Impact of COVID on Airline Delay

Jike Lu, Jiarui Chen, Jingxuan Sun, Xuanyu Shen, Yiwei Jiang

I. Introduction

Due to the outbreak of COVID-19 in 2020, many American industries got affected to varying degrees and all experienced a tough period. Particularly, COVID-19 impeded citizen's travelling and thus hit the airline industries. In this project, we compared airline operations before and during COVID-19's population. Some research had shown that COVID-19 shocked the airline economy dramatically and decreased the departure performance and thus had impact on airline on time performances [1]. In contrast, other research claimed that airlines are more punctual during COVID-19 [2]. We were curious if we could understand the COVID-19's impact on airline operations. To successfully explore the correlation between the COVID-19 and the airline performance, we utilized different machine learning techniques, such as matching, difference in difference analysis, and regression discontinuity, to demonstrate the impact of COVID on flight on time accuracy, and the correlation between the flight on time accuracy and the state's policies.

II. Literature Review

As the Covid-19 spread out rapidly, the government enforced a lot of travel restrictions to control the pandemic. The airline businesses are hit unprecedentedly. As recorded, the number of daily flights dropped by 64 percent suddenly, and the density of global flights decreased by 51 percent between 2020 and 2021[3] In addition, people also avoid unnecessary travel, which decrease the demand of flight.[4]. However, beside the agreement, there is an argument on the performance of the airline regarding the delay. Susan Hotle indicates an increase in the overall delay due to the covid-19, which might cause by the shortage of employee.[1] This point is clarify in Joseph's paper. He concluded that the total airline employment decreased by 8 percent.[5] However, Yimga argues that "Flights are departing and arriving with less delay amid the pandemic"[2]. More specifically, Yimga further concluded that the arrival delay decreased by 1 minute 42 seconds and 2 minutes for departure delay [2].

III. Data

Arrival delay	Actual arrival - scheduled arrival
Carrier delay	Carrier Delay, in Minutes
Security delay	VSecurity Delay, in Minutes
Late aircraft delay	Late aircraft delay, in minutes
TaxiIn	In Time, in Minutes
TaxiOut	Taxi out Time, in Minutes

We have two datasets in our project. These datasets contain the flight information in New York and Texas from 2019 to 2021. We collected data from the Bureau of Transportation under US Department of Transportation. For the variables, we mainly focus on time of arrival delay, time of carrier delay, time of security delay and aircraft departure delay. After we drop all the nan values, there are around 35,000 rows of data.

IV. Descriptive analysis

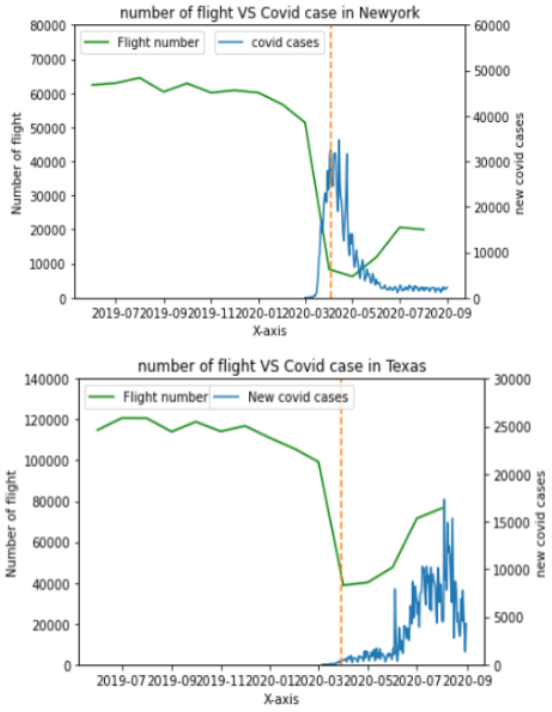

Fig. 1. Covid cases and flight number

Fig.1 shows the data from New York state and the other shows the data from Texas, with x as the time interval, left y axis as the number of flights and the right y axis as the daily covid cases. The orange dash line indicates the time of the first covid restriction issued in each state. From the graph, we can see that before the restriction, the number of flights tended to be very stable. Right after the first restriction rule was issued, the number of flights dropped immediately for both Texas and New York. New York number of flights dropped approximately 79 percent after the start of the pandemic and Texas dropped 58 percent , which indicates covid definitely have some influence on the number of flights.

Fig.2 shows the trend of arrival delay, departure delay, carrier delay, security delay, and aircraft delay respectively

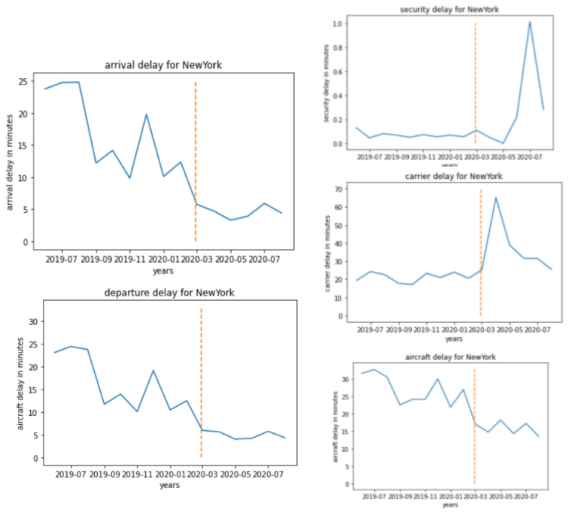

Fig. 2. New York flight description

in New York. New York state declares a state of emergency on march 7th, 2020 , which is the orange dash line on each graph. We see downward trends on arrival, departure, and aircraft delay. We also see upward trends on carrier and security delay on the graph. The conjecture will be discussed in the Difference-in-Difference Analysis sections.

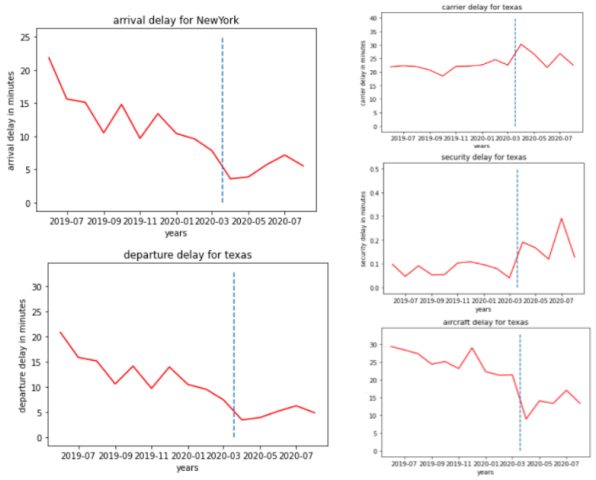

Fig. 3. Texas flight description

Texas issued a restriction similar to New York state on march 21st 2020, which prevents social gathering. Fig.3 shows the flgit information in Texas The trends are very similar to New York state with downward trend on arrival,departure and aircraft delay, and upward trend on security and carrier delay. However, compared to New York, Texas's trends are flatter. We infer that the less daily covid cases in Texas might have a smaller impact on delay compared to New York.

V. Goals and Hypothesis

In this project, we would like to investigate the casual relationship between COVID and the flight delays by using different explanatory analysis tools such as difference-in-difference analysis, regression discontinuity analysis, and statistical matching. The details of these methods will presented in the later sections. The questions that our study tries to answer include:

- Does the pandemic make airlines more punctual?
- Does the increase in severity of the pandemic lead to better flight on-time performance?
- Does the stricter of local COVID policies negatively affect the on-time performance?

Before the making any analysis, we expected that there was less delay during the pandemic due to less flights and more restrictions. We did estimated flights from or to Texas to be less impacted by the pandemic than flights from or to New York due to difference in the strictness of local COVID policies.

VI. Methods

A. *Difference-in-Difference Analysis*

The difference-in-difference analysis is a quasi experimental method that compares the change of experiment and control group before and after a treatment is imposed [7]. The analysis is based on regression using three interaction variables as the input: Time, Treatment, and their product. The dummied Time variable represents whether the observation is collected before (0) or after (1) the treatment. The dummied Treatment variable represents whether the observation comes from the control group (0) or the experiment group (1). When the product of these two variables equals to 1, it means that the observation is from the experiment group after the treatment, which is what the analysis is interested in. The assumption of the analysis is the Equal Trend Assumption, which requires that the trend of the target variable for both control group and experiment group move roughly in the same way (parallel). The advantage of difference-in-difference analysis is that it facilitates the casual inference when randomized experiment is impossible, and it takes unobservable variables that may also influence the experiment outcome into consideration and thus yields a more accurate estimation on the actual impact of the treatment or intervention [7].

We used this approach to investigate how does the pandemic influence flight delay factors, such as Carrier Delay, Tax-in Time, Tax-out Time, etc. In our study, the treatment was the first round of COVID outbreak in New York starting from March 19th, 2020. Such selection of treatment made the New York the experiment group. At the same time, there were not many daily cases in Texas, making it the control group. The data we used for this specific analysis ends on 2020-05-31, which is when the case number in Texas started to surge, making Texas no longer a valid control group. Before performing the analysis, we used visualizations to check whether our target variables behave similarly in New

York and Texas before the pandemic, and the visualizations proved that the Equal Trend Assumption held in this case. Some of the visualizations were selected to be presented in this report (Fig.4 and Fig.5).

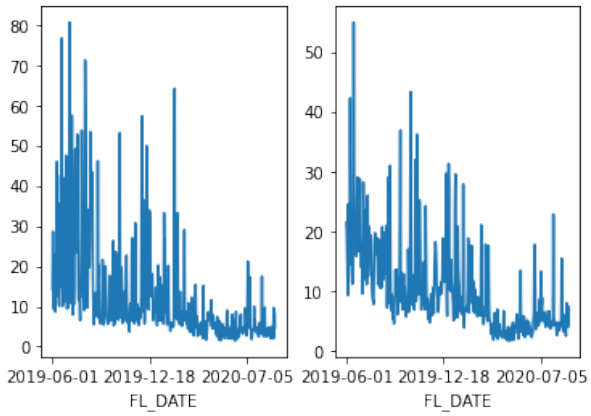

Fig. 4. Trend of Departure Delay for New York (Left) and Texas (Right)

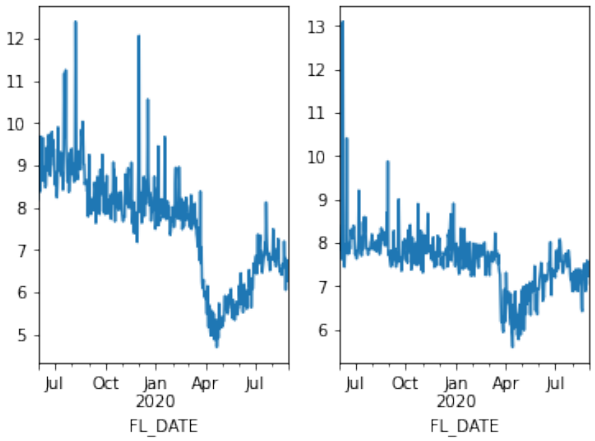

Fig. 5. Trend of Taxi In Time for New York (Left) and Texas (Right)

B. Regression Discontinuity

Regression Discontinuity is a method where we aim to determine the causal effects of interventions by assigning a threshold to see changes in the regression lines. Thresholds are assigned manually and therefore are subjective. Regression Discontinuity is usually used to determine if a certain state of change will result in different behavior in the treatment groups before and after the threshold is reached. However, it is not useful to show an overall increasing or decreasing trend of the data. In our experiment, we set different treatment groups according to the state of seriousness of the pandemic in both states. We chose to analysis the impact of COVID on flights' arrival delay(ARR_DELAY_NEW) and departure delay(DEP_DELAY_NEW) in both states, and managed to analyze the difference in patterns for both states. We were running regression discontinuity using weighted least squares regression, and we chose our thresholds to be:

- 30% maximum cases: mild stage of the pandemic in both states.
- 50% maximum cases: serious stage of the pandemic in both states.
- 70% maximum cases: very serious stage of the pandemic in both states.

By examining both sides of the regression line we analyzed if there is a causal effect that has been raised from increasing in COVID cases.

C. Propensity Score Matching

Propensity Score Matching is a statistical matching technique that implements observational data to ascertain the validity so that we can conclude that there is a potential causal link between a treatment (intervention) and an outcome(s). In addition, PSM is reliable on erasing the effect of bias and confounding variables. We used two methods to achieve PSM. They both had a similar matching process, but with different evaluation methods and models. We compared the results of the two processes in order to better understand the significance of PSM.

The first step was data preprocessing. Since TX was one of the states that had least restricted COVID-19 policies, we labeled all TX airline "0" and selected them as a control group. Then we labeled all NY airlines, marking them with "1"s and selected them as a treatment group. This can be understood as: TX has no COVID-19 policy, NY has policy. Moreover, we only chose features that are related to delay as covariates, then we dropped all rows that have NAs to refine our samples.

The first method implemented a python package "PSMPY". We randomly sampled 20000 rows and found 6957 matched rows. The main process was: 1. Instantiate the class, include all covariates and exclude others. 2. Balance the data and Calculate logistic propensity scores. 3. Using KNN matching to perform matching process. 4. Calculate effective size. Since this package used covariants to predict the intervention, we just needed to use effect size to measure the covariants' weights (As all covariants were related to delay, the one we most likely to measure was the airlines' delay time, know as DEP_DELAY_NEW) for the outcome: The smaller the covariants' effect size, the less influence on the change of the outcome (Fig.6).

For our second method, we built the propensity model by predicting the probability of receiving the treatment given the confounders. Typically, logistic regression is used for the classification model. Since propensity score tells us the probability of an individual getting the treatment given the confounders. We used logit transformation to the probability to get the propensity score.

Fig.7 shows the comparison of propensity score before and after matching.

Base on the graph (Fig.8), both groups have overlap in their propensity score before and after the logistic transformation. [8] Match records: After we got the propensity score, we were going to match the most similar control records to the treatment group. Since the propensity score is

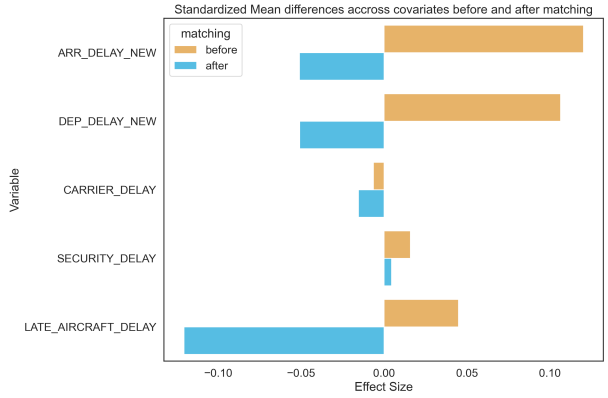

Fig. 6. Effect Size Before and After Matching

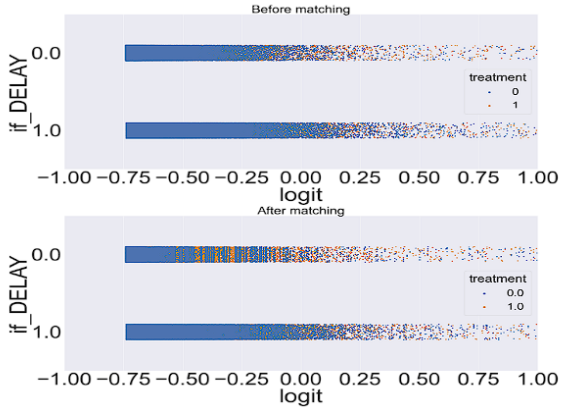

Fig. 7. Comparison of the distribution of two groups .

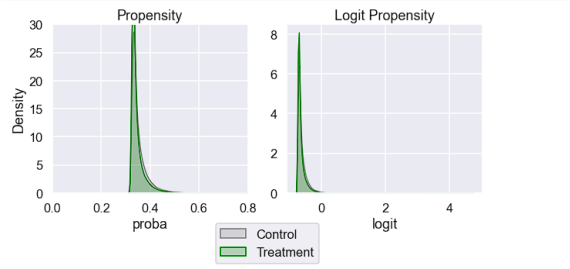

Fig. 8. Comparison of two overlaps.

a balancing score, which means if we match records based on the propensity score, the distribution of the confounders between matched records will be likely similar. The method we used here is called Nearest Neighbors algorithm. We set the caliper to be 25% of standard deviation of logit propensity score. Based on the figure8, matched elements have similar propensity scores. Matched elements might have some differences in some dimensions but their propensity score is always very close. This proves the fact that propensity scores lead to loss of information due to the compression of multiple dimensions to a single number. [9]

VII. Results

A. Difference-in-Difference Analysis

We conducted separate difference-in-difference analysis on several variables that would potentially affect the overall arrival delay, including Late Aircraft Delay, Carrier Delay, Departure Delay, Taxi In Time, and Taxi Out Time. We found that COVID increased the carrier delay and departure delay. It had barely any impact on the late aircraft delay and is statistically insignificant (with a P-Value of 0.8). In contrast, COVID decreased taxi in time, taxi out time, and arrival delay.

A closer look at the analysis output showed that Carrier Delay has the most drastic increase in magnitude: COVID increased Carrier Delay by 10.27 minutes. Carrier Delay was followed by Departure Delay, which increased by 5.28 minutes in average due to COVID. Not all airline delay factors were deteriorated by COVID. Our analysis showed that Taxi In time, Arrival Delay, and Taxi Out Time decreased by 1.52 minutes, 1.86 minutes, and 6.48 minutes respectively. These changes were ranked and presented in Fig 9.

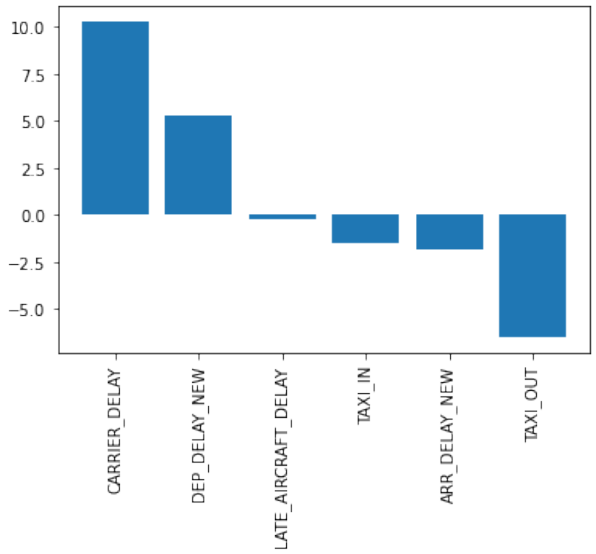

Fig. 9. Trend of Taxi In Time for New York (Left) and Texas (Right)

The findings yielded by difference-in-difference analysis could be linked together and reasonably explains each other. Carrier delay and departure delay increased because airlines were understaffed and needed more time to clean up the aircraft. After departure from the gate, there was less air traffic with fewer people traveling during COVID, contributing to shorter taxi and queuing time. The time saved during taxing was longer than the time wasted during the departure stage, thus offset the effect and made the average arrival delay decrease.

B. Regression Discontinuity

30% Threshold: mild stage of the pandemic in both states. Its results on regression discontinuity is shown in Fig.10 The arrival delay and departure delay graphs for both states

examine similar patterns. Both states show a vertical gap between the left regression line and the right, indicating both states have their flights delayed when entering the mild stage of the pandemic. However, the gap in New York state is more dramatic than it is in Texas. According to the model, NY shows a 2.6684 minutes gap in departure delay and 3.0131 minutes gap in arrival delay, while Texas shows 1.4512 minutes in departure and 1.7794 minutes in arrival, indicating that the influence of COVID cases is more impactful to flight delays in NY than in Texas. The model prediction of the intercept gap at the threshold is shown in Table I with other thresholds.

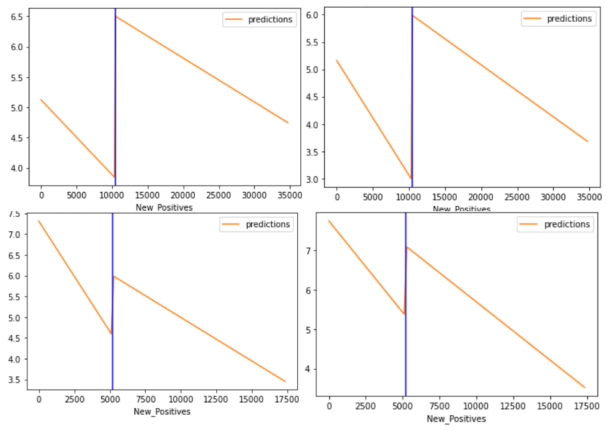

Fig. 10. Regression Discontinuity of mild stage of the pandemic. The top two graphs from left to right are regressions of departure and arrival delay on COVID cases of NY, and the bottom are the same regressions of TX.

50% Threshold: serious stage of the pandemic in both states. As shown in Fig.11, the states examine different patterns at 50% threshold. NY shows an increase in departure and arrival delay when reaching the threshold, and a decrease in delays when leaving the threshold; whereas Texas shows a continuous decreasing trend in delays both before and after the threshold. Similar to 30% threshold, the arrival delay and departure delay graphs for both states show similar patterns.

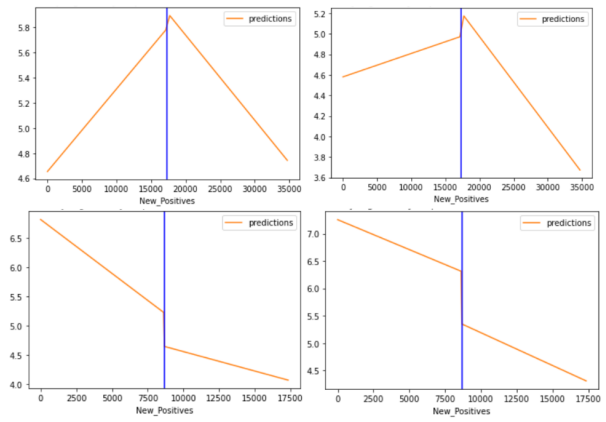

Fig. 11. Regression Discontinuity of serious stage of the pandemic.

70% Threshold: The threshold represents the very serious stage of the pandemic in both states. Regression results are shown in Fig.12. Here NY's delay drops dramatically after the threshold, while Texas again shows a general decreasing trend. At the very serious stage of the pandemic, the regression lines arrival delay and departure delay still examine similar patterns, indicating a strong correlation between the two variable at all stages of the pandemic.

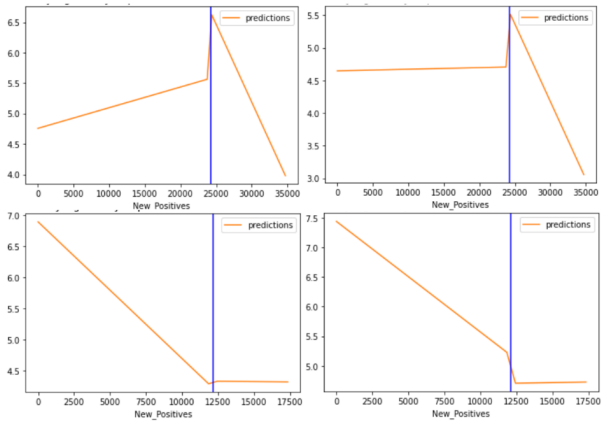

Fig. 12. Regression Discontinuity of serious stage of the pandemic.

TABLE I
REGRESSION DIFFERENCE INTERCEPT AT THRESHOLD (MINUTES)

	Dep(NY)	Dep(TX)	Arr(NY)	Arr(TX)
30%	2.6684	1.4512	3.0131	1.7794
50%	0.1229	-0.5706	0.2288	-0.9580
70%	1.0646	0.1061	0.8267	-0.4680

Result Analysis: Our results on regression discontinuity indicated that although the states have behaved differently in the mild stage, both states showed decrease in flight delays as they entered the serious stage of the pandemic. One potential reason behind the results is that as the pandemic got more and more serious, the number of flights decreased. Simultaneously, the airports became less busier and therefore reduced the waiting time for departing and arriving.

C. Propensity Score Matching

Method 1: the effect size threshold is as follows: small effect <= 0.2, medium effect <= 0.4, large effect <= 0.8. Therefore, all our effect sizes are less than 0.2, so there's little potential correlation between COVID-19 policies and flights delay.

Method 2: after we got the propensity score, as discussed before, we moved on to match the most similar control records to the treatment group. Propensity score is a balancing score, which means if we match records based on the propensity score, the distribution of the confounders between matched records will be likely similar. We used Nearest Neighbors algorithm here and we set the caliper to be 25% of standard deviation of logit propensity score(Table II).

TABLE II

EFFECT SIZE DATA BEFORE AND AFTER MATCHING

Variable	Matching Status	Effect Size
ARR_DELAY_NEW	Before	0.120
ARR_DELAY_NEW	After	-0.051
DEP_DELAY_NEW	Before	0.106
DEP_DELAY_NEW	After	-0.051
CARRIER_DELAY	Before	-0.006
CARRIER_DELAY	After	-0.016
SECURITY_DELAY	Before	0.016
SECURITY_DELAY	After	0.005
LATE_AIRCRAFT_DELAY	Before	0.045
LATE_AIRCRAFT_DELAY	After	-0.121

PSM result: Table III shows the difference between propensity scores was 0.05 which was very small. Based on the Fig.13, we can see that both groups have some overlap in their propensity scores across the range. This also proves that restriction policy during COVID in 2020 did not affect a lot of delay.

TABLE III

PSM_RESULT2

	Mean	Var	Std	Count
0	0.540802	0.248337	0.498334	123291
1	0.481892	0.249674	0.499674	123961

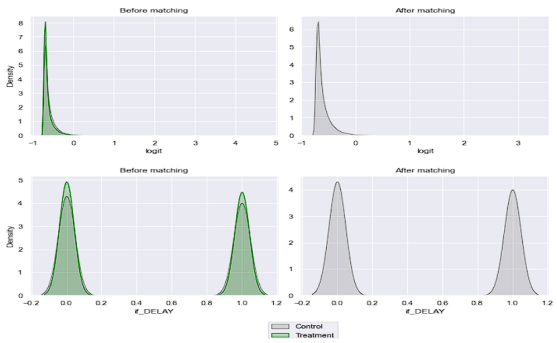

Fig. 13. Distribution after matching.

VIII. CONCLUSION

In our research we focused on looking at flight delays and its connection to the pandemic. By conducting analysis on variables that would potentially affect the overall arrival delay, we found that COVID increases the carrier delay time and departure delay time, and has barely impact on the late aircraft delay. COVID also decreased both taxi in and taxi out time, which resulted in decreased arrival delay. By conducting regression analysis, we found out that although the states behave differently in the mild stage of COVID, both states show decrease in flight delays as the pandemic became serious. In our first matching model, the average effect size is 0.0773. In the second model, the average effect size is 0.05891. Both models in our matching analysis proved that restriction policy during COVID in 2020 did not have significant affect on flight delays because matching results of both groups have overlapped in their propensity scores across the range.

IX. AMOUNT OF CHALLENGE AND DIFFICULTY

The project is a challenging one for us for the following reasons:

1) Much preprocessing on the dataset had to be performed. We combined data from multiple sources (i.e. Flight data and the pandemic data) to form the final dataset we used for our study.
2) Finding state policies related to COVID-19 control was hard. It was also hard to form a correlation between state policies since different states may have different situation regarding to the pandemic.
3) All the three methods we used in this study are not covered in the lecture. We did huge scale of self-learning to understand the methodology and apply it to get a practical solution.

REFERENCES

[1] S. Hotle and S. Mumbower, "The impact of covid-19 on domestic U.S. Air Travel Operations and commercial airport service," Transportation Research Interdisciplinary Perspectives, 09-Dec-2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198220301883. [Accessed: 15-Apr-2022].

[2] J. Yimga, "The airline on-time performance impacts of the COVID-19 pandemic," Transportation Research Interdisciplinary Perspectives, 14-May-2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198221000932. [Accessed: 15-Apr-2022].

[3] T. Suzumura et al., "The Impact of COVID-19 on Flight Networks," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2443-2452, doi: 10.1109/BigData50022.2020.9378218.

[4] M. Gupta, "Covid-19 and Economy - Researchgate," researchgate. [Online]. Available: https://www.researchgate.net/profile/Mrinal-Gupta-4/publication/340211524_COVID-19_and_economy/links/5ebc3928299bf1c09abbb530/COVID-19-and-economy.pdf. [Accessed: 02-May-2022].

[5] J.J. B. Sobieralski, "Covid-19 and airline employment: Insights from historical uncertainty shocks to the industry," Transportation Research Interdisciplinary Perspectives, 01-May-2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198220300348. [Accessed: 01-May-2022].

[6] A. Ellison, "States ranked by covid-19 restrictions," Becker's Hospital Review. [Online]. Available: https://www.beckershospitalreview.com/rankings-and-ratings/states-ranked-by-covid-19-restrictions-040821.html. [Accessed: 15-Apr-2022].

[7] "Difference-in-Differences" The World Bank. [Online] Available: https://dimewiki.worldbank.org/Difference-in-Differences [Accessed: 15-Apr-2022]

[8] Z. Luvsandorj, "Propensity score matching," Medium, 21-Apr-2022. [Online]. Available: https://towardsdatascience.com/propensity-score-matching-a0d373863eec. [Accessed: 02-May-2022].

[9] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," Multivariate behavioral research, May-2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/. [Accessed: 02-May-2022].