REPORT

GROUP PROJECT
USA FLIGHTS DATASET ANALYSIS

BIG DATA TOOLS AND ANALYTICS

MASTERS DATA ANALYTICS FOR BUSINESS

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO – UNIVERSIDADE DE LISBOA

André Viana – l54543

andreviana@aln.iseg.ulisboa.pt

Gonçalo Duarte – l48505

goncaloduarte@aln.iseg.ulisboa.pt

José Cabral – l54997

l54997@aln.iseg.ulisboa.pt

Lisbon School of Economics & Management
Universidade de Lisboa

May 2021

Abstract

Nowadays, as easy as it is for any of us to travel from one city to another, one country to another, the flight industry is expected to be extremely profitable and resourceful. With hundreds of thousands of registered flights per day, there is a big volume of data possible to be analyzed.

The year of 2020 brought change to our lives. Being hit by a pandemic changes the way all of us think and look at the world around us. Do we feel safe to travel in the next months? Can we even travel? Some questions arose last year and for many of them there are no answers yet. What we do know is that the we way travel has changed. That impacts numbers of the aviation industry.

Despite being one single country, there are more than five thousand public airports across the USA – which provides with a huge amount of data possible to be analyzed. Developing analysis on flight data means providing – or trying to provide – answers to many questions, such as: how much did the number of flights decreased from 2019 to 2020 due to COVID 19? Which airports have the highest (and lowest) traffic? Which are the top states on rate of cancellations and rate of delayed flights?...

Working with big volumes of data such as daily flight data across all US airports from 2017 to 2020 brings an additional challenge: working with tools that allow these tasks to be performed effectively. Traditional methods usually start to be inefficient when working with increased sized files. Working on a cloud is a possible solution to this problem – for example, Databricks.

There are many results and conclusions to be taken out of this project. We hope we can provide them.

Key – words

Travel, Flight Industry, COVID 19, Big Data, Cloud, Databricks.

Contents

Abstract 2
List of figures 4
List of tables 4
Introduction 5
Methodology 6
Business understanding: 6
Data understanding: 6
Data preparation: 7
Results/Discussion 8
Data Exploration 8
Network 12
Shortest Path 14
Conclusion 15

List of figures

Figure 1 - Number of flights per month per airport (2017-2020) 8

Figure 2 - Number of cancelled flights (2019) per state 10

Figure 3 - Rate of cancelled flights (2019) per state..... 11

Figure 4 - Initial network..... 12

Figure 5 - Network with coordinates..... 12

Figure 6 - Network on US map..... 13

Figure 7 - Shortest Path 14

List of tables

Table 1- Airports with highest and lowest number of flights per month in 2020.. 9

Table 2 - Shortest Path 14

Introduction

This project aims at analyzing US flights – available at the Bureau of Transportation Statistics of United States Department of Transportations database – explore its evolution over the years concerning different features, develop a network of the routes using Databricks platform and study the impact of COVID 19.

Living through a pandemic is an unprecedented situation (in the most recent decades) that surely took a tool on the travel industry. Further on the project, it will be possible to visualize the evolution of the number of flights by month and year. An intense decrease is expected when comparing 2020 to previous years. We want to base our analysis in different features – such as total number of flights, cancelled flights, rate of cancellations, delayed flights and the correspondent delay – across airports and states. One practical goal when developing this project, was to be able to effectively visualize our results. In order to do so, different visualization and interactive tools will be displayed, so that the user can go through all the work.

Lastly, we want to push our boundaries when it comes to the network part of the project. In fact, to fly from one airport to another, surely there are many different possible routes. But what if we want to only focus one the shortest trajectory between airports? A small application was built, by taking two inputs: origin and destination airports – then it will return the shortest path between the two.

Disclaimer: The full code written to develop this project (and the corresponding jupyter notebooks) is available on the **github repository**: https://github.com/ISEG-MDAB/BDTA-us-flights

When going through the repository, please start by reading the file *READEME.md* to get all information on both data (also explained on the *Methodology* of this report) and on what is the utility of each notebook (as well as in which order they should be analyzed).

Methodology

Business understanding:

In Bureau of Transportation Statistics of United States Department of Transportations website, we can find several datasets, regarding information about all kinds of transportation on the US. Analyzing this type of information can be a powerful tool to understand the current trends and prepare for the future.

When airports plan their routes, they can use this analysis to fill some gaps that exist in some airports or remove some routes that are no strictly necessary.

Data understanding:

To develop this project, we worked with **Bureau of Transportations Statistics data sets**, regarding flights from 2017 to 2020:

The data set used for the development of this project contains a row per flight and as columns:

- *FL_DATE* = Flight Date (yyyy-mm-dd)
- *OP_UNIQUE_CARRIER* = Carrier ID *ORIGIN_AIRPORT_ID* = Origin Airport ID. An identification number assigned by US DOT to identify a unique airport.
- *ORIGIN* = Origin Airport
- *ORIGIN_CITY_NAME* = Origin Airport city name
- *ORIGIN_STATE_ABR* = Origin Airport state abbreviation
- *ORIGIN_STATE_NM* = Origin Airport state name
- *DEST_AIRPORT_ID* = Destination Airport ID. An identification number assigned by
- US DOT to identify a unique airport.
- *DEST* = Destination Airport
- *DEST_CITY_NAME* = Destination Airport city name
- *DEST_STATE_ABR* = Destination Airport state abbreviation
- *DEST_STATE_NM* = Destination Airport state name
- *CRS_DEP_TIME* = Programmed departure time (local time: hhmm)
- *DEP_TIME* = Actual departure time (local time: hhmm)
- *DEP_DELAY* = Departure delay in minutes, early departures show negative numbers(DEP_TIME - CRS_DEP_TIME).
- *WHEELS_OFF* = Wheels Off Time (local time: hhmm)
- *WHEELS_ON* = Wheels On Time (local time: hhmm)
- *CRS_ARR_TIME* = Programmed arrival time (local time: hhmm)
- *ARR_TIME* = Actual arrival time (local time: hhmm)

- *ARR_DELAY* = Arrival delay in minutes, early arrivals show negative numbers.(ARR_TIME - CRS_ARR_TIME).
- *CANCELLED* = Cancelled Flight Indicator (1=Yes)
- *CANCELLATION_CODE* = Specifies The Reason For Cancellation
- *CRS_ELAPSED_TIME* = Programmed Elapsed Time of Flight, in Minutes
- *ACTUAL_ELAPSED_TIME* = Actual Elapsed Time of Flight, in Minutes
- *AIR_TIME* = Flight Time, in Minutes
- *DISTANCE* = Distance between airports (miles)
- *CARRIER_DELAY* = Carrier Delay, in Minutes
- *WEATHER_DELAY* = Weather Delay, in Minutes
- *NAS _DELAY* = National Air System Delay, in Minutes
- *SECURITY_DELAY* = Security Delay, in Minutes
- *LATE_AIRCRAFT_DELAY* = Late Aircraft Delay, in Minutes

Data preparation:

- **Data preparation for visualization**
    - Load the dataset using pyspark and aggregate it on STATE, ORIGIN_AIPORT_ID and FL_DATE;
    - Create calculated columns of our interest to use further on the analysis;
    - Convert the aggregated dataset to a pandas data frame to do the visualization analysis;
- **Data preparation for network**
    - Run a script that return the coordinates of the airport location;
    - Filter the data on the 350 routes with more flights;

Results/Discussion

When presenting and analyzing the results obtained, we will work in two different stages. Each stage corresponds to one jupyter notebook available on the **github repository** (link displayed on report introduction):

1. Data Exploration: notebook*: Data analysis.ipynb*
2. Shortest Path App: notebook: *Shortest path.ipynb*

Data Exploration

To explore the data, compare airports and states and get to know better how the aviation industry has changed over the years in the USA, three interactive plots were developed:

- Average number of flights by airport per month, per airport;
- Number of flights, cancellations, delays and average delay per year, per state;
- Rate of cancellations and delayed flights per year, per state.

Average number of flights by airport by semester (impact of COVID 19)

On the first plot, we aim at understanding how the average number of flights has changed from 2017 to 2020. It is also possible to compare the average evolution with the evolution of any specific airport. By selecting the airport on the menu, two scatter plots are displayed: a line in blue showing the changes for the airport flight numbers; and a line in orange with the overall average per airport.

To simplify our analysis on this report, we present an example by choosing the "John Wayne Airport – Orange County".

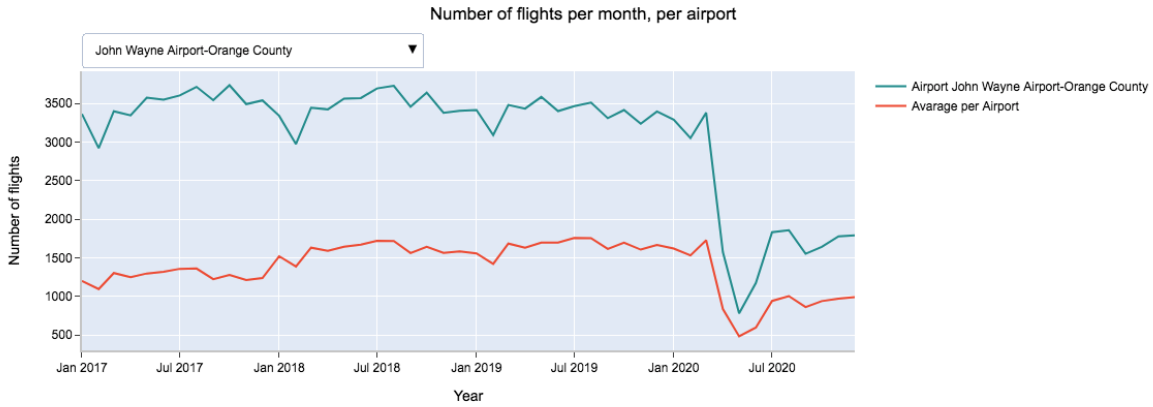

Figure 1 - Number of flights per month per airport (2017-2020)

As one of California's largest airports, it is normal that its numbers are above the average per airport. Flight records seem to be pretty stable from 2017 up until 2020.

As expected, 2020 was a rough year for aviation industry. When compared with previous years (since 2017), it registered the lowest number. COVID 19 pandemic forcing all of us to live 2020 in lockdown and social distancing to be the "new normal",

really translates into these numbers. In fact, from March 2020 to May 2020, aviation industry suffered an approximate 72.16% decrease. This is a totally different number from what has happened in recent years: from 2017 onwards, the number of flights has been overall stable (with tendency of growing).

To complete our analysis by airport we also tried to understand what were the top and bottom airports regarding the number of flights. In the notebook, you will find a table with top and bottom three on this matter, for each month from 2017 to 2020. For simplicity purposes, table 1 displays the top and bottom two airports on total number of flights for each month of 2020:

Month	Highest n. of flights	2nd highest n. of flights	Lowest n. of flights	2nd lowest number of flights
January	Hartsfield-Jackson Atlanta International	Chicago O'Hare International	Barnstable-Municipal-Boardman/Polando Field	Branson Airport
February	Hartsfield-Jackson Atlanta International	Chicago O'Hare International	Barnstable-Municipal-Boardman/Polando Field	Branson Airport
March	Hartsfield-Jackson Atlanta International	Chicago O'Hare International	Barnstable-Municipal-Boardman/Polando Field	Branson Airport
April	Dallas/Fort Worth International	Charlotte Douglas International	Barnstable-Municipal-Boardman/Polando Field	Bellingham International
May	Dallas/Fort Worth International	Denver International	Barnstable-Municipal-Boardman/Polando Field	Branson Airport
June	Dallas/Fort Worth International	Hartsfield-Jackson Atlanta International	Barnstable-Municipal-Boardman/Polando Field	Cedar City Regional
July	Dallas/Fort Worth International	Hartsfield-Jackson Atlanta International	Barnstable-Municipal-Boardman/Polando Field	Cheyenne Regional/Jerry Olson Field
August	Hartsfield-Jackson Atlanta International	Dallas/Fort Worth International	Cheyenne Regional/Jerry Olson Field	Erie International/Tom Ridge Field
September	Hartsfield-Jackson Atlanta International	Dallas/Fort Worth International	Cheyenne Regional/Jerry Olson Field	Florence Regional
October	Hartsfield-Jackson Atlanta International	Dallas/Fort Worth International	Barnstable-Municipal-Boardman/Polando Field	Cheyenne Regional/Jerry Olson Field
November	Hartsfield-Jackson Atlanta International	Dallas/Fort Worth International	Barnstable-Municipal-Boardman/Polando Field	Del Rio International
December	Hartsfield-Jackson Atlanta International	Dallas/Fort Worth International	Barnstable-Municipal-Boardman/Polando Field	Branson Airport

Table 1- Airports with highest and lowest number of flights per month in 2020

Through table 1 it is possible to have an understanding of the airports with most (and least) amount of traffic in 2020. Both Hartsfield-Jackson Atlanta International and Dallas/Fort Worth International are dominant – being the only two airports to have been the top airport in a month. When it comes to the lowest number of flights, Barnstable-

Municipal-Boardman/Polando Field and Cheyenne Regional/Jerry Olson Field were the airports that topped the podium in every month of 2020.

Number of flights, cancellations, delays and average delay per year, per state

We are now switching our analysis from individual airports to states. In order to understand the differences, we first developed a USA map from which the following analysis are possible to be made by year:

- Number of flights;
- Number of cancelled flights;
- Number of flights with delay;
- Number of flights with delay < 15min;
- Number of flights with delay >= 15min.
- Average delay per flight (in minutes)

So, by selecting one the features above for a specific year, the user will then be able to see how is the correspondent distribution across all US states. On the report, we display an example of *Number of cancelled flights in 2019*:

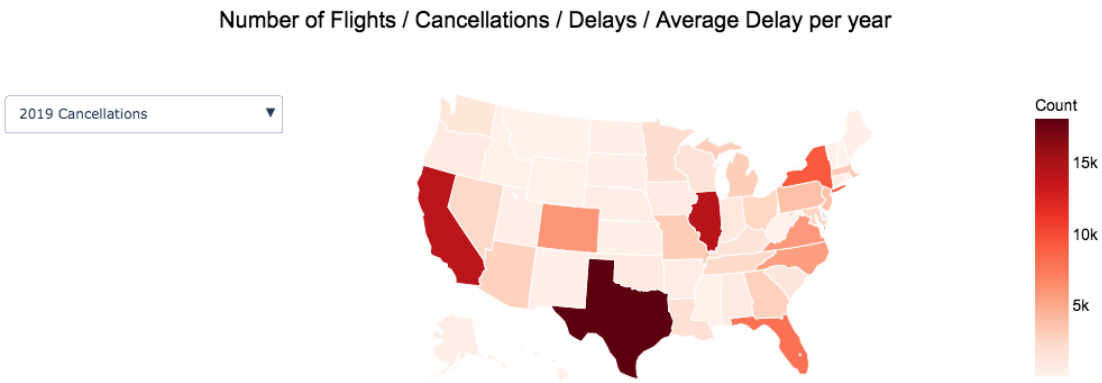

Figure 2 - Number of cancelled flights (2019) per state

We can now conclude that the states of Texas, Illinois and California, are the ones with the highest number of cancelled flights with totals of 18.119k, 14.39k and 14.101k respectively. West Virginia and Montana are amongst the states with the least aviation traffic with 180 and 135 cancelled flights respectively.

Note that this visual allows us to perform similar analysis for each of the features on the dropdown menu (mentioned above) per year – which will not be individually done on this report for extension reasons.

Rate of cancellations and delayed flights per year, per state.

As significant as it is to analyze state that in absolute numbers, it is also interesting to understand what happens when analyzing it in rate terms. On the third visual, a USA map is displayed, now with the following possible features on the dropdown menu per year:

- Rate of cancelled flights;
- Rate of delayed flights;
- Rate of flights with delay < 15min;
- Rate of flights with delay >= 15min.

Once again, to simplify this document, we are choosing a specific feature on a specific year to analyze in detail. To compare with the values obtained through the previous figure, figure 3 displays the *Rate of cancelled flights in 2019*:

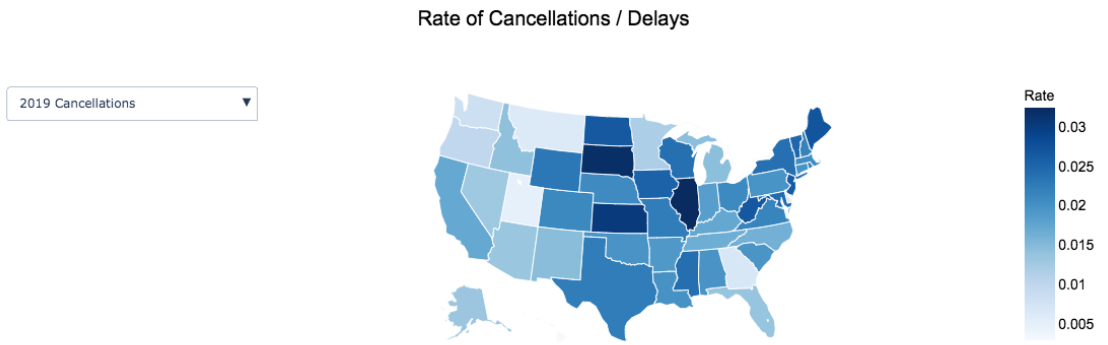

Figure 3 - Rate of cancelled flights (2019) per state

As you can see, when working with the rate of cancelled flights $\left(\frac{n^{o}\ of\ cancelled\ fligthts}{total\ n^{o}\ of\ flights}\right)$ instead of just the number of cancelled flights, the results are different: Illinois, South Dakota and Kansas are the states with the highest cancellation rates (3.248%, 3.2% and 3.05% respectively). California, which was the third state with highest number of cancellations in 2019, is not in the top 10 concerning the rate of cancellation, having registered a 1.73% rate on the same year.

On the other side, when it comes to states with the lowest cancellation rates, Utah and Montana top the list with 0.489% and 0.627% respectively.

Once again, these visuals proved to be very powerful and useful to develop our analysis, being an intuitive and interactive way of displaying many results.

Network

After developing analysis through time and states, we wanted to focus our attention on connections between airports. On the next stage of the project, the goal was to develop and display a network based on airport routes.

The process to achieve the network desired is now explained: First, a simple network – figure 4 – was presented with all airports (nodes) and links between them (edges), as it follows:

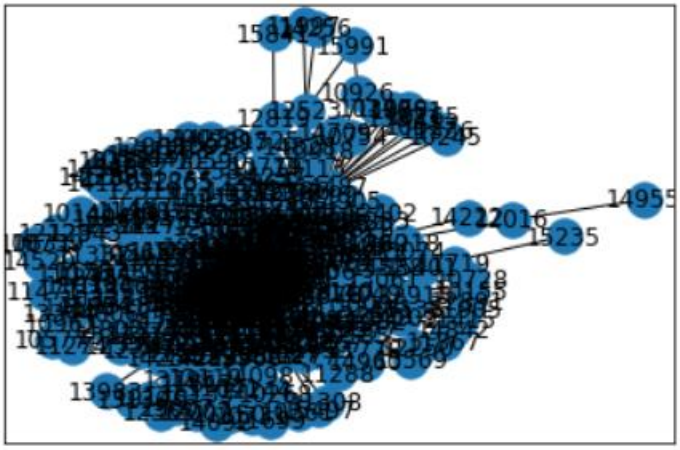

Figure 4 - Initial network

As expected, the plot obtained does not give much information since no connection to airports location was made. To try to visualize the roots on a more efficient way, airport coordinates were added to the nodes.

Note that, in order to do so, a script was built to attach both latitude and longitude of each airport on the data frame – as described on Data Preparation: Data preparation for network, above on the report. (*Notebook: Coordinates Script).*

The network obtained is displayed through figure 5:

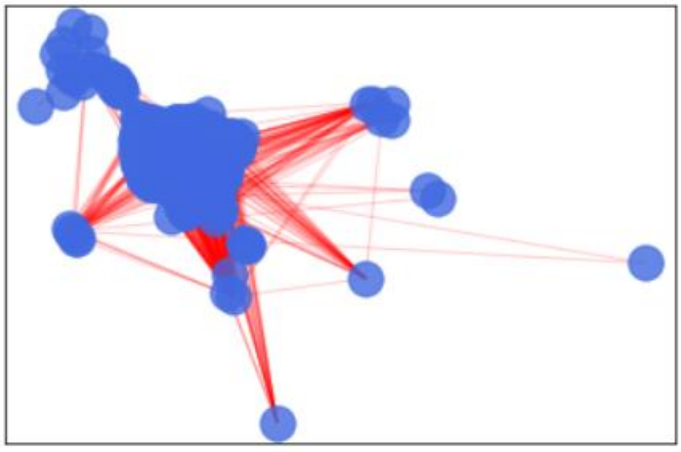

Figure 5 - Network with coordinates

After adding the coordinates, it is possible to extract more information when comparing to the first try. As you can see on the figure above – figure 5 – it is possible to understand where are the airports concentrated and which routes have the highest number of flights – analyzing red lines: the darker the line, the biggest the traffic on that route; so more transparent lines represent routes with less number of flights, comparing to darker lines.

Finally, one more step that we could take to improve this visualization: plot the network on top of a map. To develop this overlap, we used MAPBOX (a tool to build better mapping, navigation and search experiences across platforms).

Recalling the abstract of this report, there are more than five thousand airports across the USA. So, as expected the number of routes between airports is really high. This became a limitation to our plot – not being possible to display all the routes. To solve this issue, only the 350 routes with more flights were plotted. The final visualization for this stage of the project is shown on figure 6:

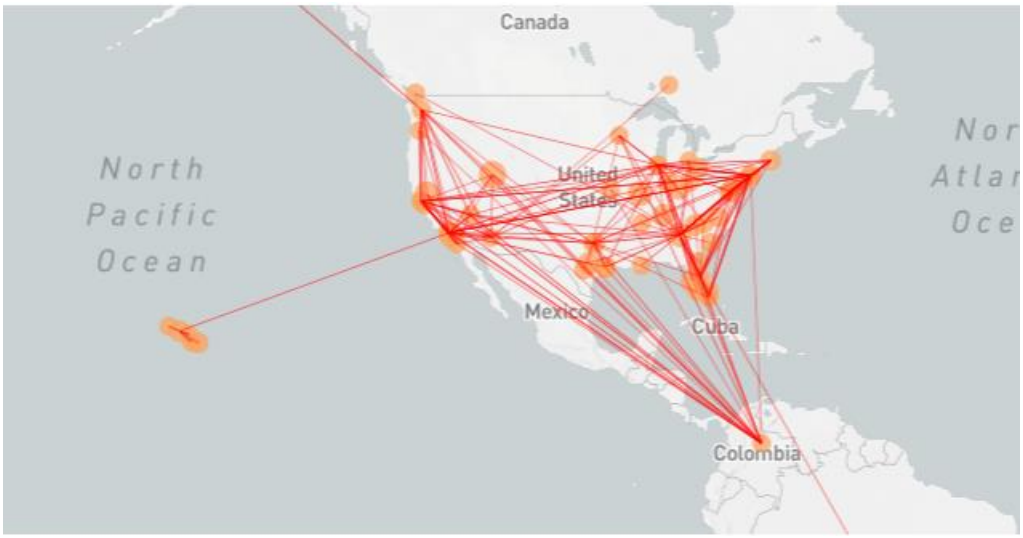

Figure 6 - Network on US map

It is now possible to have a full understanding of the US air routes. By analyzing the routes distributions, some conclusions might be taken regarding where some gaps are, the possible need to create new routes and/or terminate some that might not be so important.

- For further information on MAPBOX, all the documentation is available at: https://www.mapbox.com/

Shortest Path

As seen on the Network part of this project, many possible routes exist to connect two airports. The focus on the last stage of the project is to get the shortest way to go from one airport to another. To develop this analysis, a new notebook was created:

- *Shortest Path:* a notebook through which an "application" as developed. It takes two inputs from the user: Origin Airport and Destination Airport. Then, using the previous network, the shortest path between the two airports is determined and displayed on the US map.

To simplify the report, an example is shown. The goal is to connect the Pellston Regional Airport of Emmet County and Roswell International Air Center. The shortest path app returns the following path:

Order	Airport
0	**Pellston Regional Airport of Emmet County**
1	Cherry Capital
2	Chicago O'Hare International
3	Tulsa International
4	**Roswell International Air Center**

Table 2 - Shortest Path

Figure 7 shows the shortest path to link the airports on the US map:

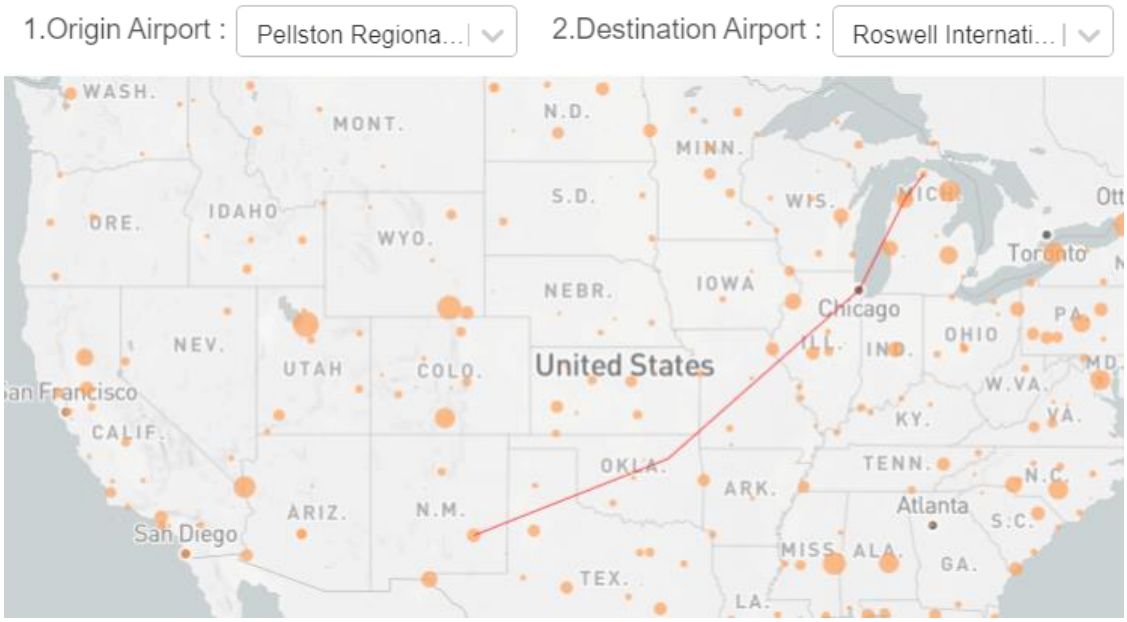

Figure 7 - Shortest Path

Red line represents the path between the airports. By choosing any other two airports, the user would get a similar process.

Conclusion

Reaching the end of our project, we are now able to understand that the world of big data comes with many challenges and so there is a need to use adapted tools for this field. If not, there would not have been possible to analyze this dataset due to its size. The solution used on this project was Databricks.

One of the main goals was to understand what the impact of the COVID19 pandemic on the aviation industry was. After developing all the analysis, it is possible to state that this industry took a big hit when the pandemic started, decreasing the average number of flights by 72.16%. We sure hope the flight numbers get back to its previous numbers in the future.

Two different analyses were conducted: data exploration and network. Through data exploration, we were able to visualize data from different perspectives, which gave us the ability to characterize it. A variety of charts were displayed on this project, for us to get more insight into the aviation industry, focusing on the number of flights, cancellations and delays. On this section, it was also possible to compare the evolution of number of flights per airport and it's also possible to compere the metrics referred earlier by state in absolute values and relative values. We found the visuals very appealing to perform this type of analysis, since one single chart can describe different results based on which features the user wishes to base the analysis on.

Lastly, we wanted to challenge ourselves and build an application that takes 2 airports as inputs from the user and then returns the shortest route between these two. Through this app, we found a way to implement the network developed earlier on the project and to add some real usage of our analysis.