Analyzing Flight Data

Using Machine Learning Models to Predict Flight Delays

Rebecca Patterson

September 8, 2021

Agenda

- Problem Statement
- Background
- Challenges
- Data
- Modeling
- Conclusions
- Future Research
- Streamlit App

Problem Statement

Use machine learning to predict if a flight will be delayed using publicly available Bureau of Transportation data

Background

- Flight delays are incredibly costly.
- It is estimated that in 2019, flight delays cost the airline industry 33 billion dollars in the US alone[1].
- They are also disruptive to passengers forcing schedule changes that can have ripple effects throughout the entire travel industry.

1 https://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf

Background

- Some airlines are rolling out programs to help passengers once they already know the flight is delayed.

- For example, in 2019 United rolled out 'Connection Saver' which allows them to hold certain planes if passengers are making a connection from a delayed flight.

- *"During the past four months, more than 14,400 customers, who would have otherwise missed their connections, were able to make their flights thanks to ConnectionSaver. Flights that were held for connecting customers were delayed an average of six minutes."*[1]

- But imagine if delayed flights could be predicted ahead of time?

1 https://hub.united.com/united--makes-connecting-easier-connectionsaver-2638762086.html

Challenges

- Data Size:
    - The bureau of transportation provided data from 2003 - Spring 2021
        - One year alone was ~ 6 million rows of data
        - Therefore, I had to subset down to a reasonable size which meant potentially losing additional predictive power
- Model Speed:
    - Because of the size of my data, the models would take hours to run, which forced me to choose and prioritize what/when to run

Data (Sourced from Bureau of Transportation Statistics[1])

Features	Target	Years
Month	**Flight Delay**	2018 - 2019
Day of the Week		
Airline		
Origin		
Destination		
Scheduled Departure Time		
Scheduled Arrival Time		
Scheduled Elapsed Time		
Distance		

1 https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ

Covid Considerations

- In 2020 there was a sharp decline in air travel. It is estimated that it fell around 60% compared to 2019 levels.[1]

- Therefore, I chose to focus on the years 2018 - 2019 for this project.

- Future research: I would like incorporate 2020 data with a flag indicating time to/from covid.

1 https://www.iata.org/en/pressroom/pr/2021-02-03-02/

Cleaning Steps

1. Each month was stored separately so I had to pull in each month and run it through a cleaning function:
    1. Remove unnamed columns
    2. Clean up column values (ex. remove city from state)
    3. Fix data types of columns (ex. turn strings to ints)
    4. Split the delay and cancel files (cancel rows had many null values)
    5. Drop appropriate nulls
    6. Concatenate the separate months into one data frame
2. My target was whether or not a flight was delayed so I had to create a binary column for this:
    1. I used 'arrival delay' to determine a flight was delayed (if the departure is delayed but the flight ends up landing on time this was not considered a delay)
    2. If the arrival delay minutes was greater than 0 - this was considered a ***delayed flight***

Additional Cleaning

My final data frame ended up with 14,340,049 rows

Additional cleaning steps:

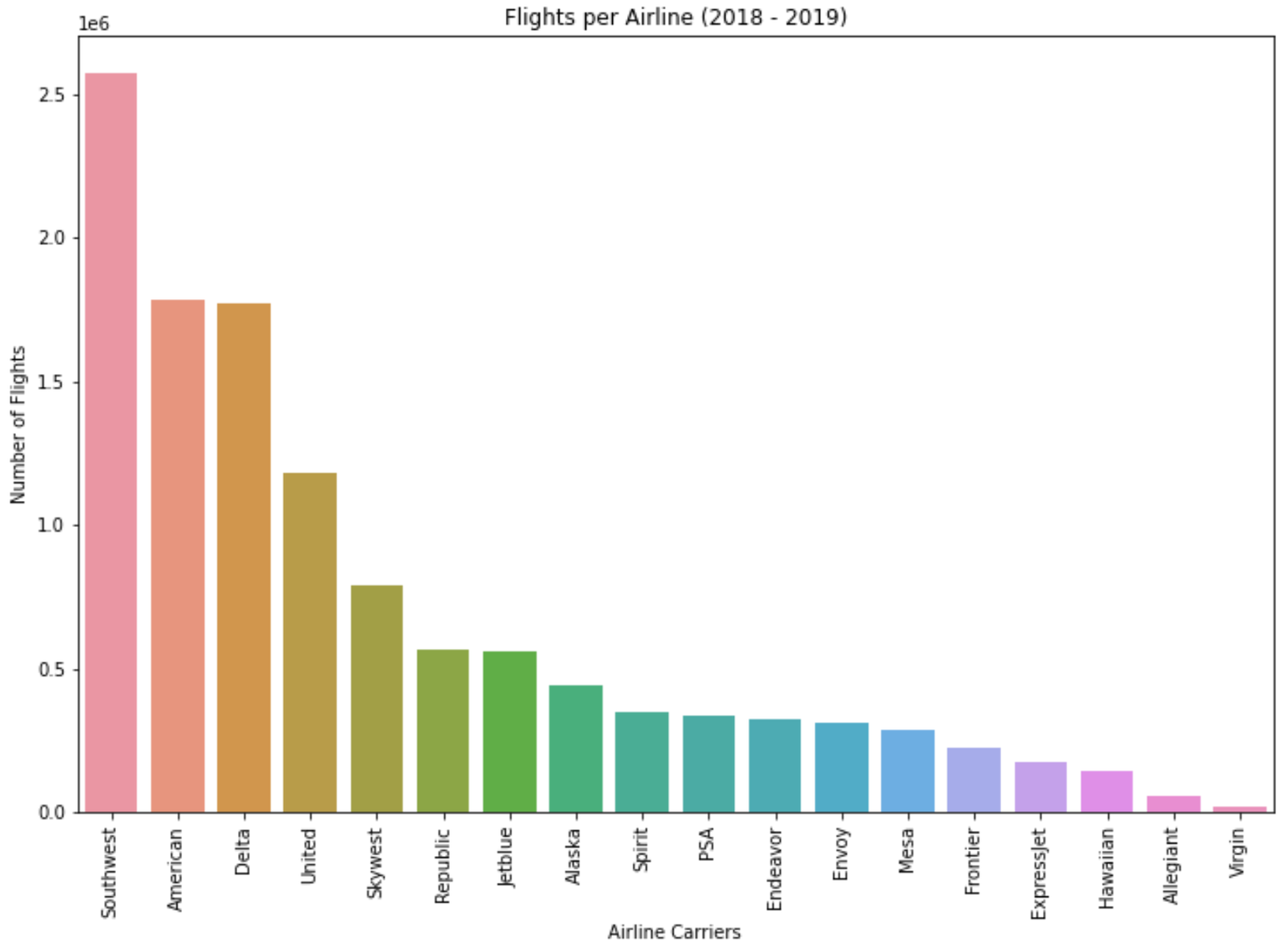
1. Rename airline carriers to be more human readable (e.g. DL → Delta)

2. Drop outliers (e.g. a flight from Jacksonville, FL to Newark, NJ taking over 26 hours)

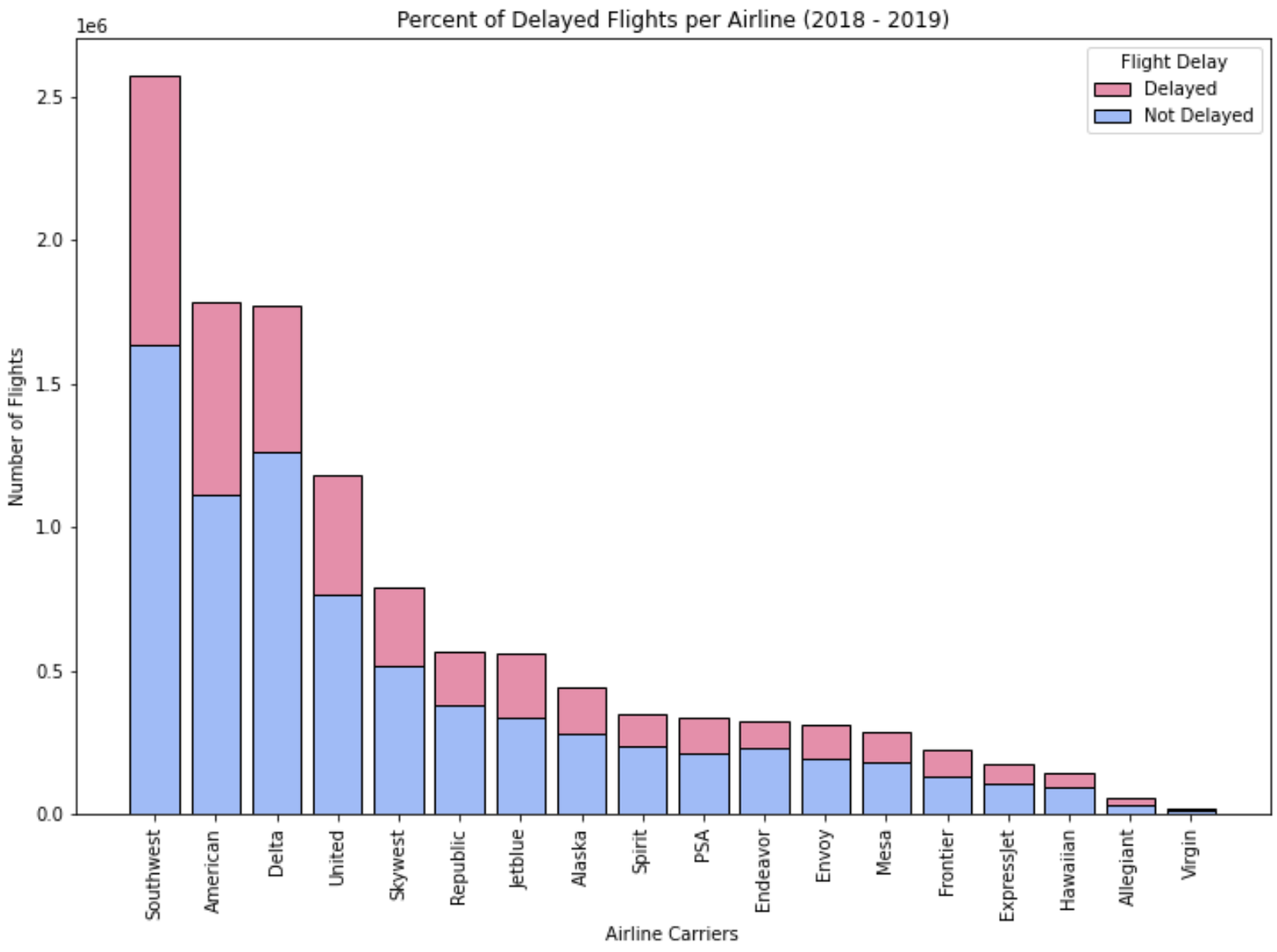
3. There were 363 total airports in the data set so I reduced it to the top 100

Final dataframe size: 11,887,768 rows

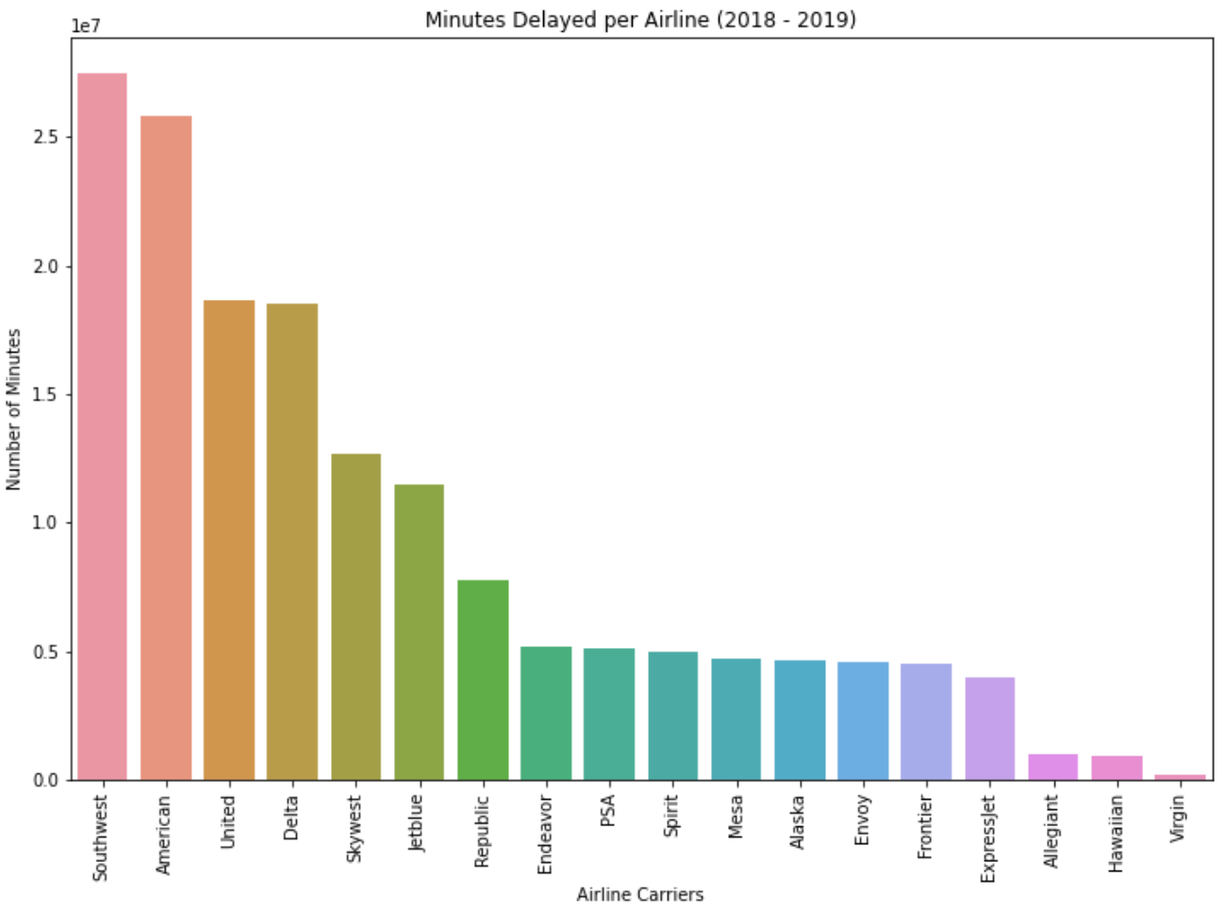
Flights per Airline

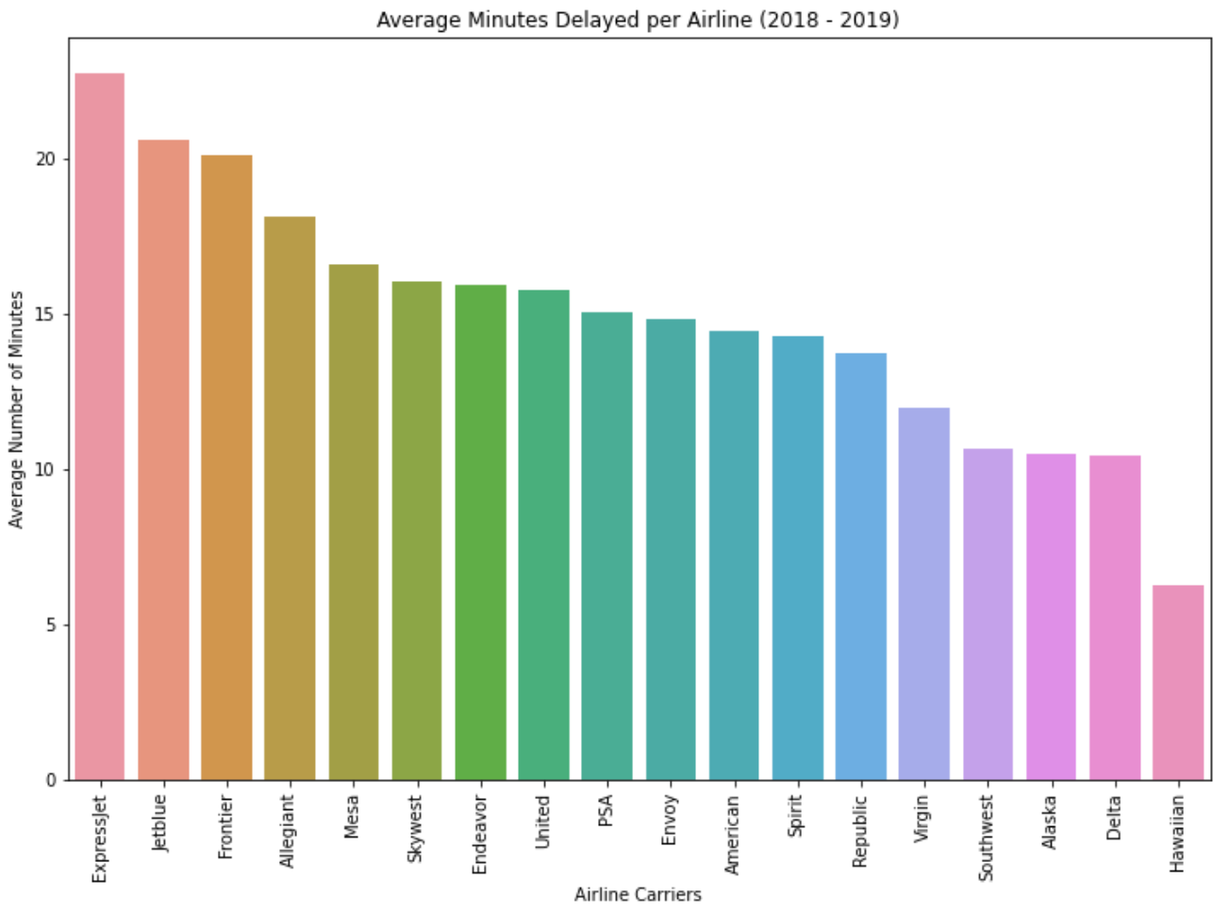


Delayed Flights per Airline



Minutes Delayed per Airline





Best and Worst Airlines to Fly

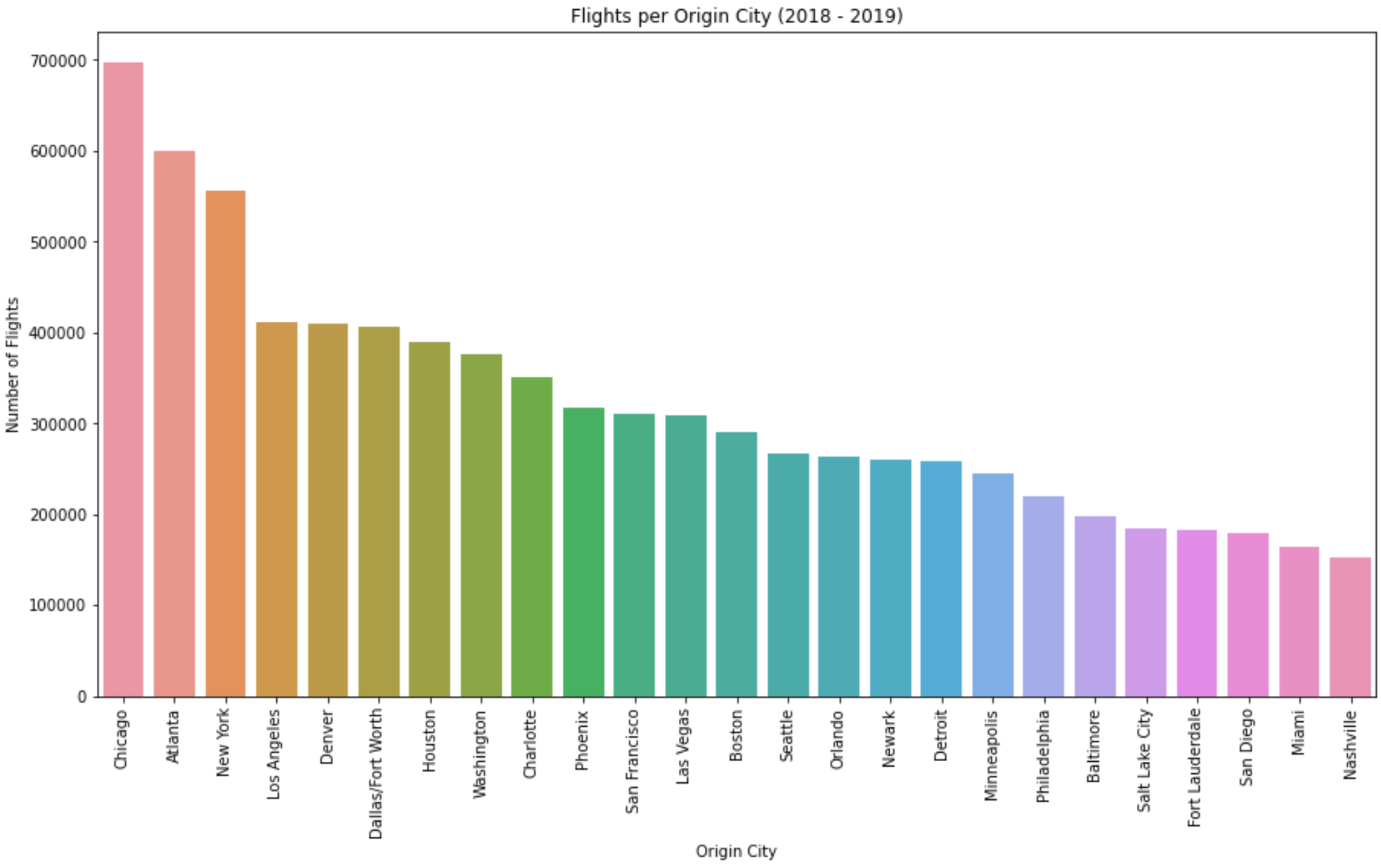
Overall Best Airline: **Delta**

- Lowest percent delayed: 29%
  - *Definition: number of flights delayed divided by total number of flights*
- Second lowest average minutes delayed: 10.45 mins
  - *Definition: total number of minutes divided by number of flights*
- Third highest number of flights flown: 1,770,079

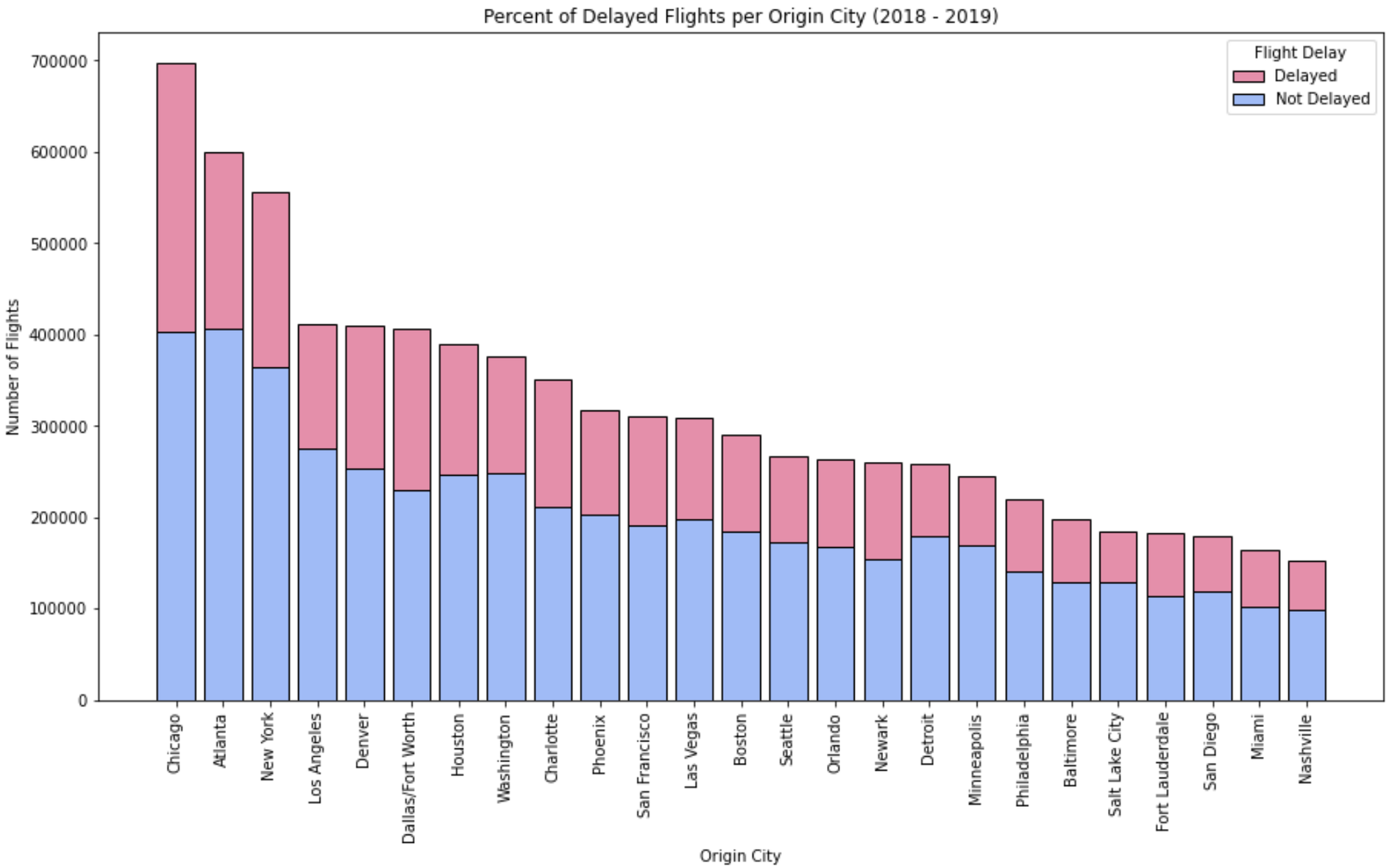
Overall Worst Airline: **JetBlue**

- Third highest percent delayed: 40%
  - *Definition: number of flights delayed divided by total number of flights*
- Second highest average minutes delayed: 20.57 mins
  - *Definition: total number of minutes divided by number of flights*
- Seventh highest number of flights flown: 559,251

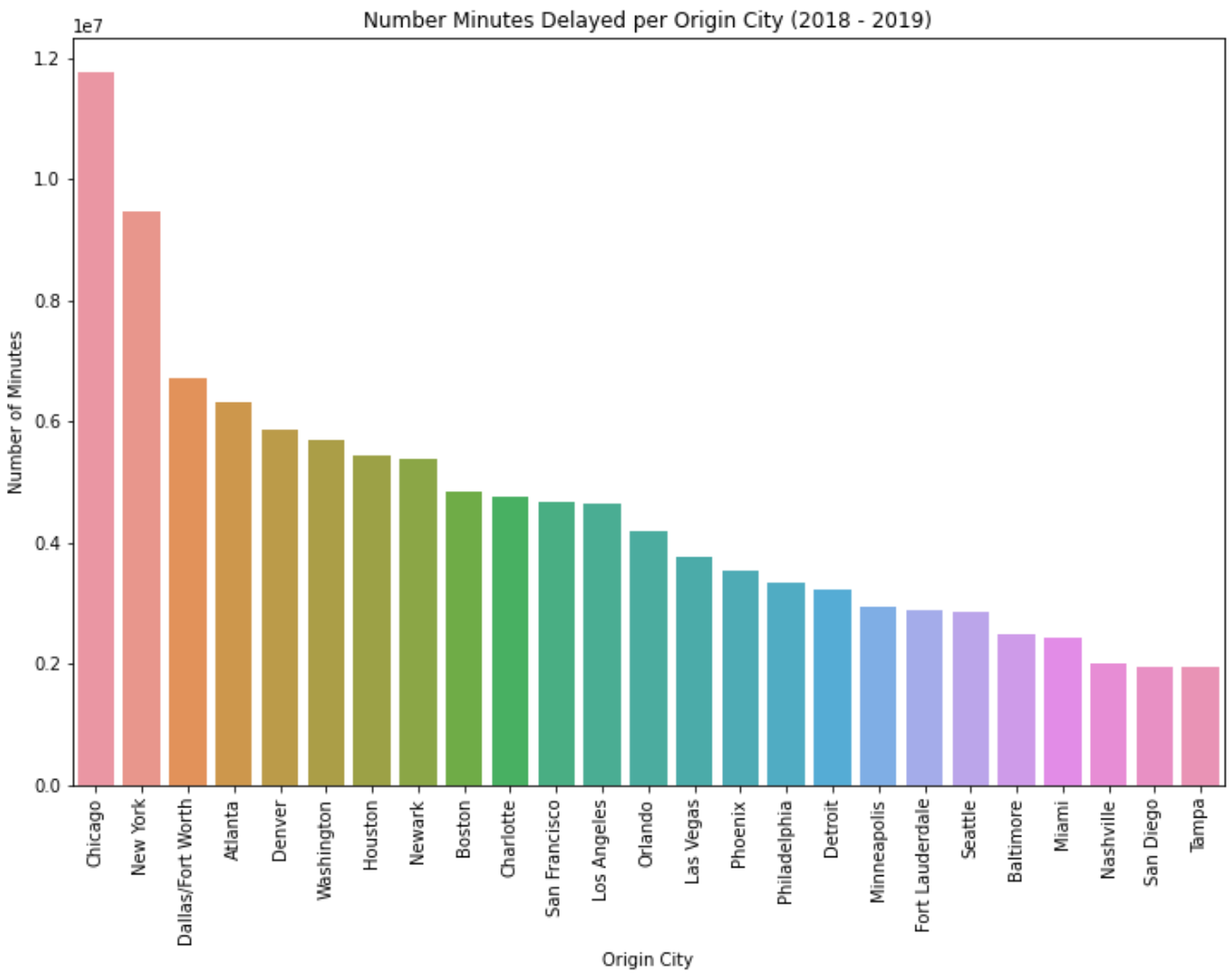
Flights per Origin City

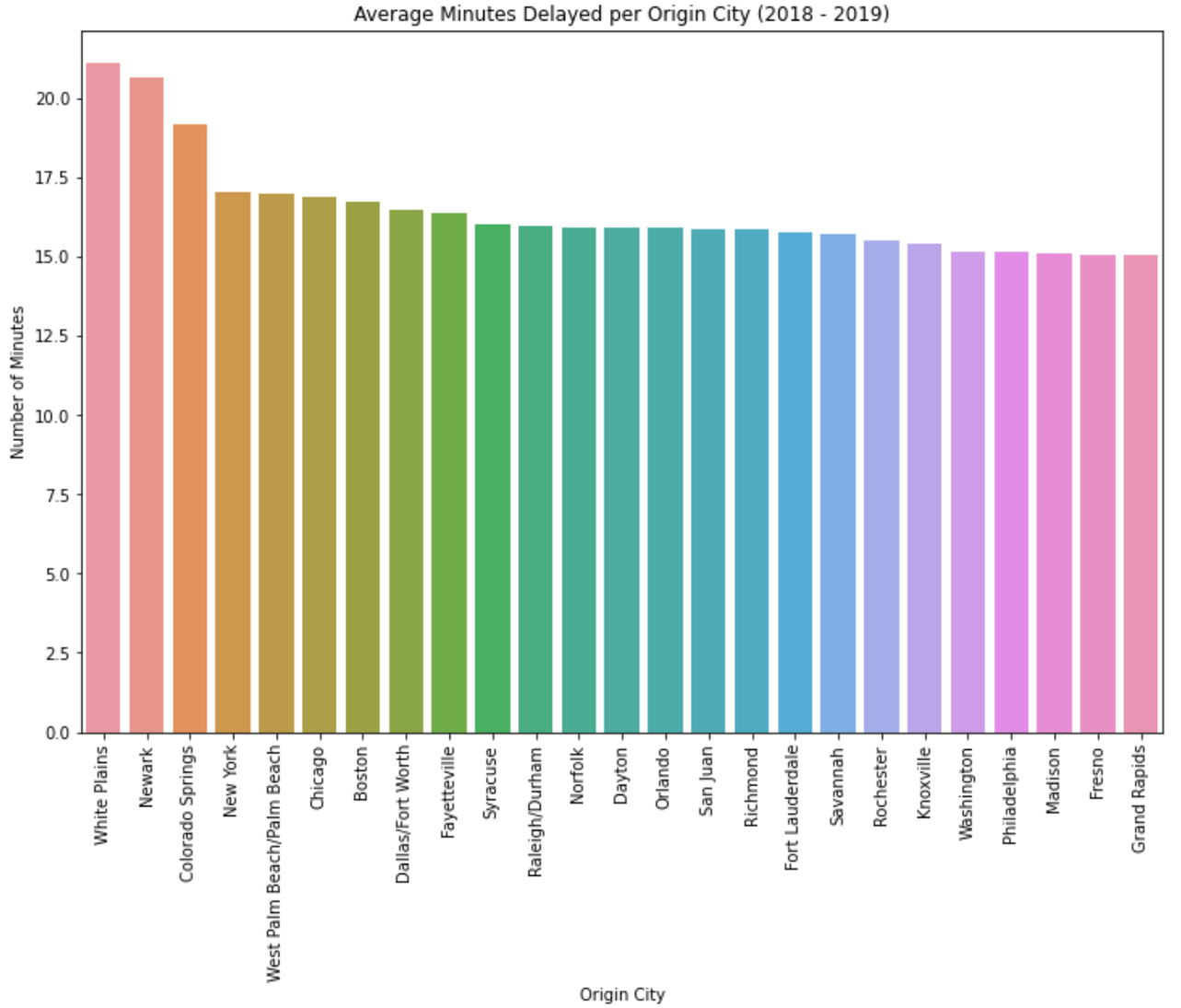


Delayed Flights per Origin City

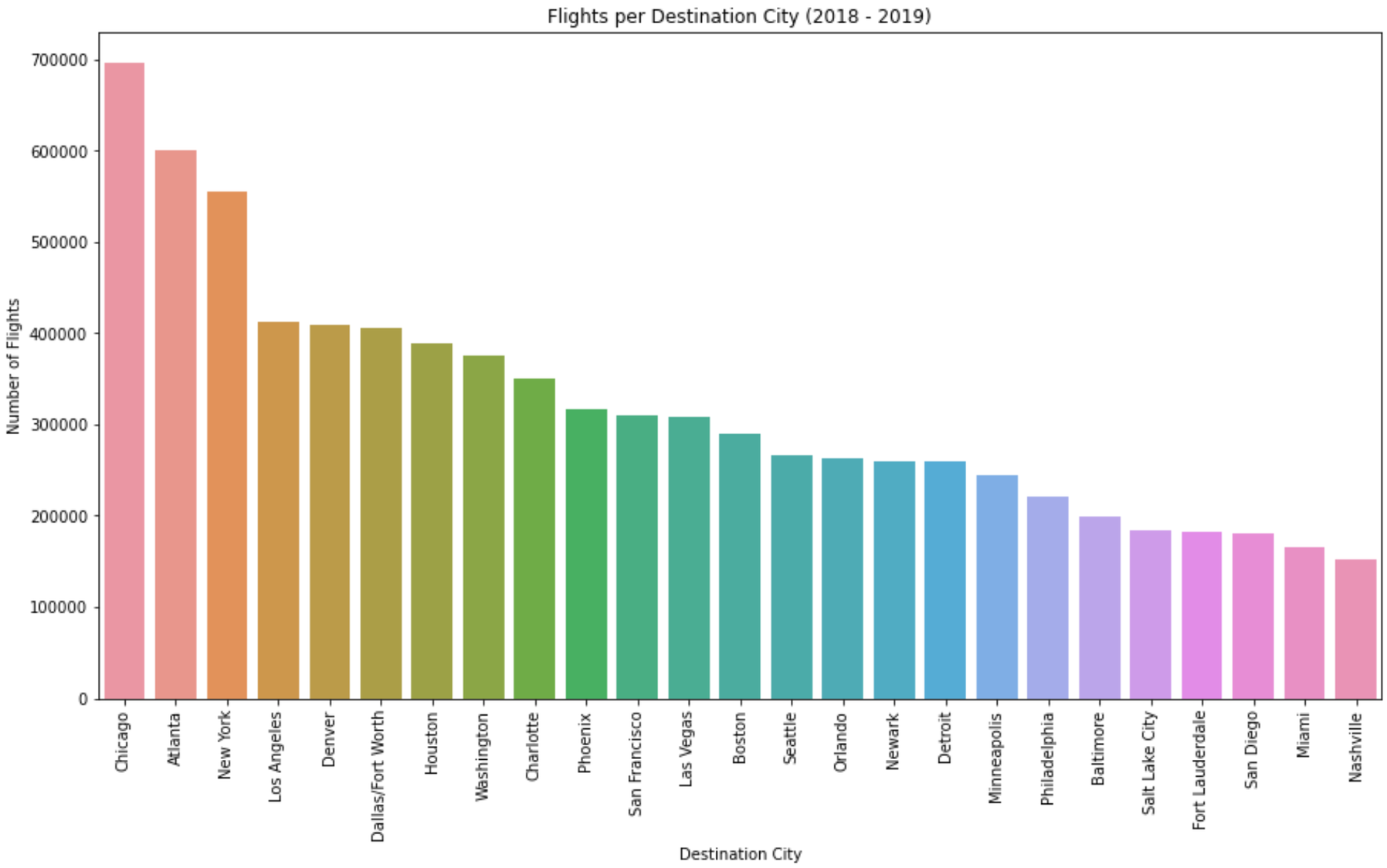


Minutes Delayed per Origin City

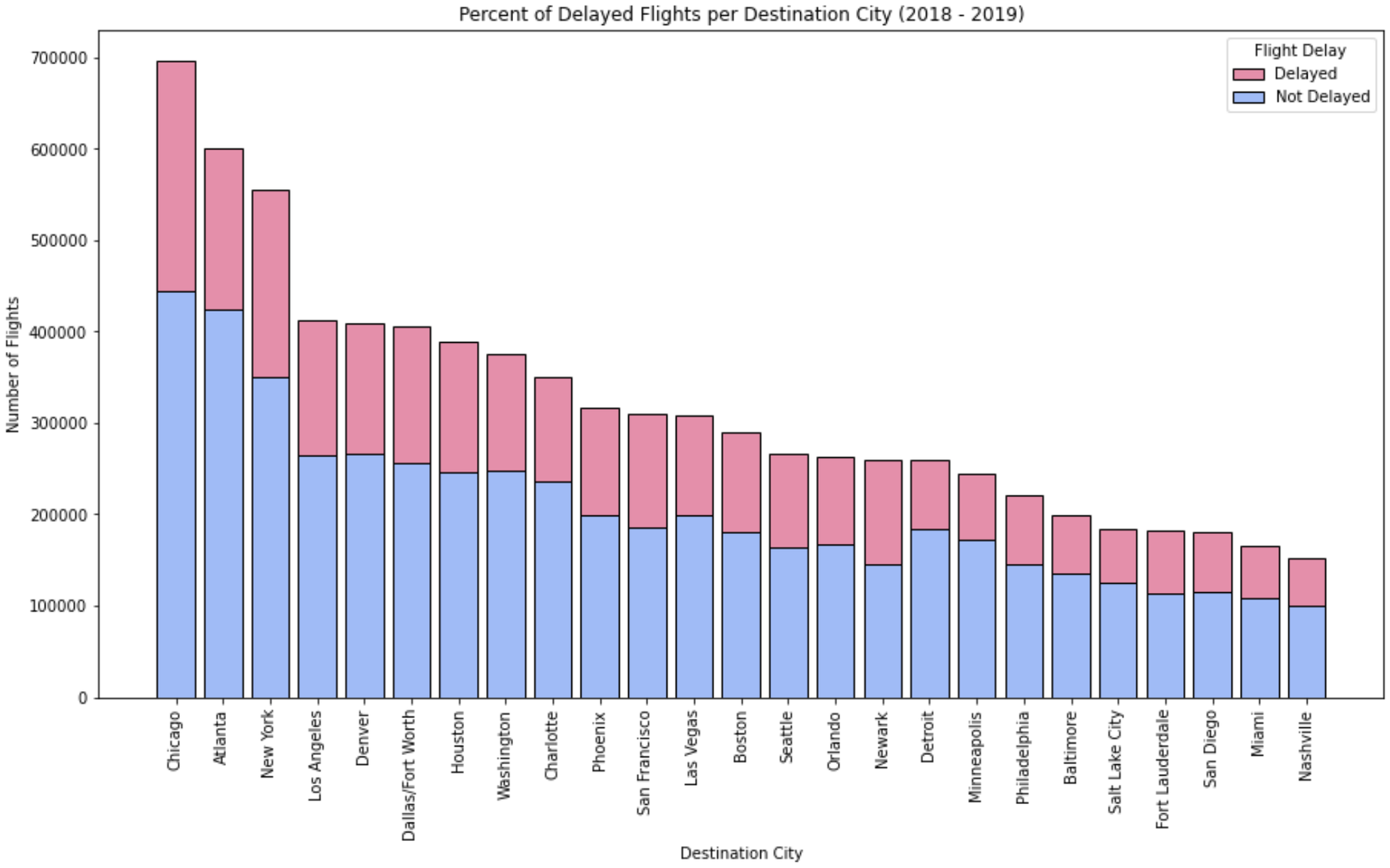




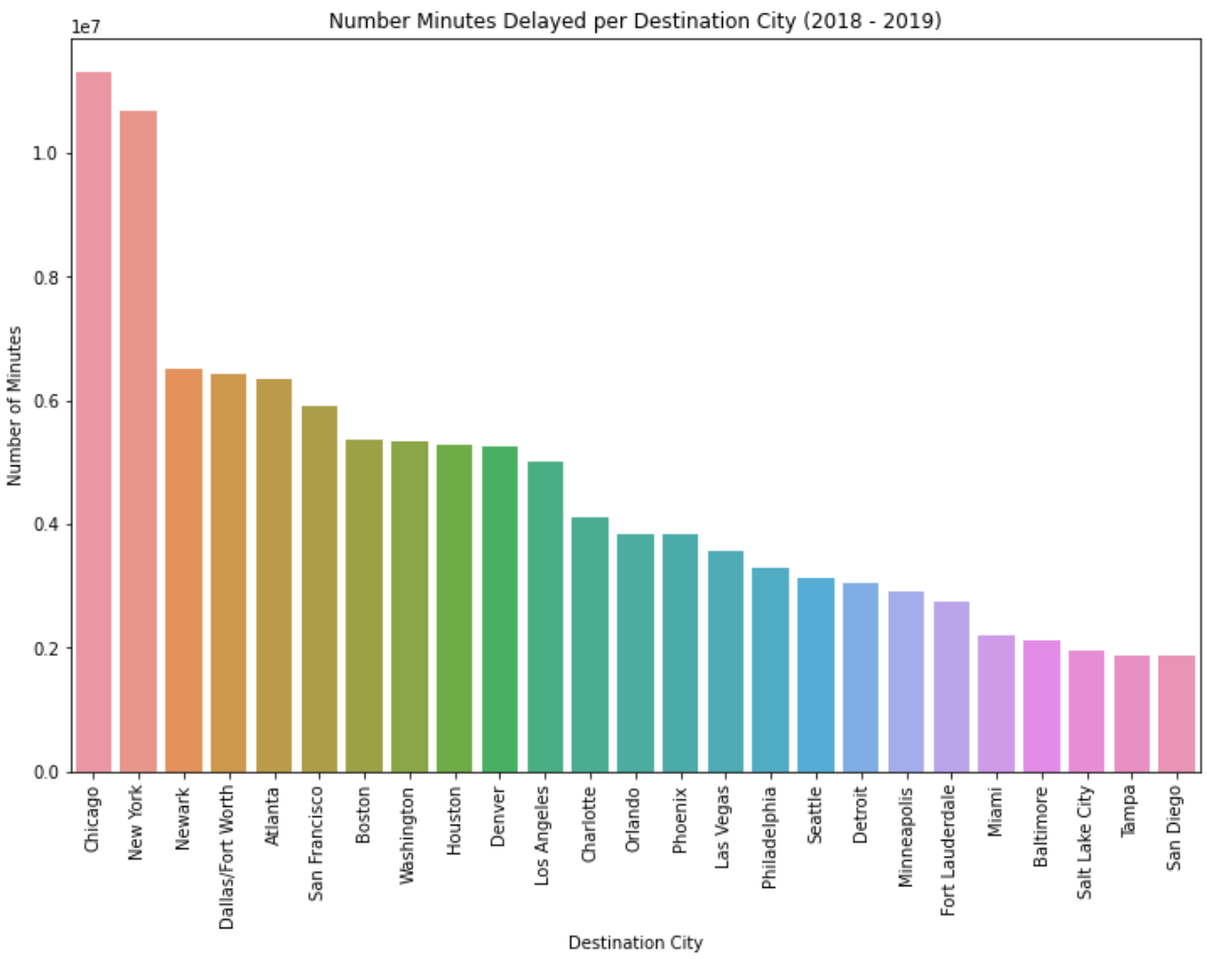
Flights per Destination City

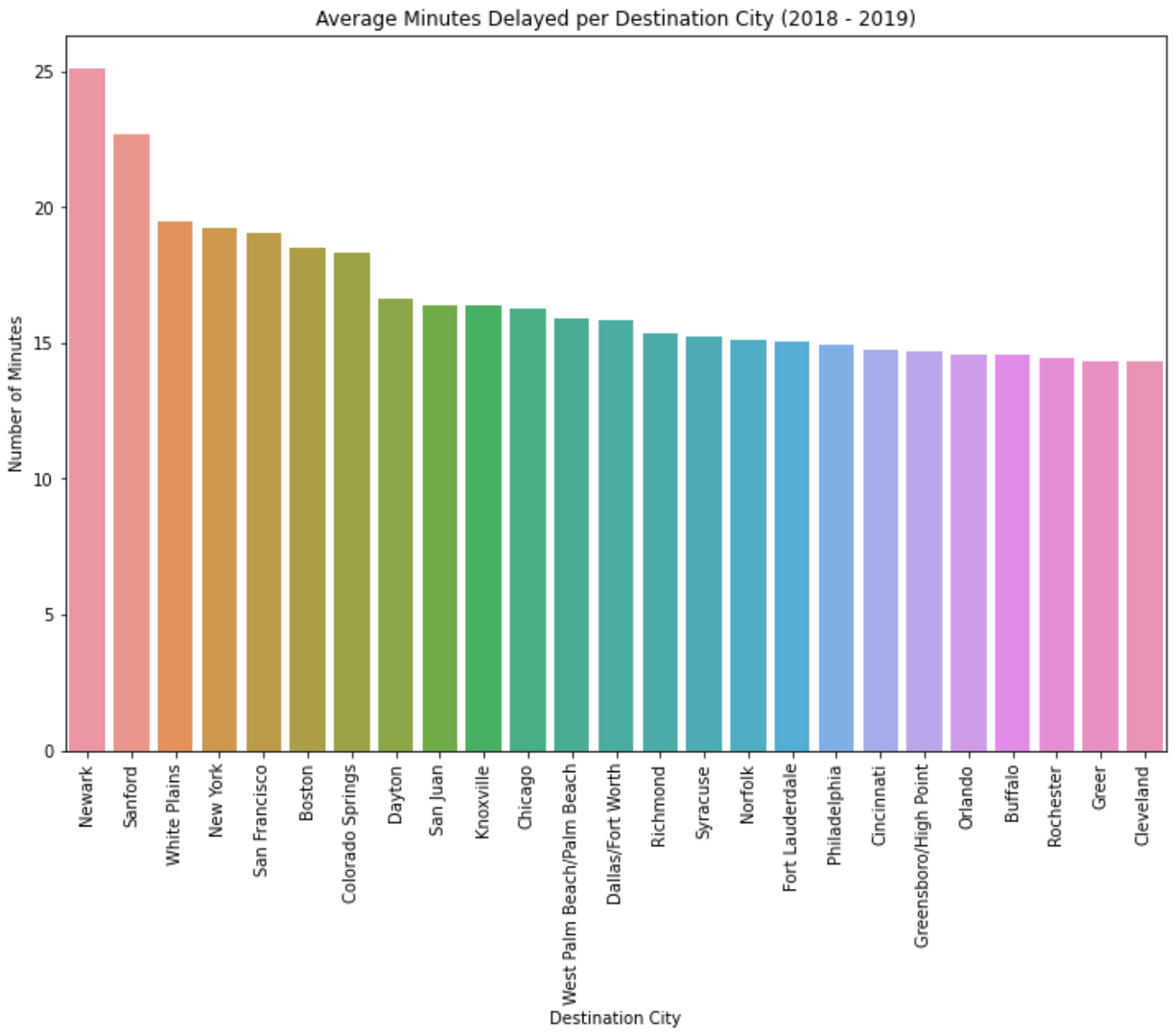


Delayed Flights per Destination City



Minutes Delayed per Destination City





Best and Worst Cities to Fly In and Out of

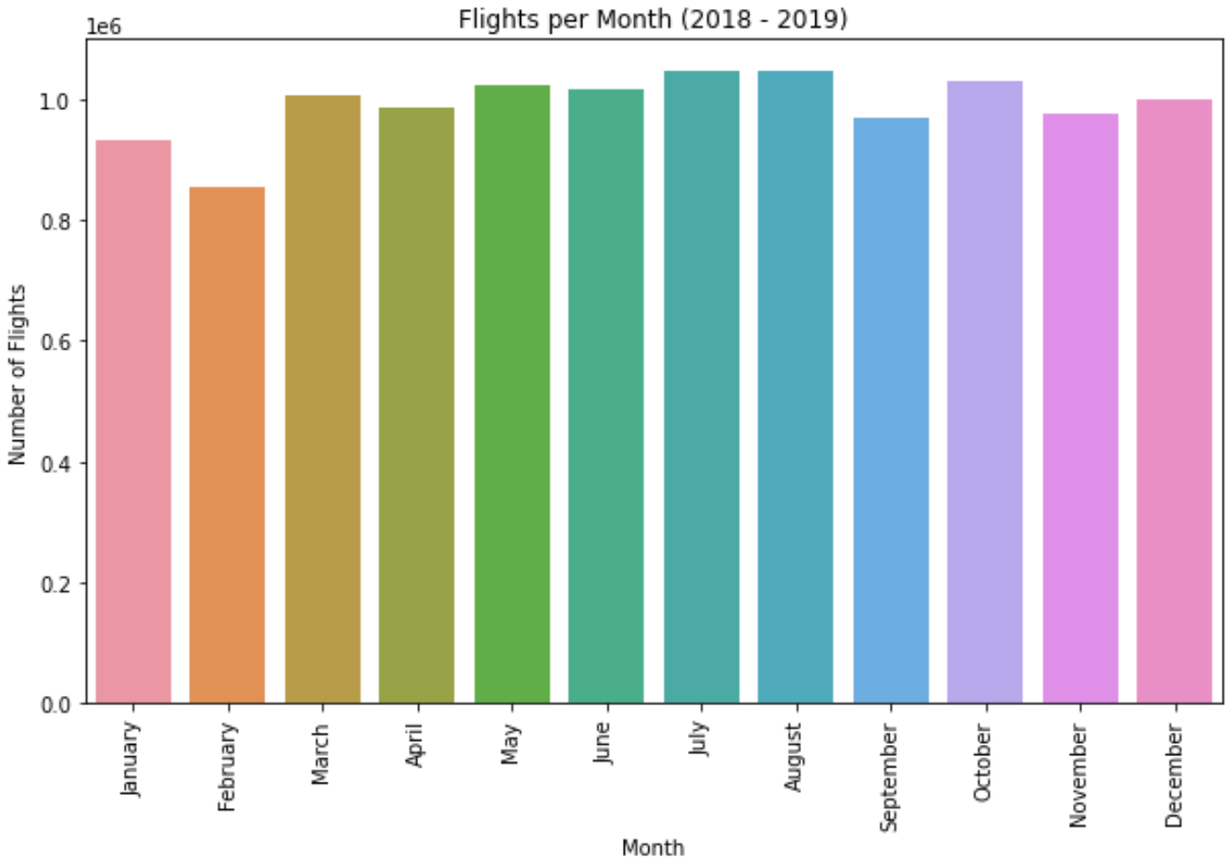
Overall Best City: **Kona**

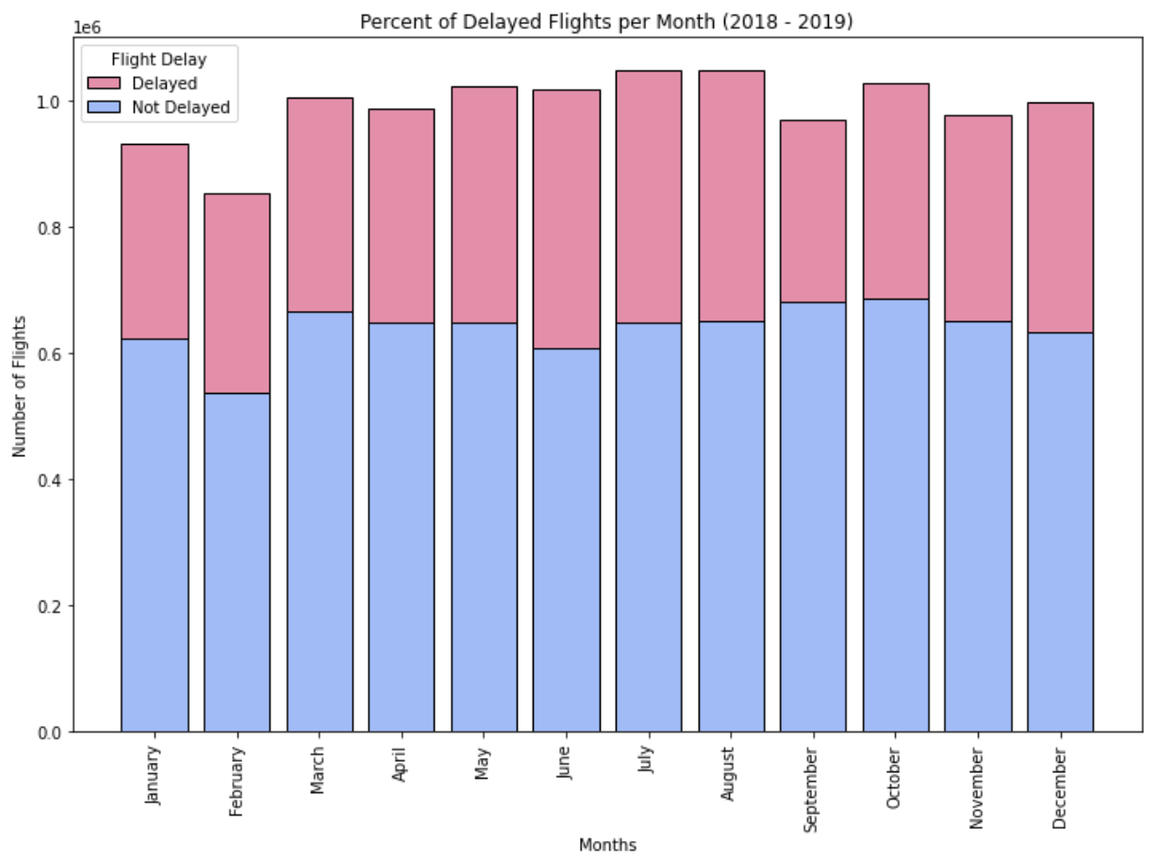
- Lowest percent delayed: 25% (departing), 28% (arriving)
  - *Definition: number of flights delayed divided by total number of flights*
- Lowest average minutes delayed: 6.22 mins (departing), 5.55 mins (arriving)
  - *Definition: total number of minutes divided by number of flights*
- Best hub: **Atlanta** (2nd highest number of flights, 4th lowest percent delayed arriving)

Overall Worst Cities: **Dallas/Forth Worth** (departing), **Newark** (arriving)

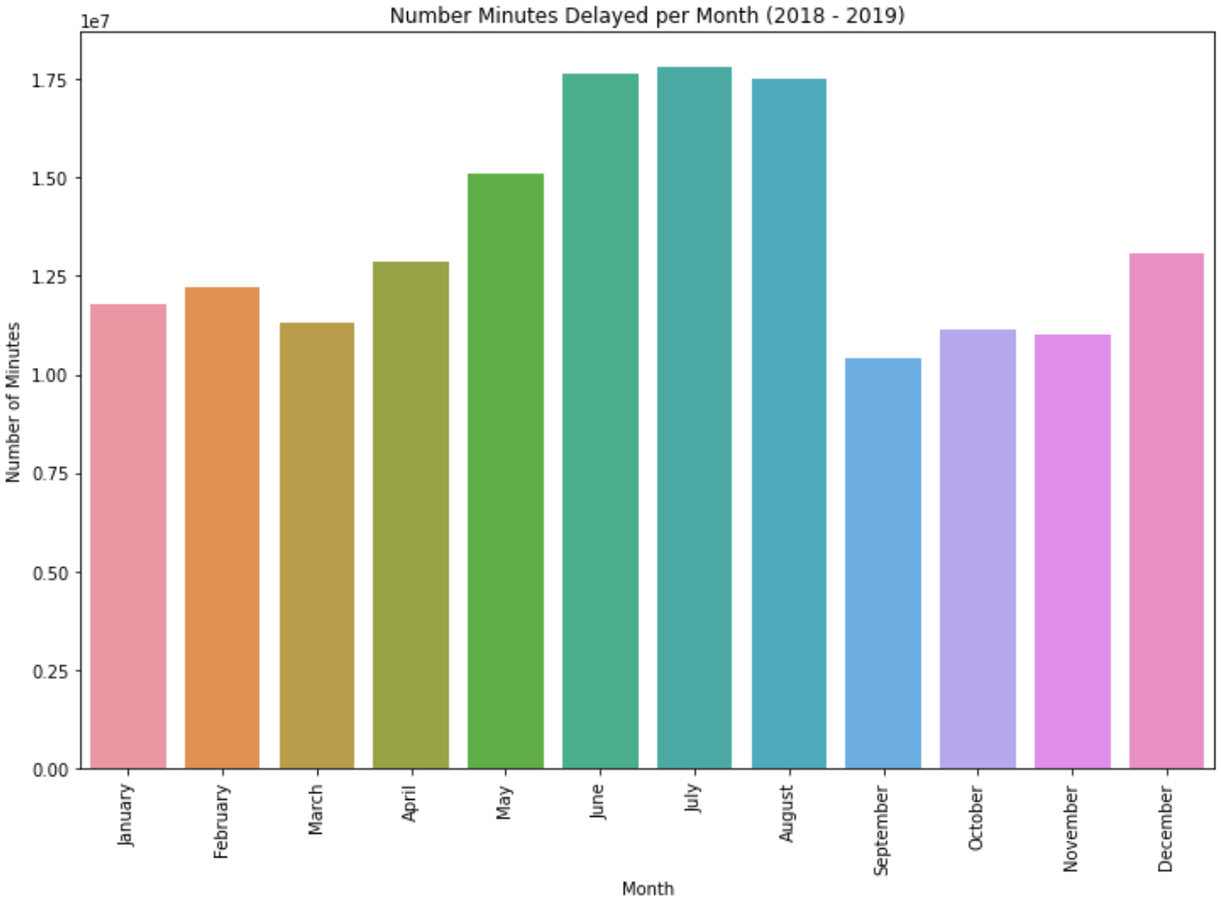
- Highest percent delayed: Dallas: 44% (departing), Newark: 44% (arriving)
  - *Definition: number of flights delayed divided by total number of flights*
- Second highest average minutes delayed: Newark: 20.63 mins (depart), 25.09 mins (arrive)
  - *Definition: total number of minutes divided by number of flights*
- Worst hub: **Chicago** (highest number of flights, 2nd highest percent delayed departing)

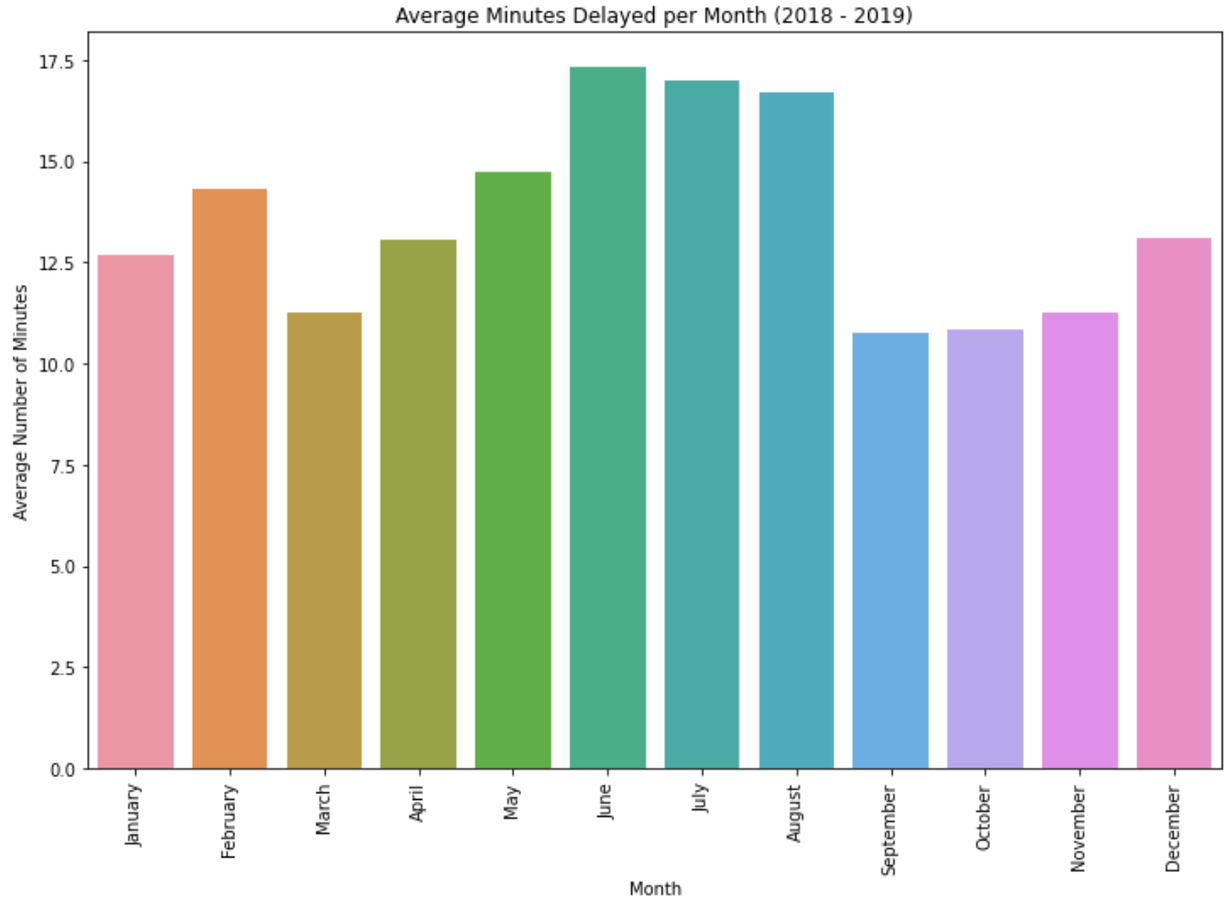
Flights per Month





Minutes Delayed per Month





Best and Worst Months to Fly

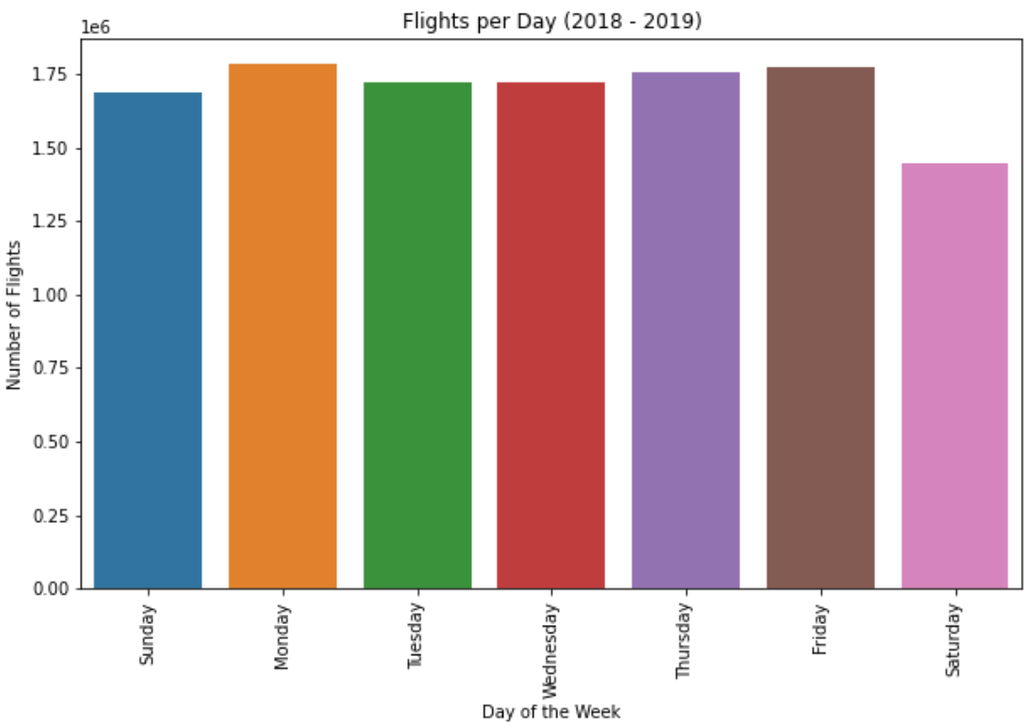
Overall Best Month: **September**

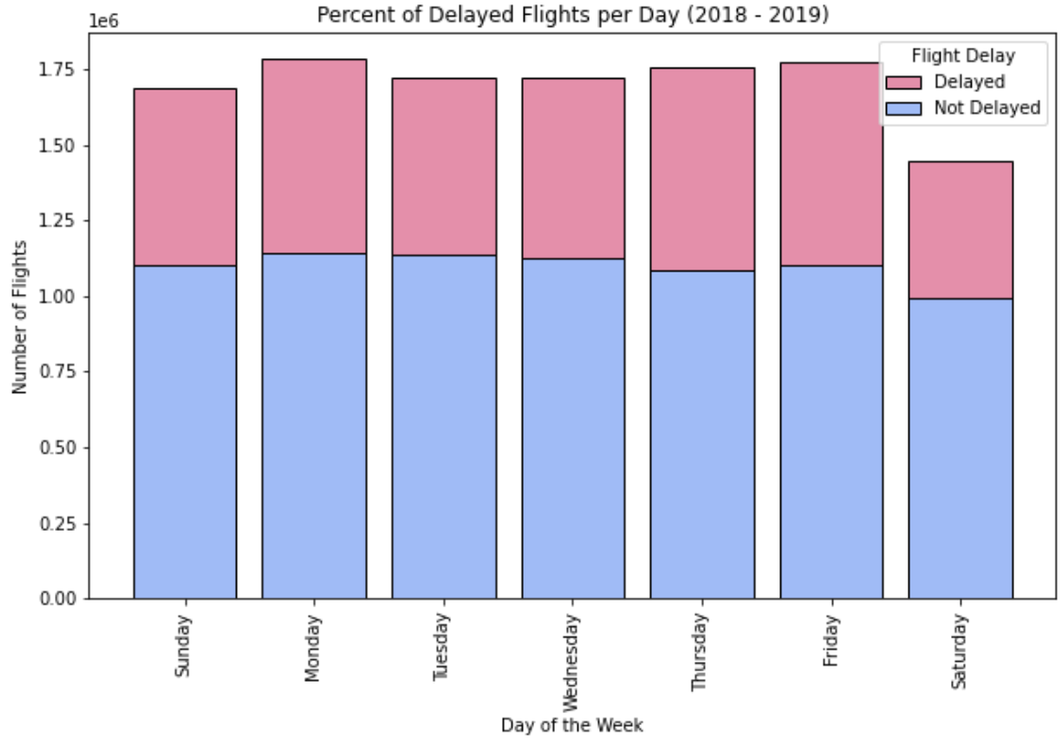
- Lowest percent delayed: 30%
  - *Definition: number of flights delayed divided by total number of flights*
- Lowest average minutes delayed: 10.76 mins
  - *Definition: total number of minutes divided by number of flights*

Overall Worst Month: **June**

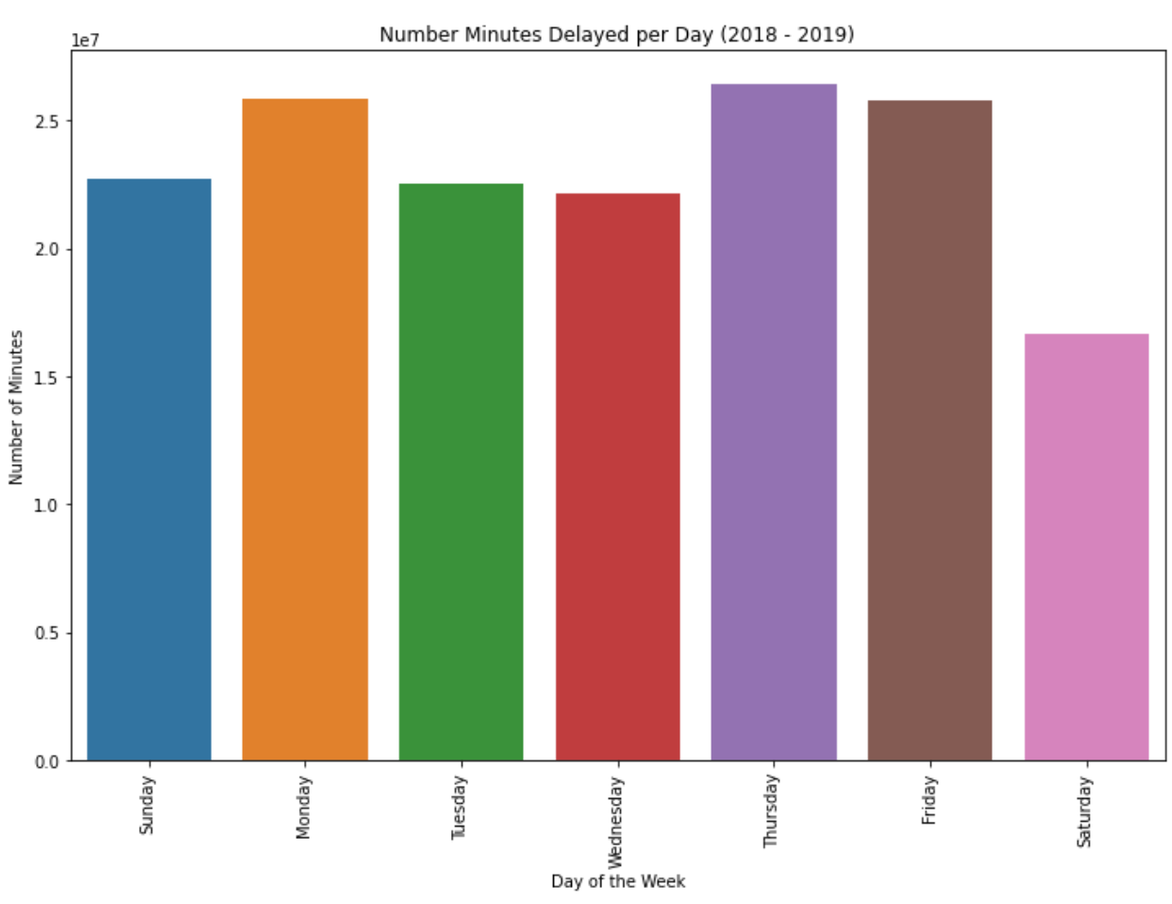
- Highest percent delayed: 40%
  - *Definition: number of flights delayed divided by total number of flights*
- Highest average minutes delayed: 17.33 mins
  - *Definition: total number of minutes divided by number of flights*
- ***Note: in general, there was not much variation between the months with the summer showing slightly higher delays***

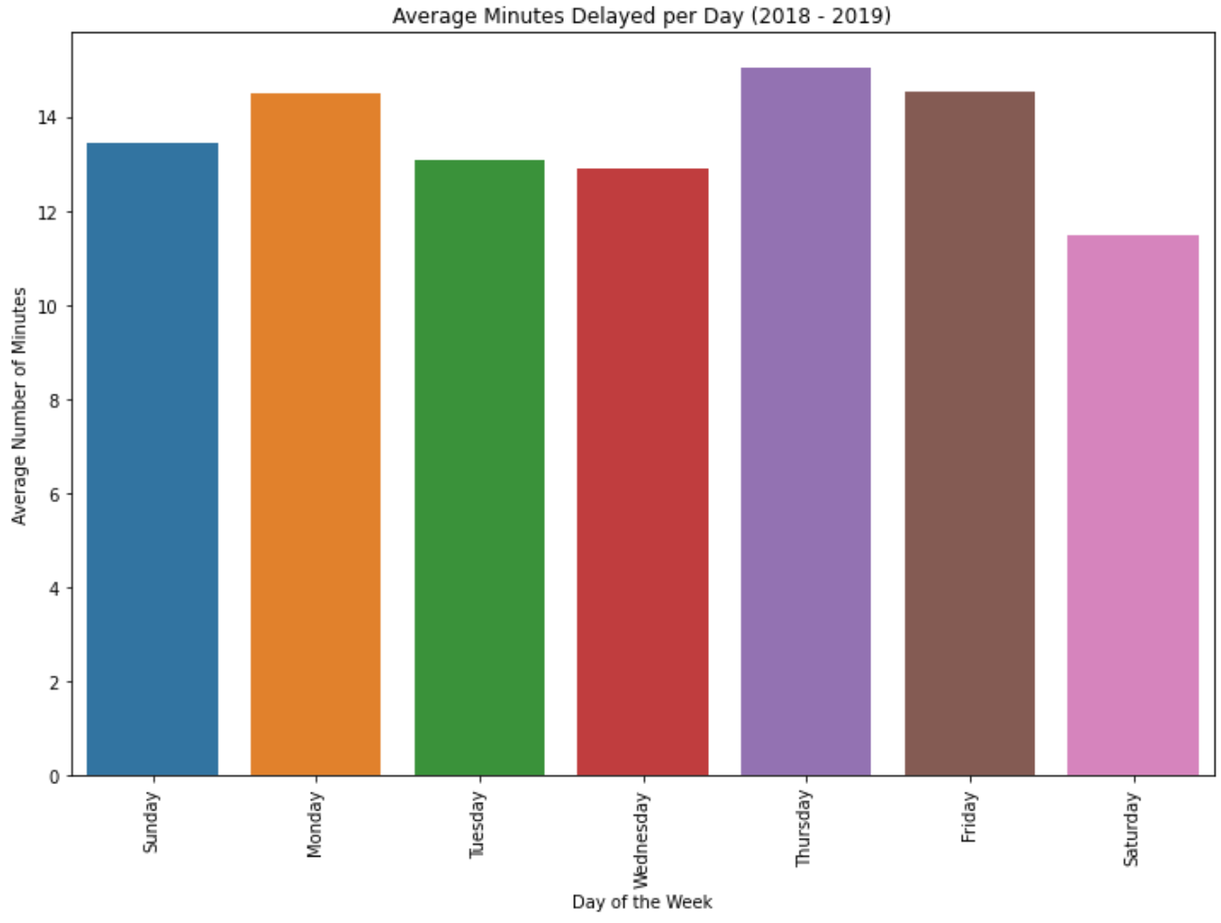
Flights per Day of the Week





Minutes Delayed per Day of the Week





Best and Worst Days to Fly

Overall Best Day: **Saturday**

- Lowest percent delayed: 31%
  - *Definition: number of flights delayed divided by total number of flights*

- Lowest average minutes delayed: 11.48 mins
  - *Definition: total number of minutes divided by number of flights*

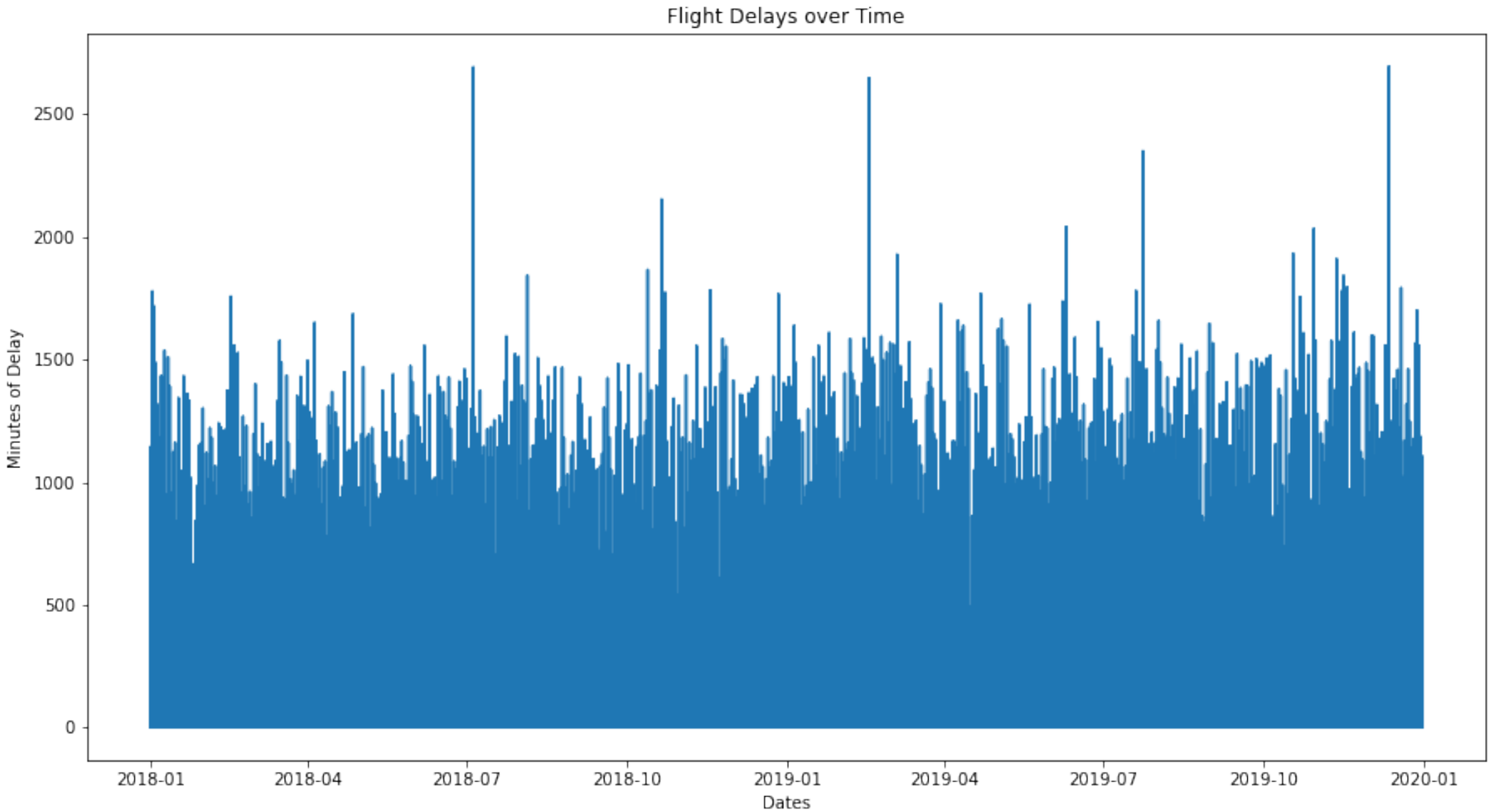
Overall Worst Day: **Thursday**

- Highest percent delayed: 38%
  - *Definition: number of flights delayed divided by total number of flights*

- Highest average minutes delayed: 15.05 mins
  - *Definition: total number of minutes divided by number of flights*

- ***Note: in general, there was not much variation between the days of the week with the weekend showing slightly less delays (but also less travel)***

Arrival Delays over Time



Not much seasonal variability identified (e.g. spikes at 7/18, 3/19, 12/19)

Modeling

Models Evaluated:

- Logistic Regression
- Random Forest
- XG Boost
- Neural Net

Baseline Model:

Not Delayed	0.647
Delayed	0.353

Modeling Metrics

Accuracy: what percent of predictions were correct

Precision: what proportion of positive identifications (flight delayed) was correct

Model Results

Model	Training Accuracy	Testing Accuracy	Testing Precision (Class = 0)	Testing Precision (Class = 1)
Logistic Regression	0.66	0.66	0.67	0.56
Random Forest	0.65	0.65	0.65	0.00
XG Boost	0.68	0.67	0.67	0.63
Neural Net (4 dense layers, early stopping at 34 epochs)	0.67	0.67	0.59	0.59

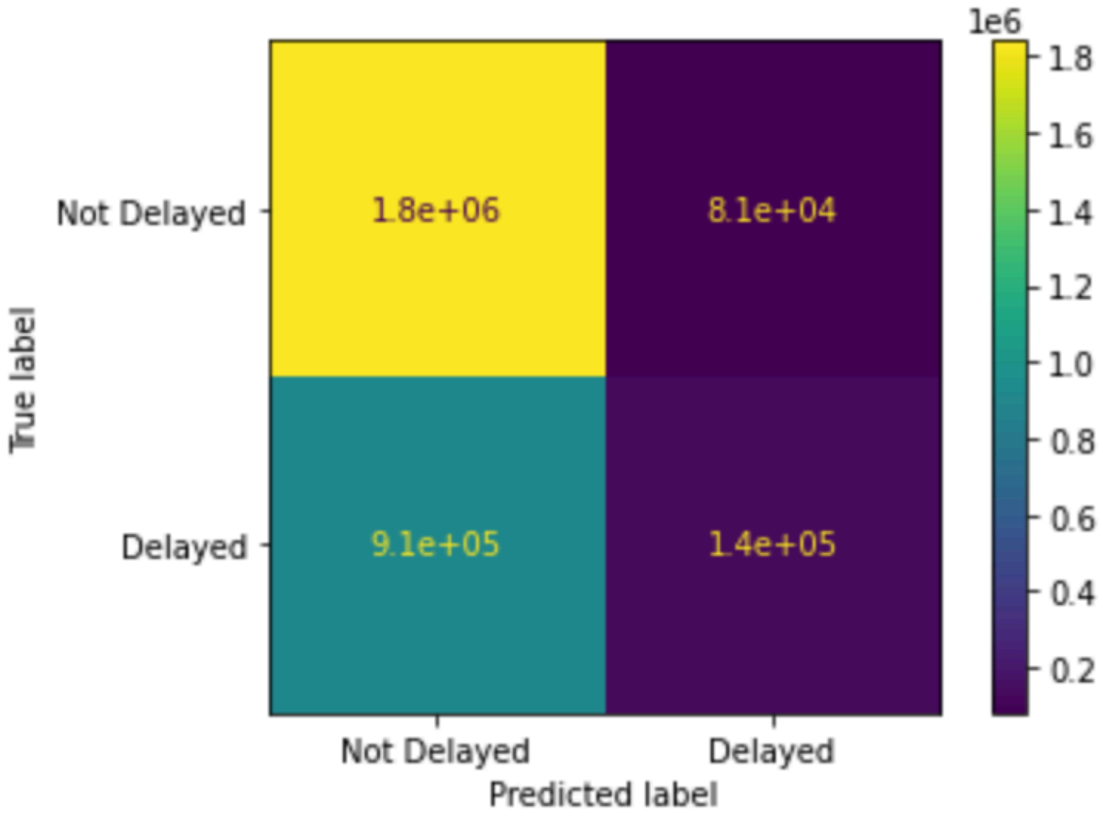
Best Model: XG Boost

XG Boost was the highest model evaluated with the following results:

- Training accuracy score: 0.68
- **Test accuracy score: 0.67**
- **Test precision score (class = 1): 0.63**

- Best hyperparameters:
  - XG max_depth: 20
  - XG n_estimators: 200

Confusion Matrix of Results



- We do see the not delayed predicted correctly as not delayed: **true negatives** (1,800,000)
- There are more delayed predicted correctly as delayed vs not delayed predicted as delayed: **true positives** vs **false positives** (140,000 vs 81,000)
- The model over predicted not delayed: **false negatives** were high (910,000)

Conclusions

- While we were able to predict slightly above baseline, in general the model results were not as strong as we would have hoped
- We found that XG Boost was slightly better than Logistic Regression and the Neural Net at predicting flight delays
- Random Forest just labeled everything as ‘not delayed’
- When analyzing the data, we did not find much variability between months or days to account for differences in flight delays

Future Considerations

- Set up a databricks cluster so that I could:
    - Run my model with many years worth of data
        - With more computing power, run more experiments and iterate faster
- Continued feature engineering around type/size of airport (looked at airport size and hub status but neither feature added any predictive power to my model)

Future Research and Improvements

- I would like to incorporate Covid data into the model and see what affect that has on the results
- Add additional data sources, e.g. weather reports
- Streamlit app: add in additional steps that only allow the user to enter actual flight parameters (for example, if Hawaiian airlines does not fly from Anchorage to Denver do not let the user enter that as a flight option)

Streamlit App

https://share.streamlit.io/becks2318/flight-delays/main/st_app.py

Questions?