US Domestic Air Travel Trends Data Analysis

Isaac Madera, Raymond Hung, and Aashray Puri
https://github.com/UC-Berkeley-I-School/Project2_Puri_Madera_Hung/

Intended Audience:

Frequent US domestic air travelers who are interested in using historical US flight data to better understand US airline/airport trends, and delay/cancelation patterns

Purpose & Overview:

This document provides a detailed exploration of US domestic air travel data by conducting an extensive exploratory data analysis (EDA) with a particular focus on flight trends, cancellations[1], and delays[2], as well as weather patterns.

To conduct our analysis, we begin by discussing the datasets used, and the data pre-processing and clean-up procedures performed to prepare the data for EDA. We then delve into the examination of flight trends over time, analyzing flight volume and identifying the most / least popular airports and airlines. Furthermore, we investigate the impact of weather on flight operations out of specific airports, exploring the influence of various weather conditions on flight delays and cancellations.

Overall, this document provides valuable insights on US domestic air travel, highlighting trends and complexities across airports, airlines, and related delays/cancellations.

Dataset Description:

The core datasets used in this analysis come from Kaggle, “an online community of Data Scientists and Machine Learning practitioners” (Girl, 2021). The original source of this data, however, is from the US Department of Transportation (USDOT), which has been collecting and sharing US domestic flight data since January 2018. The USDOT’s dataset lags reality by ~3 months (i.e February 2023 data will be available by May 2023), however the Kaggle data is static, providing data only between January 2018 and July 2022. This study intentionally focuses on Kaggle’s 2019 - 2021 datasets because the datasets for 2018 and 2022 are noticeably incomplete[3]. Moreover, 3 additional datasets are used along with the Kaggle data for this analysis: 1) airline name data[4], to adjust airline codes to names we recognize (i.e AA to American Airlines), 2) airport geographical information[4] (for mapping purposes), and 3) weather data[5] for the 10 airports with the highest number of flight departures in 2021 (to understand the impact of weather on domestic US air travel in 2021).

Question Plan:

The chosen dataset is rich in both volume and dimensions, to the point that it can be analyzed in many different ways. To narrow the scope of this project, our group decided to focus on the following 3 areas:

1. **Competitive landscape:** This dataset is a living repository of every past domestic flight in the United States. We felt inclined to use this information to better understand overarching trends in the US air travel industry.

[1] Defined as when a planned flight is called-off/terminated, and unable to depart from the Origin as scheduled.
[2] Flight delays occur when a flight departs more than 15 minutes after its scheduled departure time. Within the context of this report, we use the boolean value provided in the DepDel15 column.
[3] 2018’s dataset did not include all the airlines and 2022’s dataset only went as far as September
[4] Also found on Kaggle
[5] Data from the National Centers

- Key questions include: How did COVID impact flight volume[6]? Which airline has the most flights in the US? What are the busiest states and airports from a flight volume perspective in the US?

2. **Delay landscape:** The dataset has a multitude of layers in which to evaluate delayed flights. We will explore nuances to better understand what segments are more prone to seeing delays.
   - Key questions include: Which airline has the most/fewest delayed flights in the US? Which airports are most likely to experience a delay? In which month / day / hour-of-day are delays most common?
3. **Weather impact**: We leveraged an external datasource to merge weather information into our dataset.
   - Key questions include: Which airports are more prone to experiencing weather events (rain, snow, thunderstorms, fog)? What is the impact of thunderstorms and snowstorms on flight operations?

Sanity Checking, Data Cleansing & Exploratory Findings:

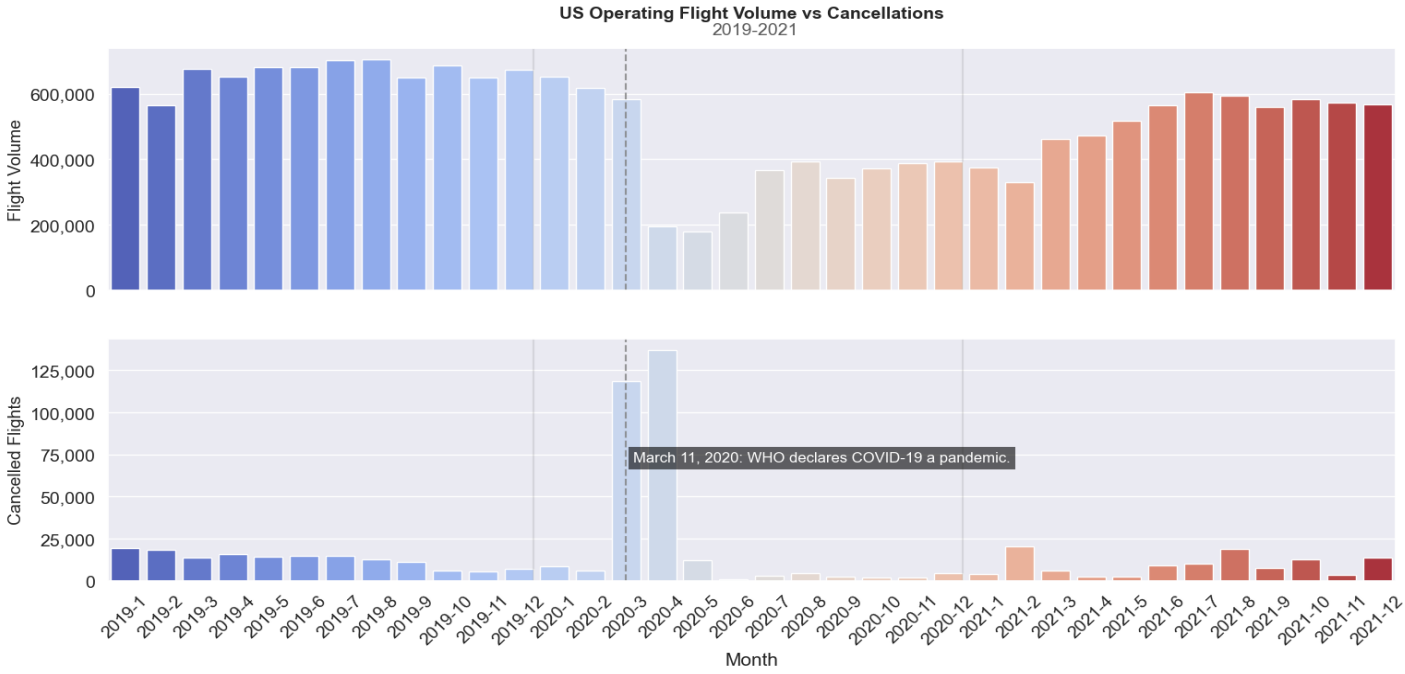
- **Shape of Kaggle 2019-2021 datasets loaded as pandas dataframes**: The shape of the datasets confirmed that all 3 years have between 5 million and 8 million records (i.e domestic flights), and 61 columns. 2019 saw the most records (i.e. 8,091,684), followed by 2021 (i.e. 6,311,871) and then 2020 (i.e. 5,022,397) - this trend can likely be attributed to the impact of COVID-19 on travel.
- **List of Column Names & Respective Data Types 2019-2021:** Seeing the column names and datatypes for the datasets (which matched for 2019-2021) confirmed that the 3 years' data could be merged into one overarching dataset. To be safe, the `.head()` and `.tail()` methods were run on each years' data as an initial top level inspection - no anomalies were found.
- **All 3 years data grouped into 1 dataframe called `combined_df`:** The combined dataframe did not have a pre-set index. Considering this project's guiding questions, it was decided that `FlightDate` would be the most appropriate field for an index. Before setting the new index (and dropping the original `FlightDate` column), the data type of `FlightDate` was changed to `datetime64[ns]`. The `.info()` method was used to confirm combined_df had been transformed correctly.
- **`combined_df.duplicated().sum()` confirmed no records were repeated.**
- **`combined_df.isnull().sum()`, shows some columns with a significant amount of null values:** This was concerning at first, but further analysis confirmed these null values were attributed to canceled/diverted flights. This is because when a flight is diverted/canceled, most of the columns (i.e. ArrTime, DepDel15 etc.) in our dataframe become irrelevant, and filled with NaN's. There were 565,702 canceled flights and 44,184 diverted flights in combined_df, and this roughly matched the total number of null records in the dataframe. As such, this did not raise any red flags.
- **`combined_df.describe().T` used to understand high level summary statistics for integer and float dtype columns:** Of the 41 numeric columns in the dataset, all have a row count between 18.8 million to 19.4 million. `DepDel15` and `ArrDel15` are the only 2 columns that show a min/max of 0/1, which suggests these columns should be boolean dtypes, even though they show up as floats. As such, these two columns were re-typecasted as booleans. The 500k+ NaN's which were present in each of these columns were accordingly converted to 'False'.
- **Of the 61 available columns, we decided to drop 41, and focus on 20 (including `FlightDate` as the index) to conduct our analysis:** Excluding columns which weren't central to our analysis helped reduce memory usage and improved data manageability. The 20 selected columns were those that were most relevant to answering the project's guiding questions. Further information of which columns were kept/excluded can be found in the appendix/jupyter notebooks.

[6] Defined as the number of flights that depart an origin airport and land at the destination airport, without being canceled or diverted, irrespective of delays.

- **Merging `combined_df` with `airlines_df` (i.e dataframe of airlines.csv):** These datasets were joined on `Marketing_Airline_Network` and `Code` respectively, so as to translate the airline's 2-letter parent airline[7] code to its airline full name (i.e. AA to American Airlines). The `head()`, `tail()` and `duplicated().sum()` methods were used to ensure the `airlines_df` was clean for merging before-hand.
- **We merged `airport_geo_df` with our combined_df via `IATA` and `Origin` airports**. We verified via `head()`, `tail()` and `duplicated().sum()` methods that `airport_geo_df` was clean for merging with combined_df. Our intent was to extract **longitude and latitude** information for each airport in the US, so we could create visual maps accordingly.
- **A subset of `combined_df` focusing on 2021 flight/airport/airline trends was created:** This dataframe is called `sub_2021_df`, and will be used to answer parts of this project's guiding questions.
- **We merged 2021 weather data (`weather_df`) with our flight data (`combined_df`) by joining on airport code and date of flight/report**. We focused on the weather data for the top 10 airports by Flight Volume in 2021. We chose to focus on 10 airports to narrow the scope of our analysis. Each airport had it's own weather related csv file, and we filtered the day-end summary of each file (`REPORT_TYPE == 'SOD'`) to maintain 1 row per day, per airport. Numeric metrics were validated via `head()`, `tail()` and `duplicated().sum()` methods.

Findings Part 1: Competitive Landscape - Identifying Flying Patterns:

What was the landscape of the airline industry 2019-2021?



Some may remember images of March/April 2020, when the COVID-19 pandemic left airports filled with grounded airplanes and empty airports. We wanted to see if we could spot this event in the data, and to what extent. Our hypothesis was quickly validated. In March and April 2020, flight volume decreased considerably (top graph). There were over 100,000 canceled flights in each of those two months (bottom graph), up from the

[7] In the airline industry, there is a distinction between the parent airline (Marketing_Airline_Network) and the operating airline (Airline). The parent airline is the company that sells and brands the aircraft, whereas the operating airline operates and provides the flight crew. Throughout this report, our results are aggregated under the parent airline.

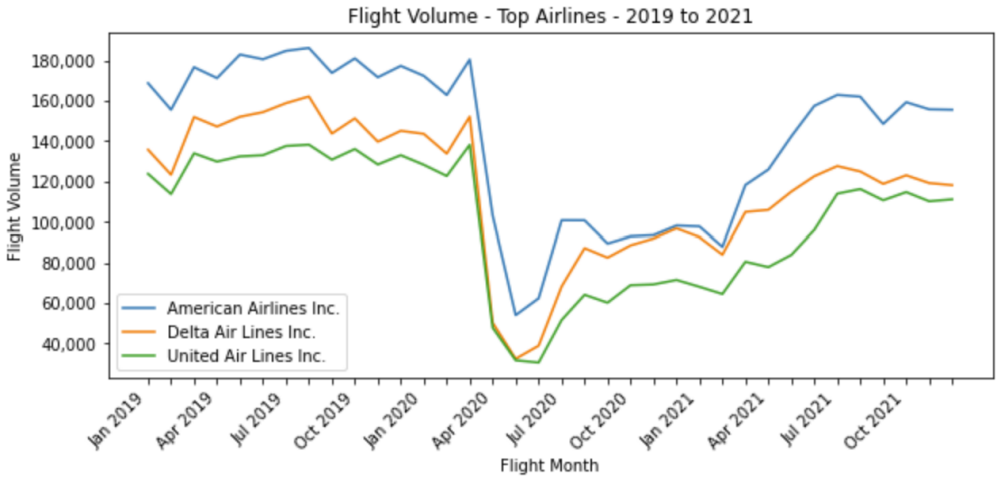
usual 10-20,000 per month. Recovery in 2021 was slow and even by the end of the year, flight volume did not recover to 2019 (pre pandemic) levels.

Between 2019 and 2021, which airlines had the highest flight volume?

	American Airlines Inc.	Delta Air Lines Inc.	United Air Lines Inc.	Southwest Airlines Co.	Alaska Airlines Inc.	JetBlue Airways	Spirit Air Lines	Frontier Airlines Inc.	Allegiant Air	Hawaiian Airlines Inc.
Flight Volume	5,096,872	4,190,263	3,605,095	3,389,862	1,085,859	644,276	531,308	363,860	319,675	198,882

Based on 2019-2021 flight volume (above), the top 3 major US airlines were: American Airlines, Delta Airlines, and United Airlines. Because these 3 airlines (American, Delta, United) are also considered the big 3 legacy airlines (CAPA, 2021) in the US, we will focus on these 3 at times in this report for deeper analysis.

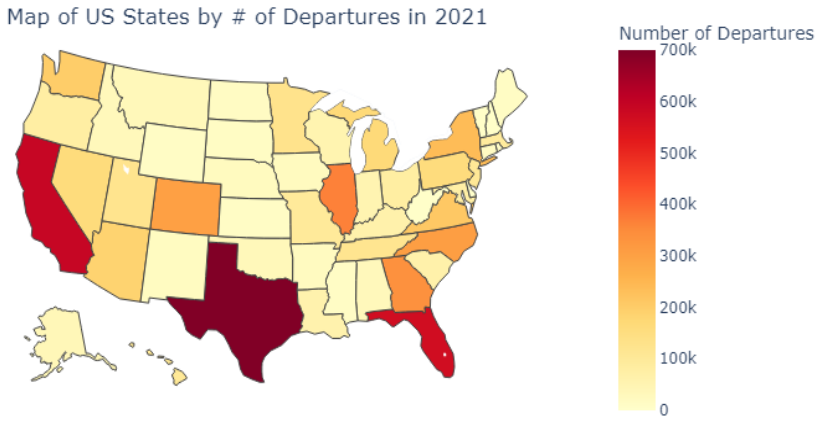
What were the flight volume trends for the big 3 legacy airlines like between 2019 and 2021?



When analyzing the flight patterns from 2019-2021 by airline, it's evident that all 3 major airlines saw a similar reduction of flights during COVID with no exceptions. Notably, American Airlines had the most flights pre-COVID, and came out with a larger lead post-COVID in flight volume. American Airlines was also more aggressive in retaining flights during the pandemic period (Mar-May 2020).

2021 Flights- A deeper look at Flight Volume by state and airport:

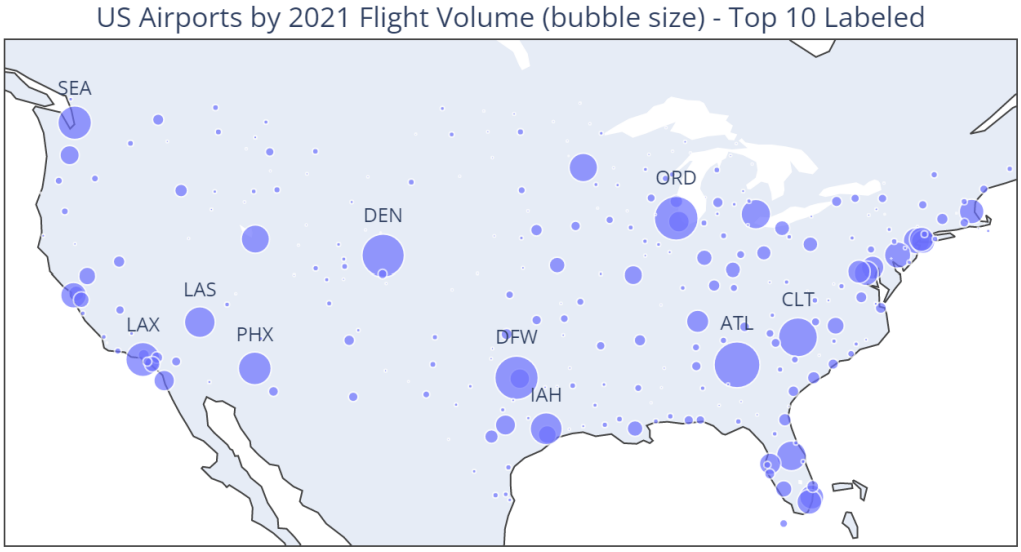
For the remainder of the report, we will focus on 2021 data due to the unusual 2020 flight patterns that all airlines went through.



In a 2021 state-level analysis, we see that flight volume generally correlates with population. Texas, California, and Florida lead the way. One surprising fact may be that the Northeastern US is outside the top 5 states by flight volume. We believe there are 3 drivers of our findings:

1. Airlines have major operations in Texas (Dallas and Houston), as seen in the next section.
2. California and Florida are popular leisure destinations, but also business travel was dormant in 2021 due to COVID (Bui, 2021).

3. The Northeast policies on COVID-19 were stricter than more conservative states, leading more travelers staying at home, forcing airlines to reduce flights into and out of the region (Wright, 2021).



	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
Flight Volume	964,289	913,567	811,024	760,594	668,731	555,774	494,333	462,777	439,336	424,150

The chart above sizes all domestic airports by flight volume (bubble size). Airports of major cities stand out, in particular ORD (Chicago), ATL (Atlanta), DFW (Dallas-Fort Worth) and DEN (Denver). As we will explore, the size of some of these airports is often driven by airline's strategies on where to establish hubs where they 'flow' their traffic through. We can see that much of the middle of America is filled with many small airports.

On the table below the map, we outlined the top 10 airports by flight volume in 2021. Moving forward, we will focus on these 10 airports to get deeper insights.

Strategic airline hub placement- A look at Flight Volume by airline & airport:

Origin	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
MajorAirline										
American Airlines Inc.	12,351	112,188	242,784	9,962	205,427	37,862	6,487	75,465	8,806	12,333
Delta Air Lines Inc.	246,399	12,188	9,889	11,487	8,576	43,770	42,620	11,467	7,537	13,749
United Air Lines Inc.	6,057	136,074	7,674	141,561	4,739	28,274	6,679	8,273	120,265	9,910

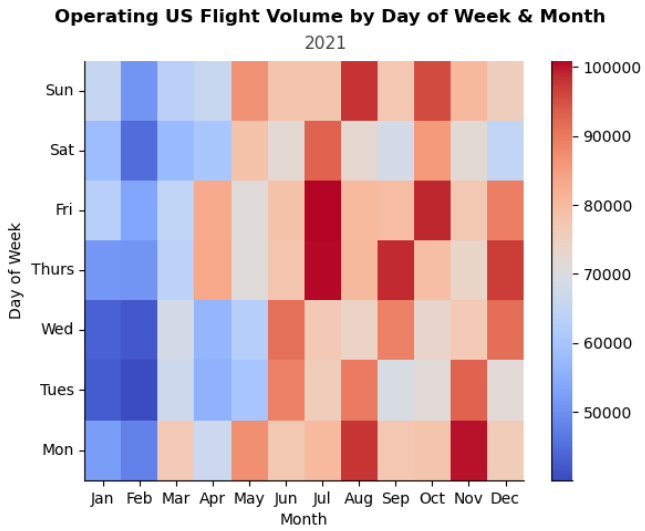
The table above shows a clear strategy in airport placement by each airline. In the airline industry, this is referred to as a 'hub', where airlines concentrate their flights to streamline operations and fill more planes. We can see a clear preference of the big 3 as follows:

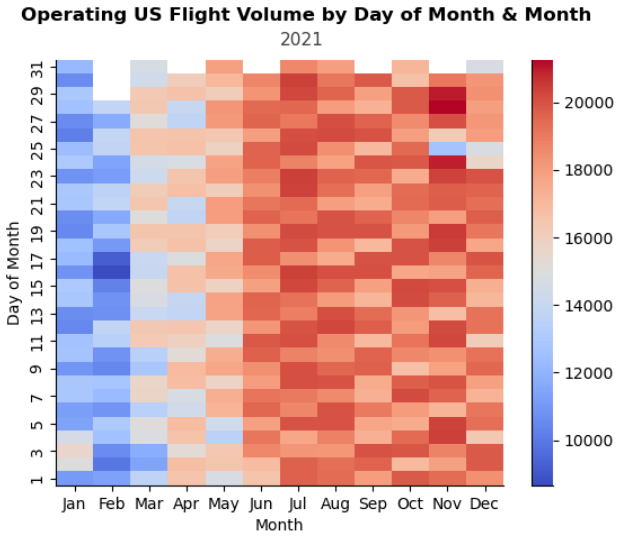
- American: DFW (Dallas Fort Worth), ORD (Chicago), CLT (Charlotte), PHX (Phoenix), LAX (Los Angeles)
- Delta: ATL (Atlanta), LAX (Los Angeles)
- United ORD (Chicago), DEN (Denver), IAH (Houston), LAX (Los Angeles)

All 3 airlines have a large presence in LAX, which seems to be a large strategic airport which no airline wants to concede. In Chicago, United and American are competing head-to-head.

What day of the week/month saw the highest Flight Volume in 2021?

The following heat maps represent the daily flight volume in 2021 based on: 1) the Day of the Week vs. Month, and 2) Day of the Month vs. Month.



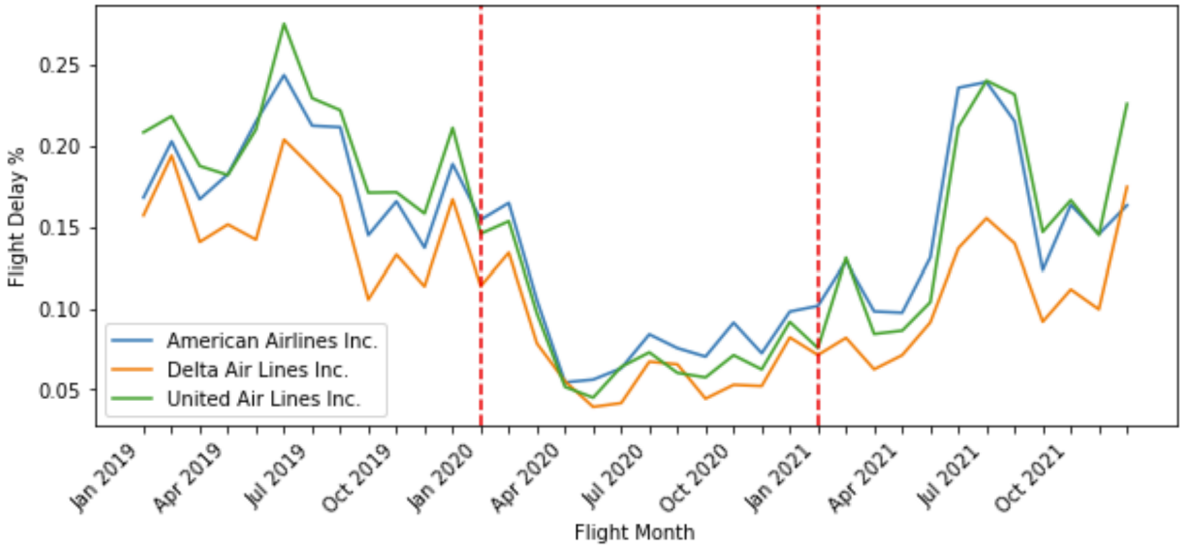


Key Observations:

- Flight volume started the year gradually, as the US population was still weary of contriving COVID, but then increased from June 2021 (reaching its peak in July) as COVID vaccines became available.
- Nov 29th 2021 (i.e Thanksgiving 2021) saw the highest number of flights taken on a given day
- The busiest travel days, in the busiest travel month (July), were Thursdays, Fridays, and Saturdays.

Findings Part 2: Delay Landscape - Evaluating Flight Delays & Cancellations

How have flight delay patterns shifted from pre vs post COVID years?



Now that we have a better understanding of the US airline market & flight patterns, next we'll be reviewing flight delay data. We wanted to understand the long term delay trends of the major airlines (American, Delta, United) in 2021. Notably, the Flight Delay %[8] significantly dropped during COVID (i.e most of 2020) as airline operations were not at full capacity (as we highlighted earlier). Unfortunately for passengers, it appears that delays slowly crept up through 2021 as airline operations resumed closer to "normal" (i.e pre COVID times).

[8] Flight Delay % = # of Flight Departures that were > 15 mins Delayed / (Total # of Flight Departures)

From a competitive standpoint, Delta is the best performing airline (Flight Delay %) of the 3 legacy airlines. United and American are in a close battle for 2nd place after Delta.

Which were the best and worst performing airlines in 2021?

	MajorAirline	2021 Avg Delay %
1	Hawaiian Airlines Inc.	8.80
2	Delta Air Lines Inc.	10.99
3	Alaska Airlines Inc.	13.24
4	American Airlines Inc.	15.93
5	United Air Lines Inc.	16.26
6	Spirit Air Lines	20.25
7	Frontier Airlines Inc.	21.81
8	JetBlue Airways	26.23
9	Allegiant Air	26.47
10	Southwest Airlines Co.	26.86

Top 3 Best Performing Airlines

1. Hawaiian Airlines
2. Delta Airlines
3. Alaska Airlines

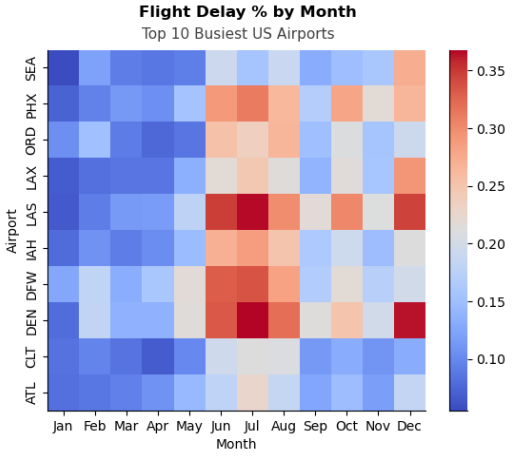
Top 3 Worst Performing Airlines

1. Southwest Airlines
2. Allegiant Air
3. JetBlue Airways

If we expand this airline analysis to all carriers in the US, we can observe that the bottom 5 performing carriers are low-cost (airlines without dedicated business cabins, cheaper airfares). In fact, Hawaiian and Alaska Airlines, sometimes considered legacy by some in the industry, also observed better on-time performance (What Are Legacy Carriers?, 2023). A passenger may want to consider this when buying their next ticket on a low-cost airline.

Which airports were more susceptible to delays in 2021?

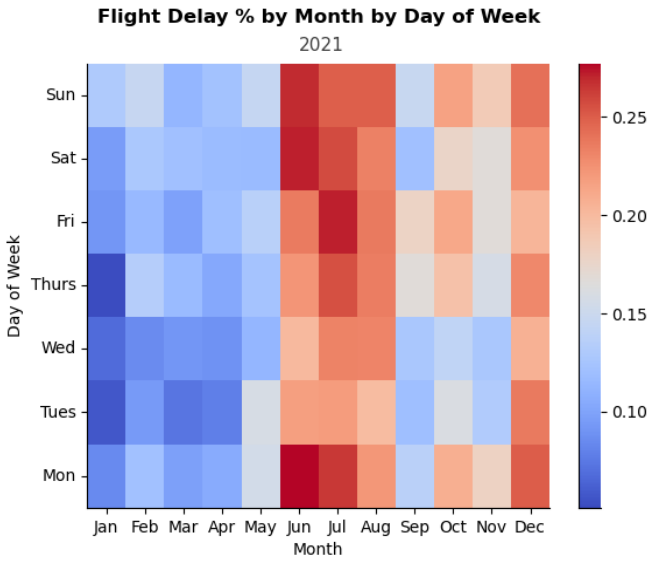
Origin	2021 Avg Delay %
DEN	22.90
LAS	22.15
DFW	20.54
IAH	19.51
PHX	18.81
ORD	18.70
LAX	17.00
SEA	15.05
CLT	14.83
ATL	13.42

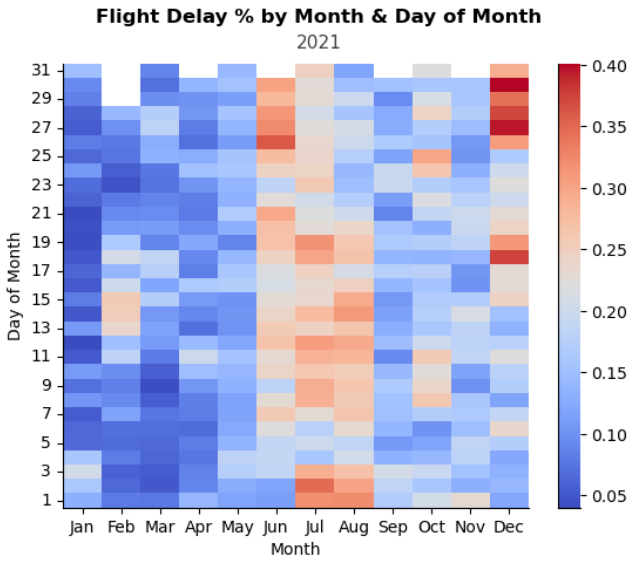


The table above shows the average delay percentage for the top 10 busiest airports in the US. DEN (Denver) had the highest percentage of flights delayed, and most airports had an increase in delays over the summer months. In 2021, the heatmap shows that LAS and DEN had the highest percentage of delays, especially in July and December.

2021 Flight Delay Distribution and Seasonality (hourly, day of week, monthly)

The following heat maps represent the percentage of delays by (1) Day of the Week and Month and (2) Day of the Month and Month. Summer months appear consistently delayed through days-of-week, while there is a higher concentration of flight delays around major US holidays.

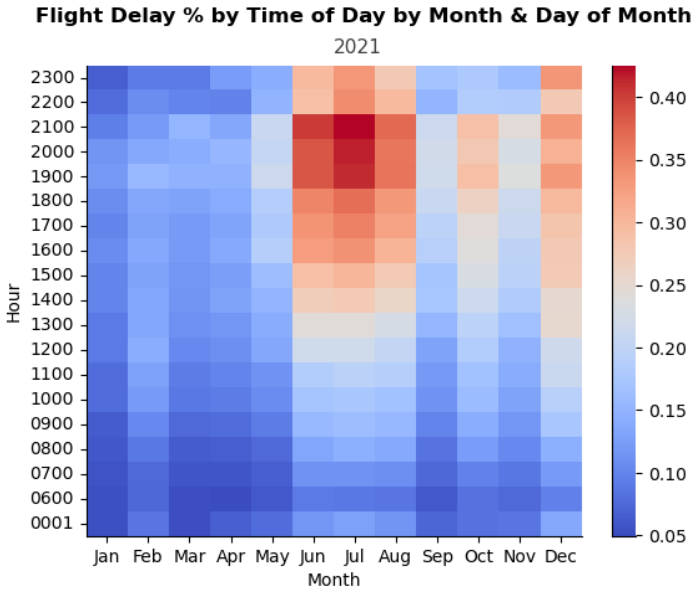




	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Flight Delay %	8.53	11.88	10.04	10.59	13.56	24.06	25.16	22.97	14.40	18.95	15.95	22.67

Key Observations

- In 2021, the weekday with the lowest percentage of delays were Thursdays in January
- The weekdays with the highest percentage of flight delays were on nearly every day in June and July. However, there were slightly less delays on Tuesdays and Wednesdays.
- There is a high concentration of flight delays during the summer months: June, July, and Aug.
- The highest percentage of flight delays were between Christmas and New Years (12/26-12/31)



Flights that departed before noon were less likely to be delayed compared to flights that departed in the afternoon and evening time (13:00-23:00).

In July, the time of day that had the highest percentage of delays was in the evening (19:00-21:00).

Findings Part 3: Weather Impacts

“Thunderstorms are nice, said no airline ever”:

airport	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
rain_flag	0.37	0.34	0.27	0.18	0.35	0.10	0.44	0.13	0.39	0.11
t_storm_flag	0.15	0.11	0.16	0.10	0.13	0.01	0.02	0.07	0.21	0.05
snow_flag	0.00	0.13	0.01	0.13	0.01	0.00	0.03	0.00	0.01	0.00
freeze_flag	0.00	0.02	0.01	0.05	0.00	0.00	0.01	0.00	0.01	0.00
fog_flag	0.06	0.04	0.04	0.08	0.07	0.11	0.03	0.01	0.16	0.00

As a bonus for this report, we wanted to understand the impact of weather events on airline operations - by downloading 10 individual files of weather reports for the top 10 airports by flight volume. The dataset provides an end-of-day summary of weather conditions for a given airport, allowing us to segment by rain, snow, thunderstorms, freezing precipitation, foggy conditions, and more. We focused on 2021.

- Denver (DEN) and Chicago (ORD) have the most days with snowfall (13% for both). They also have the most days with freezing events which can wreak havoc on airport/flight operations. Both of these airports also have a thunderstorm season. As such it seems both of these airports are exposed to poor weather conditions potentially year-round.
- Houston (IAH) has the most days with thunderstorms
- Airports in the West (LAX, SEA, LAS, PHX) avoid most of the intense weather (snow and t-storms)

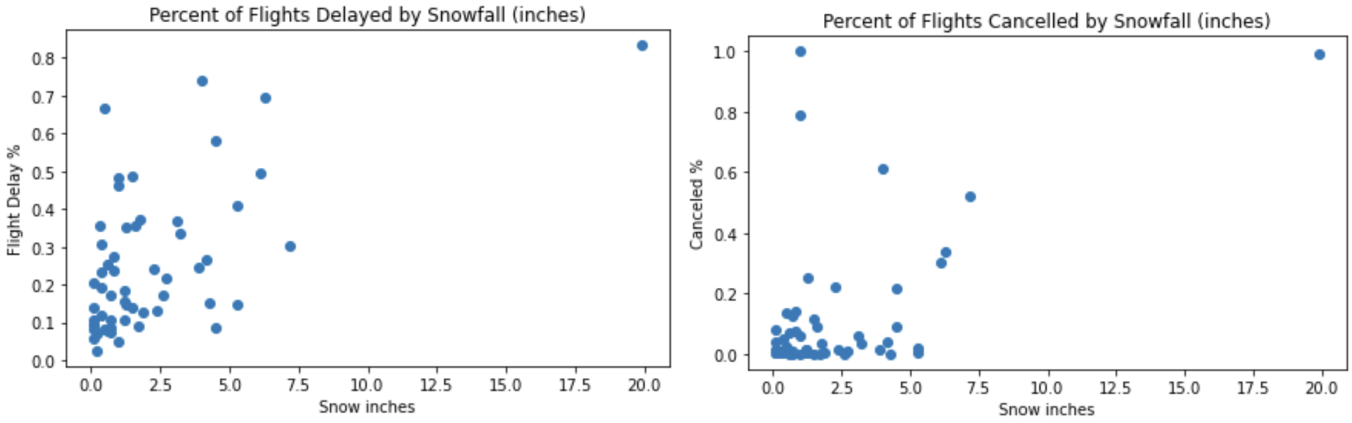
A key highlight is that **United Airlines** has a hub in three of the weather-sensitive airports we highlighted: ORD (Chicago), DEN (Denver), IAH (Houston). It is possible United’s poor on-time performance is driven by placing their hubs in airports prone to more extreme weather conditions.

Origin	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
t-storms	0.25	0.34	0.35	0.42	0.27	0.23	0.17	0.31	0.31	0.37
no t-storms	0.12	0.15	0.19	0.22	0.11	0.17	0.14	0.19	0.14	0.23

The table above shows the delay percentage on days with thunderstorms versus days without them. We focused on the top 10 airports by volume that we highlighted earlier. Most airports see a 2-3x increase in delays when there are thunderstorms. The only exceptions are west coast airports (LAX, SEA),and this is likely because thunderstorms in the West Coast are not as severe as those in the Plains, Midwest, or East Coast.

Again, unfortunately for United, we see that their hub of Denver has the highest Flight Delay % when there are thunderstorms. United’s other hubs of Chicago (ORD) and Houston (IAH) also see 30%+ delays on days with thunderstorms.

What about the impact of snow on delayed and canceled flights?



Both graphs represent days with snow > 0.0 inches.

The impact of snow (in inches) on flight delays (percent of flights delayed) is clearly evident on the left graph above. The more it snows, the more flights are delayed - and it escalates rather quickly after just a few inches. On the right, we have the percent of flights that are canceled vs snowfall amount, and we see a similar story. That said, the cancellation data is slightly more volatile.

Conclusion:

This research project looked at a core dataset of US domestic airline flight data between 2019-2021, as well as 3 additional supplementary datasets. The focus of the project was to better understand the flight trends in the US domestic air travel industry - in particular focusing on 2021, the big 3 ‘legacy’ airlines, flight delays/cancellations, flight seasonality and weather impacts.

We found many compelling answers to our questions. Whether it was understanding the impact of COVID on flight volume and cancellations, or the fact that United Airlines’ hubs are more prone to extreme weather and perhaps put United in a difficult position, we were able to extract a variety of insights from the data.

Among other things, if we could give advice to consumers looking to avoid delays, it would be the following: fly between Q1-Q2, avoid low-cost airlines, avoid DEN (Denver) and ORD (Chicago), and fly before noon.

Works Cited

Bui, Quoctrung, and Sarah Kliff. “Air Travel Is Already Back to Normal in Some Places. Here's Where.” *The New York Times*, The New York Times, 1 Apr. 2021, https://www.nytimes.com/2021/04/01/upshot/flights-rebounding-vacations.html.

Bonnie, Jin. “Airport Code and Geographical Information.” *Kaggle*, 5 Nov. 2020, https://www.kaggle.com/datasets/jinbonnie/airport-information.

CAPA - Centre for Aviation. “US Big 3 Airlines Take Varying Approaches to Recovery.” CAPA - Centre for Aviation, 10 Nov. 2021, https://centreforaviation.com/analysis/reports/us-big-3-airlines-take-varying-approaches-to-recovery-581256.

“Flight Status Prediction.” *Kaggle*, https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?resource=download.

Girl, Tensor. “Kaggle Global Trends.” *Kaggle*, 18 Apr. 2021, https://www.kaggle.com/datasets/usharengaraju/kaggle-global-trends.

“What Are Legacy Carriers?: Book Flights with Legacy Carriers.” *What Are Legacy Carriers?* | *Book Flights With Legacy Carriers*, https://www.alternativeairlines.com/legacy-carriers.

Will Wright, Juliana Kim and Troy Closson. “Covid-19 News: As U.S. Sees Encouraging News over All, Northeast Looks like an Outlier (Published 2021).” *The New York Times*, The New York Times, 18 May 2021, https://www.nytimes.com/live/2021/03/17/world/covid-19-coronavirus.

Appendix

The 10 Top most delayed airports in the US:

Origin	2021 Avg Delay %
PPG	100.00
BIH	57.14
ADK	47.92
DAL	36.31
BQN	34.95
LCK	33.16
BRW	33.14
OGD	32.32
BLV	32.15
ILG	31.90

All 61 columns and their respective dtypes for combined_df:

```
columns (total 30 columns):
Column                                  Dtype
------                                  -----
FlightDate                              object
Airline                                 object
Origin                                  object
Dest                                    object
Cancelled                               bool
Diverted                                bool
CRSDepTime                              int64
DepTime                                 float64
DepDelayMinutes                         float64
DepDelay                                float64
ArrTime                                 float64
ArrDelayMinutes                         float64
AirTime                                 float64
CRSElapsedTime                          float64
ActualElapsedTime                       float64
Distance                                float64
Year                                    int64
Quarter                                 int64
Month                                   int64
DayofMonth                              int64
DayOfWeek                               int64
Marketing_Airline_Network               object
Operated_or_Branded_Code_Share_Partners object
DOT_ID_Marketing_Airline                int64
IATA_Code_Marketing_Airline             object
Flight_Number_Marketing_Airline         int64
Operating_Airline                       object
DOT_ID_Operating_Airline                int64
IATA_Code_Operating_Airline             object
Tail_Number                             object
```

```
columns (total 31 columns):
Column                                  Dtype
------                                  -----
Flight_Number_Operating_Airline         int64
OriginAirportID                         int64
OriginAirportSeqID                      int64
OriginCityMarketID                      int64
OriginCityName                          object
OriginState                             object
OriginStateFips                         int64
OriginStateName                         object
OriginWac                               int64
DestAirportID                           int64
DestAirportSeqID                        int64
DestCityMarketID                        int64
DestCityName                            object
DestState                               object
DestStateFips                           int64
DestStateName                           object
DestWac                                 int64
DepDel15                                float64
DepartureDelayGroups                    float64
DepTimeBlk                              object
TaxiOut                                 float64
WheelsOff                               float64
WheelsOn                                float64
TaxiIn                                  float64
CRSArrTime                              int64
ArrDelay                                float64
ArrDel15                                float64
ArrivalDelayGroups                      float64
ArrTimeBlk                              object
DistanceGroup                           int64
DivAirportLandings                      float64
```

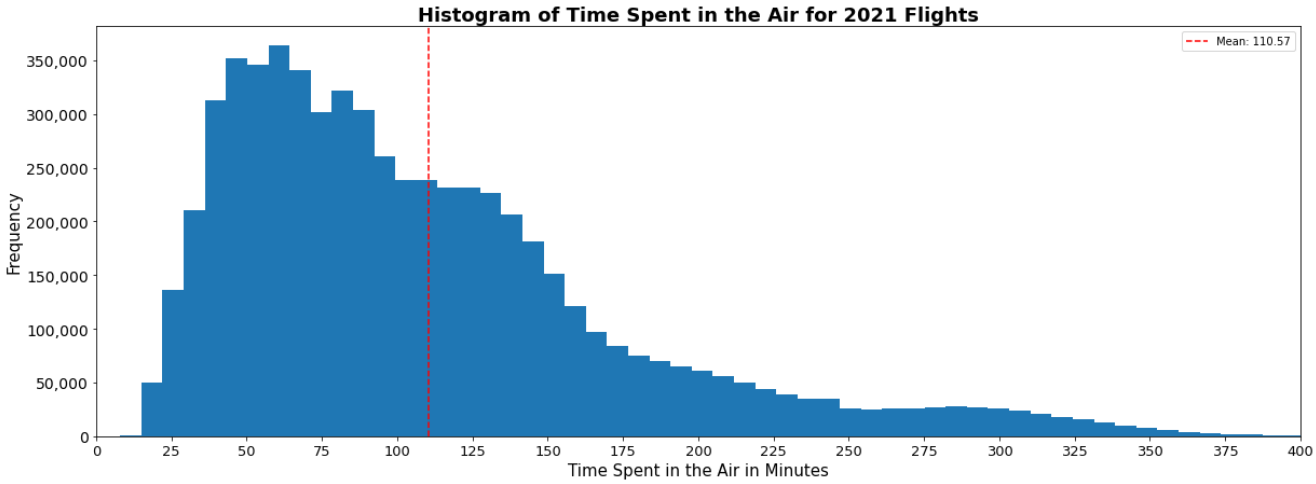
Output of combined_df.describe().T:

	count	mean	std	min	25%	50%	75%	max
CRSDepTime	19,425,952.00	1,324.37	477.93	1.00	920.00	1,317.00	1,725.00	2,359.00
DepTime	18,870,113.00	1,326.69	489.71	1.00	921.00	1,321.00	1,732.00	2,400.00
DepDelayMinutes	18,869,938.00	11.90	45.33	0.00	0.00	0.00	4.00	3,890.00
DepDelay	18,869,938.00	8.26	46.44	-131.00	-6.00	-3.00	4.00	3,890.00
ArrTime	18,854,629.00	1,471.78	517.26	1.00	1,100.00	1,506.00	1,908.00	2,400.00
ArrDelayMinutes	18,816,064.00	11.93	44.89	0.00	0.00	0.00	4.00	3,864.00
AirTime	18,816,064.00	108.36	68.41	4.00	59.00	91.00	138.00	1,557.00
CRSElapsedTime	19,425,936.00	138.18	69.91	-292.00	88.00	120.00	168.00	1,509.00
ActualElapsedTime	18,816,064.00	132.42	69.98	13.00	82.00	115.00	162.00	1,604.00
Distance	19,425,952.00	773.57	575.11	29.00	353.00	621.00	1,010.00	5,812.00
Year	19,425,952.00	2,019.91	0.86	2,019.00	2,019.00	2,020.00	2,021.00	2,021.00
Quarter	19,425,952.00	2.50	1.13	1.00	1.00	3.00	4.00	4.00
Month	19,425,952.00	6.52	3.48	1.00	3.00	7.00	10.00	12.00
DayofMonth	19,425,952.00	15.75	8.78	1.00	8.00	16.00	23.00	31.00
DayOfWeek	19,425,952.00	3.98	2.00	1.00	2.00	4.00	6.00	7.00
DOT_ID_Marketing_Airline	19,425,952.00	19,825.43	265.96	19,393.00	19,790.00	19,805.00	19,977.00	20,436.00
Flight_Number_Marketing_Airline	19,425,952.00	2,733.89	1,833.46	1.00	1,136.00	2,356.00	4,314.00	9,888.00
DOT_ID_Operating_Airline	19,425,952.00	20,009.83	386.72	19,393.00	19,790.00	19,977.00	20,378.00	21,167.00
Flight_Number_Operating_Airline	19,425,952.00	2,733.84	1,833.44	1.00	1,136.00	2,356.00	4,314.00	9,888.00
OriginAirportID	19,425,952.00	12,668.17	1,531.69	10,135.00	11,292.00	12,889.00	14,057.00	16,869.00
OriginAirportSeqID	19,425,952.00	1,266,820.76	153,169.11	1,013,505.00	1,129,202.00	1,288,903.00	1,405,702.00	1,686,901.00
OriginCityMarketID	19,425,952.00	31,757.42	1,336.68	30,070.00	30,693.00	31,453.00	32,575.00	36,101.00
OriginStateFips	19,425,952.00	27.44	16.67	1.00	12.00	26.00	42.00	78.00
OriginWac	19,425,952.00	54.83	26.19	1.00	34.00	45.00	81.00	93.00
DestAirportID	19,425,952.00	12,668.16	1,531.70	10,135.00	11,292.00	12,889.00	14,057.00	16,869.00
DestAirportSeqID	19,425,952.00	1,266,819.80	153,169.33	1,013,505.00	1,129,202.00	1,288,903.00	1,405,702.00	1,686,901.00
DestCityMarketID	19,425,952.00	31,757.46	1,336.68	30,070.00	30,693.00	31,453.00	32,575.00	36,101.00
DestStateFips	19,425,952.00	27.44	16.67	1.00	12.00	26.00	42.00	78.00
DestWac	19,425,952.00	54.83	26.19	1.00	34.00	45.00	81.00	93.00
DepDel15	18,869,938.00	0.16	0.37	0.00	0.00	0.00	0.00	1.00
DepartureDelayGroups	18,869,938.00	-0.09	2.11	-2.00	-1.00	-1.00	0.00	12.00
TaxiOut	18,862,751.00	16.56	9.24	0.00	11.00	14.00	19.00	256.00
WheelsOff	18,862,751.00	1,350.56	490.68	1.00	936.00	1,335.00	1,746.00	2,400.00
WheelsOn	18,854,610.00	1,466.41	513.49	1.00	1,056.00	1,502.00	1,902.00	2,400.00
TaxiIn	18,854,610.00	7.51	6.02	0.00	4.00	6.00	9.00	316.00
CRSArrTime	19,425,952.00	1,491.60	500.68	1.00	1,112.00	1,516.00	1,913.00	2,400.00
ArrDelay	18,816,064.00	2.25	48.41	-139.00	-16.00	-8.00	4.00	3,864.00
ArrDel15	18,816,064.00	0.16	0.37	0.00	0.00	0.00	0.00	1.00
ArrivalDelayGroups	18,816,064.00	-0.37	2.25	-2.00	-2.00	-1.00	0.00	12.00
DistanceGroup	19,425,952.00	3.57	2.26	1.00	2.00	3.00	5.00	11.00
DivAirportLandings	19,425,863.00	0.00	0.11	0.00	0.00	0.00	0.00	9.00

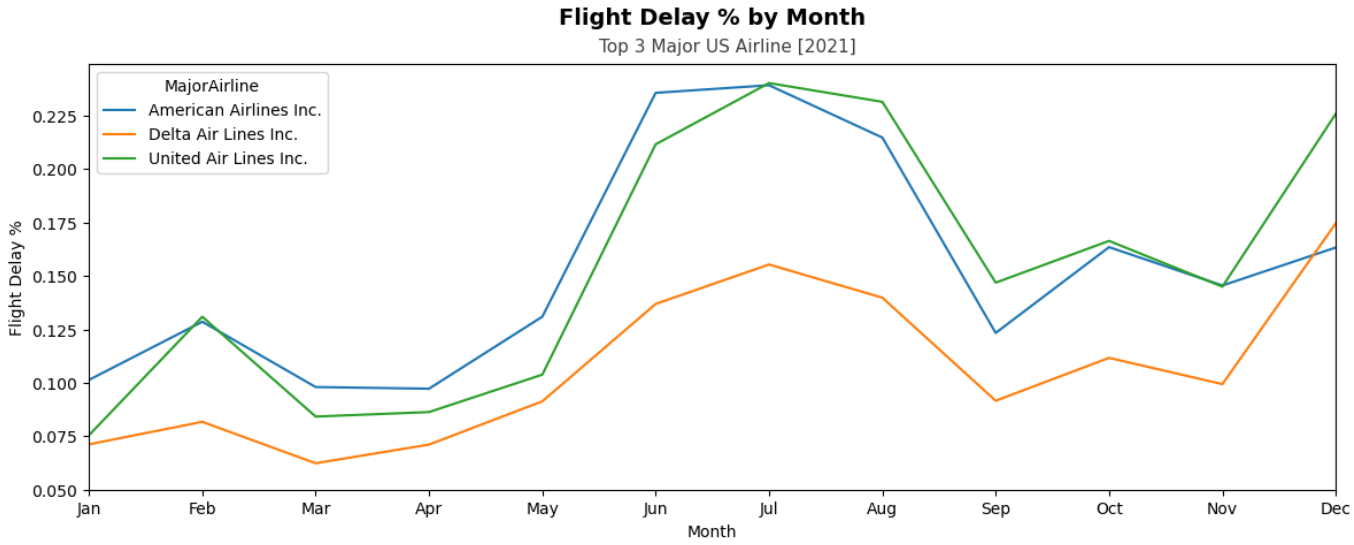
List of the 19 columns which were kept out of the 61 which came with the primary dataset (‘FlightDate’ not included here but that was kept as the index):

```
columns_to_be_kept = ['Airline', 'Origin', 'Dest', 'Cancelled',
    'Diverted', 'DepDelay', 'AirTime', 'Distance', 'Year',
    'Month', 'DayofMonth', 'DayOfWeek', 'Marketing_Airline_Network',
    'OriginCityName', 'OriginState', 'DestCityName', 'DestState', 'DepDel15', 'DepTimeBlk']
```

Histogram of Time Spent in the Air for 2021 Flights (with mean marked):



Flight Delay % in 2021 for American, Delta and United:



US Domestic Air Travel Trends & Data Analysis

Presentation by Isaac Madera, Raymond Hung and Aashray Puri

4/20/2023

Intended Audience

Datasets & Guiding Questions

Primary Dataset:

- US Domestic Flight Data - Kaggle/USDOT

Supplementary Datasets:

- Airline Names - Kaggle
- Weather Data - NOAA
- Airport Geographical Information - Kaggle/NOAA

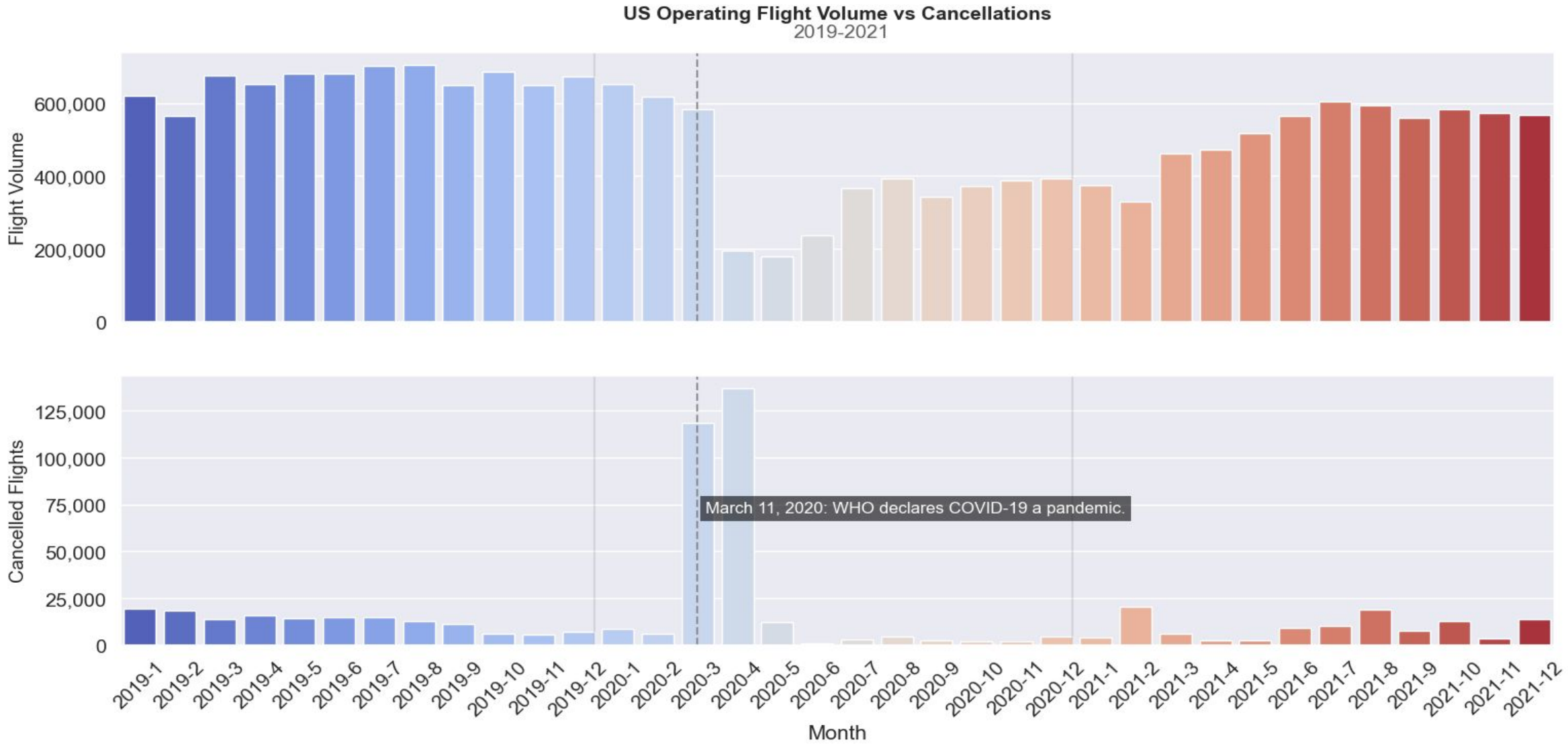
1. **Competitive landscape:**
   - How did COVID impact flight volume?
   - Which airline has the most flights in the US?
   - What are the busiest airports in the United States?
2. **Delay landscape:**
   - Which airline has the fewest delayed flights in the US?
   - Which airports are most likely to experience a delay?
   - In which month/day/hour of day are delays most common?
3. **Weather data**:
   - Which airports are more prone to experiencing weather events (rain, snow, thunderstorms, fog)?
   - What is the impact of thunderstorms and snowstorms on flight operations?

Data Cleaning & Initial Exploration

- Shape Check for consistency
- Column name and dtype check
- Grouping 2019-2021 data into 1 dataframe (combined_df).
- combined_df.duplicated().sum()
- combined_df.isnull().sum()
- combined_df.describe().T
- Dropping 41 columns and focusing on 20 columns (including new index)
- Merging combined_df with other datasets
- Making a subset of combined_df focusing on 2021

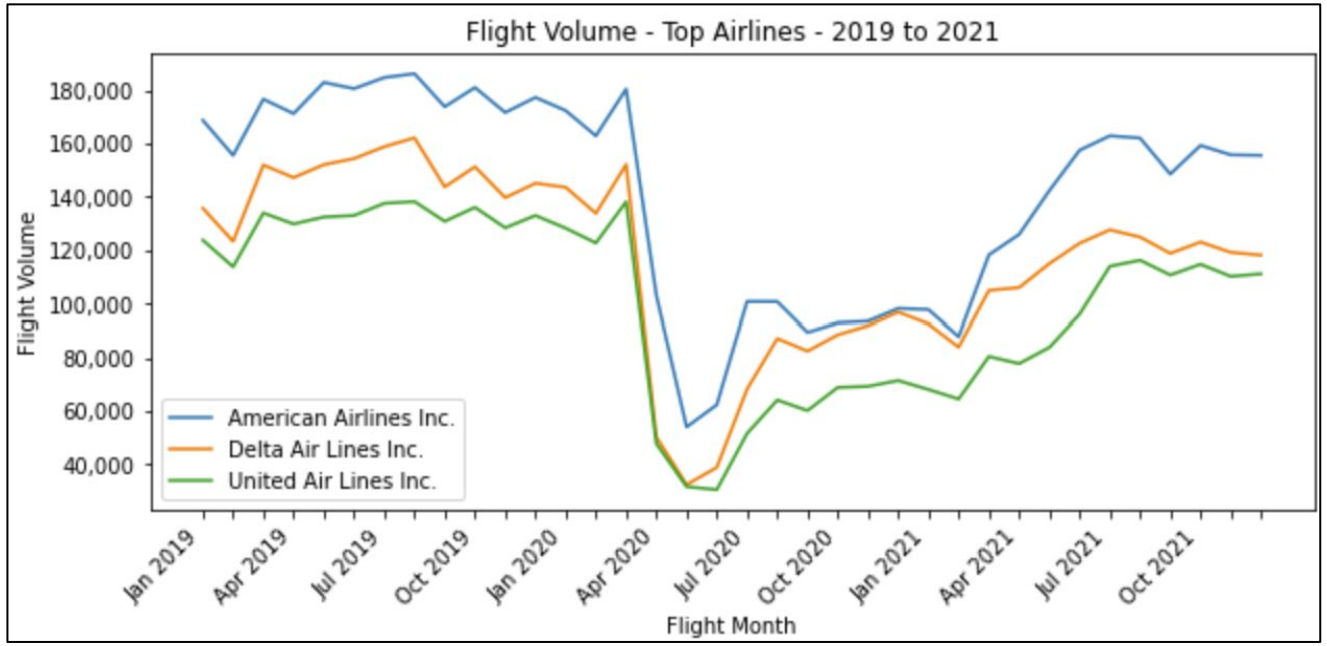
Data Cleaning

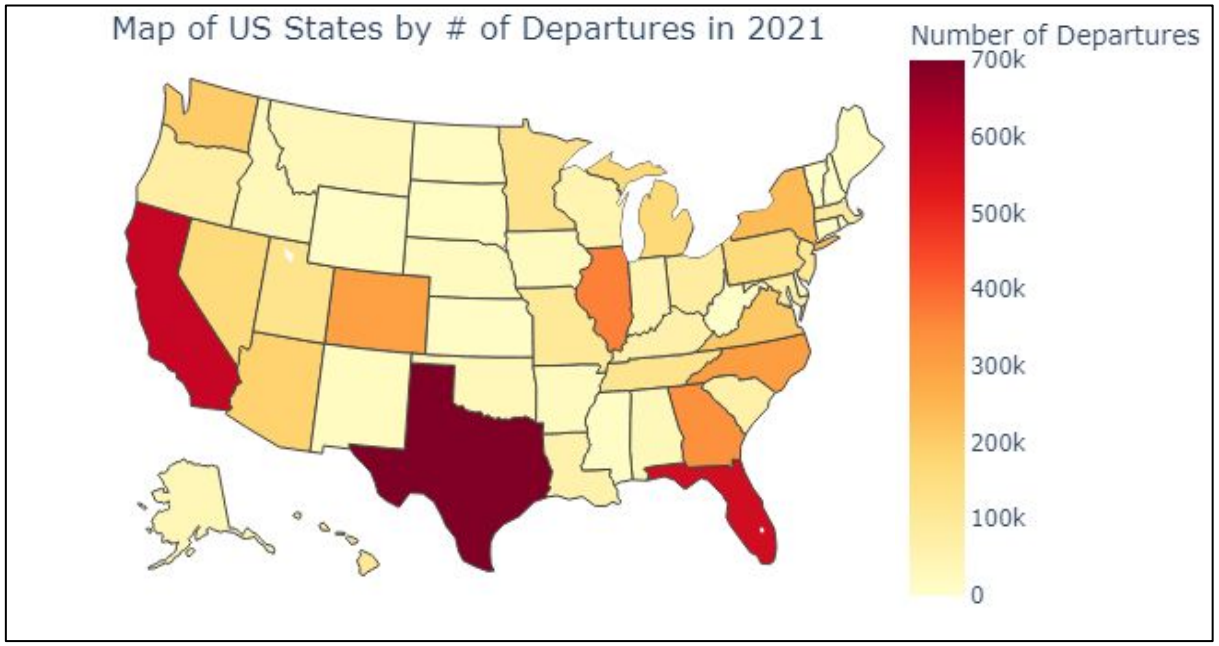
How did COVID impact Flight Volume?



Which airline has the most flights in the US?

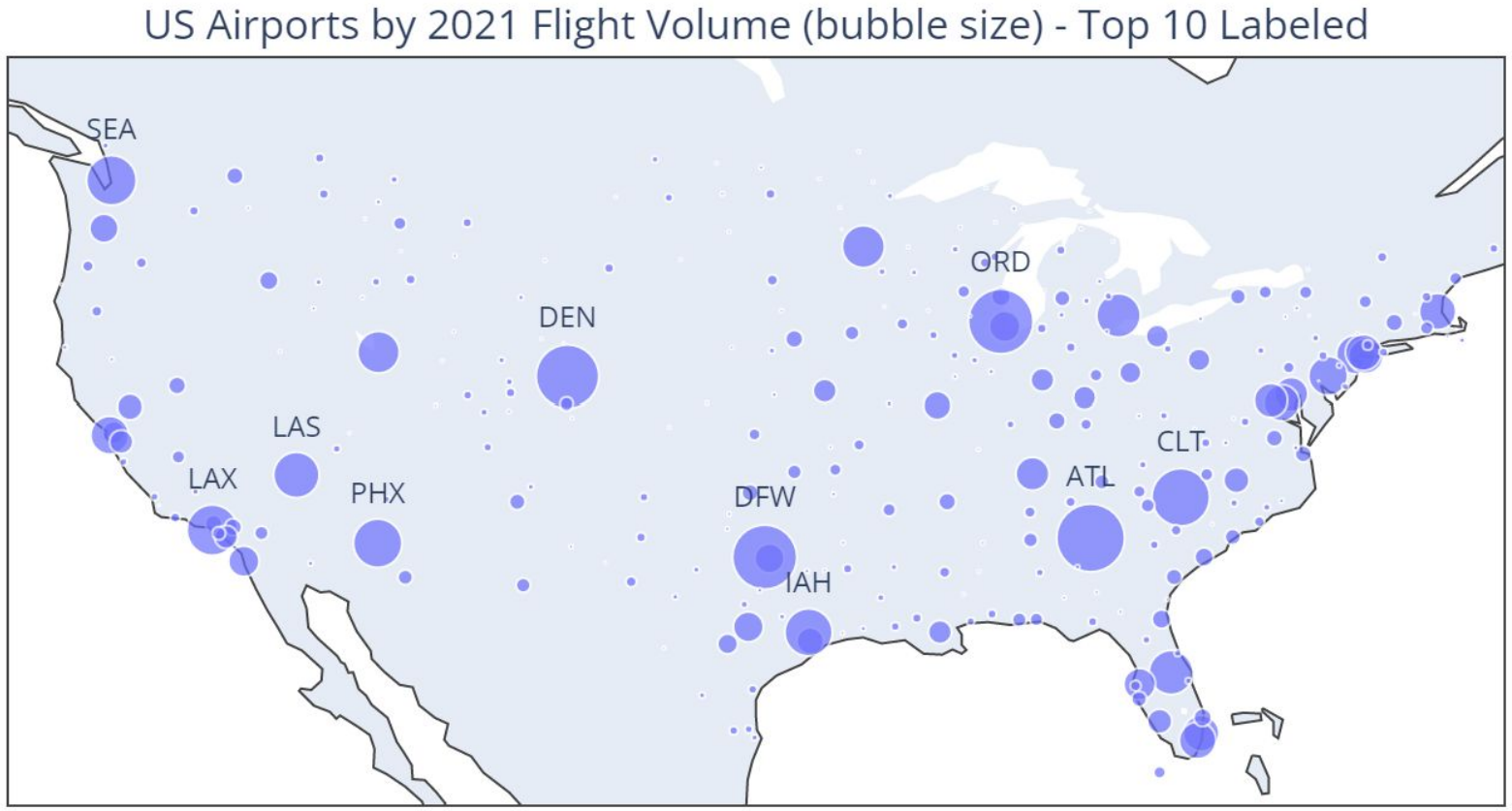
	American Airlines Inc.	Delta Air Lines Inc.	United Air Lines Inc.	Southwest Airlines Co.	Alaska Airlines Inc.	JetBlue Airways	Spirit Air Lines	Frontier Airlines Inc.	Allegiant Air	Hawaiian Airlines Inc.
Flight Volume	5,096,872	4,190,263	3,605,095	3,389,862	1,085,859	644,276	531,308	363,860	319,675	198,882





- All 3 major legacy airlines saw Flight Volume fall because of the COVID-19 pandemic.
- American Airlines saw the Highest Flight Volume 2019-2021, despite COVID.
- In 2021, Texas, California and Florida saw the highest number of flight departures

What are the busiest airports in the US?



- In 2021, **ATL (Atlanta)** was the busiest airport in the US
- A lot of traffic is concentrated on the **East** side of the country
- The size of some airports is largely because of airlines' strategic decision on where to set up their base
- Table below provides insights into airlines claiming their territory (**'hubs'** in the airline industry)

Strategic airline hub placement- A look at Flight Volume by airline & airport:

Origin	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
MajorAirline										
American Airlines Inc.	12,351	112,188	242,784	9,962	205,427	37,862	6,487	75,465	8,806	12,333
Delta Air Lines Inc.	246,399	12,188	9,889	11,487	8,576	43,770	42,620	11,467	7,537	13,749
United Air Lines Inc.	6,057	136,074	7,674	141,561	4,739	28,274	6,679	8,273	120,265	9,910

Which airline has the fewest delayed flights in the US?

	MajorAirline	2021 Avg Delay %
1	Hawaiian Airlines Inc.	8.80
2	Delta Air Lines Inc.	10.99
3	Alaska Airlines Inc.	13.24
4	American Airlines Inc.	15.93
5	United Air Lines Inc.	16.26
6	Spirit Air Lines	20.25
7	Frontier Airlines Inc.	21.81
8	JetBlue Airways	26.23
9	Allegiant Air	26.47
10	Southwest Airlines Co.	26.86

Bottom 3 Airlines

Southwest®

allegiant®

jetBlue

Top 3 Airlines

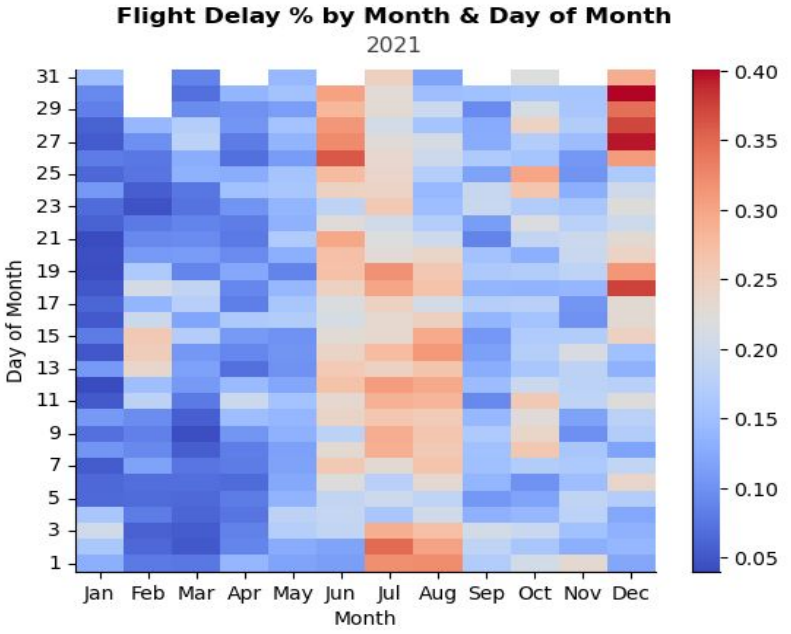
HAWAIIAN AIRLINES®

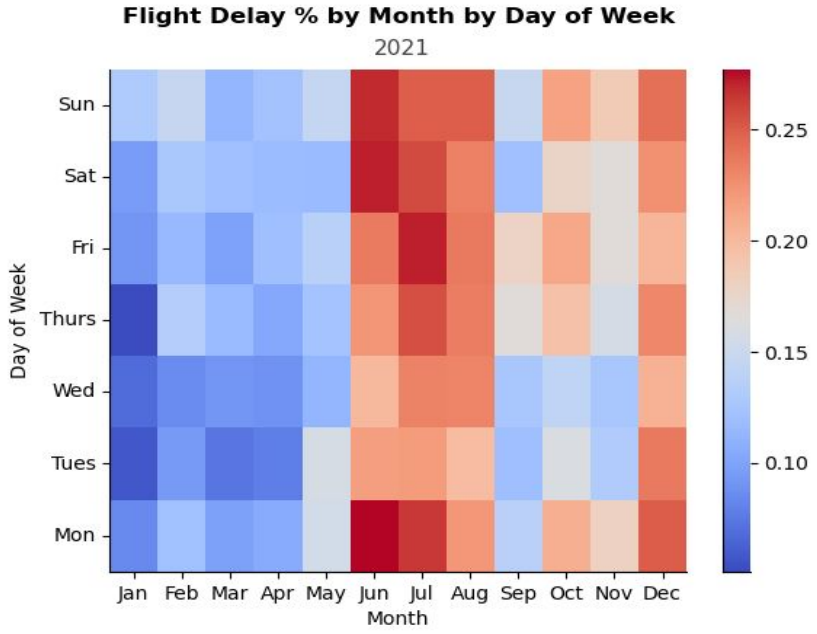
DELTA

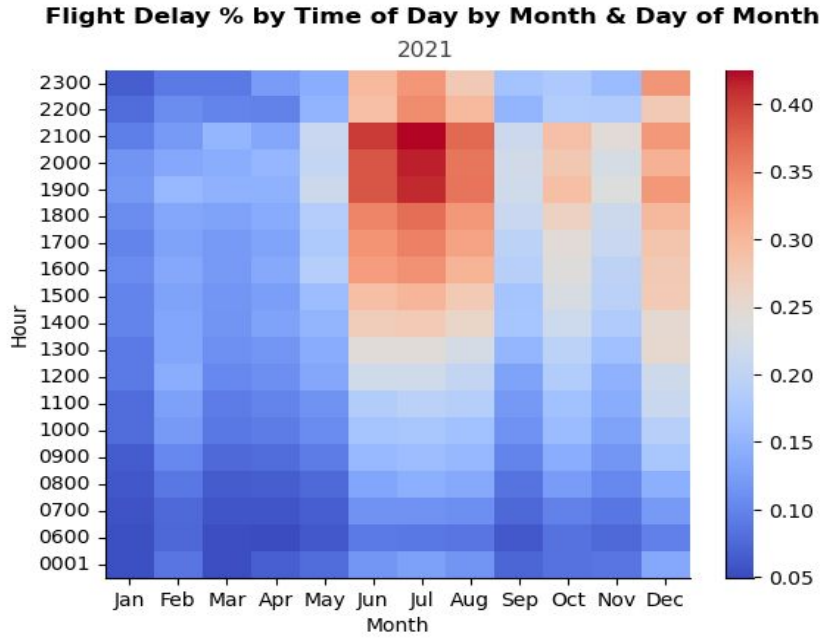
Alaska AIRLINES

- Flight Delay % = (# of Flight Departures that were > 15 mins Delayed) / (Total # of Flight Departures)
- Low cost airlines had the worst flight delay % in 2021 among all airlines, while legacy carriers performed above-average

In which months/day/hour of the week are delays most common?







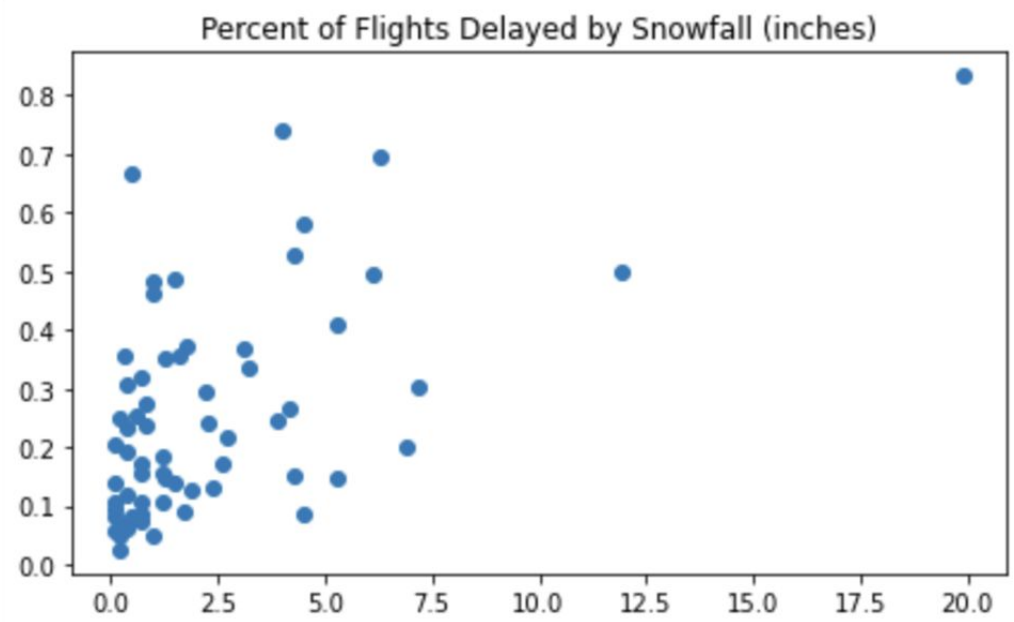
- **Summer months** had the highest Flight Delay % overall
- The worst days to fly in 2021 were **Fridays, Saturdays, Sundays, and Mondays**
- The worst week to fly in 2021 was **between Christmas and New Years**
- The worst time to fly during Summer 2021 was during the **late afternoon - evening hours**

Which airports are more prone to experiencing weather events?

airport	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
rain_flag	0.37	0.34	0.27	0.18	0.35	0.10	0.44	0.13	0.39	0.11
t_storm_flag	0.15	0.11	0.16	0.10	0.13	0.01	0.02	0.07	0.21	0.05
snow_flag	0.00	0.13	0.01	0.13	0.01	0.00	0.03	0.00	0.01	0.00
freeze_flag	0.00	0.02	0.01	0.05	0.00	0.00	0.01	0.00	0.01	0.00
fog_flag	0.06	0.04	0.04	0.08	0.07	0.11	0.03	0.01	0.16	0.00

- We leveraged airport weather data to understand likelihood of various conditions in the data
- **ORD (Chicago) and Denver (DEN)** have the most volatile weather, prone to both snowstorms and thunderstorms.
- Rain is not necessarily indicative of delays as much as thunderstorms (see Seattle)

What is the impact of snowstorms and thunderstorms on flight operations?



Flight Delay % on days with Thunderstorms vs Without

Origin	ATL	ORD	DFW	DEN	CLT	LAX	SEA	PHX	IAH	LAS
t-storms	0.25	0.34	0.35	0.42	0.27	0.23	0.17	0.31	0.31	0.37
no t-storms	0.12	0.15	0.19	0.22	0.11	0.17	0.14	0.19	0.14	0.23

- The scatterplot shows a **positive correlation** between snow (inches) and flight delays.
- Most airports are severely impacted by thunderstorms. Often **2-3x more delays when there are thunderstorms**. Only West coast airports (LAX, SEA) are not significantly affected

In Conclusion.....

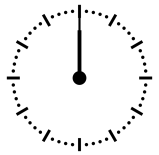
If you are a consumer who wants to maximize odds of not getting delayed:

Book a flight for Q1 or early Q2!

Avoid holiday travel!

Avoid low-cost airlines!

Stay away from Denver & Chicago!



Fly before Noon!

Summary of Findings

Top 3 Airlines by Flight Volumes 2021: American Airlines, Delta, United

Top 3 States by Flight Volumes 2021: Texas, California & Florida

Top 3 Airports by Flight Volumes 2021: ATL, ORD, & DFW

Top 3 Airlines with least Flight Delay %: Hawaiian Airlines, Delta Airlines, & Alaska Airlines

Bottom 3 Airlines with most Flight Delay %: Southwest Airlines, Allegiant Air, JetBlue Airways

Top 3 Airports with lowest Flight Delay %: ATL, CLT, SEA

Bottom 3 Airports with the highest Flight Delay %: DEN, LAS, DFW

Best Month to Fly 2021: January

Best Days of the Month to Fly 2021: 4th/5th

Best Days of the Week to Fly 2021: Tuesday/Wednesday

Best Hour of the Day to Fly 2021: 6 am

Airports in 2021 most impacted by poor weather: Denver/Chicago

Impact of Snow & Thunderstorms on Flight Operations: Both delay Flights, but Thunderstorms are worse than Snow

Works Cited

1. CAPA - Centre for Aviation. “US Big 3 Airlines Take Varying Approaches to Recovery.” CAPA - Centre for Aviation, 10 Nov. 2021, https://centreforaviation.com/analysis/reports/us-big-3-airlines-take-varying-approaches-to-recovery-581256.
2. Bonnie, Jin. “Airport Code and Geographical Information.” *Kaggle*, 5 Nov. 2020, https://www.kaggle.com/datasets/jinbonnie/airport-information.
3. “Flight Status Prediction.” *Kaggle*, https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?resource=download.
4. Girl, Tensor. “Kaggle Global Trends.” *Kaggle*, 18 Apr. 2021, https://www.kaggle.com/datasets/usharengaraju/kaggle-global-trends.

Image Links

https://www.rd.com/article/how-high-do-planes-fly/

https://www.houstoniamag.com/travel-and-outdoors/2018/07/flight-delay-tips

https://www.istockphoto.com/photo/man-waiting-for-flight-in-airport-lounge-gm862026202-143149629

https://www.kaggle.com/

https://www.noaa.gov/

https://www.transportation.gov/

https://www.vecteezy.com/vector-art/6094871-data-cleaning-in-flat-outline-icon-editable-vector