Communicate Data Findings

Giancarlo Marchesi
Data Analyst Nano Degree
03.21.21.

American Airport Delays During the Last Decade

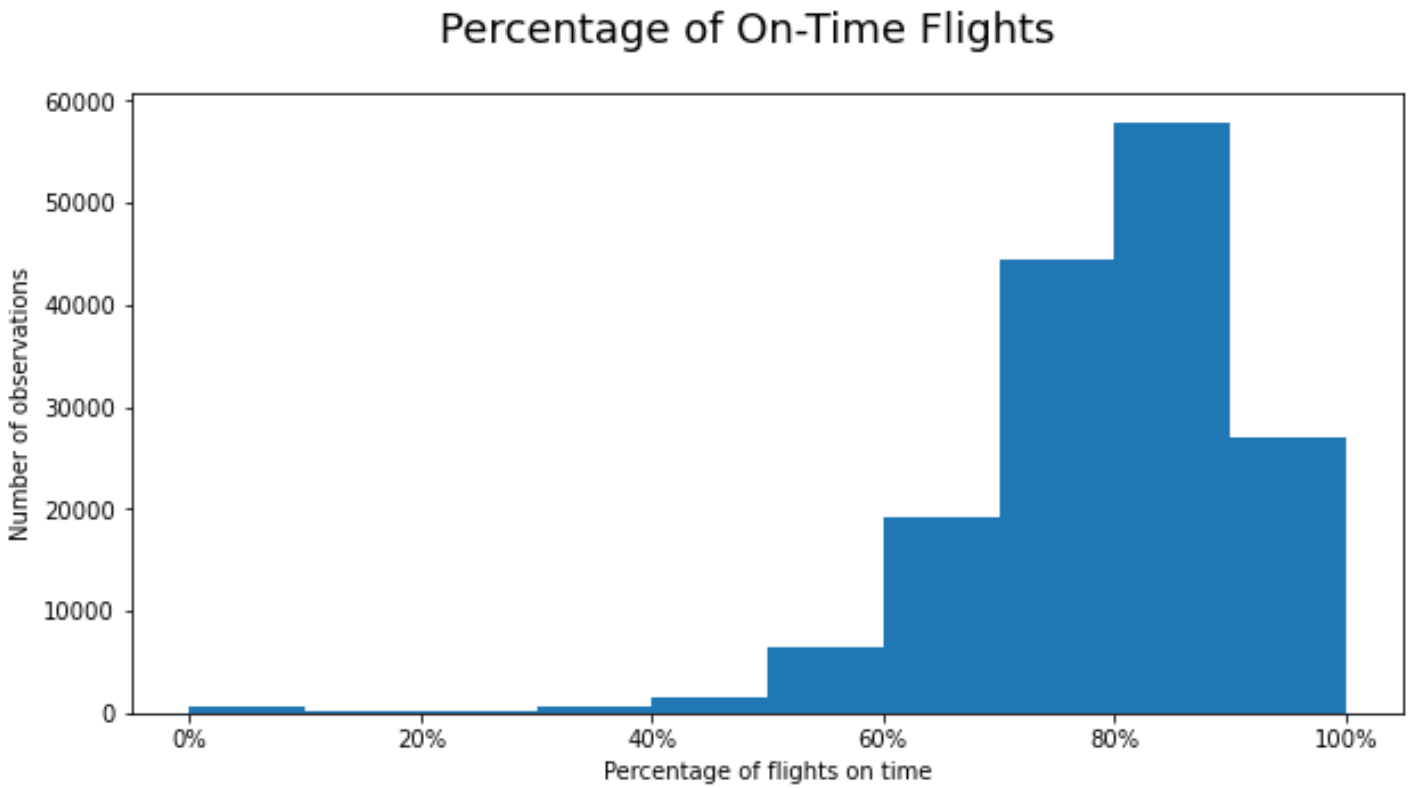
Investigation Overview

In this project, I explore the characteristics and causes of commercial airplane flight delays in the United States. I analyze time trends of flight delays; what months delays are most likely to occurred and which airline carriers and airports have the best and worst track records. Finally, I want to understand the reasons behind the delays.

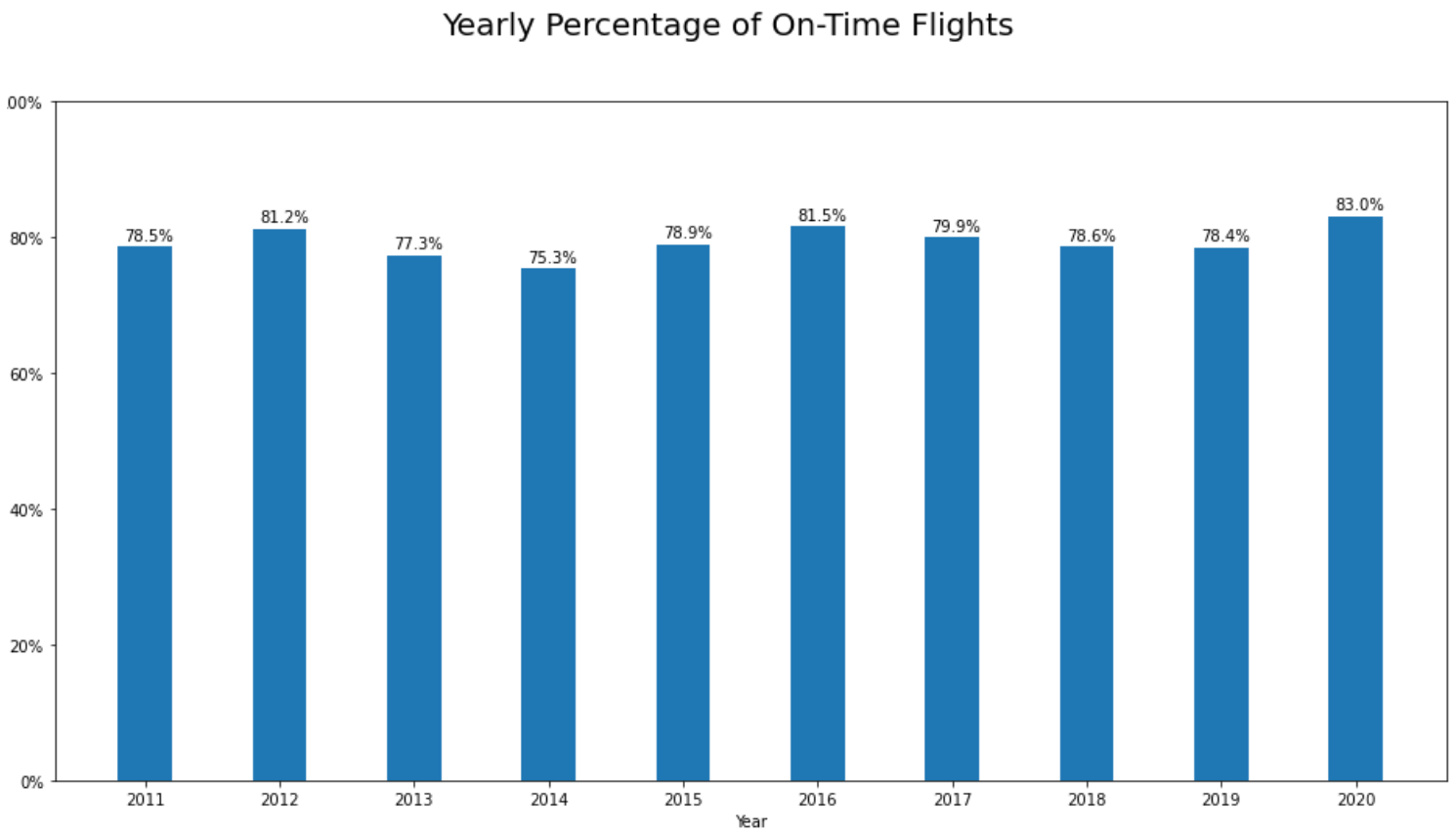
Dataset Overview

- The dataset corresponds to airline flight statistics and delay causes for airports in the United States, by month, for the period between January 2011 and December 2020. The total number of observations for the clean dataset is 158,102.
- The data centers around delays (measured in number of flights and total minutes delayed) and its five causes. There are four cross sectional dimensions to the delay measurements: time (measured in years), seasonality (measured in months), airline carrier and airport.
- Transformations were done to create the variable "ontime" to measure timely flights, as the complement to delayed, cancelled and diverted flights. On-time and delayed flights add to more than 98% of all fligths. Hence, cancelled and diverted flights are not representative. All analysis is performed on the former two categories.

By and large, the percentage of flights that are on time is above 70%, with 90% being the mode



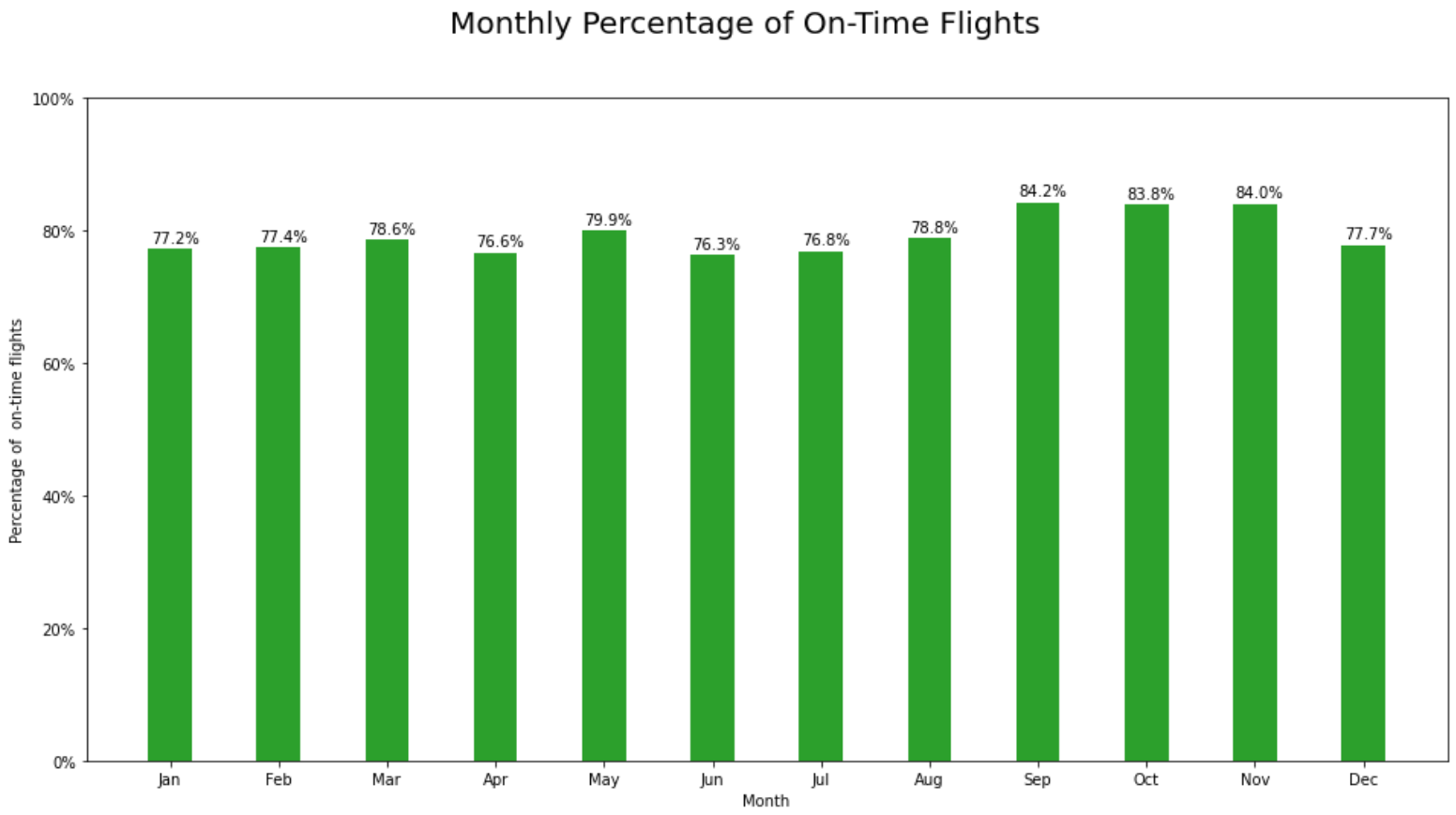
Percentage of on-time flights by year



My initial hypothesis was that some technological or logistical improvement could have increased the number of on-time flights during the decade. Yet, there seems to be no trend of yearly improvement.

For instance, 2014 had the worst performance, 2020 the best, while 2016 was better than any year between 2017 and 2019. It seems that other factors might be at stake, like the weather, which can introduce a level of randomness on a year-to-year basis.

Percentage of on-time flights by month

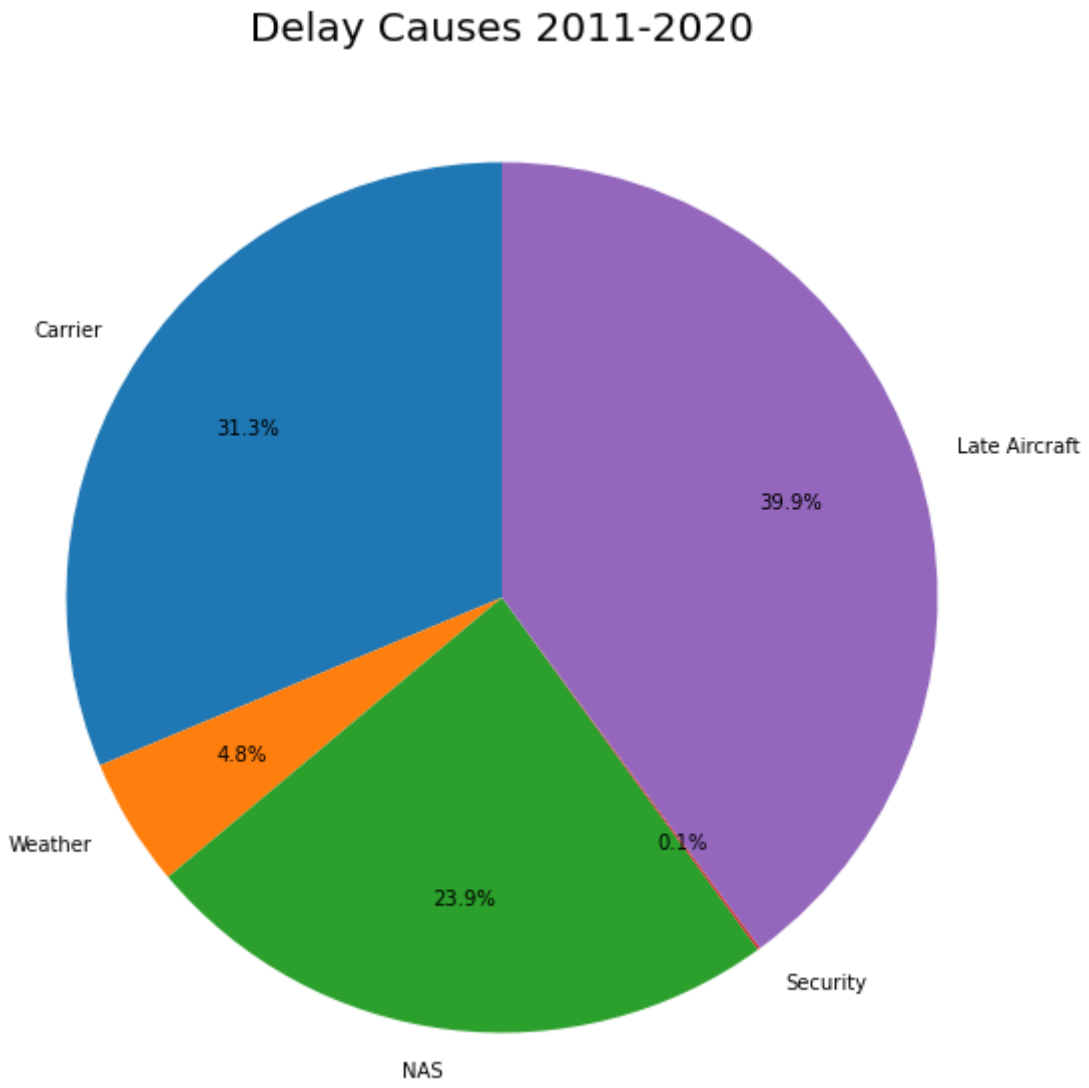


The most likely months for delays are April, June and July. This was expected, specifically for summer, because extreme weather like rainfall and, perhaps excess passenger traffic because of vacations, can cause delays.

Main causes for delays

Carrier delays and late aircraft arrival are the most common causes for delays.

Carrier delays are due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

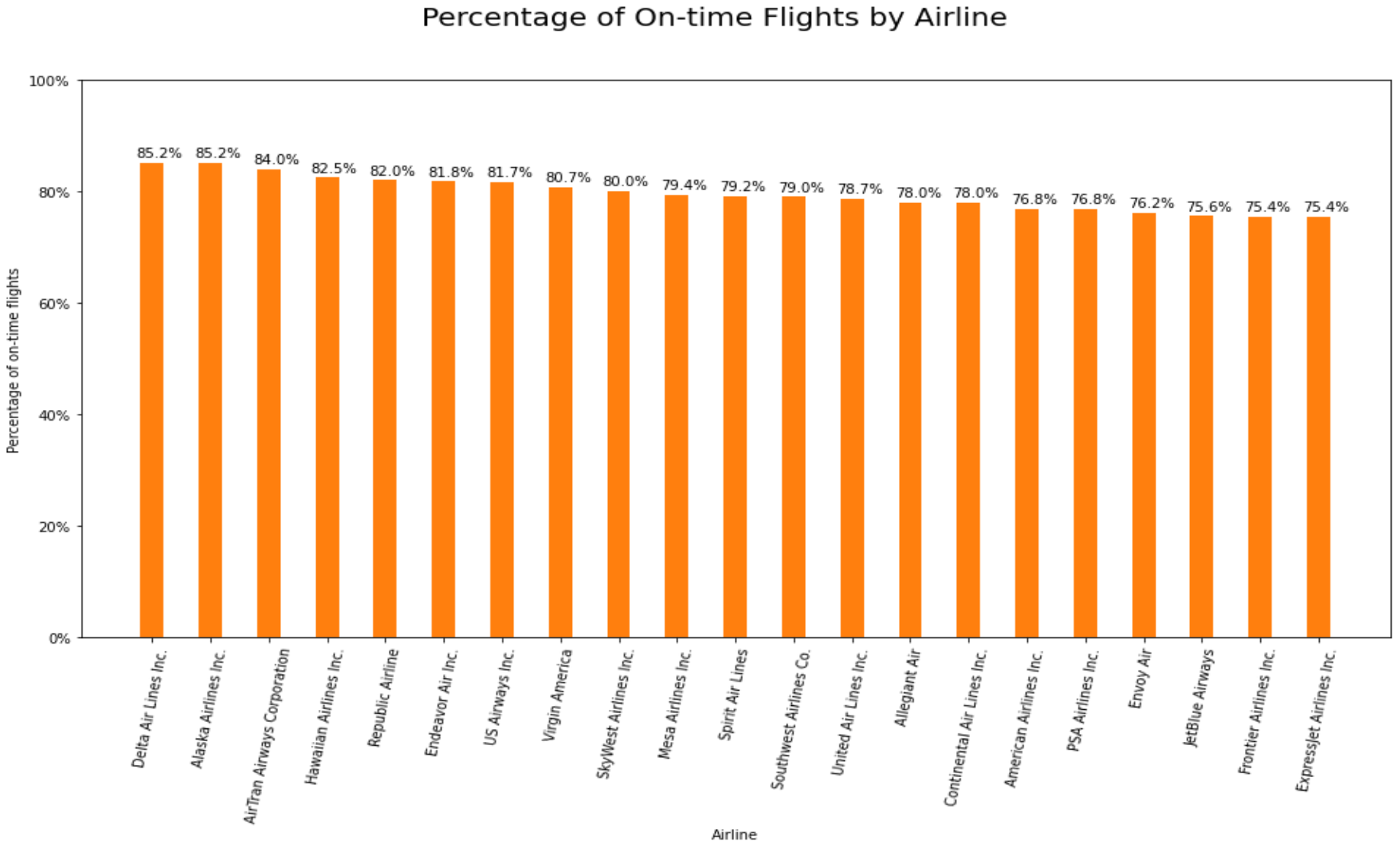


Ambiguity in the definition of weather delay

A portion of NAS and late arrival delays are caused by weather. When these corrections are considered, BTS estimates that slightly less that 40% of all delays can be attributed to weather. There is only a point estimate on the BTS website for NAS delays correction due to weather for the month of December 2020 and no estimate for late arrival correction.

Hence, a transformation for the whole database is not possible. I will analyze carrier delays at the airline level and NAS delays at the airport level.

Analyzing delays at the Airline Level

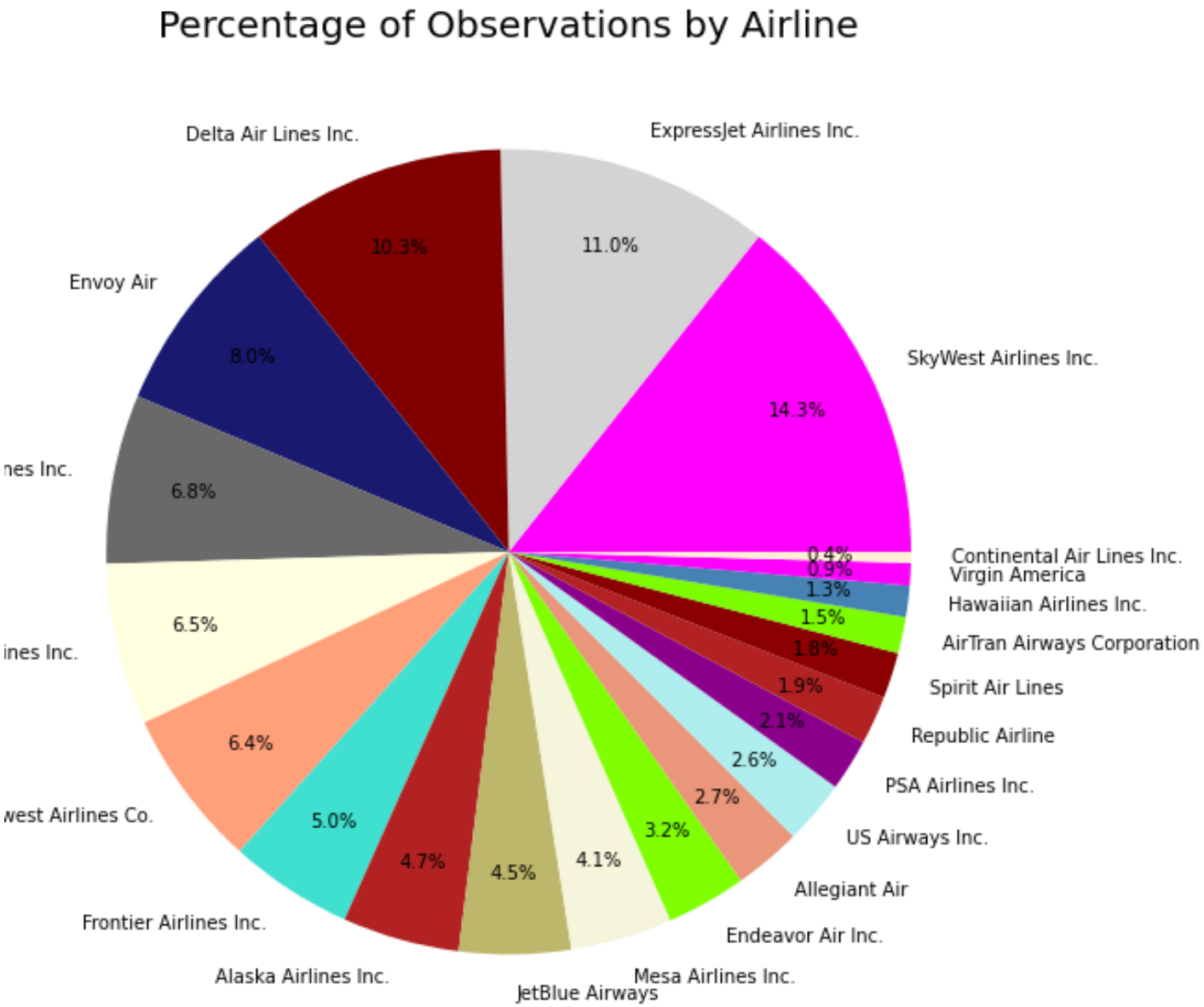


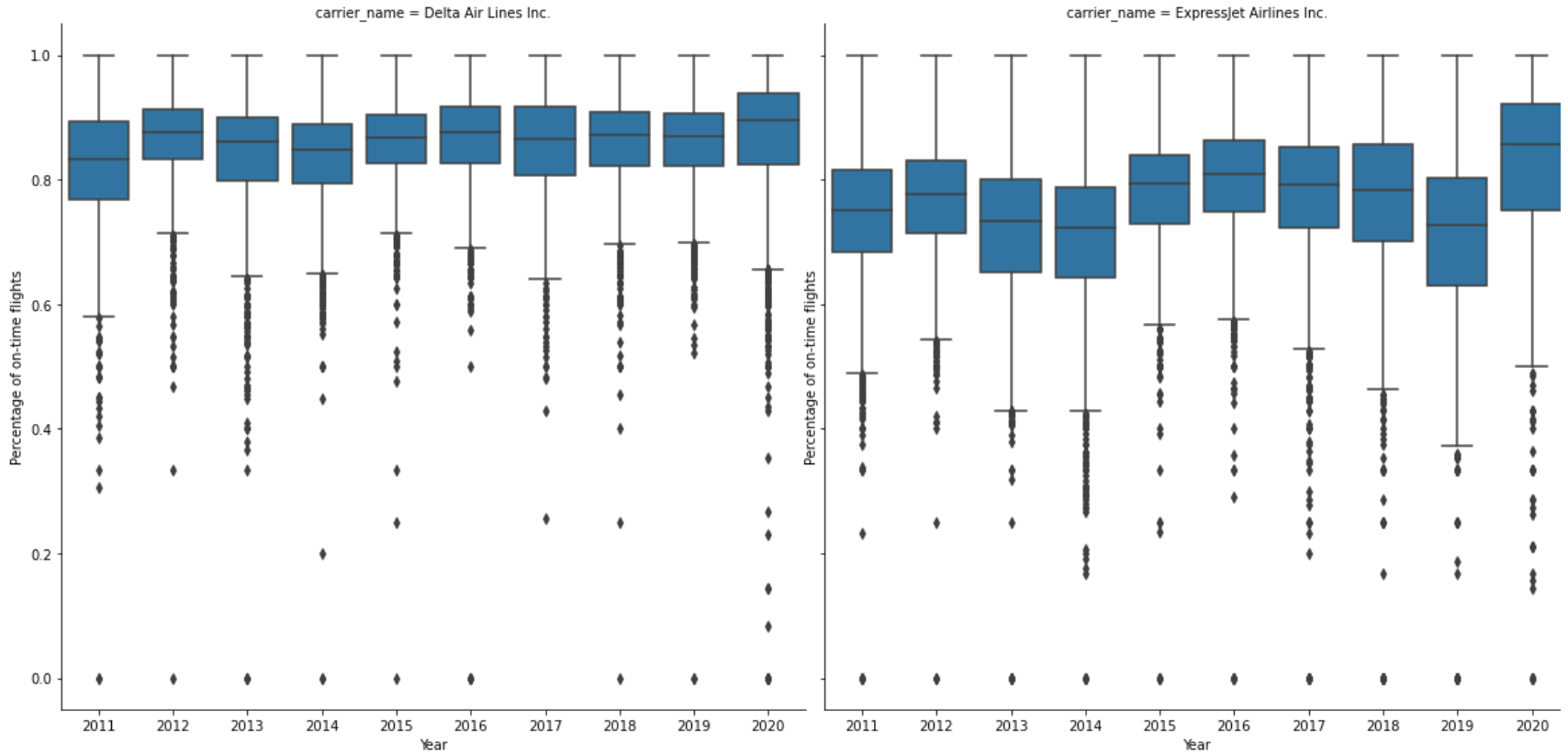
Delta is the best performer among all airlines in terms of on-time flight arrivals, while Express Jet is the worst.

There is nearly a 10% difference between the two airlines.

Selecting the top 5 Airlines for a closer look

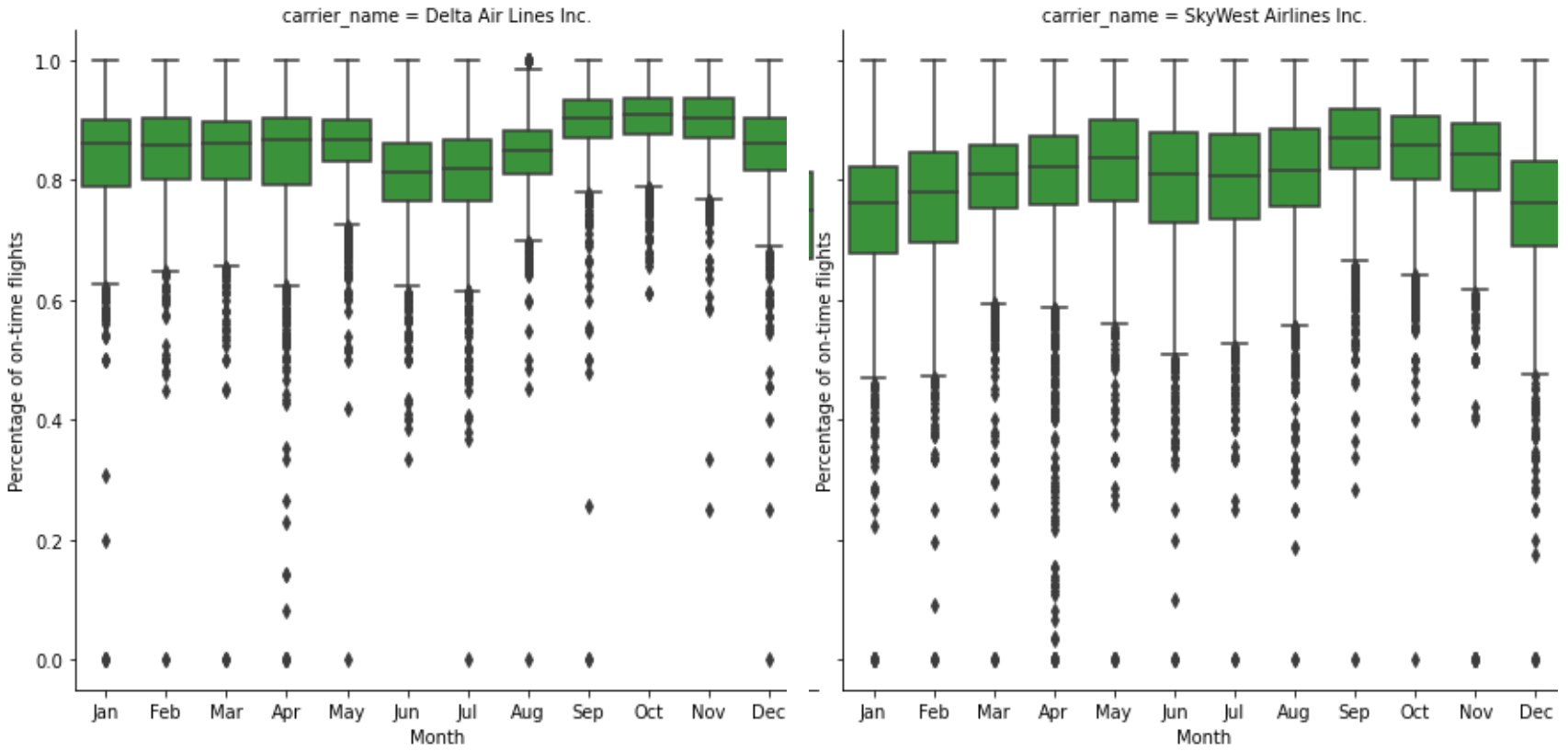
- I zoomed in on a subset of the major airlines. I chose SkyWest, ExpressJet, Delta, American and United, which acount for 50% of all fligths. Envoy does not have data for all years in the sample. Hence, I replaced it for United.





Among the 5 major airlines, Delta is more likely to be on time

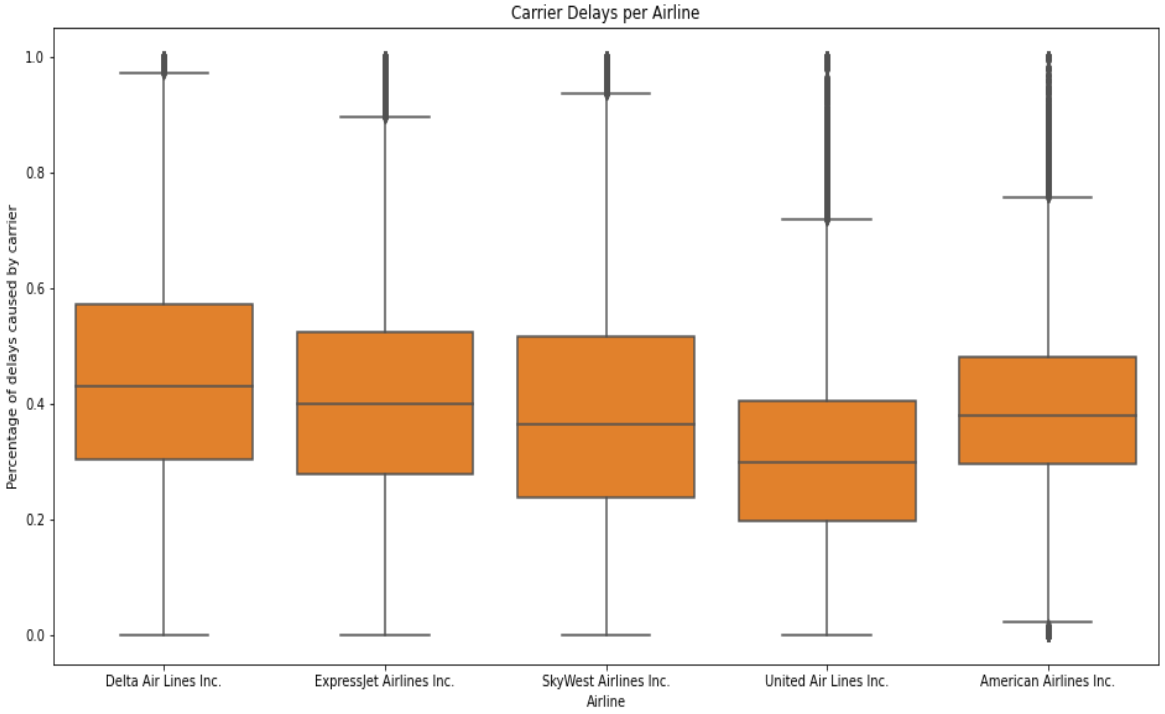
Delta is the most likely to be on time and its performance is fairly even across years, whereas ExpressJet is the least likely to be on time and its performance is uneven across years.

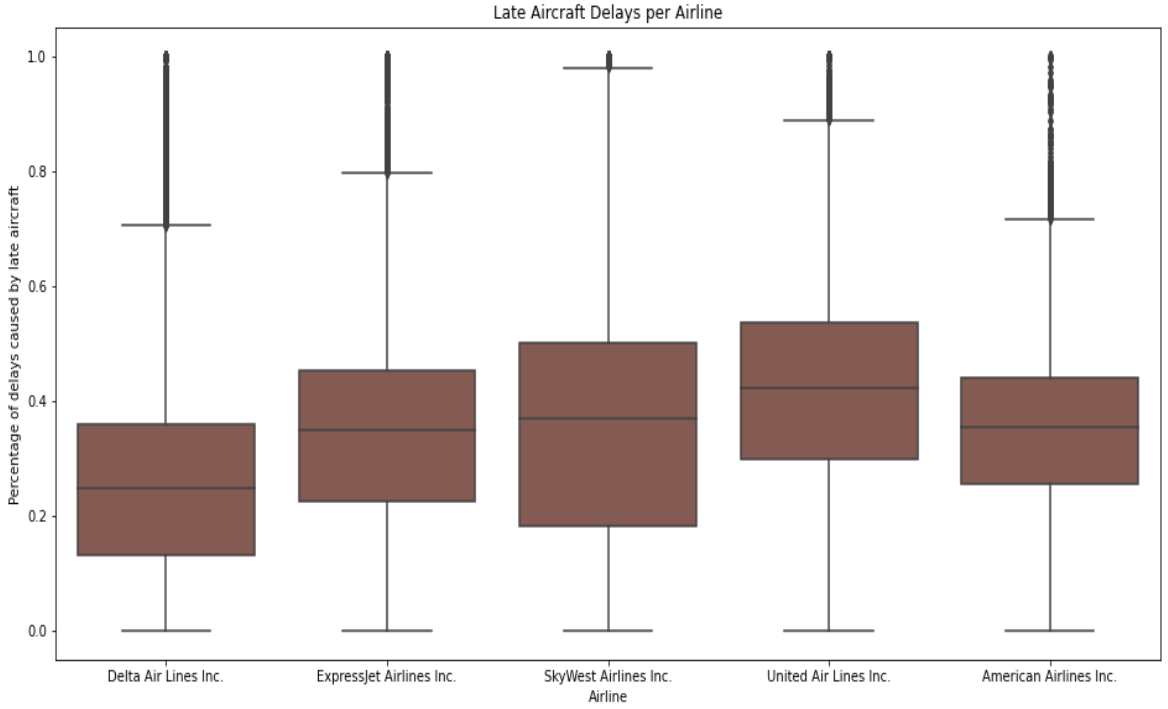


Seasonality is a little different among the top 5 airlines

Delta, like American, United and ExpressJet, is more likely to have flights delayed during the months of June and July, whereas SkyWest is more likely to be delayed during the winter months of December, January and February.

United has the best performance on carrier delays but the worst on late arrivals



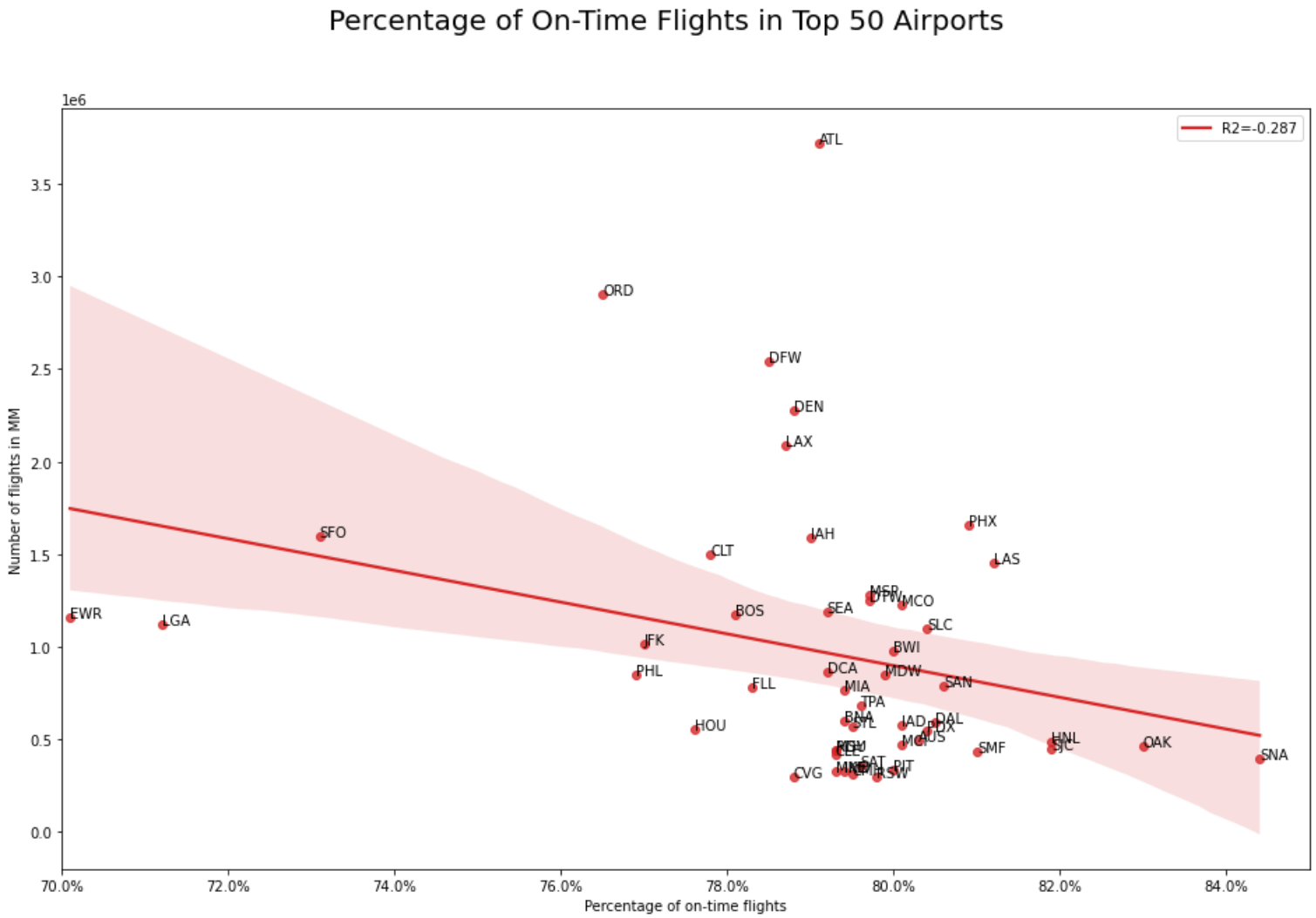


…but let’s remember that late arrivals also have a weather component that is not mapped out in the data

Analyzing delays at the airport level

At the top-50 airport level, there is negative correlation between number of flights and percentage of on-time flights, but the relationship is not that strong, (R2: -0.289).

The percentage of on-time flights for most of airports is in the range between 78% and 80% on time. Santa Ana (SNA), Oakland (OAK), Honolulu (HNL) and San Jose (SJC) are the most likely to be on time. Conversely, Newark (EWR), La Guardia (LGA) and San Francisco(SFO) are the least likely to be on time.



NAS delays at the airport level

NAS delays are a proxy for airport efficiency, as BTS measures factors that are mostly under their control.

When NAS delays are analyzed against number of flights, we find a positive relationship with an R2 of 0.442. SFO, EWR and LGA are the least efficient while ATL, ORD, DFW, DEN and LAX overperform their expected level, despite being among the five busiest airports.