Project Report

Chengxiang Xiao

Abstract

This project is trying to find the relationship behind airplane delay in the USA. The data contain the reason and number of delays in a month for each airline company at each airport. The period is from Jan 2015 to Feb 2021. This project use Matplotlib and MetPy to generate histogram plot, pie plot, line plot and surface plot at points on the USA map.

Data description

The data is downloading from https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp as a zip file. The shape of the data is (87860,22). The column shows the reason of delay, the airport, airline company, and the location of the airport. The row shows the number of each airline company delay at each month and airport for each reason.

Code description

The code includes 3 parts. The first part is loading original file and import necessary tools.

The second part is pull out the necessary data and regroup them for plotting. The third part is finding the best way to plot the prepared data.

A list of python libraries is showing below:

i. Numpy – This package is using for support large matrices calculation.
ii. Pandas -This package is using for data analysis in this project.
iii. Matplotlib – This package is using for line plot, histogram plot, pie plot in this project.
iv. Metpy -This package is using for surface plot and interpolation plot in this project.

Result and code details

The surface plot on metpy

Data preparation: use the *groupby* to add all number of delay base on the airport. Then use the *pd.merge* to combine the coordinate of each airport with the delay number. Then use metpy to decide the axis of map and surface plot by using stationplot.

On the Figure 1, it shows the name of each airport and number of delays.

Because the Figure 1 is plotting too much information, the figure 2 is focusing on number of delays in Texas airports. Figure 2 shows, it has 29870 delay airplanes in San Antonio in the past five years. And the DFW has the largest delay number in the past five years. The number is 214376, due to its large capacity.

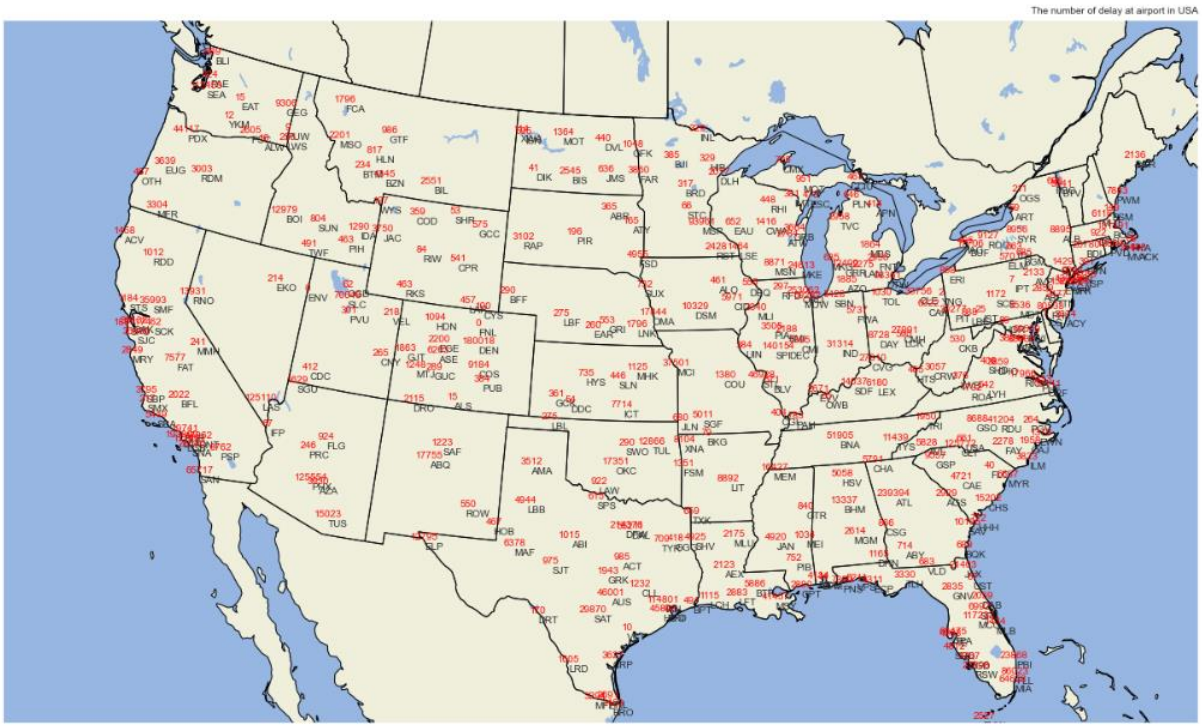


Figure 1: The number of delays at each airport in USA in past 5 years

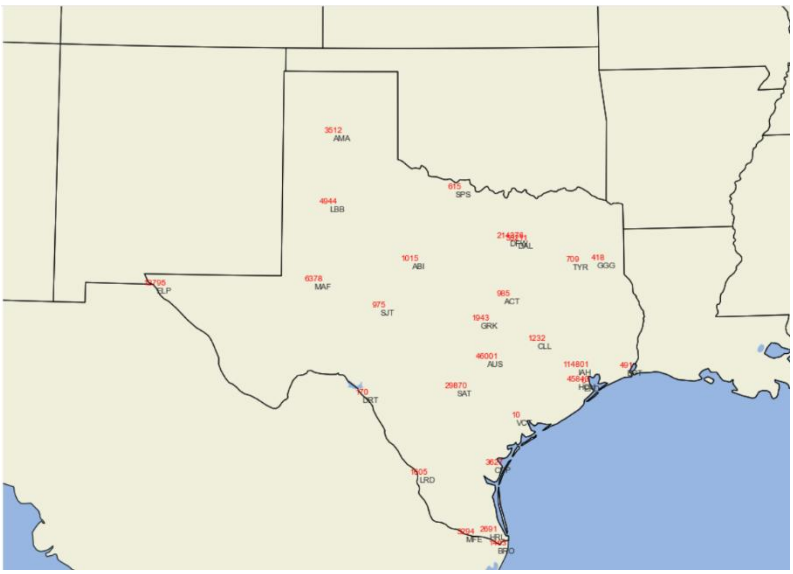


Figure 2: The number of delays at Texas airport in the past 5 years.

Histogram plot

The histogram plot is powerful and straight forward. The histogram plot shows the DFW airport has largest delay in Texas and its twice than IAH. SAT is the sixth large delay airport in Texas.

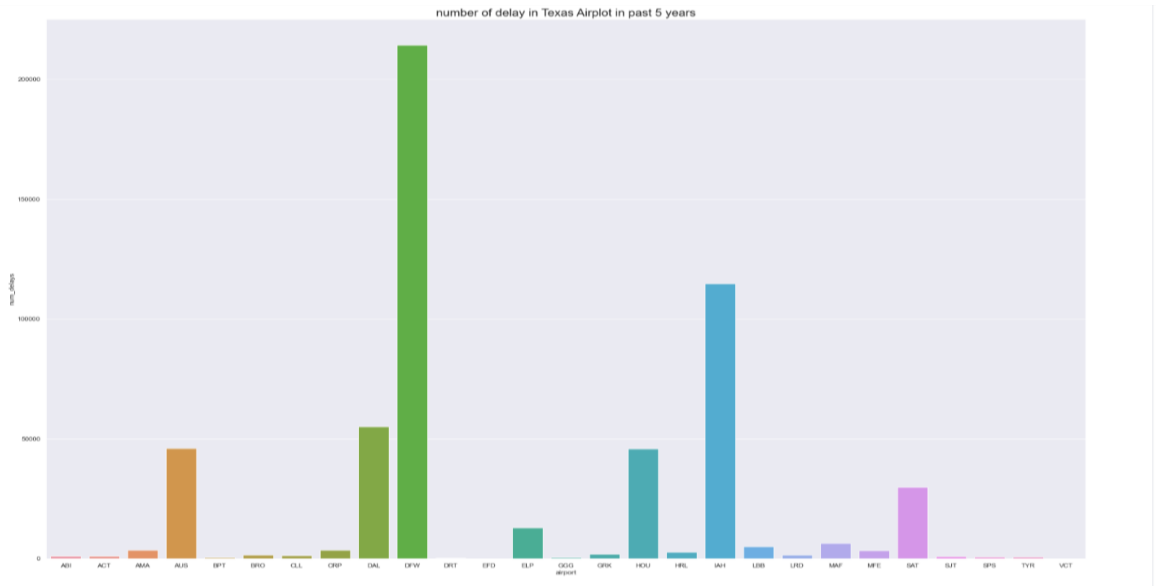
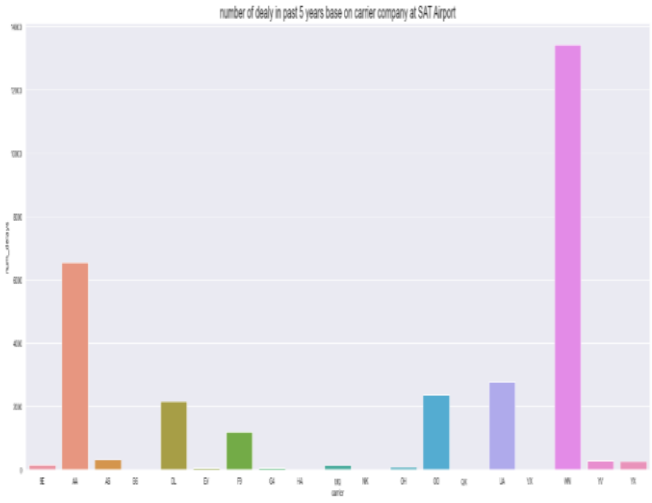


Figure 3: Histogram plot for different Texas airport delays.

Figure 4 shows the number of delays at SAT base on the carrier company in the past years. WN, Southwest Airlines has the largest delay and AA has the second delay number. And the pie plot shows the percentage of WN delay at SAT is 45%, and AA is 22%.



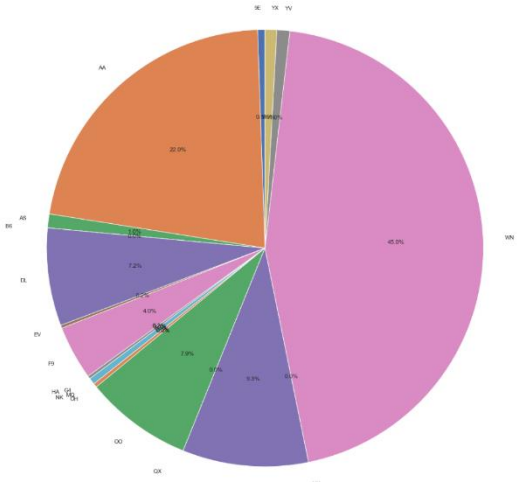


Figure 4: Number of delays at SAT base on carrier company

Figure 5 show the number of delays for all airport base on different carrier company. Also, only looking at the total number of delays can't really tell us what airline company have the less delay. Because the large airline company will have bigger delays even with small proportion of delay. The proportion is calculating as total number of delays divide the total arrival fight. The figure 5 is trying to analysis both the total number of delay and the proportion of delay.

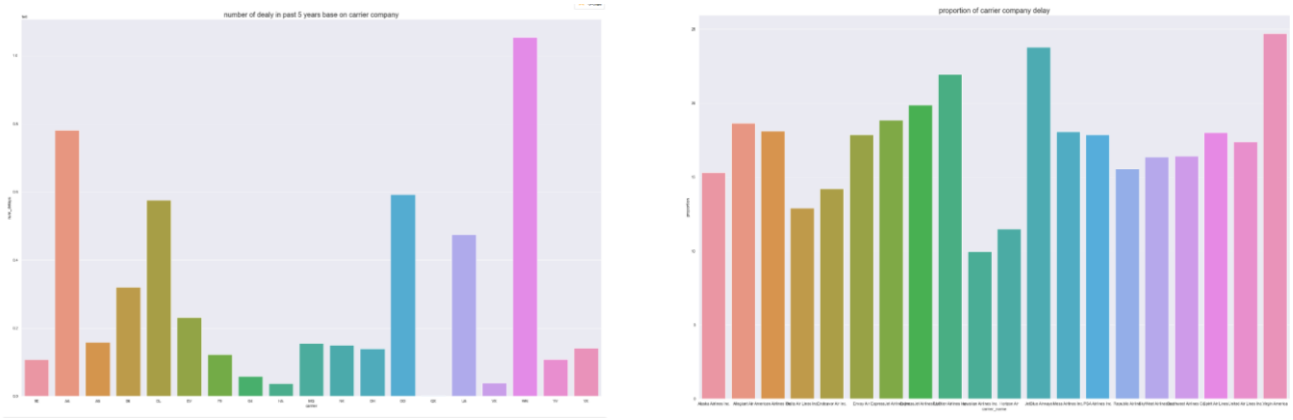


Figure 5: Total delay and proportion of delay for each carrier company.

The Southwest Airline and American Airline still have the larger total number of delay fight, but for proportion, Virgin America and JetBlue Airways has the higher proportion delay.

Line plot

Line plot is good to showing the trace of data change, which is suitable for delays VS time period.

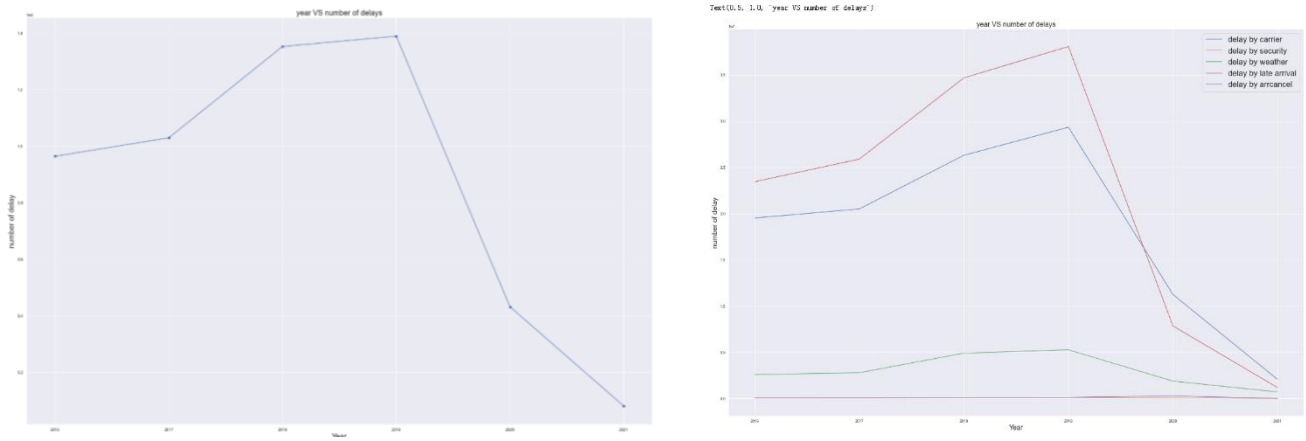


Figure 6: Total number of delays comparing to year

The figure 6 show the 2019 has the highest delay numbers, but because the covid-19. In 2020, the delay of flight drops rapid. Because the number of fights is decrease. The right figure shows the reason and number of delay in the past five years.

Figure 7 shows the total number of delays monthly. In the summer and winter time, it has higher delay than spring and autumn. The plot for 7.2 is made by for loop.

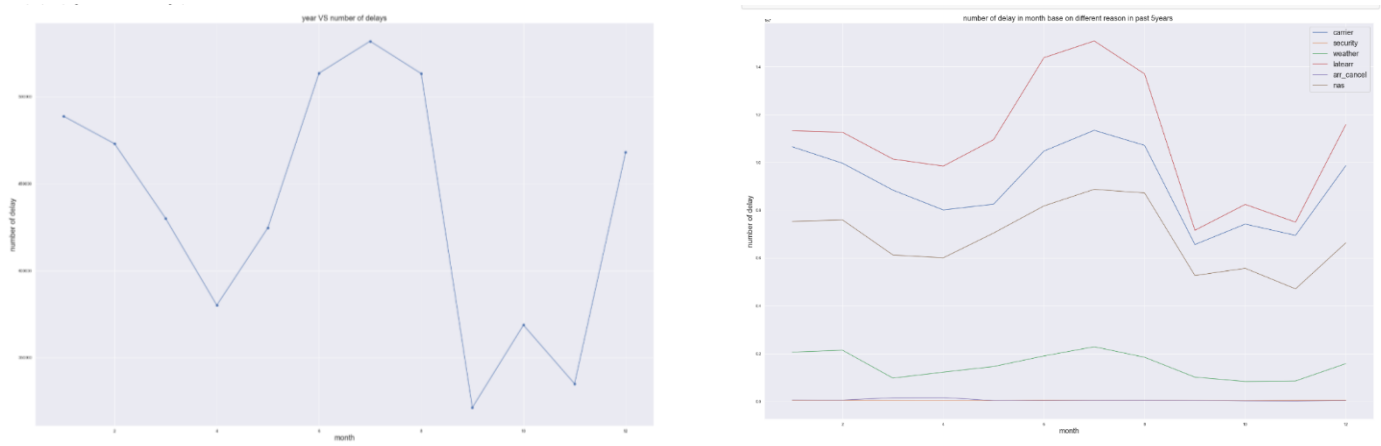


Figure 7: Delay base on month

Linear regression

The linear regression is not showing a good result, comparing the month and delay number, because the data set is small to analysis

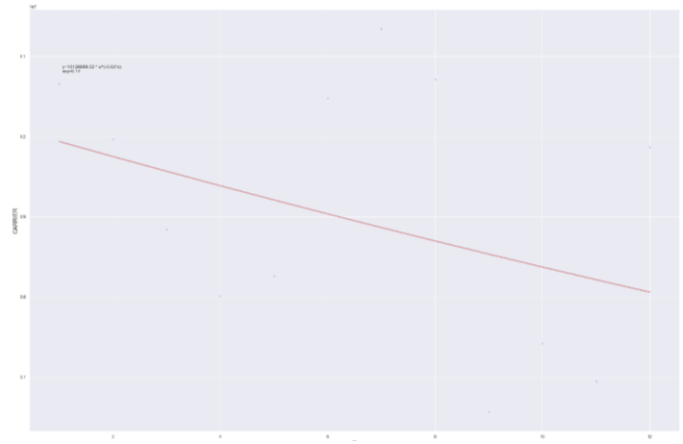


Figure 8: Linear regression comparing month and number of delays

Correlation

The Figure show the correlation between delay reason. As the figure show, delay by carrier company, weather, late arrival, and National airspace system have strong relation.

Out[517]:

	carrier	security	weather	latearr	arr_cancel	nas
carrier	1.000000	0.636227	0.905779	0.930269	-0.009102	0.928197
security	0.636227	1.000000	0.583451	0.563872	-0.488996	0.537891
weather	0.905779	0.583451	1.000000	0.845957	-0.132641	0.910473
latearr	0.930269	0.563872	0.845957	1.000000	-0.040225	0.955980
arr_cancel	-0.009102	-0.488996	-0.132641	-0.040225	1.000000	-0.082344
nas	0.928197	0.537891	0.910473	0.955980	-0.082344	1.000000

Figure 9: Correlation between delay reason

Number of delays comparing to temperature in Jan 2016.

The figure 10 shows the location of delay and number of delays in Jan 2016 and also the temperature. The temperature of the plot is using the packing on the metpy website and base grid interpolation.

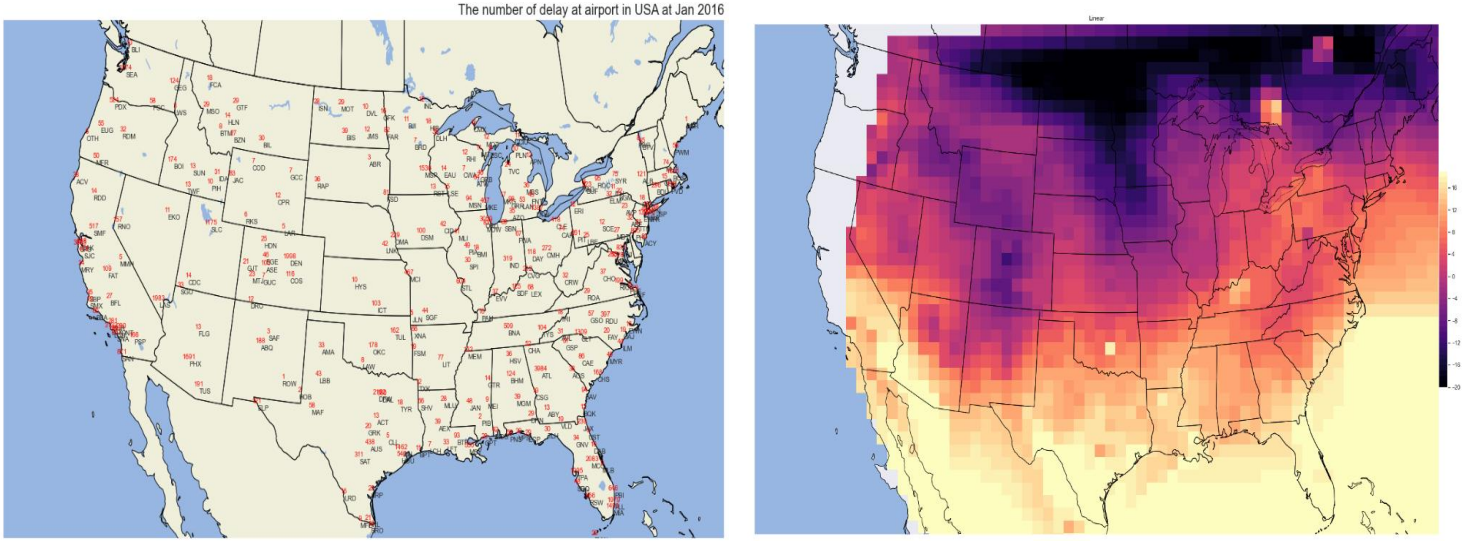


Figure 10: Delay at Jan 2016 and the temperature.

Conclusion

The project is using surface station plot, histogram plot, line plot to show delay of fight in American in the past five years. The total number of delay fights decrease rapidly after the covid-19 breakout. And the number of delays is change significantly with season. Also, the large airport, like DFW has higher delay number comparing to other airports. SW and AA have the highest total delay fight, but VG and JBU has the highest delay proportion.