IS5126 HANDS-ON WITH BUSINESS ANALYTICS

Project Report

Supervisor: DR TUAN Q. PHAN

Submitted by

Cho Zin Tun (Student ID:A0098996W, Email:e0230036@u.nus.edu)

Chua Hian Choon (Student ID:A0176643X, Email:e0232247@u.nus.edu)

Ngoh Chang Chiat Vincent (Student ID:A0176587J, Email:e0232191@u.nus.edu)

Toh Pei Xuan (Student ID:A0000584R, Email:e0229629@u.nus.edu)

Ye Honghai (Student ID:A0176590W, Email:e0232194@u.nus.edu)

Flying Should Be An Enjoyable Experience
Can TSA Pre✓ really help to save time and reduce flight delays?

Abstract

Since 911 attack, security has been tightened at the airports and longer time is required for screening before flights. In Dec 2011, Transportation Security Administration (TSA) has introduced TSA Pre✓ service at some airports to expedite the screening process. This project aims to find out if the introduction of this program will help to reduce flight delays in the participating airports.

This project will work with airports and flight information datasets. Methodology such as Difference-In-Differences(DiD) is used to establish causality between flight delays and the TSA Pre✓ service.

Introduction

A flight delay in taking off and/or landing later than its scheduled time can be costly to airlines and passengers besides the inconvenience and frustrations encountered.

Flights delay is of great cost to both airlines and passengers. It disrupts airlines' and airports' operations, resulting in compensation, higher carrier costs and airfares. According to the Federal Aviation Administration of United States, flights delay cost could be of significant amount to airlines.[1] To passengers, flight delay resulting in flights arriving late at final destinations which could cause them to miss scheduled events and commitments. In short, flight delay is a huge deadweight loss in terms of economic, social and environmental costs on the economy.

It is of great value for airports and airlines to optimize their operation and reduce the flight delays. In 2016, the average cost of aircraft block (taxi plus airborne) time for U.S.

[1]Ann Brody Guy, Flight delays cost $32.9 billion, passengers foot half the bill, 18 October 2010 http://news.berkeley.edu/2010/10/18/flight_delays/

passenger airlines was $62.55 per minute. On top of this direct operating cost, it cost the airlines billions to compensate the passengers due to the delayed flights which includes accommodation, meals and flight changes. It is estimated that the average value of a passenger's time is $48.5 per hour[2]. For a aircraft with a capacity of 300 passenger, it will save the airline approximately $300 by reducing the flight delay for 1 minute.

It is widely believed that airports and airline could reduce the flight delays by improving their technologies and existing operations. As for travellers, there are also many tips available to avoid or mitigate flight delays. For example, flying morning flights, book direct flights and don't fly during peak days like weekends. However, a more scalable practice is needed for improving the overall flight on-time performance, not for the individual airlines or airport.

In this research, a different perspective was taken to assess the impact of public policies on flight delays. Airport and airline management could use the information to decide if they should enroll in some programs which could help them reduce the flight delays. For passenger, it is also helpful for them to enroll these programmes or fly from the airports with such programmes.

[2] U.S. Passenger Carrier Delay Costs http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/

TSA Pre✓™

TSA Pre✓ (Read as Pre-Check) is a U.S. government program that allows eligible and low risk US travellers to pass through an expedited security screening at certain U.S. airports. To qualify for **TSA Pre✓**, travellers must enroll on the TSA website whereby they will need to provide proof of US. citizenship and undergo a background check. They will only qualify if they had passed the criteria.

Travellers would just need to present boarding pass number when purchasing air tickets and would get the **TSA Pre✓** printed on the pass. With the **Pre✓**, travellers could proceed through the **TSA Pre✓** Lane for expedited screening, whereby travellers not required to go through standard screening such as bags, laptops and jackets. The **TSA Pre✓** lanes in turn have with shorter line, wait time and faster screening time. The **TSA Pre✓** was first launched in Year 2011, however the took up rate was never been good till Year 2014, with an increase by almost 200 folds till 0.8 millions[3]. In Year 2014, amongst the 13 airlines in 320 airports and 50 major cities across states includes California, Florida, Alaska, Hawaii, only 102 airports were registered with **TSA Pre✓** services. The remaining 218 airports had not had any **TSA Pre✓** services and lanes then.

[3] Reflections on Risk-based Security in 2014
https://www.tsa.gov/blog/2015/01/15/reflections-risk-based-security-2014

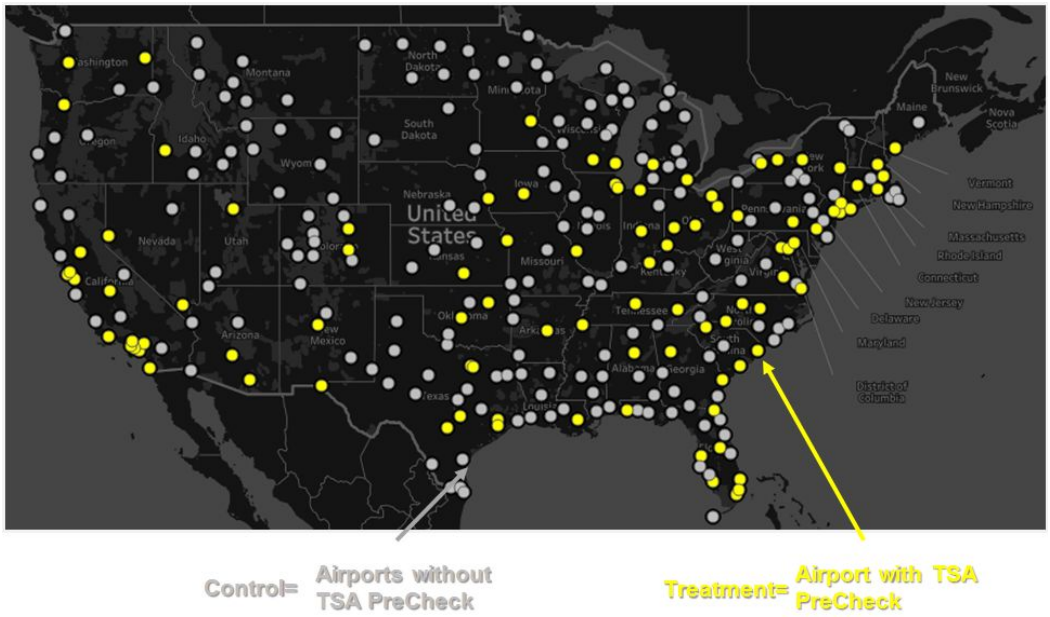


Figure: Airports in US with **TSA Pre✓** *(yellow) and without* **TSA Pre✓** *(grey)*

Hence, this project attempts to analyse whether the **TSA Pre✓** services can cause a reduction in flight delays, given the reduction in security screening time. As mentioned above, this service only picked up in 2014, hence analysis was done based on 2013 (assuming airports had no such service) and 2014 (assuming only the 102 airports had such service).

Dataset

The primary dataset used for this project is the monthly Air Travel Consumer Report published by The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics from 2013 to 2015. The report tracks the on-time performance of domestic flights operated by 13 major airlines in over 320 airports. The dataset includes summarized information on the flight's planned schedule, actual time and period of

delay due to different factors such as weather, carrier, NAS, security and late aircraft. The second major dataset which tracks the number of passengers and flights of each airport by month is also from Bureau of Transportation Statistics. Table of data sources that were used in the project can be found in *Table 1*.

Method of Data Crawling and Processing

Main Dataset: Domestic Flight On-Time Performance

Bureau of Transportation Statistics (BTS) database contains very detailed record on every flight On-Time performance. There are about 500,000 domestic flight records every month with 111 data features available. Upon studying the domain and features, 31 features were select as the based dataset. Over 11 millions records of flight data were download in CSV format and loaded into R for consolidation. Due to hardware limitation, features selection was performed to further reduce the number of feature used. In the final dataset, 9 features (*Table 2 in appendix*) were selected from this dataset for analysis.

Airport Related Statistic Data

Airport related statistic data, including passenger and number of scheduled flight departure, were crawled from BTS website. 2 years of data from over 320 airports were crawled, starting from Feb 2013 to Jan 2015, using Python. Beautifulsoup and requests libraries were utilized to achieve the purpose. Since the webpage for loading data is

handled with AJAX to attain single page app style, the url does not change. Therefore, native method of crawling data with GET method while changing the URL accordingly did not work. As a consequence, POST method was adopted and form data,with airport name and month-year, was sent through POST method as Json object. Features from airport related statistic data can be found in *Table 3*.

TSA Pre✓ Airports (2014)

Airfarewatchdog Website hosted a PDF file on the list of airports and airlines participating in **TSA Pre✓** program as of Feb 2014. Due to inconsistent table sizing used in the file direct scraping of the airport codes was not possible, thus a conversion from PDF to Word document was performed. From the converted document, column on participating airports was extracted into a CSV file. As the airport codes are of IATA format, no further matching is required.

Method of Post-processing

The main analysis was done using two sets of data. Main flight datasets for Feb 2013 to Jan 2015 were read into R. Cleaning of data was then done to remove irrelevant data-points (like flight cancellation or diversion) and columns. The data was also aggregated in R based on origin airport, month and year, as analysis was conducted based on a monthly basis. The 2nd dataset, “Airport related statistic”, contained information based on each airport and manipulation and cleaning was done in Excel to account for various issues like inconsistent representation of blank fields in the raw

data. Both datasets were then combined in R using the merge function and was merged based on month, year and flight origin. Refer to *Table 4* for Data transformation and merging information.

A separate set of additional analysis was also done based on flight routes instead of only the airports. Similarly, only two sets of data were used. The first remained the same as the initial analysis except that aggregation was done based on month, year, origin airport and destination airport. The other dataset used was generated from the 2nd dataset in the initial analysis using vlookup in excel. It was created to contain data based on routes (both origin and destination airports) instead of only the origin airports. Similarly, both datasets were combined in R using the merge function and was merged based on month, year, flight origin and flight destination.

Both analysis used the list of **TSA Pre✓** airport given in "*TSA Pre✓ Airports (2014)"* for assignment of treatment and control groups.

Method of Analysis

Causal analysis was done using difference-in-difference (DID) method. DID studies the differential effect of a treatment on a 'treatment group' versus a 'control group' and calculates the effect of a treatment on an outcome by comparing the average change over time in the outcome variable for the treatment group, compared to the average change over time for the control group.

DID Analysis Based on Arrival Delay Time and Origin Airport

Arrival delay time was defined in the model as the number of minutes a flight arrive later than scheduled. The arrival delay time for flights that arrived earlier than scheduled was taken as 0. After aggregation explained in method of post-processing above, each data-point was then assigned to be in the treatment group if the origin airport has the **TSA Pre✓** program on February 2014 or the control group if the origin airport do not have the **TSA Pre✓** program on February 2014. A time label was also assigned to each data-point, based on before and after February 2014 as the cutoff point of the treatment. Analysis was done based on the equation given below.

$$ARR\_DELAY = \beta_0 + \beta_1(time) + \beta_2(tr) + \beta_3(time) * (tr) + \beta_4 TAXI\_OUT + \beta_5 Passenger\_Departure + \beta_6 Scheduled\_Flights\_Departure + \beta_7 Scheduled\_Flights\_Departure\_Rank + \beta_8 Passenger\_Depature\_Rank$$

Interpretation

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.232e+00  5.246e-01   4.254 2.12e-05 ***
time                            1.016e+00  2.149e-01   4.725 2.34e-06 ***
tr                              6.188e-01  3.488e-01   1.774 0.076098 .
TAXI_OUT                        7.499e-01  2.868e-02  26.147  < 2e-16 ***
Passenger_Departure            -1.671e-07  9.110e-08  -1.834 0.066702 .
Scheduled_Flights_Departure     4.168e-06  8.577e-06   0.486 0.627034
Scheduled_Flights_Departure_Rank  5.996e-03  1.547e-03   3.877 0.000107 ***
Passenger_Depature_Rank         -2.034e-03  2.647e-03  -0.768 0.442311
time:tr                         -1.013e+00  3.741e-01  -2.708 0.006775 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: DID result for arrival delay time based on origin airport

Results from the analysis showed that the flights that took off from airports with **TSA Pre✓** program had a significantly lesser flight delay time from as compared to flying off from a airport without **TSA Pre✓** program. Refer to ***Graph 1*** for visualization of DID results.

Comparing the two groups of airports after 2014 (taking only significant terms), flights that took off from airports with **TSA Pre✓** program has an average flight delay reduction of 1.013 minute. Although a minute may seem insignificant, when translated to the total time saved by all passengers, and the number of flights that fly a month, the time difference can mean a lot.

Robustness Test

Different robustness tests were performed to evaluate if the treatment actually reduces the flight delay. The tests contained Propensity Score Matching (PSM), random assignment of treated and control group, and placebo test.

Propensity Score Matching (PSM)

To reduce the imbalance of data within control group and treatment group, the propensity score matching technique was utilized to create control group with airports that are similar to those under treatment group in terms of control variables. Observations were selected based on similar propensity score matching with the

treatment group. After replacing original dataset by propensity matched dataset, the result became more statistically significant (refer to *Figure* below) and coefficient was also similar compared to the baseline model. All the robustness test performed hereafter used propensity matched dataset instead of original dataset.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.158e+00  5.500e-01   3.923 8.87e-05 ***
time                            1.028e+00  2.239e-01   4.591 4.52e-06 ***
tr                              8.903e-01  3.290e-01   2.706  0.00684 **
TAXI_OUT                        7.176e-01  2.743e-02  26.166  < 2e-16 ***
Passenger_Departure            -1.324e-07  6.834e-08  -1.937  0.05275 .
Scheduled_Flights_Departure     2.573e-06  6.440e-06   0.400  0.68954
Scheduled_Flights_Departure_Rank  3.570e-03  2.137e-03   1.670  0.09493 .
Passenger_Depature_Rank         3.704e-03  3.964e-03   0.934  0.35023
time:tr                        -1.033e+00  3.172e-01  -3.255  0.00114 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: DID result for arrival delay time based on origin airport after PSM

Random Assignment of treated and control groups

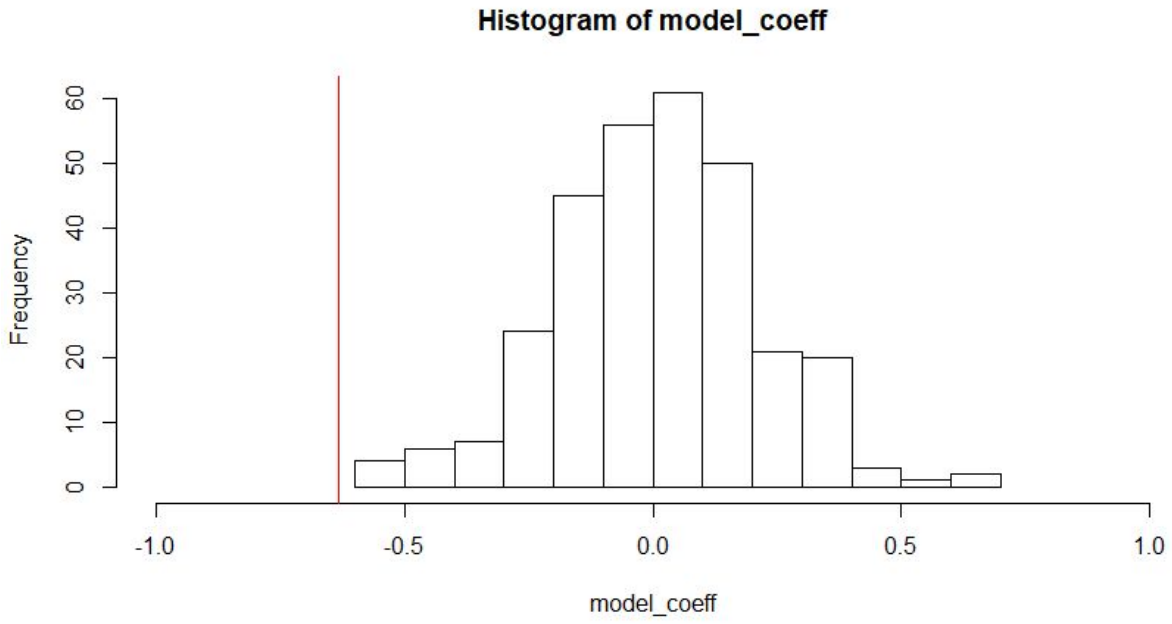
Treatment dummy was randomly assigned to airports regardless of the fact whether they were enrolled to **TSA Pre✓** or not. After that, DID model was applied to the new dataset and found out that the interaction between treatment dummy variable and time was not statistically significant. The result aligned with the expectation, which could be interpreted as the treatment only affected the treatment group and not the control group.

Placebo Test

It can be seen from the DID model that the treated group responds strongly to the **TSA Pre✓**. However, this cannot be firmly concluded yet that the decline in flight delay is

due to the **TSA Pre✓**. Thus, a placebo test was conducted to shuffle the treatment point, as well as control and treatment groups.

Quarters in between 2013 Q1 to 2014 Q1 were iteratively assigned as the treatment point, and with each assigned quarter, treatment dummy was randomly assigned to airports to form control and treated groups.



From the histogram on all the coefficients from the placebo test iterations, it can be observed that the mean was close to zero. Compare with the actual model, there were no overlapping between the true coefficient and coefficients from the placebo test. Hence, this approach validated that the decline in flight delay was mainly due to the **TSA Pre✓**.

Further analysis

Further analysis was also done to explore the **TSA Pre✓** program in terms of time saved.

DID analysis based on arrival early time and origin airport

Further analysis was done based on the number of minutes a flight arrives ahead of schedule. Dependent variable in this model is given by the number of minutes a flight arrives ahead of schedule; if the flight was late, the time was taken as 0 (this is as opposed to the initial analysis which uses flight delay time).

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      6.469e+00  1.859e-01  34.790  < 2e-16 ***
time                            -7.050e-01  7.618e-02  -9.254  < 2e-16 ***
tr                              -3.542e-01  1.236e-01  -2.865 0.004186 **
TAXI_OUT                         2.011e-02  1.017e-02   1.978 0.047939 *
Passenger_Departure             -1.088e-07  3.229e-08  -3.368 0.000761 ***
Scheduled_Flights_Departure      9.554e-06  3.040e-06   3.142 0.001682 **
Scheduled_Flights_Departure_Rank -5.402e-04  5.482e-04  -0.985 0.324476
Passenger_Depature_Rank          6.048e-03  9.384e-04   6.445 1.23e-10 ***
time:tr                          7.902e-01  1.326e-01   5.960 2.64e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: DID result for arrival early time based on origin airport

Result shows that flights which took off from airports with **TSA Pre✓** program arrived earlier than flights that took off from airports without **TSA Pre✓** program. This result corresponds with the initial analysis where flights had shorter delay time when the origin was an airport with **TSA Pre✓** program.

DID Analysis Based on Arrival Delay Time and Different Groups based on Origin and Destination

Further analysis was done by categorizing flights into 4 different groups based on whether the origin and destination airport are participating airports in **TSA Pre✓** program (in 2014) or not.The 4 groups are:

1. Participating airports to Participating airports
2. Participating airports to Non-participating airports
3. Non-participating airports to Participating airports
4. Non-participating airports to Non-participating airports

Treatment Group	Control Group	Significance of interaction term
✓ ✈ ✓	✓ ✈ ✗	Yes
✓ ✈ ✓	✗ ✈ ✓	Yes
✓ ✈ ✓	✗ ✈ ✗	No

✓ --> Airport with TSA Pre✓
✗ --> Airport without TSA Pre✓

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.112e+00  1.877e-01  21.902  < 2e-16 ***
time                           -2.328e-01  7.051e-02  -3.302 0.000961 ***
tr2                            -1.631e+00  1.781e-01  -9.156  < 2e-16 ***
TAXI_OUT                        4.559e-01  7.694e-03  59.256  < 2e-16 ***
Passenger_Departure            -3.360e-07  1.593e-08 -21.096  < 2e-16 ***
Scheduled_Flights_Departure     3.647e-05  1.445e-06  25.243  < 2e-16 ***
Scheduled_Flights_Departure_Rank  7.355e-02  5.536e-03  13.285  < 2e-16 ***
Passenger_Depature_Rank        -3.735e-02  5.884e-03  -6.348  2.2e-10 ***
Passenger_Arrival              -9.032e-09  4.676e-09  -1.932 0.053403 .
Passenger_Arrival_Rank          8.152e-03  9.700e-04   8.405  < 2e-16 ***
time:tr2                        4.847e-01  1.800e-01   2.694 0.007071 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure: DID result for ✓ ✈ ✓ **VS** ✓ ✈ ✗*

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.622e+00  1.759e-01  26.280  < 2e-16 ***
time                            -2.213e-01  7.133e-02  -3.102  0.00192 **
tr3                             -3.550e-01  1.826e-01  -1.945  0.05184 .
TAXI_OUT                         4.716e-01  7.788e-03  60.549  < 2e-16 ***
Passenger_Departure             -2.897e-07  1.657e-08 -17.482  < 2e-16 ***
Scheduled_Flights_Departure      2.294e-05  1.504e-06  15.257  < 2e-16 ***
Scheduled_Flights_Departure_Rank 1.263e-02  1.376e-03   9.178  < 2e-16 ***
Passenger_Depature_Rank         -5.902e-03  2.221e-03  -2.657  0.00788 **
Passenger_Arrival                5.906e-08  5.528e-09  10.682  < 2e-16 ***
Passenger_Arrival_Rank           2.917e-02  2.118e-03  13.773  < 2e-16 ***
time:tr3                         1.136e+00  1.820e-01   6.243  4.3e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure: DID result for ✓ ✈ ✓ **VS** ✗ ✈ ✓*

Based on results from figures above of interaction term, 0.5 to 1 minute per flight can be saved when flying from and to a **TSA Pre✓** as compared to either of the origin or destination is not a **TSA Pre✓** airport. This results make sense since flights (passengers) would usually also fly back from (initially) Destination to (initially) Origin airports, hence destination airports with **TSA Pre✓** would also have a similar effect as Origin Airports with **TSA Pre✓**.

By performing the DID validation at origin and destination with **TSA Pre✓**, the analysis could address potential endogeneity problem that may occur with the airports that had also upgraded their infrastructure and systems during the period of analysis.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5.672e+00  1.771e-01  32.028  < 2e-16 ***
time                             -2.134e-01  6.726e-02  -3.172  0.00151 **
tr4                              -8.845e+00  6.223e-01 -14.214  < 2e-16 ***
TAXI_OUT                          4.245e-01  7.863e-03  53.991  < 2e-16 ***
Passenger_Departure              -2.622e-07  1.670e-08 -15.703  < 2e-16 ***
Scheduled_Flights_Departure       2.322e-05  1.489e-06  15.598  < 2e-16 ***
Scheduled_Flights_Departure_Rank -1.303e-03  3.321e-03  -0.392  0.69479
Passenger_Depature_Rank           1.839e-02  4.090e-03   4.497  6.9e-06 ***
Passenger_Arrival                 6.861e-09  5.595e-09   1.226  0.22009
Passenger_Arrival_Rank            1.588e-02  1.753e-03   9.056  < 2e-16 ***
time:tr4                          8.480e-01  6.140e-01   1.381  0.16728
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: DID result for ✓ ✈ ✓ **VS** *✗ ✈ ✗*

Possible reasons for such results could be that non-participating airports tend to be smaller scale and have less flights. The average number of routes (data-points) between participating airports was about 3000 while routes between non-participating airports was only around 40. Also, given the graph below, it could also be concluded that these non-participating airports had less delays. Hence, the reduction in flight delays due to both airports participating might not be significant when compared against the both non-participating airports groups which already had much less flight delays.

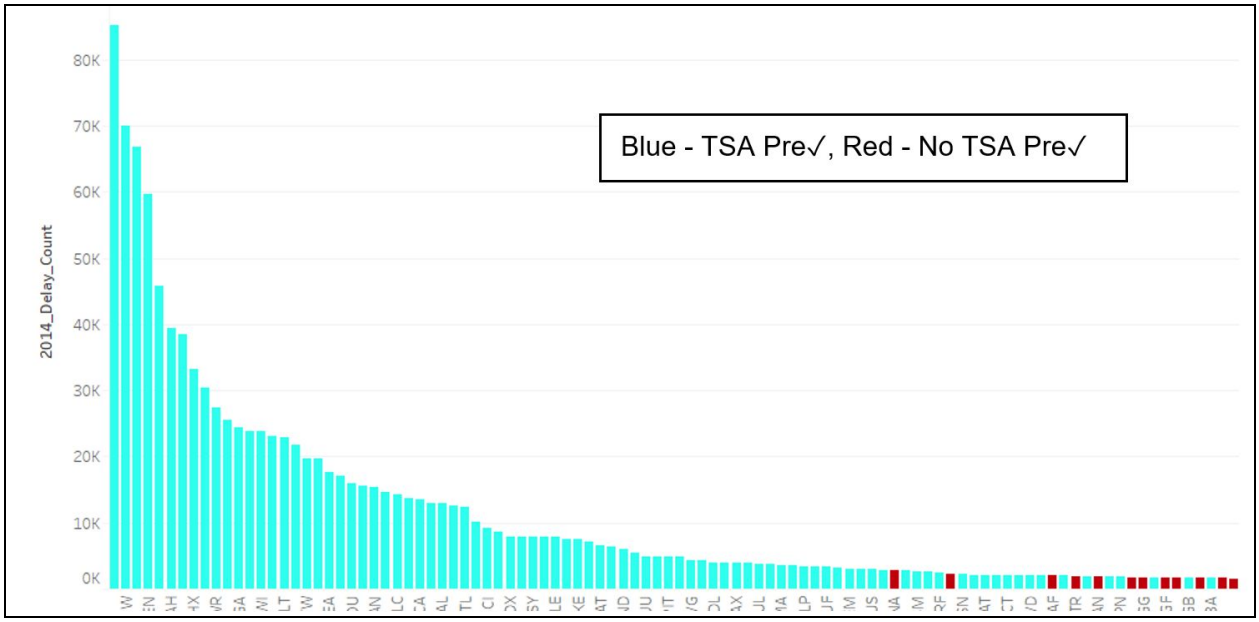


Figure: Graph showing the delay counts based on participating and non-participating airports

Conclusion

In this project, the impact of **TSA Pre✓** on flight delays were examined through a DID model. Passengers who are travelling from an airport with **TSA Pre✓** program are expected to experience less time of flight delay for about 1 minute compared to those who are travelling from an airport without **TSA Pre✓** program. In addition, 0.5 to 1 minutes can be saved when both connecting airports are enrolled in **TSA Pre✓** program compared to either of departure or arrival airport is not enrolled in **TSA Pre✓** program.

TSA Pre✓ program can help travellers save time on the security checks and have a better experience. Moreover, it helps to reduce flight delays and avoidable cost due to

the interruption of the associated business activities caused by the flight delays. **TSA Pre✓** is expected to have a greater impact with more airports and travellers enrolled in the program. Besides US domestic flights and travellers, passengers who are travelling on international flights to US can also enjoy expedited screening process by applying Global Entry program. The impact of Global Entry program on international flight delays could be examined in future works.

Limitations

This analysis was done based on assumptions that all airports did not have the **TSA Pre✓** in 2013, while only the list of airports listed in the dataset participated in the program in 2014 (assumed no other airports joined from March 2014 to Jan 2015). This was due to limited information regarding the dates the airports had joined the **TSA Pre✓**.

There may also be some form of selection bias as it seemed that airports that had more delays are those that had participated in this program in 2014. Reason may be that there was no need to participate in the program if the airports were less busy.

Future Works

The **TSA Pre✓** take up rate is not as encouraging as it much expected. There were only 5.6 millions US citizen enrolled in Year 2017, a great fall apart from 29 millions goal

that TSA set for Year 2019[4]. There are many articles reported reasons include overestimation of the enrollment, cut the screening staffs, inadequate promotions and high membership fees. In addition, it has reported that $75.7 billion has been spent in upgrades through 2019 across USA airports. Hence, the team could continue to work **TSA Pre✓** contribution to flight delays through sourcing of airports and airlines registered with **TSA Pre✓** with the partition off of the confounding effects from the airport system upgrade.

[4] U.S. airports call for $75.7 billion in upgrades over the next four years
http://www.latimes.com/business/la-fi-airports-need-upgrades-20150317-story.html

Appendix

Data	URL	Method
Air Travel On Time Report	www.transtats.bts.gov/Fields.asp?Table_ID=236	Direct Download
Airport Related Statistic Data	www.transtats.bts.gov/Data_Elements.aspx	Web Crawling
TSA Pre✓ Airports (2014)	http://i.slimg.com/sc/aw/photo/t/ts/tsa_precheck_chart_feb_2014_update_chart.pdf	Direct Download, PDF conversion to Excel

Table 1: Data Sources

Feature	Description	Data Type
ORIGIN	Origin Airport the flight started from	Factor
DEST	Airport of which the flight landed	Factor
YEAR	The year which the flight took off	Factor
MONTH	The month which the flight took off	Factor
TAXI_OUT	Time taken for flight to taxi out, in minutes	Numeric
TAXI_IN	Time taken for flight to taxi in, in minutes	Numeric
ARR_DELAY	Number of minutes a flight reach after scheduled time; 0 if the flight is early	Numeric
EARLY	Number of minutes a flight reach ahead of schedule; 0 if the flight is late	Numeric
DISTANCE	Distance between origin airport and destination	Numeric

Table 2: Features from main dataset

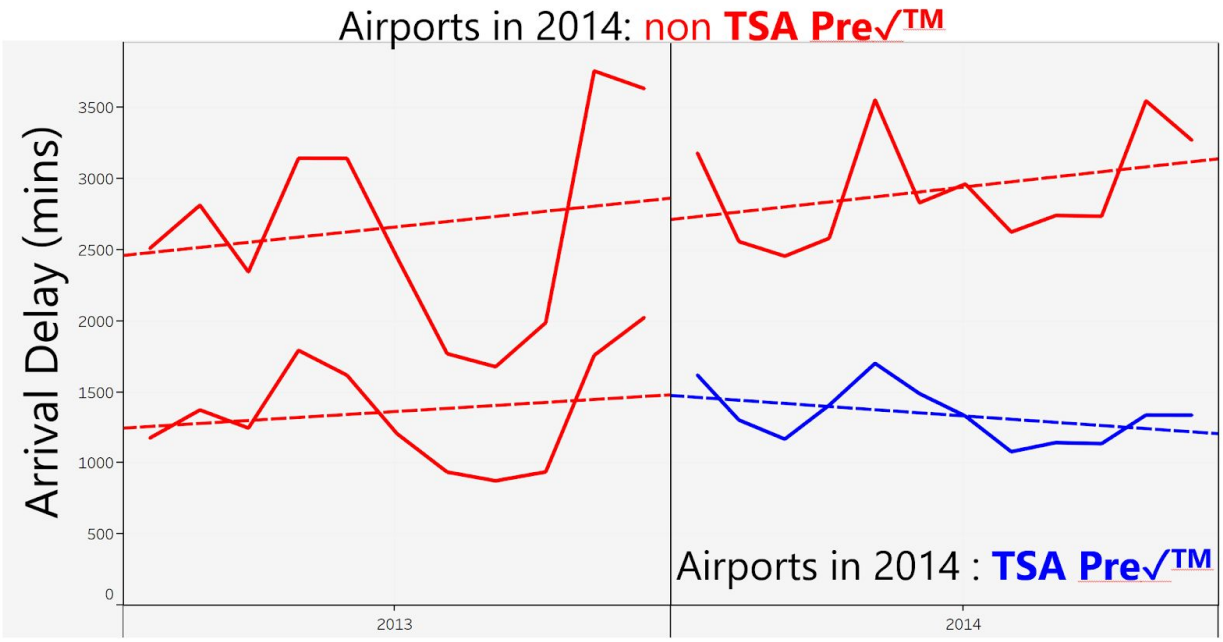
Feature	Description	Data Type
Airport	IATA airport code	Factor
Year	Year of collected data	Factor
Month	Month of collected data	Factor
Passenger_ Arrival	Number of passenger arrived at the airport	Numeric
Passenger_ Departure	Number of passenger departed from the airport	Numeric
Scheduled_ Flights_ Departure	Number of scheduled departure flight from the airport	Numeric
Passenger_ Arrival_ Rank	Airport ranking based on number of passenger arrived.	Numeric
Passenger_ Depature_ Rank	Airport ranking based on number of passenger departed.	Numeric
Scheduled_ Flights_ Departure_ Rank	Airport ranking based on number of scheduled departure flight.	Numeric

Table 3: Features from airport related statistic dataset

Primary	Join method	Keys	Secondary	Remarks
Main dataset	Inner Join	Airport Code, Year, Month	Airport related	No missing data

			statistic data	
Main dataset	Inner Join	Airport Code	Airport Infrastructure	No missing data
Main dataset	Left Join	Airport Code	TSA Pre✓ Airports (2014)	Only data after Feb 2014 were joined and assigned with treatment indicator as 1, else 0 will be assigned

Table 4: Data transformation and merging



Graph 1: DID Result Visualization