Air Travel: Pre-Pandemic vs. Post-Pandemic

Team #6 - Brian Monter and Mitchell Breeden

Dr. Huajun Zhang

George Mason University

AIT 580: *Analytics: Big Data to Information*

Abstract

The objective of our data analysis project is to show how the air travel industry as a whole was impacted by the Covid-19 pandemic, which brought the world to a halt in the middle of March 2020. We will take a look at how many passengers were still flying, how many passenger flights were still in the air, how many cargo flights were still in the air, and how each airline was impacted alone. Through different data visualization and data analysis techniques, we will model and display the data to show that the air travel industry was negatively impacted during March 2020.

Introduction

In March 2020 the world as we knew it came to a halt due to a virus that was slowly spreading across the globe, called Covid-19. As you can imagine, every bit of life as we knew it was impacted in some way or another. Our data analytics project will take a look at and analyze the impact of Covid-19 on air travel. We will compare metrics before the pandemic and during the pandemic.

Covid-19 is a disease caused by a virus called SARS-CoV-2 that spreads when an infected person breathes out droplets and very small particles that contain the virus (CDC). A few of the main reasons why it has had such a large impact on our society is because it was a new virus that no one knew anything about, it spread rapidly throughout the world, and it was more deadly than other common viruses such as influenza. Because of this, industries came to a stop with some even shutting down completely. Individuals were also given advice to stay inside and try to come in contact with the least amount of people as possible in order to mitigate the spread. A few years later and we are still going through the pandemic, but with the amount of spread and knowledge we have gained about it, normal daily life is back in full swing, even if

normal as we knew it is slightly different. This leaves Data Analysts, CEO's, Vice Presidents, and many others with questions about how to go about this new normal and how to best improve their practices for success.

Objectives

A key question many data analysts ask themselves before a data analysis project is "What do we want to show?". As mentioned above, the objective of our project will be to show how much the air travel industry was impacted by taking a look at air travel metrics before the pandemic compared to metrics during the pandemic. Specifically, we will look at:

- How many people were flying before and after the pandemic started.
- How many total flights were taken on average per day before and after the pandemic started.
- How many cargo flights were taken before and after the pandemic started.
- How specific airlines were impacted.
    - Which airlines were impacted the most in terms of flights taken compared to others.

From this, we will be able to pull insights and conclusions on how much air travel was impacted and why it was impacted.

The Dataset(s): Selection, Description, Schema, Pre-processing

Once a data analyst figures out what they are trying to show, the next key question is "What data will we use?". For our project, we chose a dataset which contains the data for all flights seen by the network's members since January 1st, 2019 and to enhance the readability of this data, we will also utilize an airline dataset, an airport dataset, and an aircraft dataset that will

be joined with our flight data. Earlier on in this course, we talked about how to choose a dataset based on a variety of questions and a few key measures called the 5 V's;

- Volume - How much data is there?
- Velocity - How fast does the data accumulate and is accessed?
- Variety - Are there different types of data?
- Veracity - How accurate is the data?
- Value - Can you convert the data into something of value?

The data that we have chosen satisfies the 5 V's. The amount of data we are analyzing is very large, once the datasets are cleaned we end up with a table with 2,113,701 rows by 18 columns. Everyday the data is accumulating as flights are being taken. There are many different parts of the data we are analyzing such as the number of passengers, number of flights, what aircrafts were impacted, and so on. The data is accurate and can be analyzed to produce insights into how to better manage the air travel industry during the pandemic.

The Crowdsourced air traffic data from The OpenSky Network 2020 | Zenodo we have chosen to use has various elements to it that will be useful to our project. The proposed system uses various features:

- **callsign:** the identifier of the flight displayed on ATC screens (usually the first three letters are reserved for an airline: AFR for Air France, DLH for Lufthansa, etc.).
- **number:** the commercial number of the flight, when available (the matching with the callsign comes from public open API).
- **icao24:** the transponder unique identification number.
- **registration:** the aircraft tail number (when available).
- **typecode:** the aircraft model type (when available).

- **origin:** a four letter code for the origin airport of the flight (when available).
- **destination:** a four letter code for the destination airport of the flight (when available).
- **firstseen:** the UTC timestamp of the first message received by the OpenSky Network.
- **lastseen:** the UTC timestamp of the last message received by the OpenSky Network.
- **day:** the UTC day of the last message received by the OpenSky Network.
- **latitude_1, longitude_1, altitude_1:** the first detected position of the aircraft.
- **latitude_2, longitude_2, altitude_2:** the last detected position of the aircraft.

As mentioned, to enhance the readability of this data, we also utilized an airline dataset, an airport dataset, and an aircraft dataset that were joined with our flight data. The features are as follows:

Airlines

- **Name:** Name of the airline.
- **IATA:** 2-letter IATA code (identifier).
- **ICAO:** 3-letter ICAO code, if available.
- **Callsign:** Airline callsign.
- **Country:** Country or territory where airline is incorporated.

Airports

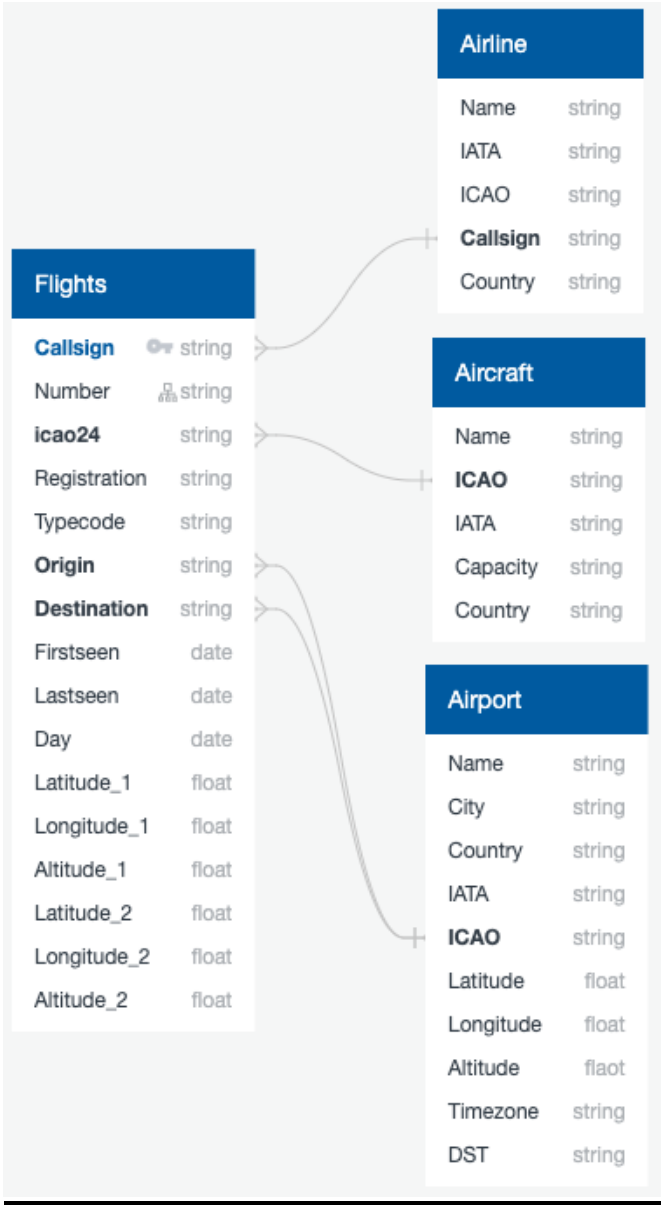
- **Name:** Name of airport. May or may not contain the City name.
- **City:** Main city served by airport. May be spelled differently from Name.
- **Country:** Country or territory where airport is located.
- **IATA:** 3-letter IATA code (identifier).
- **ICAO:** 4-letter ICAO code.

- **Latitude:** Decimal degrees, usually to six significant digits. Negative is South, positive is North.
- **Longitude:** Decimal degrees, usually to six significant digits. Negative is West, positive is East.
- **Altitude:** In feet.
- **Timezone:** Hours offset from UTC.
- **DST:** Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown).

Aircrafts

- **Name:** Name of the aircraft.
- **ICAO:** 4-letter ICAO code, if available.
- **IATA:** 3-letter IATA code (identifier).
- **Capacity:** Seating capacity (empty for cargo aircrafts).
- **Country:** Country or territory where aircraft maker is incorporated.

Our system relied on the flight dataset for all of our analytical information while the airline, airport, and aircraft datasets will provide supplemental information for readability by the end user.



Involving four different datasets required some cleaning and joining to create a dataset that fit the scope of what we were trying to figure out and predict, our steps went as followed:

1. Import two packages, pandas and datetime.
2. Adjust the DataFrame width.
3. Import the flight data and drop unneeded columns
4. Get the Airline symbol from callsign, shorten the day to YY-MM-DD, and fill nulls.

5. Import the Airline, Airport (origin), Airport (destination), and Aircraft data and name/drop unneeded columns and duplicates.
6. Merge flightsDF with Origin_AirportsDF to df.
7. Merge df with Destination_AirportsDF to df.
8. Merge df with airlinesDF to df.
9. Merge df with aircraftsDF to df.
10. Replace unmatched Airlines with Callsign, replace ENY with Envoy Air (Major airline missing from airline file)
11. Drop, reorder, and rename columns.
12. Fill nulls and convert to string, then fill nulls and convert to int.
13. Convert to First_Seen to datetime.
14. Add Count column.
15. Find Unique Airlines.
16. Find Cargo Airlines.
17. Add Cargo Airlines to the list.
18. Filter df for Cargo Airlines.
19. Filter df for Passenger flights.
20. Run the code.

After the datasets were cleaned and joined, we were left with a dataset that had 2,113,701 rows by 18 columns as mentioned above. Those columns include: Day, Airline, Aircraft, Airline_Country, Aircraft_Capacity, First_Seen, Origin_Lat, Orgin_Long, Origin, Origin_Name, Origin_City, Origin_Country, Last_Seen, Destination, Destination_Name, Destination_City,

Destination_County, and Count. This data presents all of the flights taken within the month March 2020.

The System: Architecture, Data Processing, Data Analytics Algorithm(s), SW/HW Development Platforms

As explained above, our system relied on the flight dataset for all of our analytical information while the airline, airport, and aircraft datasets will provide supplemental information for readability by the end user. Our pre-processing involved cleaning and joining these datasets into one dataset that can be used to gain insights about the questions we are posing.

Once the data was cleaned, we began visualizing and analyzing the data through Python. Below I will take you through the steps of the code for each visualization and analysis we did. First the data visualization:

1. Create a map to show all flights where they were first seen at different points during the pandemic, the steps are as followed;
    1.1. Import packages GeoPandas and KeplerGl.
    1.2. Define GeoPandas DataFrame.
    1.3. Remove unnecessary columns.
    1.4. Define the map.
    1.5. Display the map.
2. Create a line chart to show how the total number of flights changed over time, the steps are as followed;
    2.1. Import package matplotlib.pyplot.
    2.2. Set the figure size.
    2.3. Initialize a dictionary.

2.4. Create daily count DataFrame.

2.5. Update dictionary with the daily flight count.

2.6. Define X and Y.

2.7. Format and plot the chart.

3. Create two bar chart races for the number of flights at top airlines and the number of flights at top airports pre-pandemic and post-pandemic (looking at March 2020), the steps are as followed;

3.1. Import packages numpy, bar_chart_race, and IPython.display.

3.2. For loop for Airlines and Airports.

3.2.1. Find counts for each airline/airport.

3.2.2. Shrink to airline/airport and count.

3.2.3. Remove N/A.

3.2.4. Sort for top 10 airlines and airports.

3.2.5. Convert to list.

3.2.6. Filter flightsDF for top airlines/airports.

3.2.7. Pivot data.

3.2.8. Change origin_name to airport.

3.2.9. Create and display bar chart races.

Next the analysis:

1. Output the daily average of passenger flights, cargo flights, and passengers flown for March 2020, the steps are as followed;

1.1. Define the function.

1.1.1. Define daily count dictionary.

1.1.2. Use an if-else statement.

1.1.2.1. Find the number of daily flights and update dictionary (if)

1.1.2.2. Find the daily capacity and update dictionary (else).

1.1.3. Initialize variables.

1.1.4. Split data based on before or after March 15th, 2020 and count the number of days.

1.1.5. Find daily average

1.1.6. Find the percent difference.

1.1.7. Print outputs.

1.2. Call function for each DataFrame and print outputs.

2. Compare how each Airline was impacted and output the difference in flights pre-pandemic versus post-pandemic, the steps are as followed;

2.1. Filter flights data frame to just US based airlines.

2.2. Split the new function to before March 15th and after March 15th.

2.3. Count the number of flights pre-pandemic and format.

2.4. Count the number of flights post-pandemic and format.

2.5. Merge the pre-pandemic data and the post-pandemic data.

2.6. Calculate difference between the two.

2.7. Calculate the percent difference between the two.

2.8. Filter to only airlines with more than 100 flights pre-pandemic.

2.9. Sort by difference and percent difference.

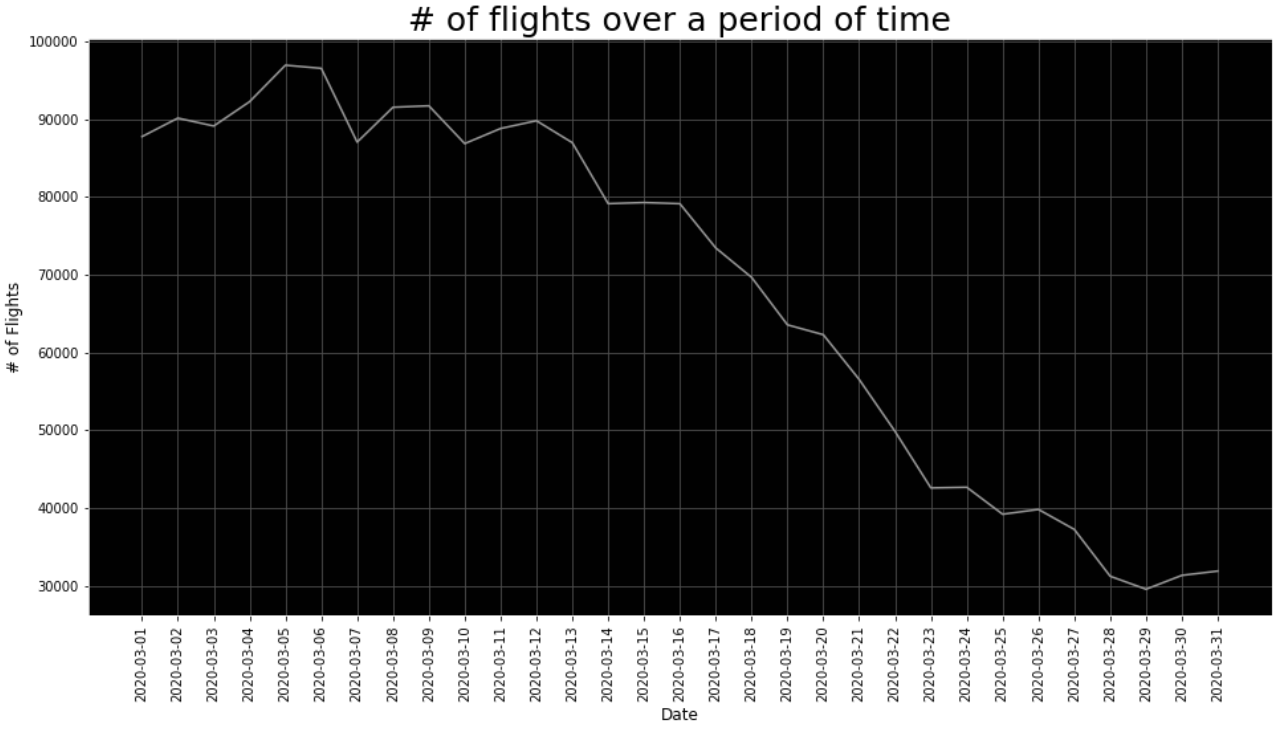
2.10. Display.

3. Compare the number of flights in March 2020 versus the number of flights predicted for March 2020 and 30 days after, the steps are as followed;
    3.1. Import packages linear_model, train_test_split, LinearRegression, and metrics.
    3.2. Set figure size.
    3.3. Format datetime column.
    3.4. Create time series.
    3.5. Initiate X and y.
    3.6. Create test and train sets.
    3.7. Create and fit model.
    3.8. Print intercept and slope.
    3.9. Plot flights.
    3.10. Find predicted values.
    3.11. Initiate X and y.
    3.12. Plot the line of best fit/prediction.
    3.13. Plot settings.
    3.14. Display.

Experimental Results and Analysis: Explore and Present Analysis of the dataset, Prepare relevant analysis and visualizations, Interpret the results

As mentioned, our main goal is to show how much the air travel industry was impacted by Covid-19. We produced four different data visualizations in Python in order to obtain a better understanding of the data.

To start, we created a line chart by importing matplotlib.pyplot to see how the total number of flights changed over the month of March 2020.



As you can see in the chart above, during the month of March 2020 the amount of flights decreased by almost 70,000 flights when you compare the peak to the minimum. This gives us a glimpse into the overall impact Covid-19 had on the industry instantly, with there being a sharp decrease in the amount of flights it is safe to predict there were less total number of people flying too.

The next way we chose to visualize pieces of our data was by showing a map of all flights by importing GeoPandas and KeplerGl where they were first seen at different points during the pandemic, illustrating the number of flights taking place on a given day (shown in appendix).

Finally, we created two bar chart race videos (shown in appendix). One of which showed the total number of flights for top airlines around the world during March 2020 and the other

which showed the number of flights at top airports around the world during March 2020. The bar chart race of the total number of flights for the top airlines helped show which airlines were able to continue with business as usual and which airlines were severely impacted by the pandemic instantly. As you watch the video, all of the top airlines begin to decrease the amount of flights each day starting on March 18th. Before March 18th, the airlines seemed to have a normal fluctuation of flights per day but on March 18th is where you see the first significant drop for the top airlines around the world. This is fascinating because the day that the impact started to be felt was March 15th as you will see later on. The bar chart race for the total number of flights at top airports shows you the same thing, whether or not top airports were able to continue on business as usual or were impacted by the pandemic. As you could guess, top airports were also impacted on about March 18th. Although the decrease for the top airports seems to be at a slower rate compared to the top airlines. This helps us answer the question of how specific airlines and the overall industry was impacted. The visualizations helped paint a picture and give us an idea of what we can expect for our results, but in order to provide meaningful information to the appropriate people, more analysis in Python is required.

For our descriptive analytics, we took a look at the number of passenger flights taken on average per day before and after the pandemic started, the number of cargo flights taken on average per day before and after the pandemic started, and the average number of daily passengers before and after the pandemic started. After comparing the numbers for before versus after, we wanted to show how much or how little of a change there was by providing the percent change. Our data provided these results:

Descriptive Analytics	**Average daily number of passenger flights**	**Average daily number of cargo flights**	**Average daily number of passengers**
Before March 15th, 2020	88,899.67 flights	1,214.93 flights	6,083,030.93 passengers
After March 15th, 2020	48,762.88 flights	1,264.50 flights	3,141,443.12 passengers
Percent Change	-45.15%	4.08%	-48.36%

As you can see in the table above, the total number of passenger flights and total number of passengers decreased at about the same rate. It was expected that the start of Covid-19 would have a negative impact on the air travel industry, but it is interesting to see that the total number of passenger and passenger flights essentially cut in half. There are not many situations or events in the world that could cause this type of impact on an industry, but as shown, a deadly virus can do just that. Another interesting take away from the table is that cargo flights actually increased in the month of March after the pandemic started. There could be many reasons for this such as more aircrafts being available to use, online shopping increasing, more need to get supplies related to the pandemic to people, and since cargo flights primarily carry cargo, there was not a substantial risk to human health.

For our inferential analytics, we took a look at how specific airlines were impacted during March 2020 in terms of how many flights were taken, and compared those numbers to each other and the industry as a whole. Instead of taking a look at all of the airlines, we narrowed it down to just take into account US airlines and airlines that had greater than 100 flights pre-pandemic. We took the pre-pandemic count of flights taken by those airlines and compared them to the post-pandemic count to find the total difference and the percent difference. Our pre-

pandemic count consists of all the flights taken between March 1st, 2020 to March 15th, 2020, while our post-pandemic count consists of all the flights taken between March 16th to March 31st, 2020.

Airline	Pre_Count	Post_Count	Difference	Percent
Delta Air Lines	45737.0	28192.0	-17545.0	-38.36%
American Airlines	44630.0	29432.0	-15198.0	-34.05%
United Airlines	31512.0	18798.0	-12714.0	-40.35%
Southwest Airlines	53973.0	41785.0	-12188.0	-22.58%
SkyWest	35277.0	29421.0	-5856.0	-16.60%
JetBlue Airways	14526.0	9808.0	-4718.0	-32.48%
Republic Airlines	14673.0	11251.0	-3422.0	-23.32%
Mesa Airlines	9794.0	7263.0	-2531.0	-25.84%
Allegiant Air	5673.0	3292.0	-2381.0	-41.97%
Atlantic Southeast Airlines	6187.0	4282.0	-1905.0	-30.79%
Frontier Airlines	5867.0	4081.0	-1786.0	-30.44%

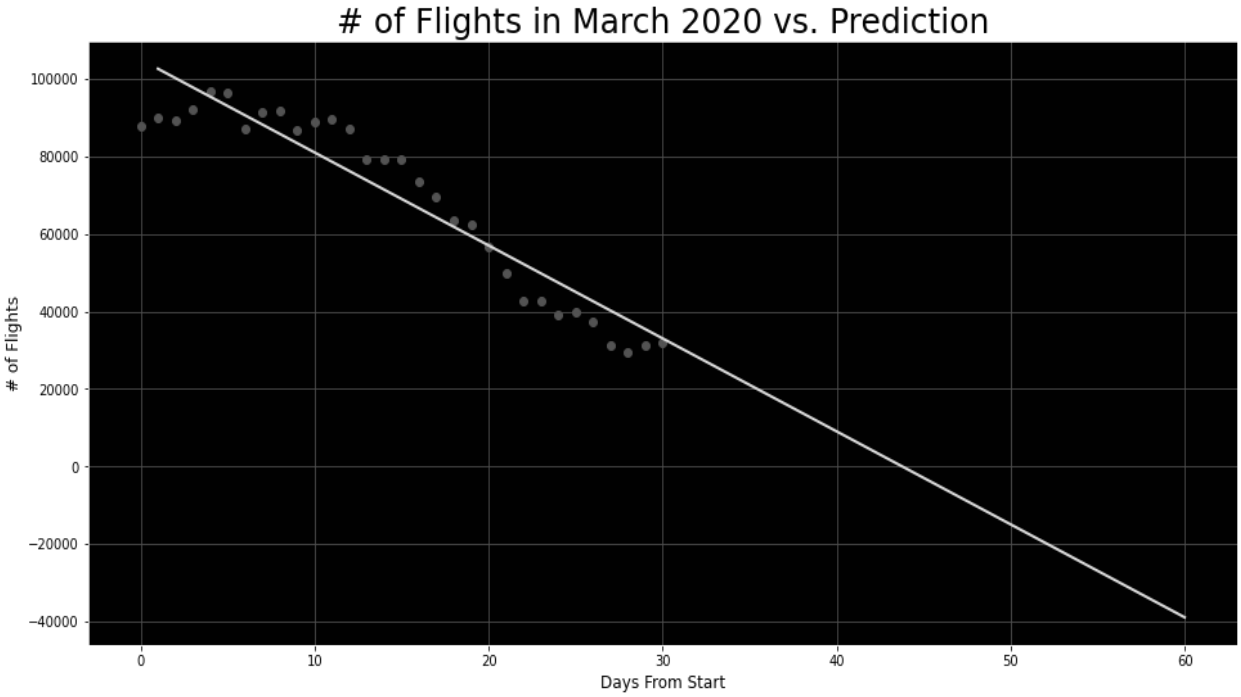
Airline	Pre_Count	Post_Count	Difference	Percent
Air Wisconsin	4531.0	3201.0	-1330.0	-29.35%
Spirit Airlines	10129.0	8978.0	-1151.0	-11.36%
Horizon Air	5209.0	4140.0	-1069.0	-20.52%
Trans States Airlines	3121.0	2182.0	-939.0	-30.09%
Piedmont Airlines (1948-1989)	5268.0	4356.0	-912.0	-17.31%
CommutAir	2506.0	1682.0	-824.0	-32.88%
Sun Country Airlines	1452.0	850.0	-602.0	-41.46%
Hawaiian Airlines	862.0	607.0	-255.0	-29.58%
Cape Air	1658.0	1455.0	-203.0	-12.24%
Arctic Circle Air Service	186.0	49.0	-137.0	-73.66%
Air Mobility Command	350.0	220.0	-130.0	-37.14%
Harbor Airlines	107.0	67.0	-40.0	-37.38%
Amerijet International	222.0	206.0	-16.0	-7.21%

Airline	Pre_Count	Post_Count	Difference	Percent
Bemidji Airlines	169.0	167.0	-2.0	-1.18%
Alaska Central Express	331.0	342.0	11.0	3.32%
Omni Air International	145.0	157.0	12.0	8.28%
Alpine Air Express	127.0	157.0	30.0	23.62%
ABX Air	605.0	666.0	61.0	10.08%
Atlas Air	1521.0	1642.0	121.0	7.96%
Kalitta Air	693.0	845.0	152.0	21.93%

What we found was that out of the 30 airlines that fit into our criteria (US-based and greater than 100 flights pre-pandemic), 24 of them showed a decrease in the amount of flights after March 15th compared to 6 of them that showed an increase in the amount of flights after March 15th. One reason those 6 flights may have shown an increase in the number of flights is because those particular airlines were mainly used for cargo flights. For example, out of the 30 airlines, the airline with the highest percent increase was Alpine Air Express. Alpine Air Express is known as one of America's largest all cargo airlines. If we take a look at the airline that increased the most in terms of the total count, that airline would be Kalitta Air. Kalitta Air is also one of the largest cargo airlines in the United States. Now let's take a look at the 24 airlines that showed a decrease in the amount of flights, as you go through the list the main thing that pops out is that many of

the larger well-known airlines decreased at a high rate. You also see that although it impacted the whole air travel industry, there were big differences in how each airline specifically was impacted. Two of the largest airlines we have in the United States are Southwest Airlines and United Airlines, both of which saw about a 12,000 flight decrease when comparing the first two weeks of March 2020 to the second two weeks of March 2020. But when you look at the percent change of those two airlines it is vastly different due to the volume of flights each of the airlines usually has. Southwest Airlines decreased by 22.58%, whereas United Airlines decreased by 40.35%. By comparing those two percentages, you can see that the way a specific airline took action had an impact on their overall business.

For our advanced analytics, we did a linear regression line and took a look at the number of flights in March 2020 versus what was predicted for the number of flights in March 2020 and the following 30 days.



As you can see in the graph above, the predicted number of flights agrees with the actual number of flights throughout March 2020. Where this model goes wrong is that the number of flights could never be negative, as you can see a few weeks after March this is what our line is telling us.

Conclusions

The objective of our data analysis project was to show how the air travel industry as a whole was impacted by the Covid-19 pandemic which brought the world to a halt in the middle of March 2020. We took a look at how many passengers were still flying, how many passenger flights were still in the air, how many cargo flights were still in the air, and how each airline was impacted alone. What we found was not at all surprising, the air travel industry had been severely impacted in a negative way, although one part of it continued to thrive. The amount of passengers taking flights places almost cut in half throughout March 2020. The average daily number of flights decreased by even more than that. What came as a surprise was that cargo flights kept on with business as usual. Lastly, most airlines were negatively impacted and some airlines were able to go on with business as usual, those airlines as you could probably guess, were mainly cargo flights. Overall, the pandemic negatively impacted the air travel industry in ways we have not seen before.

References

Coronavirus (COVID-19) frequently asked questions | CDC

Sirangelo, Cristina. http://www.lsv.fr/~sirangel/teaching/dataset/index.html

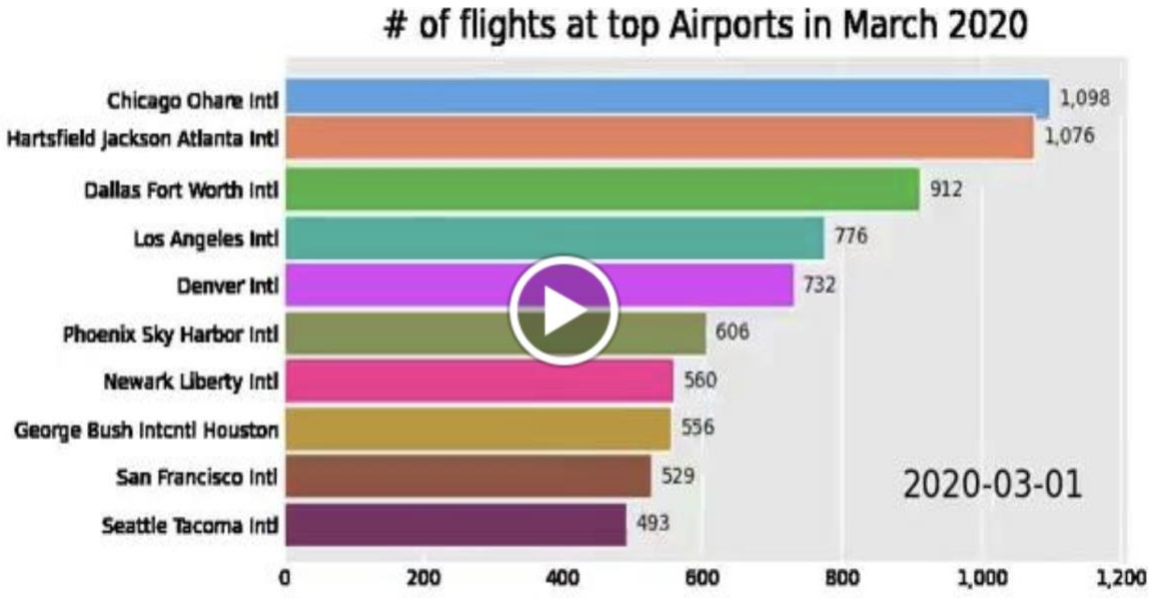
Xavier Olive, Martin Strohmeier, & Jannis Lübbe. (2022). "Crowdsourced air traffic data from The OpenSky Network 2020" (v22.01) [Data set]. Zenodo, Crowdsourced air traffic data from The OpenSky Network 2020 | Zenodo.

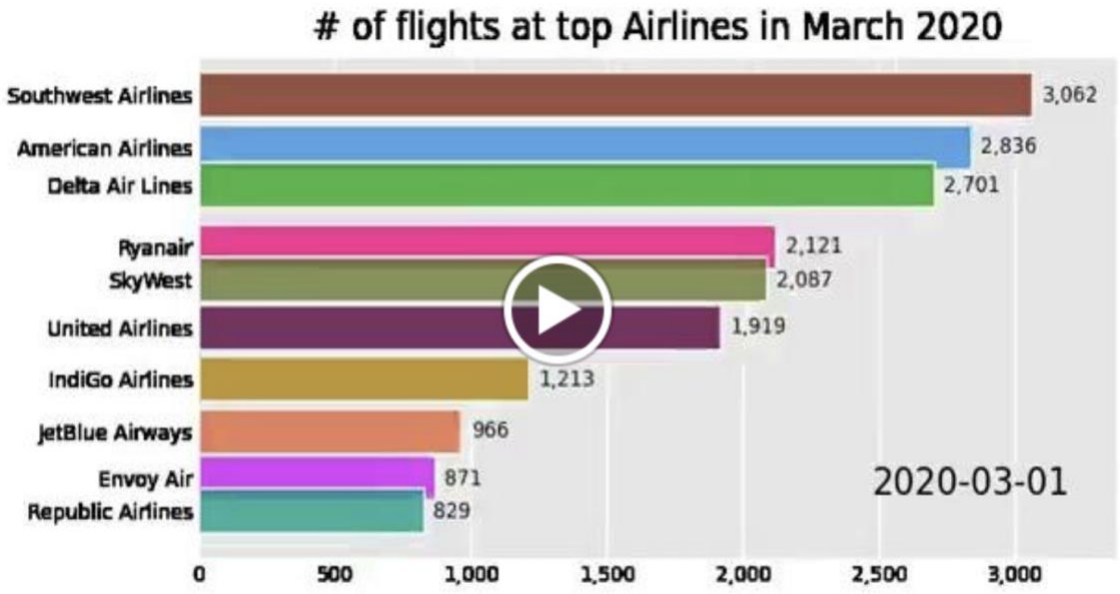
Alpine Air Express | United States (alpine-air.com)

Kalitta Air LLC

Appendix


03/02/2020
8:34 PM
03/03/2020
5:01 AM
First_Seen
Y Axis
March
Mar 08
Mar 15
Mar 22
Mar 29
EN
© kepler.gl | © Mapbox | © OpenStreetMap | Improve this map


# of flights at top Airports in March 2020
Chicago Ohare Intl
Hartsfield Jackson Atlanta Intl
Dallas Fort Worth Intl
Los Angeles Intl
Denver Intl
Phoenix Sky Harbor Intl
Newark Liberty Intl
George Bush Intcntl Houston
San Francisco Intl
Seattle Tacoma Intl
1,098
1,076
912
776
732
606
560
556
529
493
2020-03-01
0
200
400
600
800
1,000
1,200


# of flights at top Airlines in March 2020
Southwest Airlines
American Airlines
Delta Air Lines
Ryanair
SkyWest
United Airlines
IndiGo Airlines
JetBlue Airways
Envoy Air
Republic Airlines
3,062
2,836
2,701
2,121
2,087
1,919
1,213
966
871
829
2020-03-01
0
500
1,000
1,500
2,000
2,500
3,000