and test to maintain the class ratio in both the subsets like the original dataset. We proceeded to engage in the process of constructing models for the data, employing techniques including Logistic Regression, Gaussian Naive Bayes, Perceptron, and Random Forest Classifier. Upon conducting thorough analyses of all the models, we determined that the Random Forest Classifier emerged as the most suitable choice for our dataset.

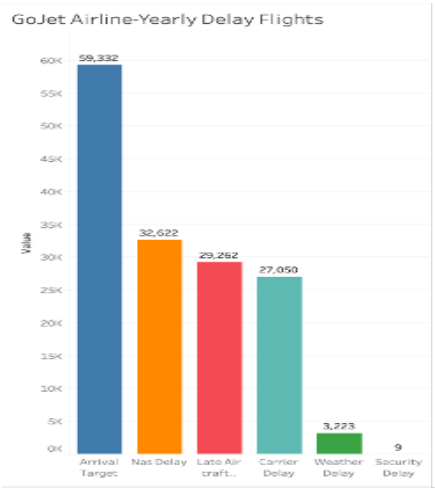
Data Visualization:



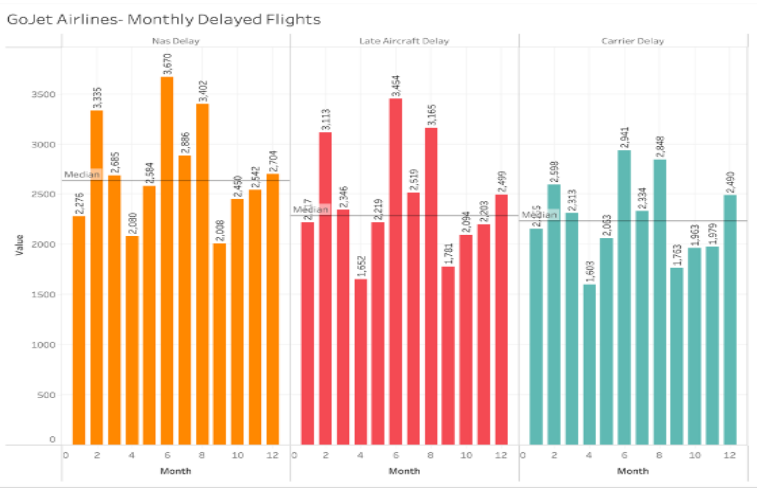
In our research on the 13 shortlisted small carriers, we found that, aside from Peninsula Airways, Empire Airlines, GoJet Airlines, and CommuteAir, all the other carriers had substantial analytics teams. GoJet Airlines lacks an analytics team based on LinkedIn, while CommuteAir has a small one. Peninsula Airways was acquired by Ravn Alaska, resulting in incomplete data for our analysis. Empire Airlines temporarily ceased operations in 2020 during a brief Chapter 11 bankruptcy period in March but later emerged successfully and has since operated as usual. Empire Airlines also acquired FedEx feeder West Air.

Business Analysis

GoJet Air

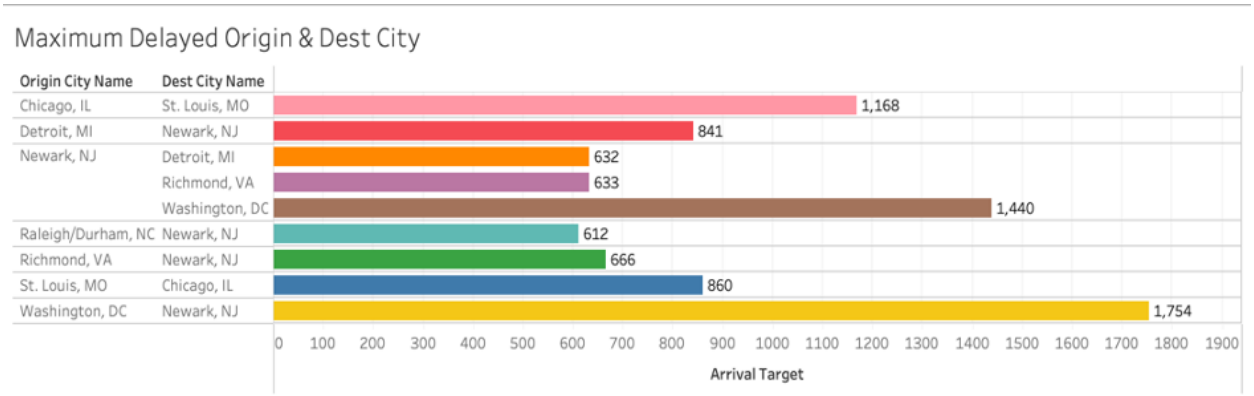


Graph 1



Graph 2

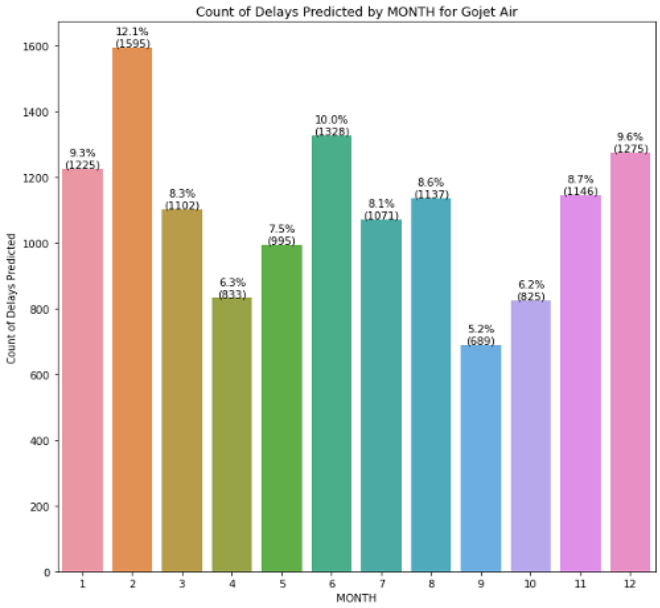
Graph 1 illustrates the annual arrival delay data for GoJet Airlines, categorized by the underlying causes of the delays. Notably, NAS Delay, Late Aircraft Delay, and Carrier Delay emerge as the primary factors responsible for these delays. In Graph 2, we examine the monthly breakdown of delays attributed to the top three reasons. Notably, February, June, August, and December stand out as the months when delays are most prominent.



Graph 3

Graph 3 displays the primary cities of origin and destination where GoJet flights experience the most substantial arrival delays. The data indicates that flights traveling from Washington, DC to

Newark, NJ, as well as those going from Newark, NJ to Washington, DC, encounter the most significant arrival delays.

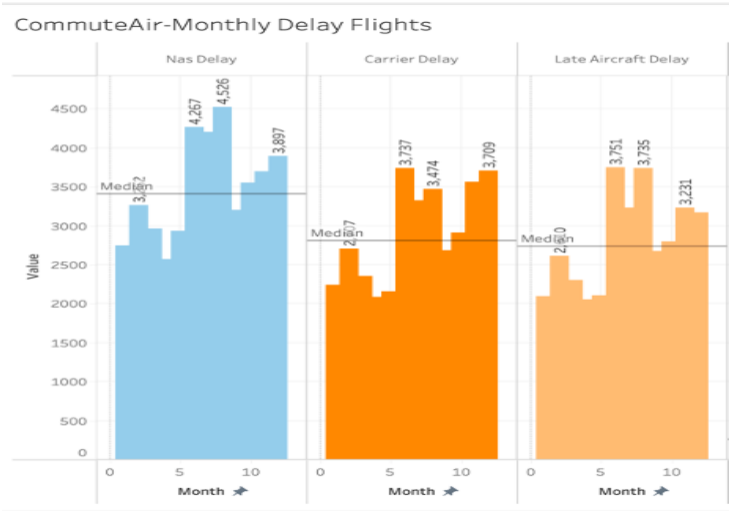


Graph 4

Graph 4 shows Random Forest model's monthly delay forecast for GoJet Air indicates that February is the month with highest number of delays at 12.1%. In second place is June, with December closely trailing as the third month with the highest number of delays.

CommuteAir:

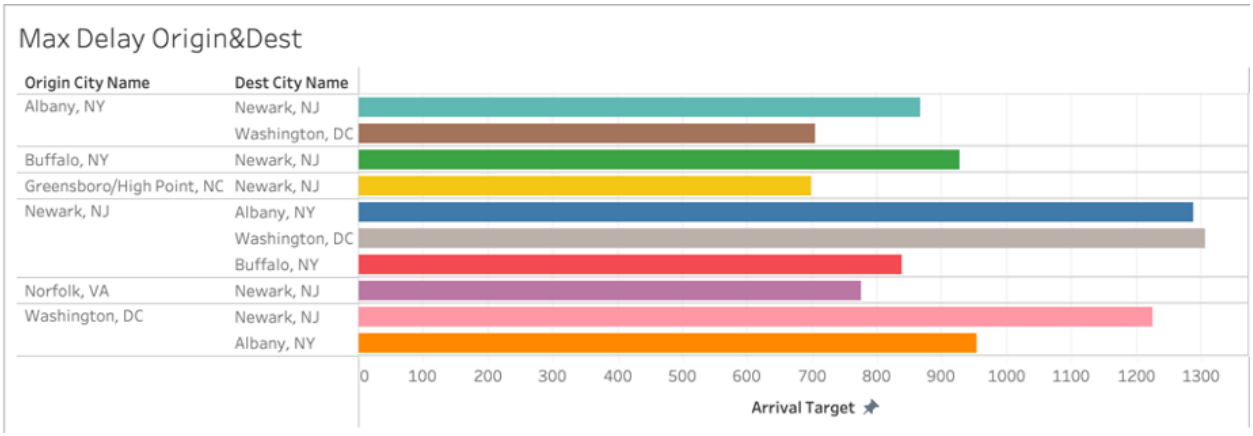
Graph 5



Graph 6

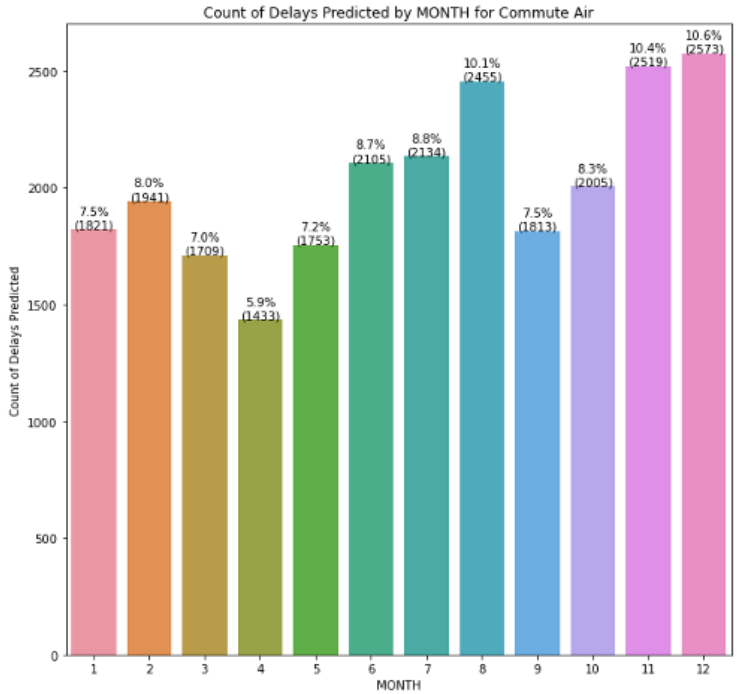
The annual statistics for Commute Airlines' arrival delays are shown in Graph 5 along with a breakdown of their underlying reasons. It's important to note that the main causes of these delays are NAS Delay, Late Aircraft Delay, and Carrier Delay. We examine the monthly distribution of

delays linked to the top three reasons in Graph 6. However, the months with the highest frequency of delays are June, August, February, and December.



Graph 7

Graph 7 displays the primary cities of origin and destination where CommuteAir flights experience the most substantial arrival delays. The data indicates that flights traveling from Newark, NJ to Washington, DC, as well as those going from Newark, NJ to Albany, NY encounter the most significant arrival delays.



Graph 8

In Graph 8, the monthly delay predictions generated by the Random Forest model for CommuteAir are depicted. It becomes evident that December records the highest delays at 10.6%. Following closely behind are November in second place and August as the third month with the most significant delay occurrences.

Conclusion

GoJet Airlines and CommuteAir are both experiencing significant flight delays, with recurring issues primarily attributed to factors like NAS Delay, Late Aircraft Delay, and Carrier Delay. These delays have a noticeable impact on their operations, leading to substantial financial costs.

GoJet Airlines: Significant areas for development are identified by analyzing the flight data of GoJet Airlines. It is a significant issue with a median of 1,101 delayed flights every month. These delays, which last an average of 50 minutes each late flight, influence both operational expenses and passenger happiness. There might be significant cost cuts with estimates of $2.89 million per month if half of the problems are fixed. GoJet could save almost $1.16 million every month even with a 20% resolution rate, demonstrating the financial benefits of dealing with these problems.

CommuteAir: The analysis of flight data by CommuteAir gives important operational insights. Optimizing planning and avoiding delays are crucial with a monthly median of 2,021 delayed flights lasting roughly 51 minutes each. The potential financial savings from addressing these problems are significant; even at a 20% resolution rate, the airline could save roughly $2.1 million per month. Resolving 50% of the problems may result in monthly savings of $5.4 million. These numbers highlight how cost-effective it is to take on these tasks.

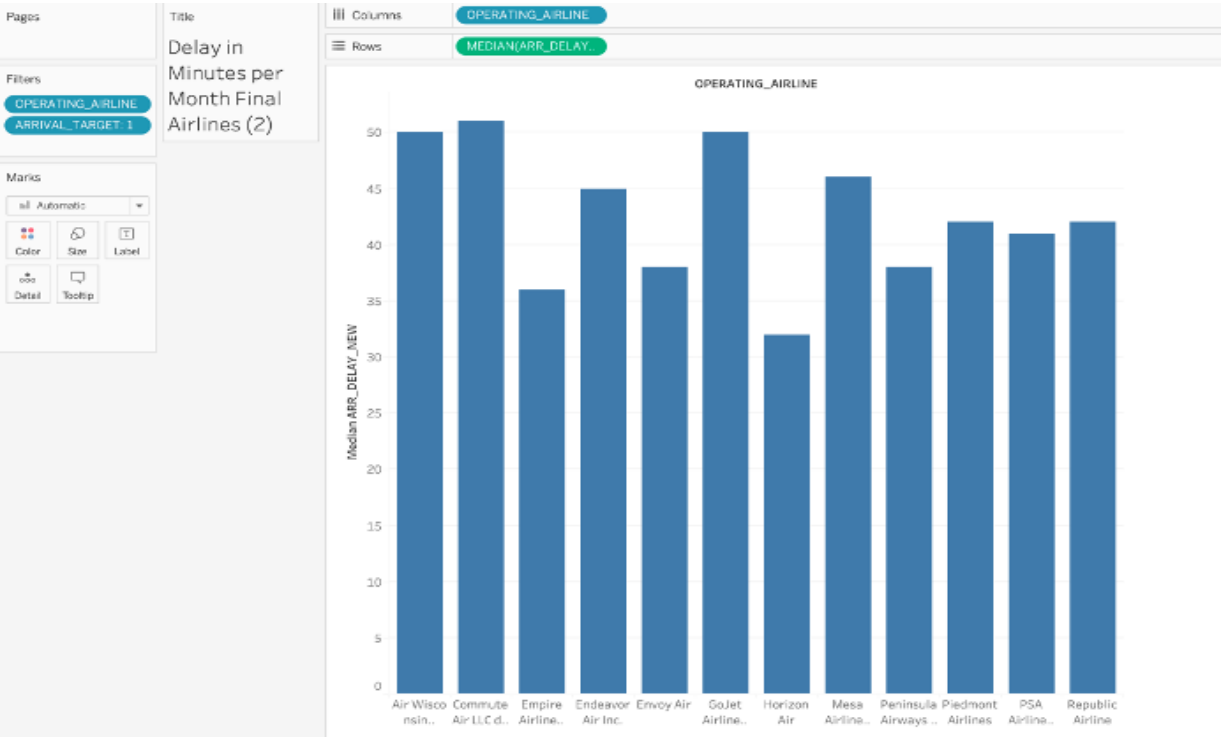
These findings highlight the significant financial impact of flight delays on both GoJet Airlines and CommuteAir. Therefore, it is recommended that Freya Systems consider targeting interventions and solutions specifically for these two airlines to reduce delays and associated costs. Such interventions could include operational improvements, better scheduling, and possibly predictive maintenance to address the recurring issues leading to delays. By implementing effective strategies to minimize delays, these airlines have the potential to achieve substantial cost savings and improve overall performance.

References

- **https://www.zippia.com/gojet-airlines-careers-1478688/revenue/#**
- https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=
- **https://www.linkedin.com/company/gojet-airlines/people/**
- **https://www.zippia.com/champlain-enterprises-careers-18793/revenue/**
- **https://www.linkedin.com/company/commuteair/people/**
- **https://www.flightaware.com**
- **https://www.linkedin.com/company/empireairlinesinc/**
- **https://www.macrotrends.net/stocks/charts/MESA/mesa-air/revenue**
- **https://investor.mesa-air.com/static-files/5782de63-b0d6-4573-a711-61ff08fbd3a3**
- **https://www.linkedin.com/company/mesa-airlines/people/**
- **https://www.zippia.com/air-wisconsin-airlines-careers-13933/revenue/#**
- **hhttps://www.linkedin.com/company/air-wisconsin-airlines/people/**
- **https://growjo.com/company/Envoy_Air**
- **https://www.linkedin.com/company/envoyair/**
- **https://growjo.com/company/Republic_Airways**
- **https://www.linkedin.com/company/republic-airways/**
- **https://growjo.com/company/Republic_Airways**
- **https://www.linkedin.com/company/republic-airways/**
- **https://growjo.com/company/Horizon_Air**
- **https://www.linkedin.com/company/horizon-air/**
- **https://growjo.com/company/Endeavor_Air**
- **https://www.linkedin.com/company/endeavor-air/people/**

Appendix

MONTHLY MINUTES DELAY(MEDIAN) FOR DELAYED FLIGHTS



PENINSULA AIRWAYS & EMPIRE AIRLINES

PENINSULA AIRWAYS

According to data from FlightAware, **Ravn Alaska** had an average of 2,705 flights per month in the USA in 2022. The airline's busiest month was June, with an average of 3,100 flights. The least busy month was January, with an average of 2,300 flights.

The airline's most popular route in 2022 was **Anchorage to Fairbanks**, with an average of 1,200 flights per month. Other popular routes included Anchorage to Juneau, Anchorage to Ketchikan, and Fairbanks to Bethel.

Ravn Alaska's flights are typically operated by Saab 340B and ATR 42-500 aircraft. The airline also operates a small fleet of Boeing 737-400F freighters.

240 Employees

EMPIRE AIRLINES

Empire Airlines ceasing operations in 2020 is because the airline was briefly placed under Chapter 11 bankruptcy protection in March of that year. However, the airline emerged from bankruptcy just a few months later and has been operating normally ever since

Airline has continued to operate and even expand in recent years. In December 2021, Empire Airlines acquired fellow **FedEx feeder** West Air, expanding its footprint to include all of California.

209 Employees

EMPIRE AIRLINES – DATA – FLIGHTAWARE.COM

Month	Total Flights
Jan-22	14,982
Feb-22	13,973
Mar-22	15,264
Apr-22	16,355
May-22	17,446
Jun-22	18,537
Jul-22	19,628
Aug-22	20,720
Sep-22	21,811
Oct-22	22,902
Nov-22	23,993
Dec-22	25,084

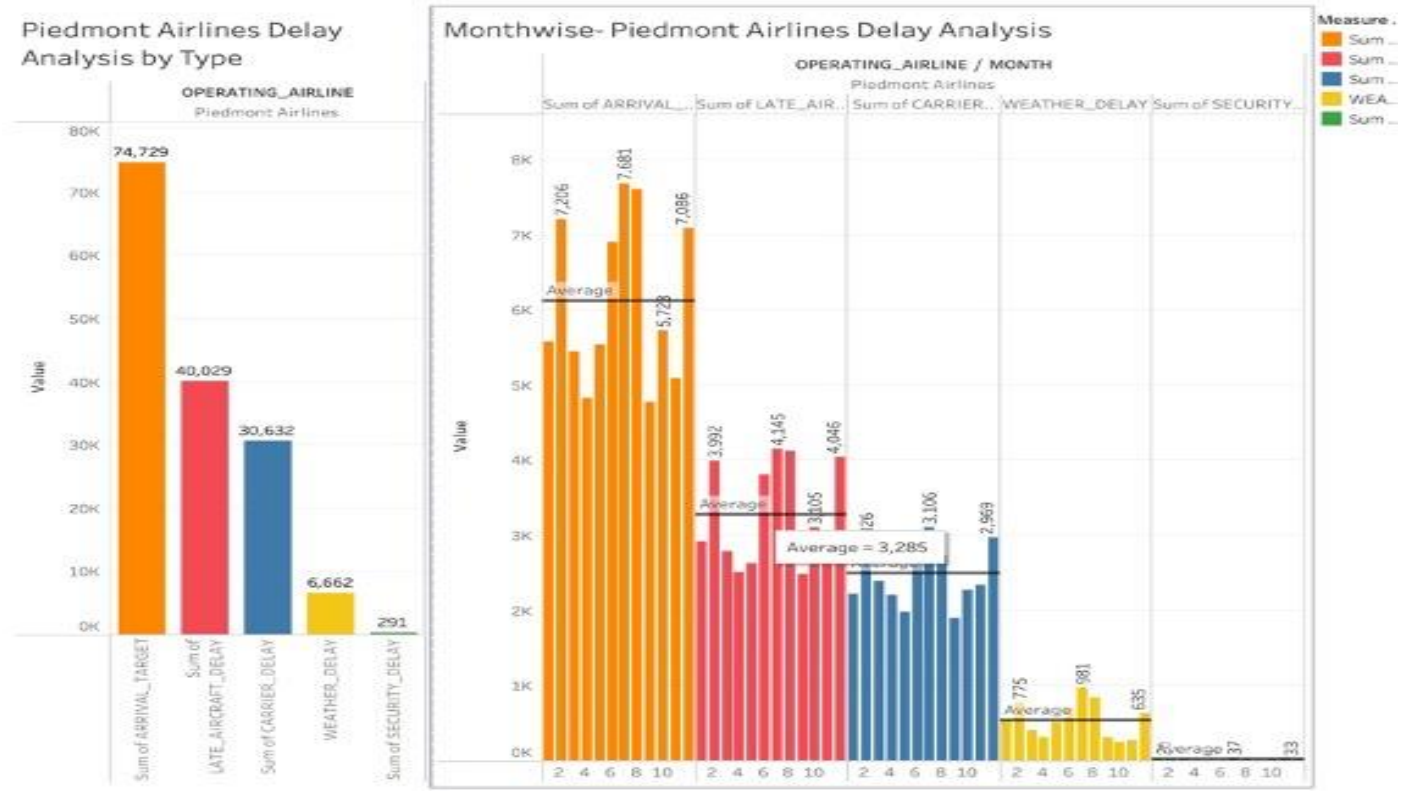
In total, there were 13,434 delayed flights for Empire Airlines that were more than 15 minutes in 2022 and 2023. The average delay time for these flights was 27 minutes.

The most common reason for delays for Empire Airlines is weather. Other reasons for delays include air traffic control delays, mechanical problems, and airline staffing issues.

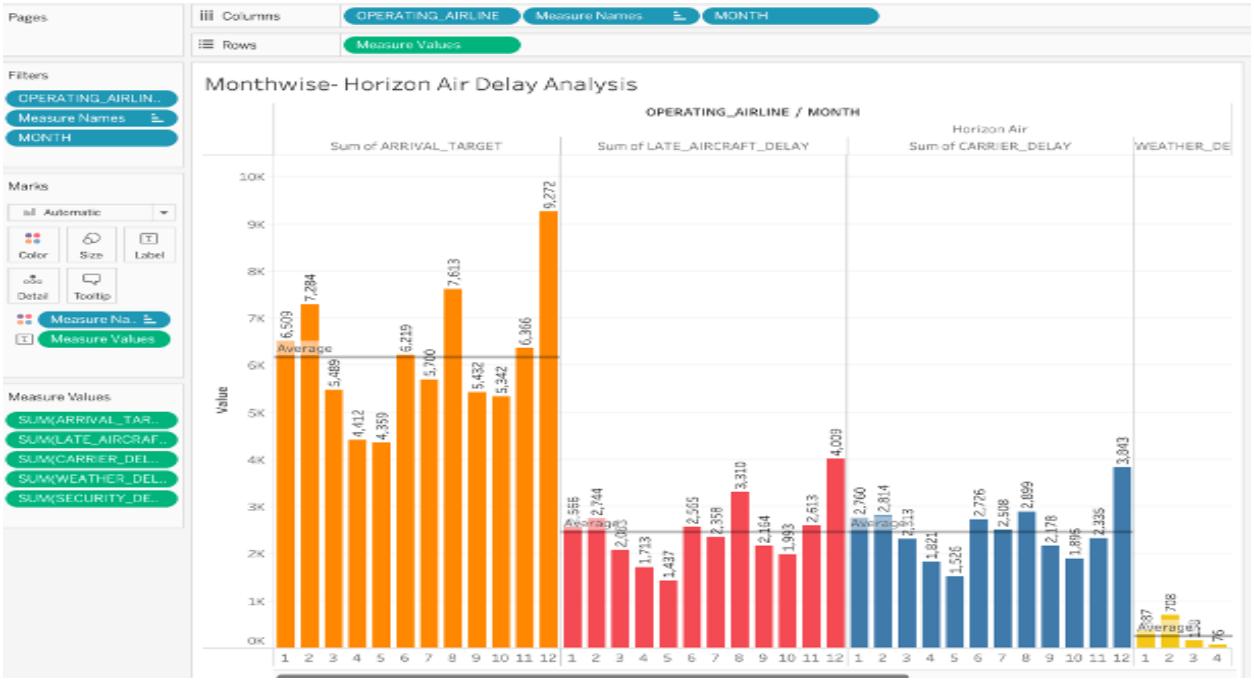
OTHER AIRLINES DETAILS

COMPANY	MESA AIRLINES	AIR WISCONSIN	ENVOY AIRLINES	REPUBLIC AIRWAYS	PIEDMONT AIRWAYS	HORIZON AIR	ENDEAVOR AIR
REVENUE	$ 531 M	$ 530 M	$2.2 Billion	$ 3 Billion	$ 3 Billion	$ 3 Billion	$ 924 Million
No. of Employees	2181	1213	5442	3502	3043	1898	2488
Team of Data Analyst	Good team	Good team	Good team	Big Team	Big Team	Good team	Good team
Specific Details	Phoenix-based Mesa Airlines operates as American Eagle from hubs in Phoenix and Dallas/Fort Worth and as United Express from Washington Dulles and Houston.	New partner American Airlines as American Eagle. In early 2023, AWA will be transitioning from previous partner, United Airlines.				Wholly-owned subsidiary of the Alaska Air Group	

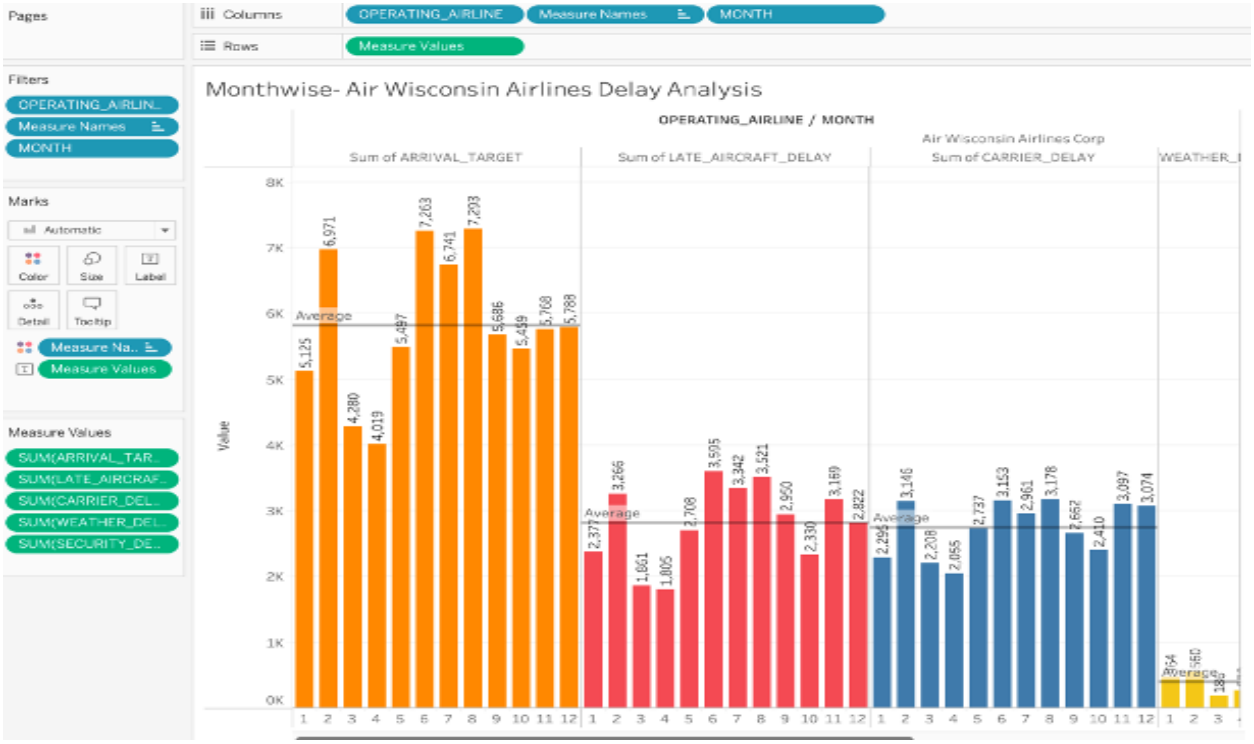
PIEDMONT AIRLINES DELAY ANALYSIS



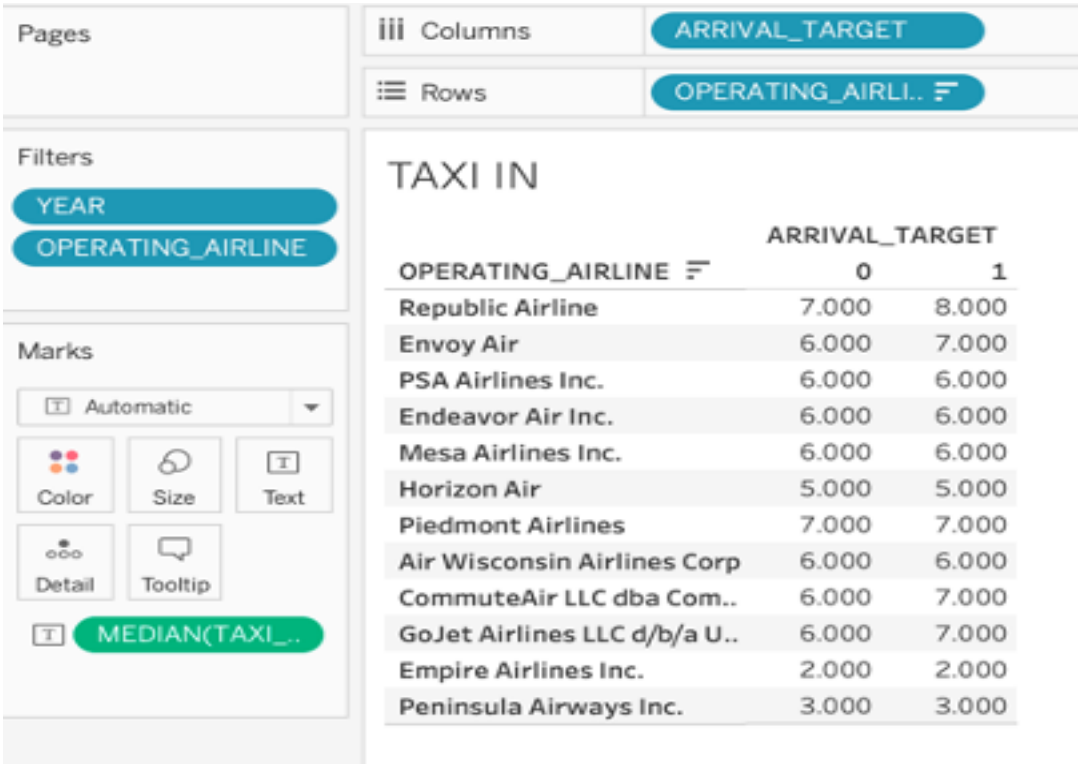
HORIZON AIR DELAY ANALYSIS



AIR WISCONSIN ALINES DELAY ANALYSIS

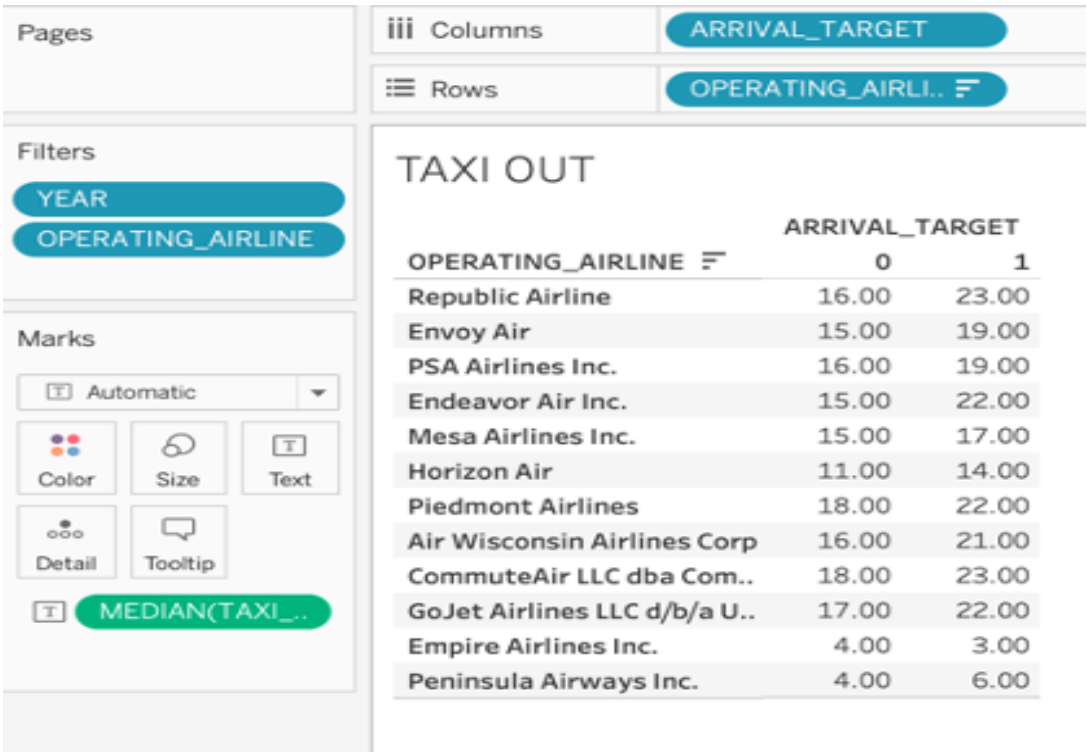


RELATIONSHIP



TAXI IN

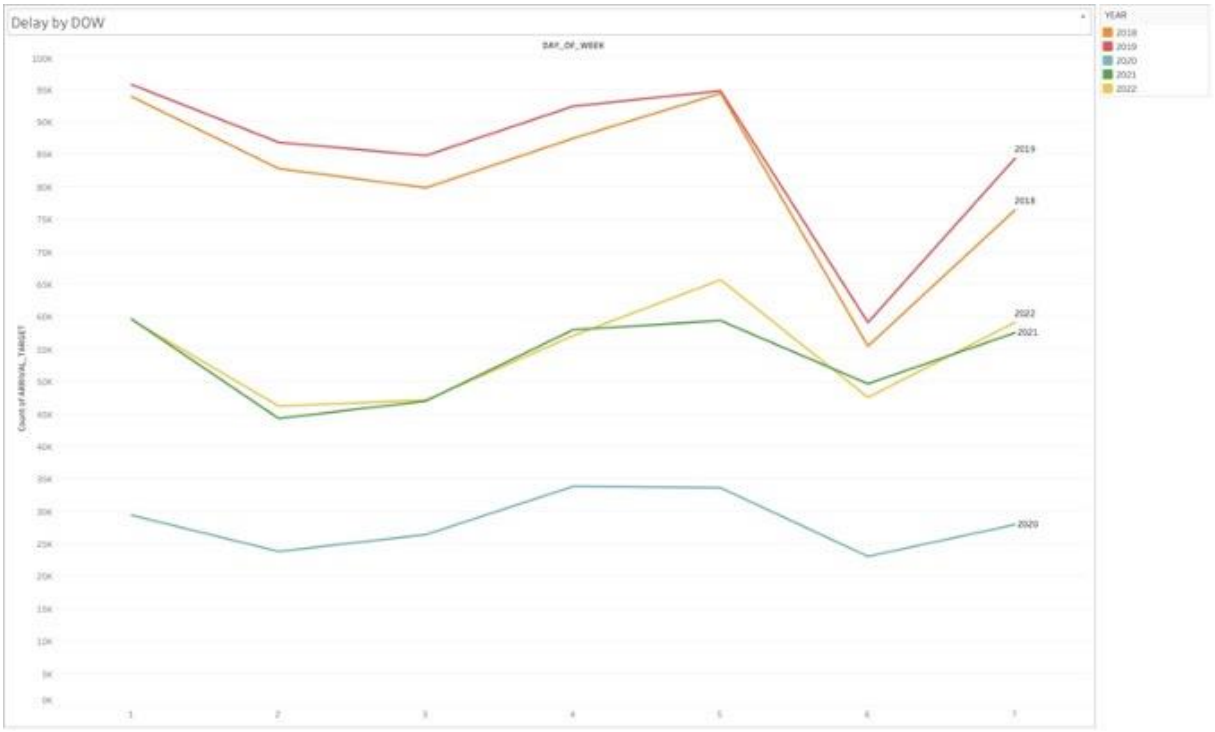
	ARRIVAL_TARGET	
OPERATING_AIRLINE	0	1
Republic Airline	7.000	8.000
Envoy Air	6.000	7.000
PSA Airlines Inc.	6.000	6.000
Endeavor Air Inc.	6.000	6.000
Mesa Airlines Inc.	6.000	6.000
Horizon Air	5.000	5.000
Piedmont Airlines	7.000	7.000
Air Wisconsin Airlines Corp	6.000	6.000
CommuteAir LLC dba Com..	6.000	7.000
GoJet Airlines LLC d/b/a U..	6.000	7.000
Empire Airlines Inc.	2.000	2.000
Peninsula Airways Inc.	3.000	3.000



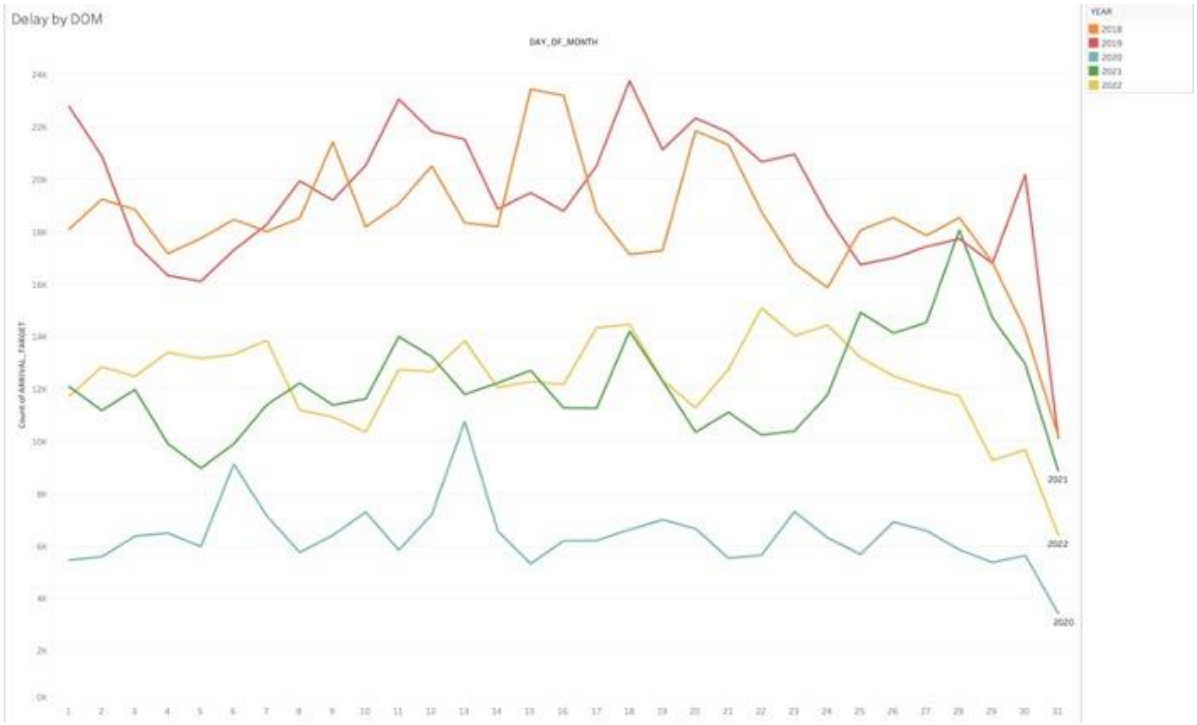
TAXI OUT

	ARRIVAL_TARGET	
OPERATING_AIRLINE	0	1
Republic Airline	16.00	23.00
Envoy Air	15.00	19.00
PSA Airlines Inc.	16.00	19.00
Endeavor Air Inc.	15.00	22.00
Mesa Airlines Inc.	15.00	17.00
Horizon Air	11.00	14.00
Piedmont Airlines	18.00	22.00
Air Wisconsin Airlines Corp	16.00	21.00
CommuteAir LLC dba Com..	18.00	23.00
GoJet Airlines LLC d/b/a U..	17.00	22.00
Empire Airlines Inc.	4.00	3.00
Peninsula Airways Inc.	4.00	6.00

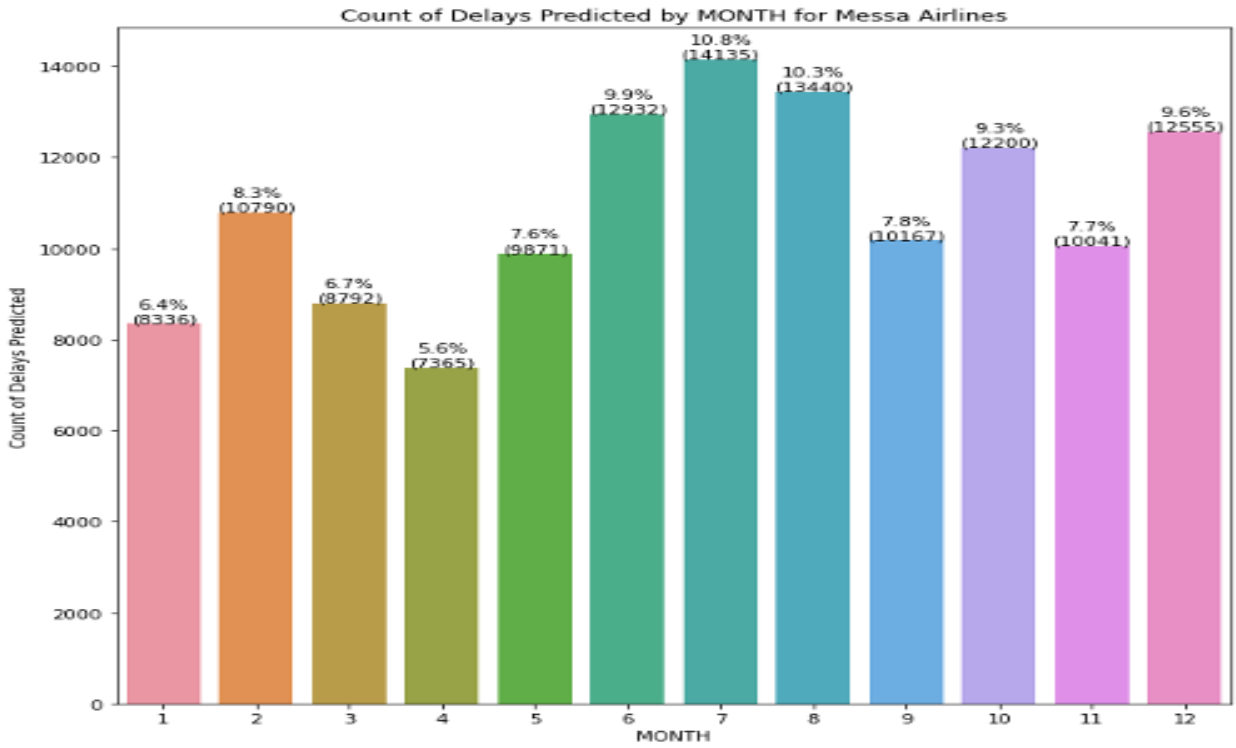
Number of Arrival per Day of the WEEK



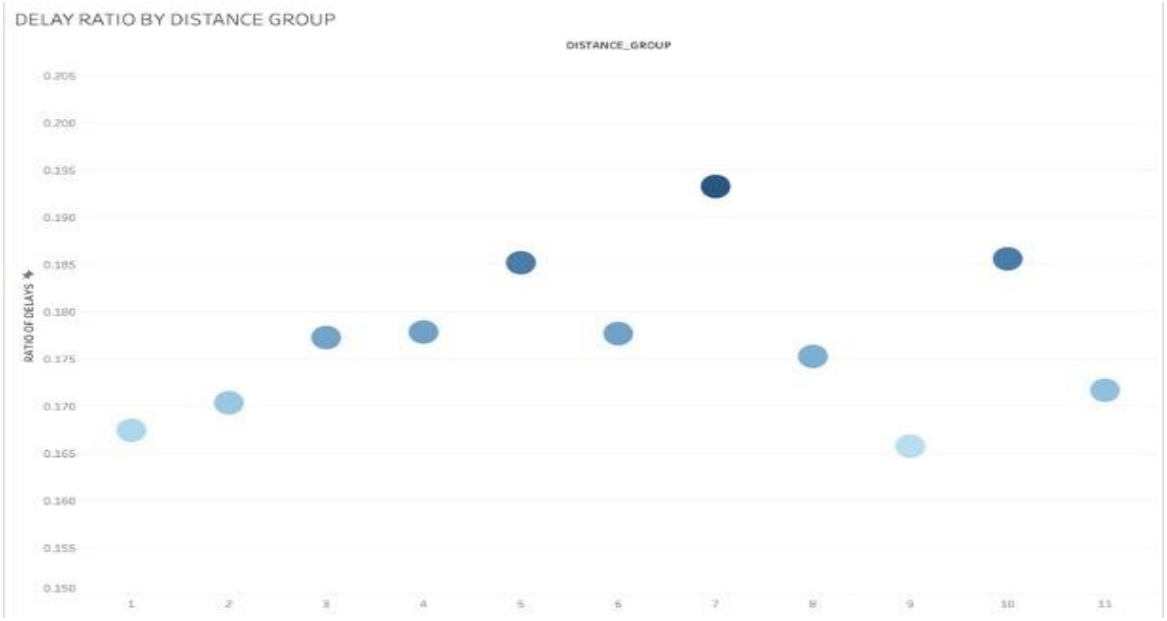
Number of Arrival Delays per day of the month



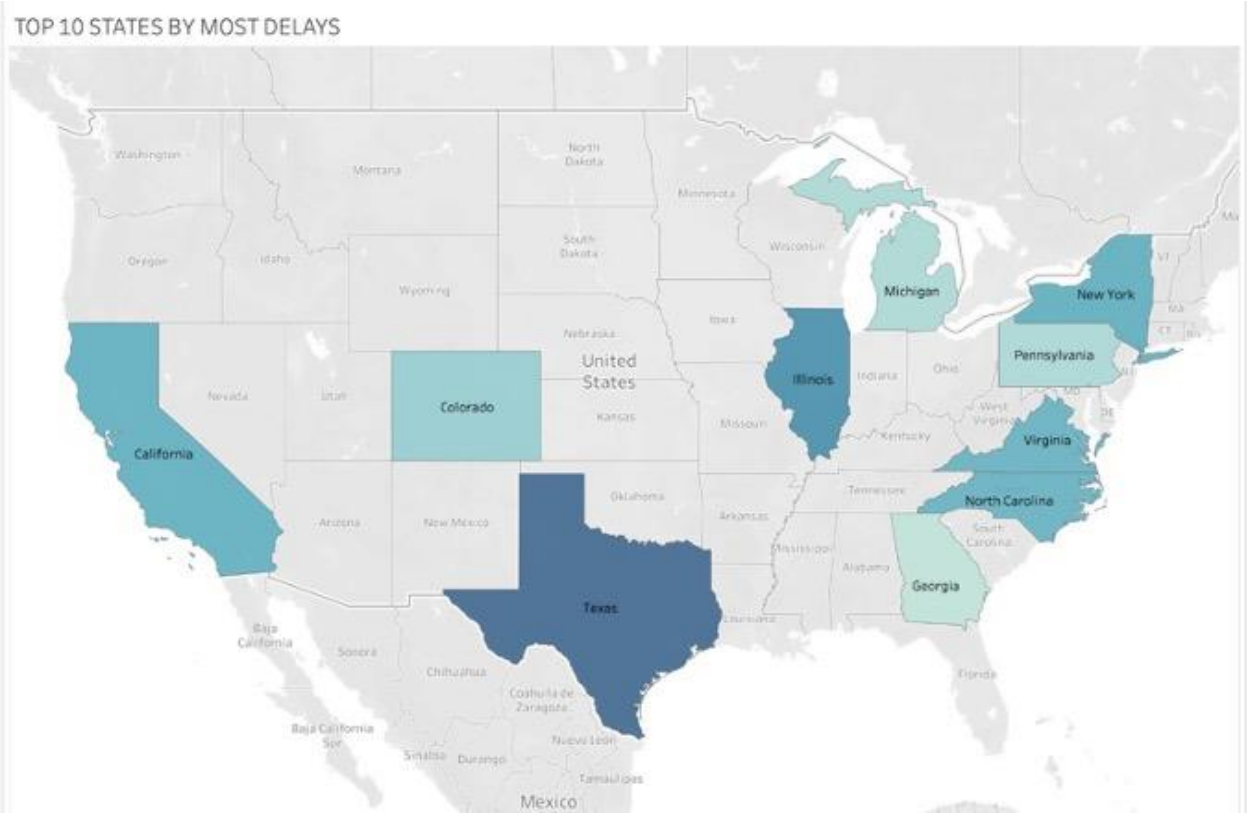
MONTHLY DELAY PREDICTIONS (RANDOM FOREST MODEL) - MESA AIRLINES



DELAY RATI0 BY DISTANCE GROUP



GEOGRAPHY - TOP 10 STATES



DATA CLEANING & TRANSFORMATION

- REMOVED ALL THE NULL VALUES FROM FINAL DATAFRAME
- Converted all float(numeric) values into integer.
- Converted Categorical into dummy variables.

```
# Iterate through columns
for col in final_df.columns:
    if final_df[col].dtype == 'float64':
        final_df[col] = final_df[col].dropna().astype('int64')
```

```
In [24]: categorical_columns = []
         for i in final_df.columns:
             if final_df[i].dtype == "object" and i not in ["MARKETING_AIRLINE","SCHEDULED_DEP_TIME", "SCHEDULED_ARR_TIME", '
                 categorical_columns.append(i)

In [25]: categorical_columns
Out[25]: ['OPERATING_AIRLINE', 'ORIGIN_CITY_NAME', 'DEST_CITY_NAME']

In [26]: categorical_df = final_df[categorical_columns]

In [27]: categorical_df_encoded = pd.get_dummies(categorical_df)

In [28]: categorical_df_encoded.shape
Out[28]: (7905957, 679)
```

```
YEAR                  0
MONTH                 0
DAY_OF_MONTH          0
DAY_OF_WEEK           0
MARKETING_AIRLINE     0
OPERATING_AIRLINE     0
ORIGIN_CITY_NAME      0
ORIGIN_STATE_NM       0
DEST_CITY_NAME        0
DEST_STATE_NM         0
SCHEDULED_DEP_TIME    0
DEP_DELAY_NEW         0
DEP_DEL15             0
TAXI_OUT              0
TAXI_IN               0
SCHEDULED_ARR_TIME    0
ARR_DELAY_NEW         0
ARRIVAL_TARGET        0
AIR_TIME              0
DISTANCE              0
DISTANCE_GROUP        0
CARRIER_DELAY         0
WEATHER_DELAY         0
NAS_DELAY             0
SECURITY_DELAY        0
LATE_AIRCRAFT_DELAY   0
DEP_HOUR              0
ARR_HOUR              0
dtype: int64
```

Create COVID Column

```
# Define a function to determine the 'covid effect'
def create_covid_effect(year):
    if year in [2020, 2021, 2022]:
        return 1
    else:
        return 0

# Create the 'covid effect' column using the apply function
final_df['COVID_EFFECT'] = final_df['YEAR'].apply(create_covid_effect)
```

Modeling

Attached herewith are the links to the codes for all the different models we attempted to implement on the data from various small carriers."

Gojet_Airlines_LLC.html

Commute Airlines-2.html

Air_Wisconsin_Airlines.html

Empire_Airlines.html

Peninsula_Airways.html

Piedmont_Airlines.html