Predicting Flight Delays

Data-Driven Insights for Operational Efficiency

Created By: Anusha Bhat, Kirthi Rao, Manas Vemuri, and Sumasree Ragi

Agenda

1. **Business Problem**
   Defining the challenge and project goals

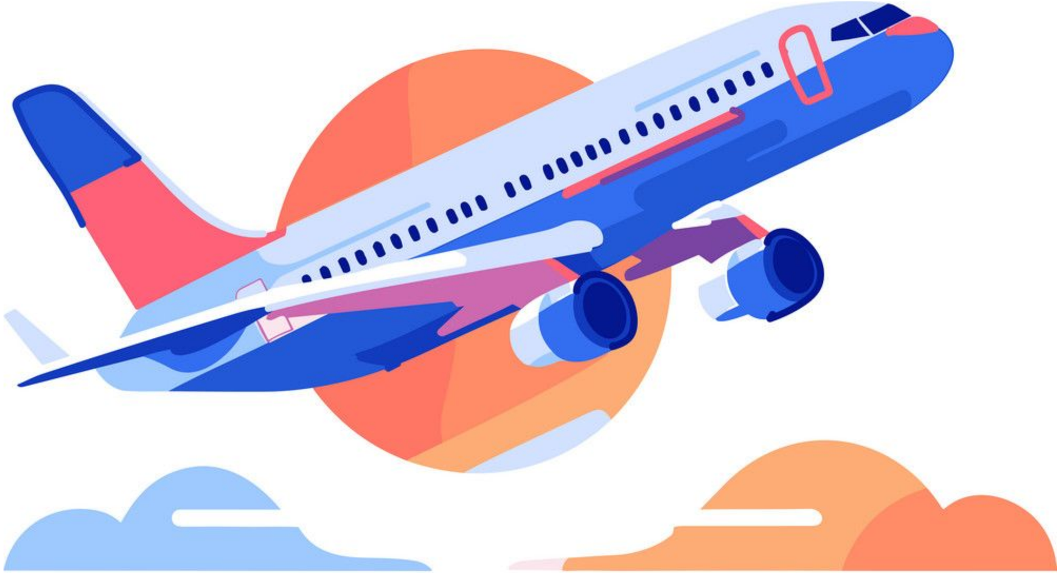
2. **Data Cleaning & Preparation**
   Ensuring data quality for reliable analysis

3. **EDA**
   Uncovering data quality and patterns in the data

4. **Model Development**
   Building a Binomial Logistic Regression model to predict delays

5. **Model Evaluation**
   Assessing model performance and accuracy

Business Problem



- Flight delays disrupt operations and customer satisfaction, requiring airlines to manage schedules proactively.
- Using **binomial regression,** we aim to **predict the probability of a flight delay** based on factors like departure time, distance, and origin-destination pairs.
- This model will enable data-driven decisions to **reduce delays and improve airline reliability.**

Data Summary

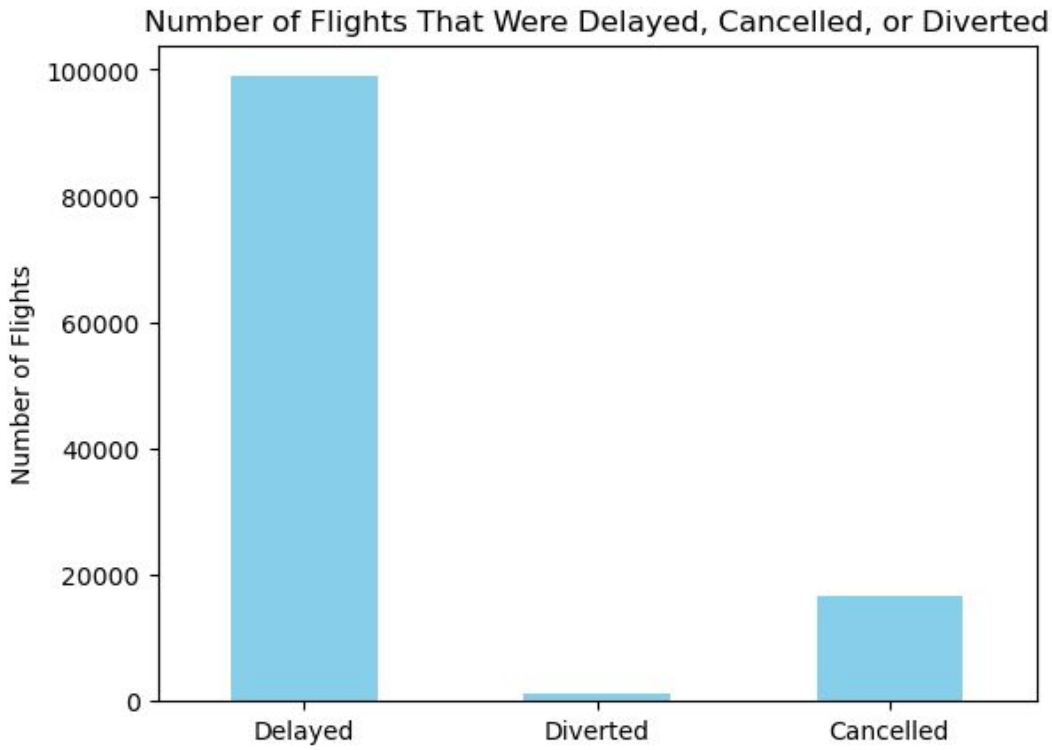
- Dataset containing **time and travel information of all domestic US flights in January 2019**
  - Total of 583,985 flights
  - 346 airports
- Collected by Bureau of Transportation Statistics, sourced from Kaggle
- Flights during all times of day, from high, medium, and low volume airports, and both budget and non-budget airlines
- 21 original features, 8 retained
- **Departure delay is our target column**
  - 17.42% of flights were delayed

Data Cleaning

- Originally had 21 columns on day of month, day of week, different carrier codes, arrival and departure airports, tail number, flight number, departure and arrival time, diverted, cancelled, and distance
- Only retained 8 columns
  - **Independent variables: Day of month, day of week, carrier type, departure and arrival airport type, departure time block, and distance**
  - **Dependent Variable: Delay.**
- Reclassified carrier type from IATA codes to the categories budget v.s. non-budget
- Reclassified departure time block from hourly intervals to time of day categories
  - Pre-dawn, early morning, morning, noon, afternoon, evening, night
- Created a new columns for arrival and departure airport type
  - Originally had airport codes
  - New columns have 3 categories: high, medium, and low volume flight traffic

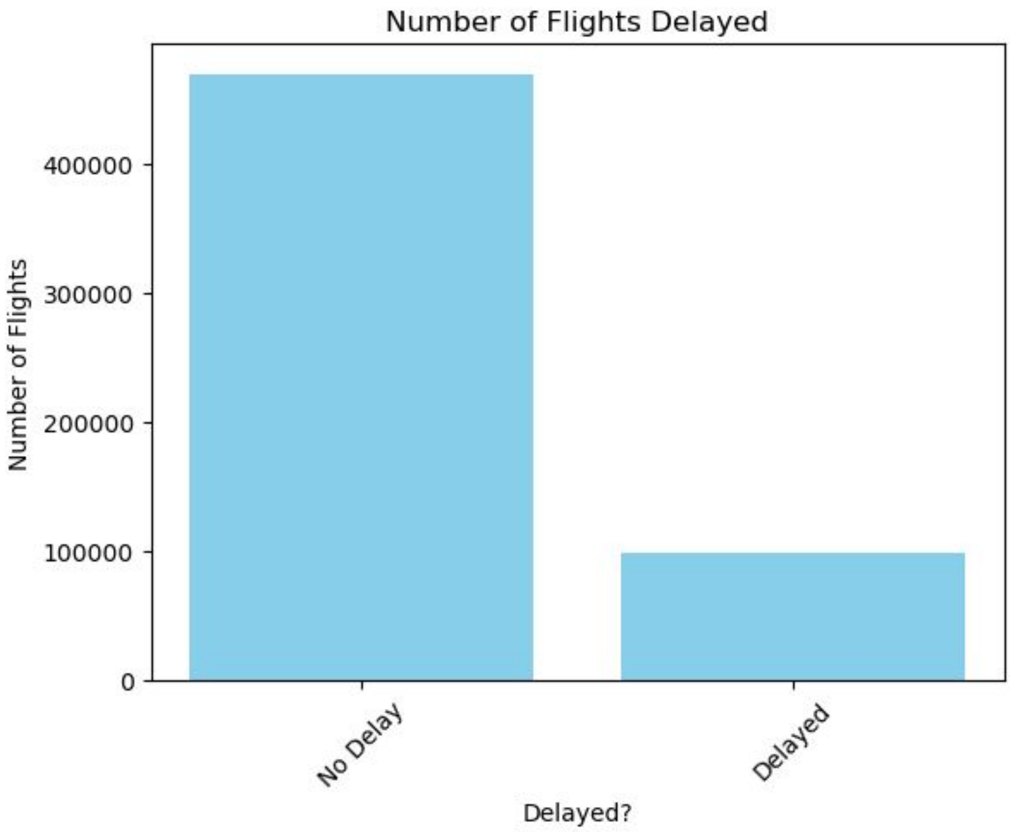
More flights were delayed compared to the number of flights that were cancelled or diverted

- Due to the discrepancy in count, we only used departure delay as the dependent variable.



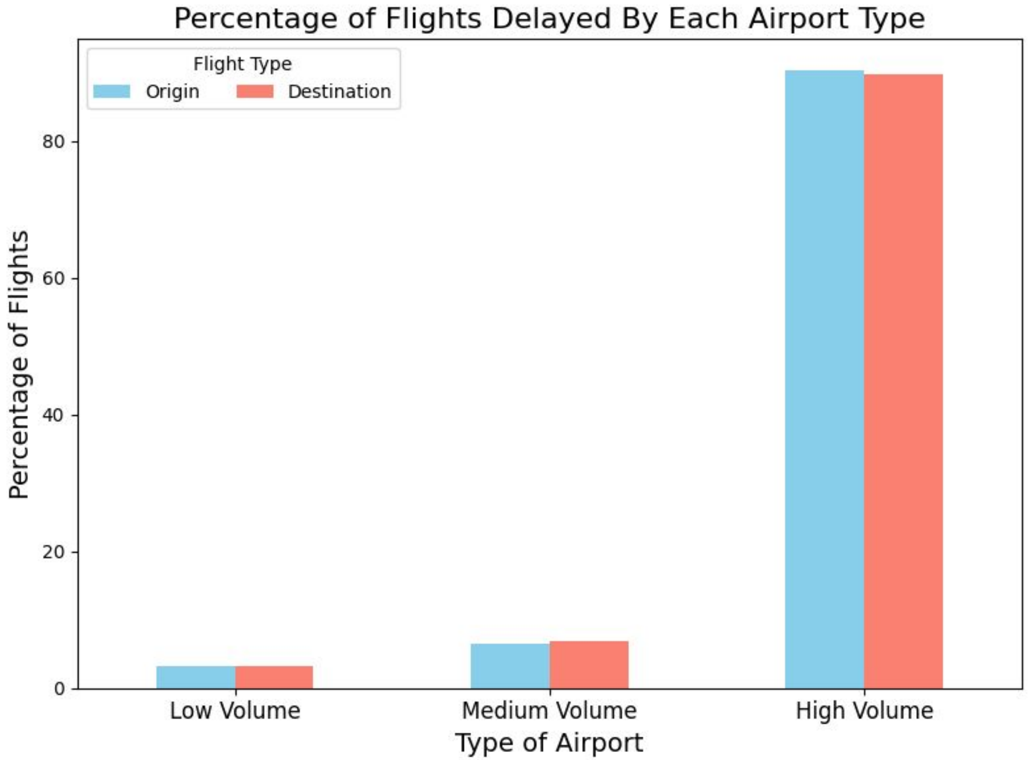
Most Flights Experienced No Delay

- Indicates potential need to undersample for a balanced model.

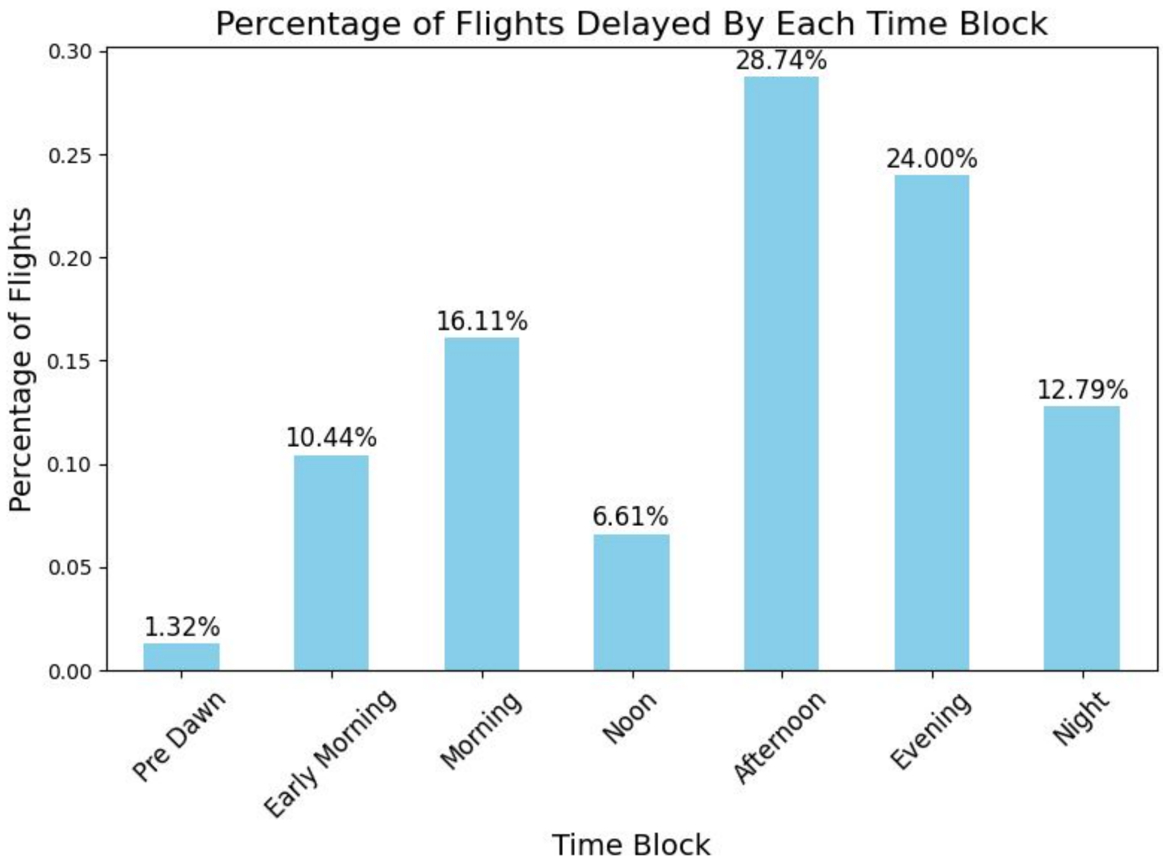


Most of the Delayed Flights Were From High-Volume Airports

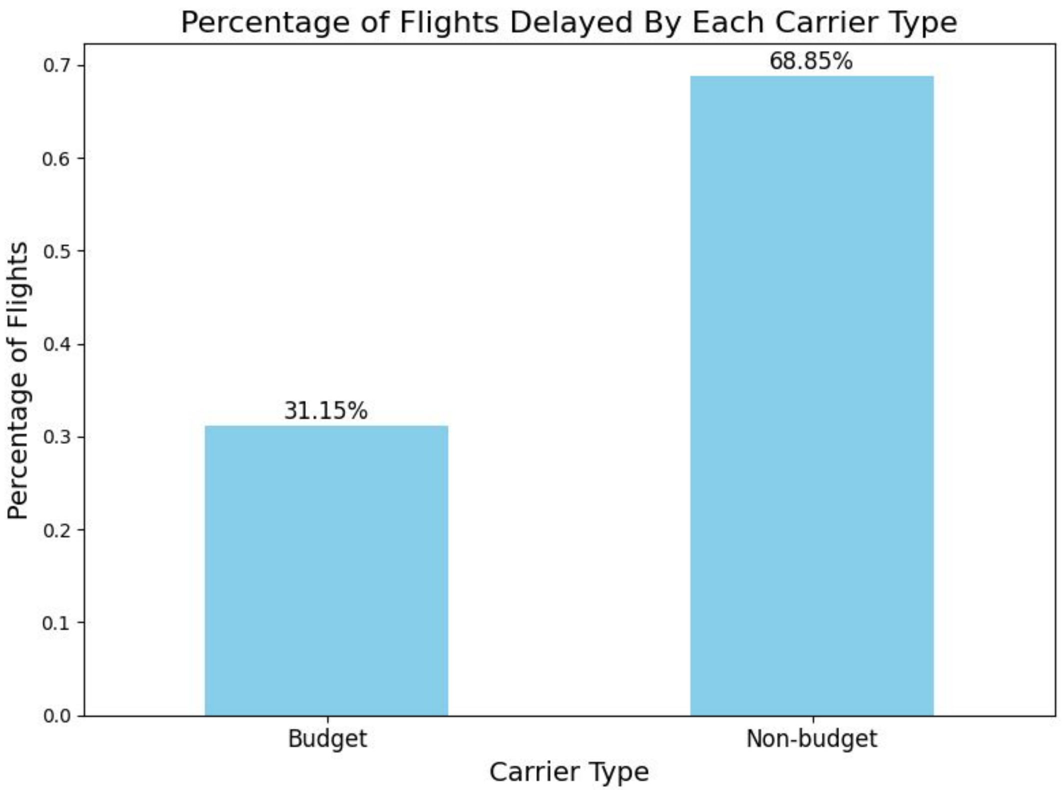
	Origin	Destination
Low volume	3.20%	3.30%
Medium volume	6.43%	6.959%
High volume	90.37%	89.752%



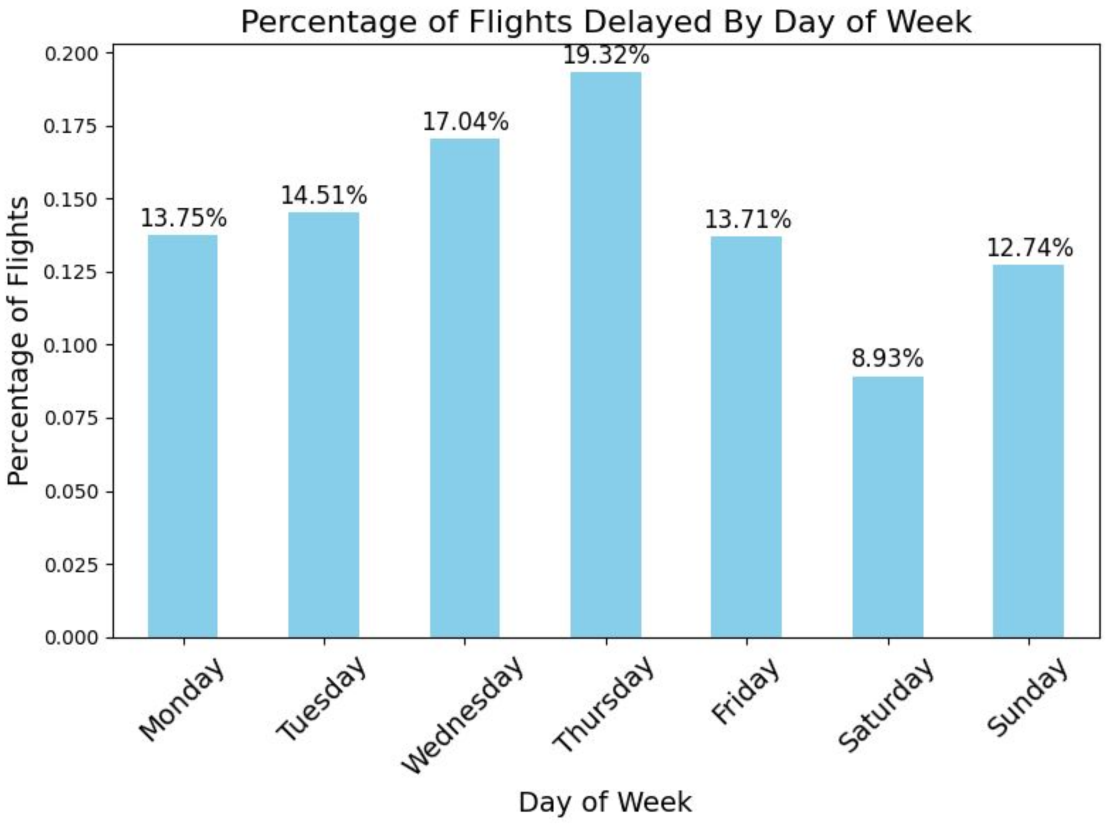
Most Delayed Flights are Later in the Day, after 12:00 pm



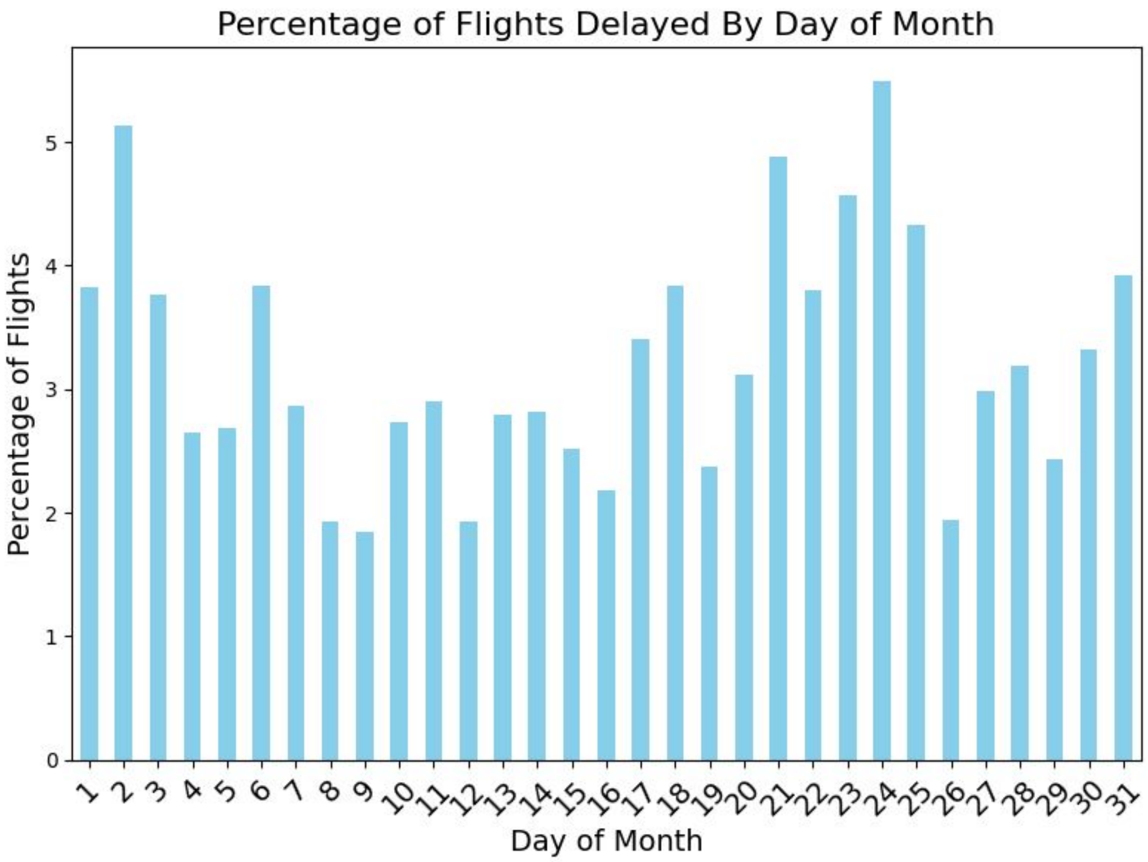
Most Delayed Flights Were From Non-Budget Airlines (Delta, United Airlines, American Airlines, and Subsidiaries)



Thursday Has the Highest Percentage of Delayed Flights, while Saturday has the Lowest Percentage



The Beginning of the Month and Days 20-25 Have Higher Proportions of Delayed Flights



Model Choice

Model	Notes
Linear Regression	● Works for continuous outcome variables and for our case we are working with a categorical outcome ● Would Work if: predicting how long a flight was delayed for
Count Regression	● Since we are not attempting to count the occurrences of something, this model was not a good fit ● Would Work if: predicting number of delays
Multinomial Regression	● Since we are only predicting two categories this model was not used ● Would Work if: predicting more outcomes like cancelled and diverted flights
Binomial Logistic Regression	● Optimal choice since we are predicting two outcomes: delay or no delay ● Using the logit link function we can output the percentage chance of a delay, which can allow for adjusting the model depending on user's risk tolerance

Logistic Regression - Logit Function

Logistic Regression is a form of Binomial Regression that uses the logit link function.

- The **logit function** maps the probability of an event to the log-odds

$$logit(p) = ln(\frac{p}{1-p})$$

- Allows for a linear model to be applied to a binary outcome

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Logistic Regression - Logistic Function

- **Logistic function**: inverse of logit function
- Used to **convert the log-odds back into probability values**
- This value tells you the probability of the response variable occurring
    - Threshold prediction of 1 if $p \geq 0.5$

$$\Rightarrow \quad p = \frac{1}{1 + e^{-logit(p)}}$$

Model Estimation Method

Maximum Likelihood Estimation (MLE) via Iteratively Reweighted Least Squares (IRLS)

- Iterative method that combines the Newton-Raphson method with weighted least squares to **find the maximum likelihood estimates** of the model parameters
- MLE is a statistical method aimed at finding the best-fit parameters for a model
- IRLS is a specific algorithm used to achieve this goal in the context of GLMs

Generalized Linear Model Regression Results

Dep. Variable:	DEP_DEL15	No. Observations:	397341
Model:	GLM	Df Residuals:	397327
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1.7909e+05
Date:	Fri, 06 Dec 2024	Deviance:	3.5818e+05
Time:	17:57:22	Pearson chi2:	3.98e+05
No. Iterations:	5	Pseudo R-squ. (CS):	0.02318
Covariance Type:	nonrobust		

$$l(\beta; y, x) = \sum_{i=1}^{N}\left[y_i log(p_i) + (n_i - y_i) log(1 - p_i) + log\binom{n_i}{y_i}\right]$$

where:

$$p_i = \frac{e^{x_i^T \cdot \beta}}{1 + e^{x_i^T \cdot \beta}}$$

Model Creation

Undersampling Delays	Run Model with All Combinations	Evaluate Metrics
The dataset contained only 17% delayed flights. To allow the model to predict more delays we undersampled the on-time flights by randomly sampling 60% of them	Created an intercept only model. Then using itertools we train the GLM with a logit link function for each combination of the available features	For each model we examine the BIC. Using this goodness of fit measure we are able to see what set of features perform the best

Intercept-Only Model

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              DEP_DEL15   No. Observations:               397341
Model:                            GLM   Df Residuals:                   397340
Model Family:                Binomial   Df Model:                            0
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -1.8375e+05
Date:                Fri, 06 Dec 2024   Deviance:                   3.6750e+05
Time:                        11:30:14   Pearson chi2:                 3.97e+05
No. Iterations:                     4   Pseudo R-squ. (CS):          3.331e-16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.5563      0.004   -372.062      0.000      -1.565      -1.548
==============================================================================
```

Interpretation:

- Intercept represents the log-odds of a flight being delayed **when all predictor variables are zero**
- Negative intercept suggests a **low probability of delay**

Final Model Construction Using BIC

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              DEP_DEL15   No. Observations:               397341
Model:                            GLM   Df Residuals:                   397327
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -1.7909e+05
Date:                Fri, 06 Dec 2024   Deviance:                   3.5818e+05
Time:                        17:57:22   Pearson chi2:                 3.98e+05
No. Iterations:                     5   Pseudo R-squ. (CS):            0.02318
Covariance Type:            nonrobust
====================================================================================================
                                       coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                               -1.4679      0.014   -101.611      0.000      -1.496      -1.440
DAY_OF_MONTH                         0.0090      0.000     19.079      0.000       0.008       0.010
DISTANCE                          8.229e-05   7.26e-06     11.330      0.000    6.81e-05    9.65e-05
OP_UNIQUE_CARRIER_Non-budget        -0.1462      0.009    -15.745      0.000      -0.164      -0.128
DEST_AIRPORT_TYPE_Medium Volume      0.0033      0.017      0.195      0.845      -0.030       0.037
DEST_AIRPORT_TYPE_Low Volume        -0.1036      0.024     -4.325      0.000      -0.150      -0.057
ORIGIN_AIRPORT_TYPE_Medium Volume    0.0976      0.018      5.577      0.000       0.063       0.132
ORIGIN_AIRPORT_TYPE_Low Volume       0.0704      0.024      2.901      0.004       0.023       0.118
DEP_TIME_BLK_Early Morning          -0.9758      0.015    -66.328      0.000      -1.005      -0.947
DEP_TIME_BLK_Evening                 0.1863      0.012     15.569      0.000       0.163       0.210
DEP_TIME_BLK_Morning                -0.3504      0.013    -26.899      0.000      -0.376      -0.325
DEP_TIME_BLK_Night                   0.0784      0.015      5.389      0.000       0.050       0.107
DEP_TIME_BLK_Noon                   -0.1431      0.018     -7.851      0.000      -0.179      -0.107
DEP_TIME_BLK_Pre Dawn               -1.0224      0.035    -28.812      0.000      -1.092      -0.953
====================================================================================================
```

Interpretation:

- As both day of month and distance increase, **delay odds slightly increase**
- Early morning, morning, noon, and pre-dawn departures have **lower delay odds** compared to afternoon departures
- Evening and night departures have **higher delay odds** compared to afternoon departures

Sample Predictions

Flight 1:

- **Origin:** Long Beach (LGB)
- **Destination:** Boston (BOS)
- **Airline:** JetBlue
- **Date:** Thursday, 01/31/2020
- **Time:** 8:33 PM
- **Distance:** 2,602 miles

Predicted Delay Odds: **31.2%**

Flight 2:

- **Origin:** Lubbock, TX (LBB)
- **Destination:** Dallas, TX (DAL)
- **Airline:** Southwest
- **Date:** Wednesday, 01/02/2020
- **Time:** 5:27 AM
- **Distance:** 293 miles

Predicted Delay Odds: **6.3%**

Model Evaluation Metrics

Confusion Matrix

True Positives **87,846**	False Positives **52,724**
False Negatives **13,854**	True Negatives **15,865**

F1 Score: 32%

- Poor balance between precision and recall
- 32% in training data as well

Accuracy: 61%

- Our model correctly predicts delays 61% of the time
  - Using a threshold of 20% for predicting delays
- Accuracy in training data is also 61%

Future Improvements

01	Collecting More Possible Predictors	• Many common factors for flight delays are not included in this dataset. For example weather and the previous flight are important factors in predicting flight delay. • With adding this data, we would likely be able to better predict delays.
02	Getting more data on cancelled and diverted flights	• In the current dataset these types of flights made up less than 1% of the data. • As a result we could not create a meaningful multinomial model.
03	Optimizing data flow and processing to collect data over wider time range	• For storage and code runtimes we used just one month of data. • With data over the course of multiple years we could better account for seasonality.

Key Takeaways

Predictive Insights

➢ The final model effectively identifies key factors influencing flight delays, improving classification accuracy to 61%.

Significant Features

➢ Distance, day of month, departure times, airline types, and airport volume emerged as the most impactful predictors of delays.

Operational Insights

➢ Early morning and pre-dawn flights are the most reliable, while evening departures face higher risks of delays.

Business Implications

➢ Non-budget carriers and low-volume destination airports offer better on-time performance.

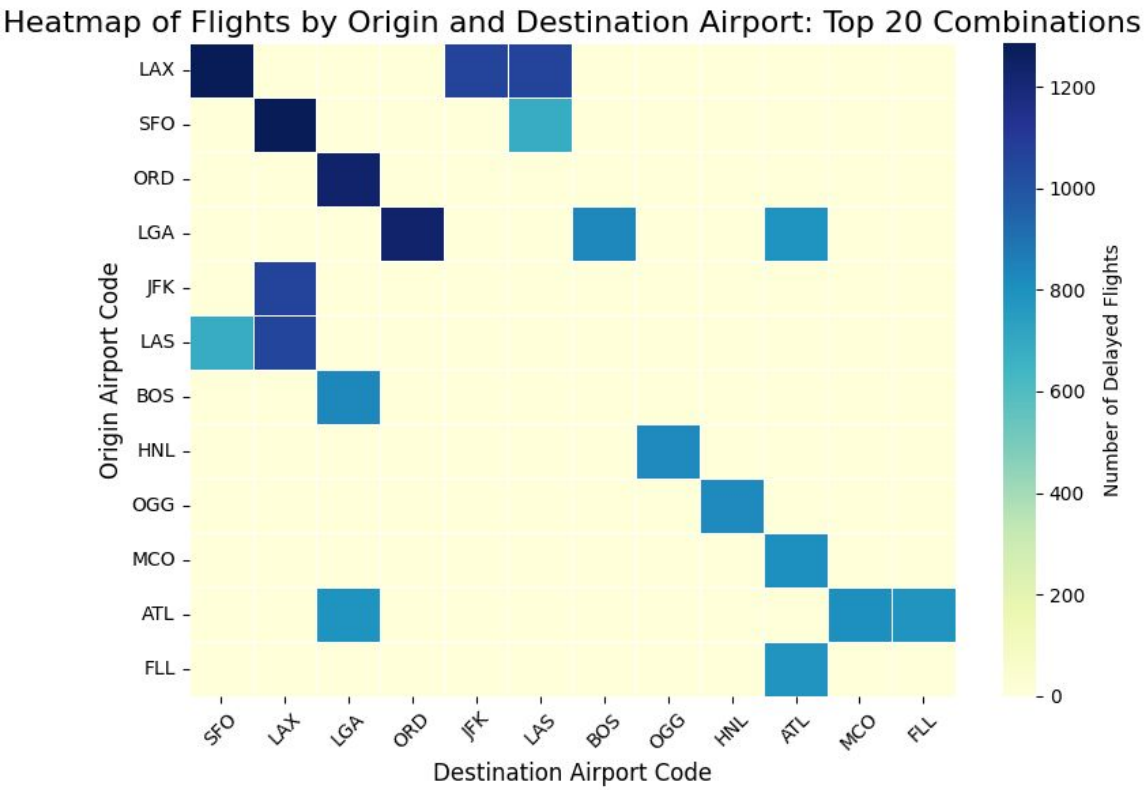
Thank you!

Appendix

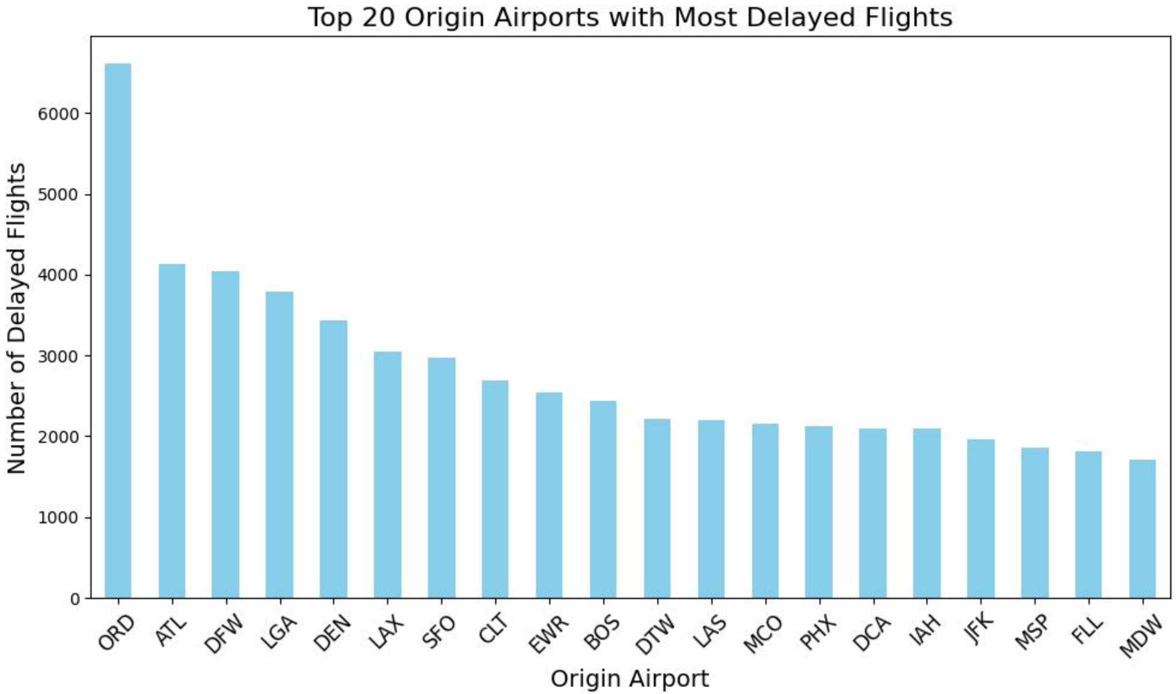
Link to Dataset

https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction

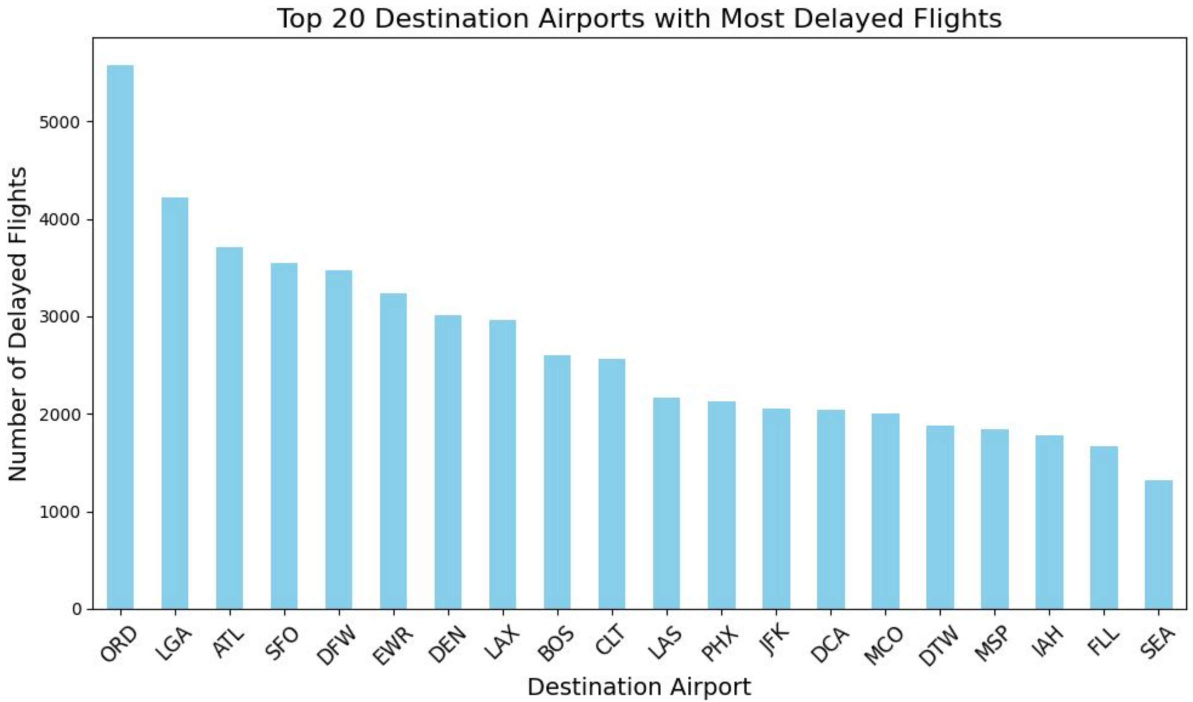
Combinations between LAX-SFO, ORD-LGA, JFK-LAX, and LAS-LAX have the highest number of flights to and from each airport.



ORD, ATL, DFW, LGA, and DEN Have the Most Delayed Flights For Origin Airports



ORD, LGA, ATL, SFO, and DFW Have the Most Delayed Flights For Destination Airports



Combinations between LAX-SFO, ORD-LGA, and BOS-LGA, have the highest number of delayed flights to and from each airport