ORIE 3120
Final Group Project

Flight Delay Analysis and Prediction

Introduction

Flight delay is an annoying thing when people are traveling. Thus, both airlines and passengers always tried their best to avoid delays, and so did us. This project will explore a dataset of 2015 flight delays and cancellations in the United States. Our goal is to find the relationship between delay and multiple factors, such as airlines, airports, time of day, seasons, etc. We will also predict the airport delays by the time and anticipate the delay on some specific airlines. Finally, we will advise how to avoid delays when we choose the flight based on our findings in the project.

Data Cleanup/Pre-processing

The dataset we use is the 2015 Flight Delays and Cancellations, which summarized information of the number of on-time, delayed, canceled, and diverted flights, and their corresponding delay time in 2015. The data is from the U.S. Department of Transportation's statistics, which tracks the on-time performance of domestic flights operated by large airlines. The data consists of flights.csv, airports.csv, and airlines.csv, which categorizes flight information, airline companies, and airports. We mainly used the flights.csv for our analysis since it contains the departure and arrival delays we focused on and tried to predict. Upon analyzing the dataset, we prompted to answer the following questions:

1. Does the arrival delay time closely relate to the departure delay time?
2. How do the predictors, such as airport, time within a day, seasons affect the delay time?
3. How can we better predict the delay time of flight? What should we focus on?

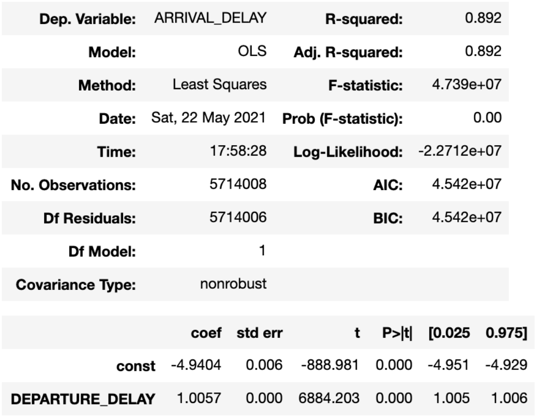
A standardized data is convenient and, thus, critical for our further analysis. We first import the three CSV files to a python notebook and make dataframes for them. There are 5,819,079 records and 31 variables in total in our original dataset. Variables are indicative of flight date, airline, flight number, departure and destination airports, scheduled departure and arrival times, actual departure and arrival times, delay time, delay reasons, and more. The terms are listed below:

YEAR, MONTH, DAY, DAY_OF_WEEK, AIRLINE, ORIGIN_AIRPORT, DESTINATION_AIRPORT, SCHEDULED_DEPARTURE, SCHEDULED_ARRIVAL, DEPARTURE_TIME, ARRIVAL_TIME, DEPARTURE_DELAY and ARRIVAL_DELAY, DISTANCE.

Null and non-related values exist as well, which would negatively impact our later data mining. Thus, we cleaned up the data by extracting the columns we needed for the later analysis and then discarded rows containing the null values. We also changed the columns involving time to type DateTime and time duration to float/integer for later use. After standardization above, our dataframe became more elegant and neat.

Initial Analysis and Data Visualization

Our objective is to analyze the relationship between flight delay vs. single/multiple factors. The delay usually refers to delay in arrival because it will affect passengers' follow-up arrangements, such as transfer to another flight, pick up by friends, or attend conferences. However, after we built a linear regression model on 'arrival delay' and 'departure delay,' as shown in Fig.1, we found a high correlation, which means the more extended departure delay flight has, the longer arrival delay the flight will have. In that case, only focusing on the analysis of the departure delay can reduce the complexity of the model, containing less number of predictors, so that we can analyze better. Therefore, this paper will focus on the analysis of the departure delay.



Dep. Variable:	ARRIVAL_DELAY	**R-squared:**	0.892
Model:	OLS	**Adj. R-squared:**	0.892
Method:	Least Squares	**F-statistic:**	4.739e+07
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	0.00
Time:	17:58:28	**Log-Likelihood:**	-2.2712e+07
No. Observations:	5714008	**AIC:**	4.542e+07
Df Residuals:	5714006	**BIC:**	4.542e+07
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-4.9404	0.006	-888.981	0.000	-4.951	-4.929
DEPARTURE_DELAY	1.0057	0.000	6884.203	0.000	1.005	1.006

Fig.1

By reading and analyzing descriptions of all predictors, we found that these four variables significantly impact departure delays, which are airlines, departure airport, departure date, and departure time.

- Airlines:

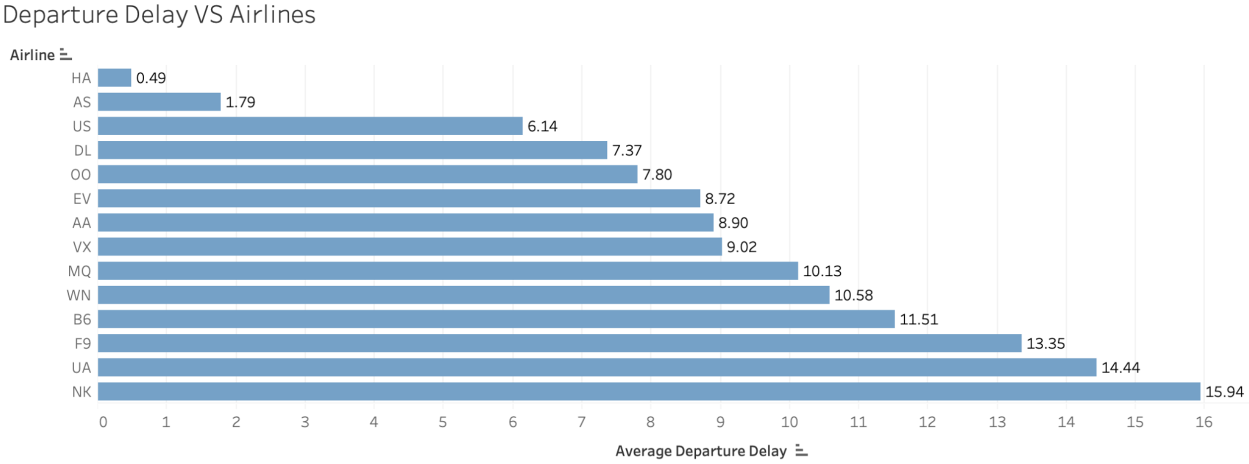


Fig.2

When it only contains airlines as the predictor, their mean departure delay is quite different. Hawaiian Airlines has the shortest departure delay, while Spirit Airlines has the longest departure delay. After our research, we found that it mainly depends on the management level or method of different airlines.

- Departure Airport:

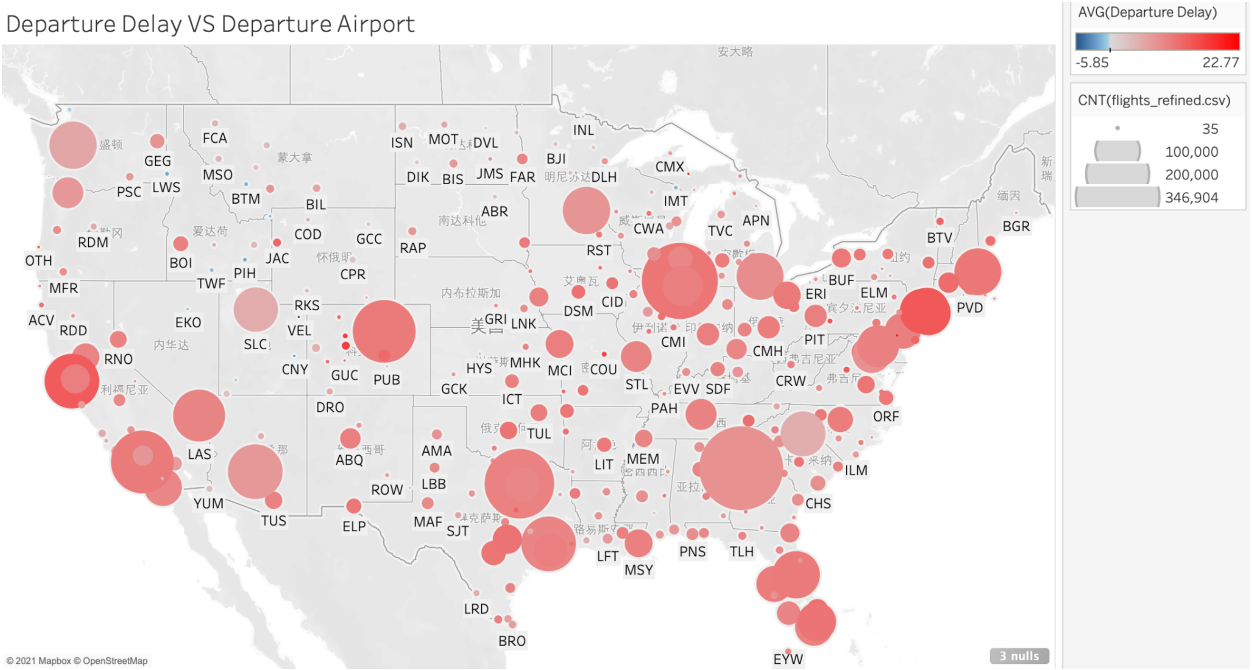
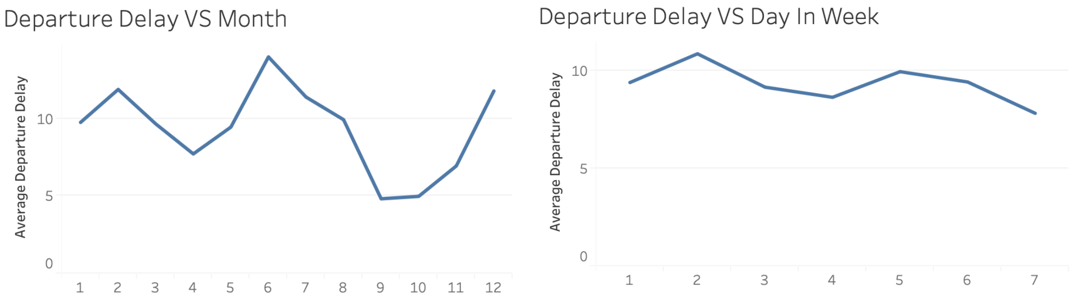

Departure Delay VS Departure Airport

Fig.3

This plot visualized the number of departure flights in every airport. The redder in color means the longer departure delay, while the bluer color indicates the shorter departure delay(in advance). We can see that most large airports have departure delays between 5 to 15 mins. For example, SFO, EWR, and JFK have an average departure delay of about 15 mins.

- Departure Date：


Departure Delay VS Month

Departure Delay VS Day In Week

Fig.4 / 5

When we analyzed the departure delay on the predictor data, we explored the relationship in different periods, year, and a week. Winter and summer often have longer delays, and we guess it may be caused by rain and snow, although the climate is different in different regions of the U.S. During spring and autumn, the weather is better, and flight is less prone to delay. It also has some relationship between delay and day in week: Sunday always has the least departure delay.

- Departure Time：

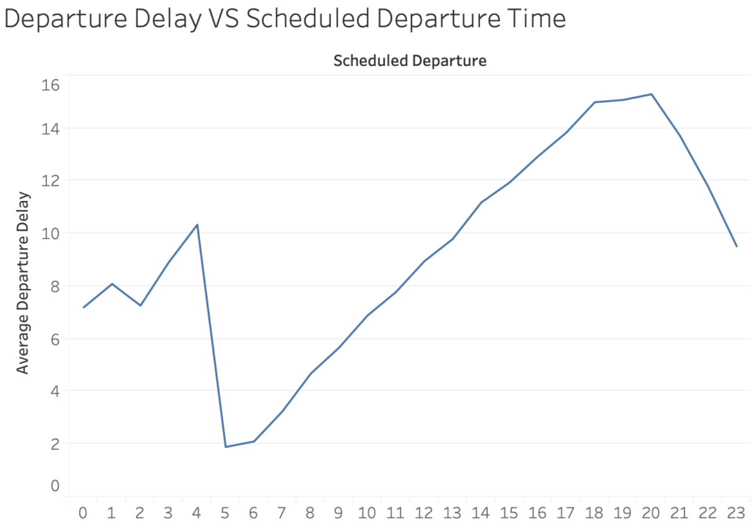


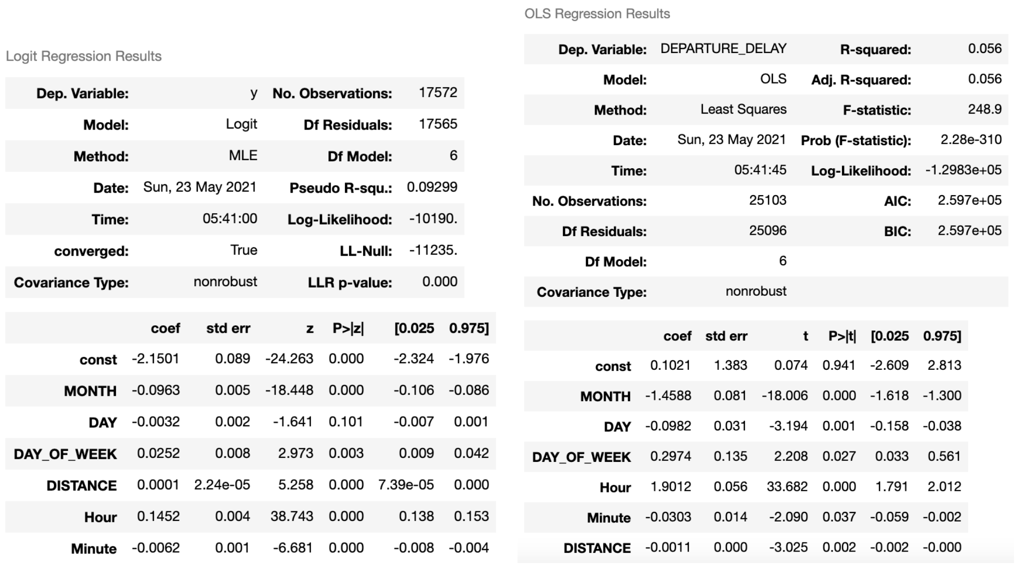
Fig.6

Each day, the time of departure delay is also related to the scheduled departure time. Typically, the flights departing in the morning have the least delay time, increasing with time until 8 p.m. That's because one aircraft often has multiple flights each day. If previous flight delays, it will affect future flights. In that case, evening flights have a longer accumulated delay time, so that they have a longer average departure delay.

Data Prediction

Our goal is to predict the departure delay for the convenience and welfare of airlines and passengers. Avoiding delays can benefit all of the parties and maintain our decent respect with time. Airline operation is complex, so we strive to take multiple factors as predictors to find whether we could forecast the occurrence of a delay. We referred to the Departure Delay VS Airlines to choose the Delta Airlines, which embrace the lowest delays while well-known, and we chose JFK because it's a big airport and has a decent amount of airplanes for Delta.

We first tried the logistic regression with some essential features to predict whether the flight has a departure delay. We set all flights which delay more than 15 minutes as 1(delayed), else they will be 0(Non-delayed). However, the model doesn't provide a high enough R-square, indicating its performance is not good enough, as shown below. Then we tried standard linear regression. Unfortunately, the same issue has been detected.



Logit Regression Results

Dep. Variable:	y	**No. Observations:**	17572
Model:	Logit	**Df Residuals:**	17565
Method:	MLE	**Df Model:**	6
Date:	Sun, 23 May 2021	**Pseudo R-squ.:**	0.09299
Time:	05:41:00	**Log-Likelihood:**	-10190.
converged:	True	**LL-Null:**	-11235.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-2.1501	0.089	-24.263	0.000	-2.324	-1.976
MONTH	-0.0963	0.005	-18.448	0.000	-0.106	-0.086
DAY	-0.0032	0.002	-1.641	0.101	-0.007	0.001
DAY_OF_WEEK	0.0252	0.008	2.973	0.003	0.009	0.042
DISTANCE	0.0001	2.24e-05	5.258	0.000	7.39e-05	0.000
Hour	0.1452	0.004	38.743	0.000	0.138	0.153
Minute	-0.0062	0.001	-6.681	0.000	-0.008	-0.004

OLS Regression Results

Dep. Variable:	DEPARTURE_DELAY	**R-squared:**	0.056
Model:	OLS	**Adj. R-squared:**	0.056
Method:	Least Squares	**F-statistic:**	248.9
Date:	Sun, 23 May 2021	**Prob (F-statistic):**	2.28e-310
Time:	05:41:45	**Log-Likelihood:**	-1.2983e+05
No. Observations:	25103	**AIC:**	2.597e+05
Df Residuals:	25096	**BIC:**	2.597e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1021	1.383	0.074	0.941	-2.609	2.813
MONTH	-1.4588	0.081	-18.006	0.000	-1.618	-1.300
DAY	-0.0982	0.031	-3.194	0.001	-0.158	-0.038
DAY_OF_WEEK	0.2974	0.135	2.208	0.027	0.033	0.561
Hour	1.9012	0.056	33.682	0.000	1.791	2.012
Minute	-0.0303	0.014	-2.090	0.037	-0.059	-0.002
DISTANCE	-0.0011	0.000	-3.025	0.002	-0.002	-0.000

Fig. 7 / 8

After trying these two approaches, we thought there might be some other approaches to predict the delay. We then tried to focus on one airline(still Delta) and one factor each time for a more precise prediction. In such a way, we could factor out the non-related factors that hinder us from the goal. We want to determine whether there is a periodic trend within each day or season for the departure delays.

After determining the airlines and factors we want to dive deep into, we figured that the departure airport impacts the delays. As shown below in Fig9, with the airports on the rows and airlines on the columns, some airports have more departure delays than others. Thus, choosing a specific airport can help us predict more concisely. Therefore, our goal is to predict departure delays in a particular airport with a particular airline. We also use JFK in this prediction.

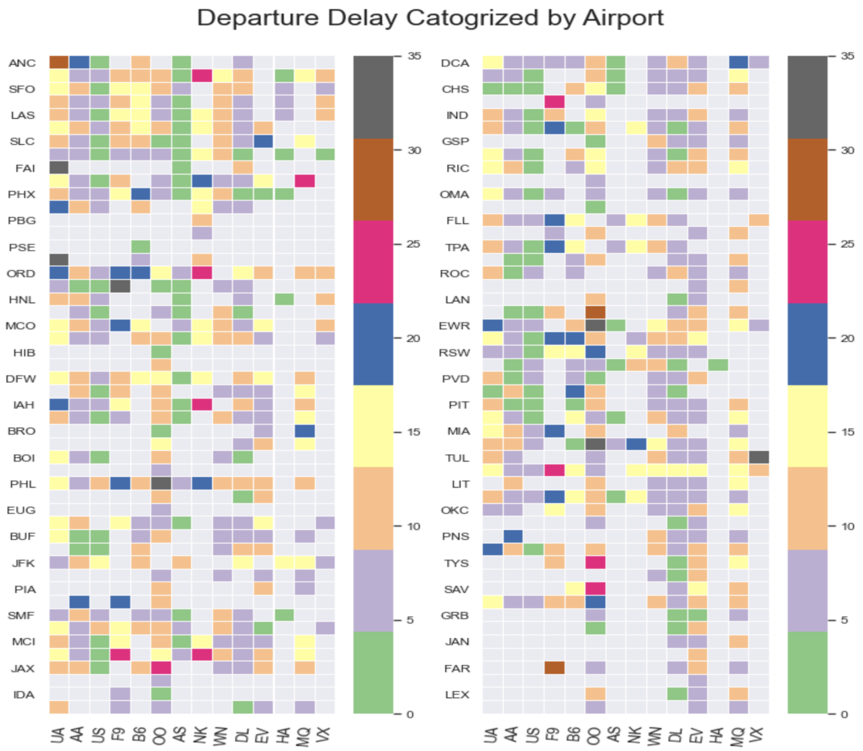
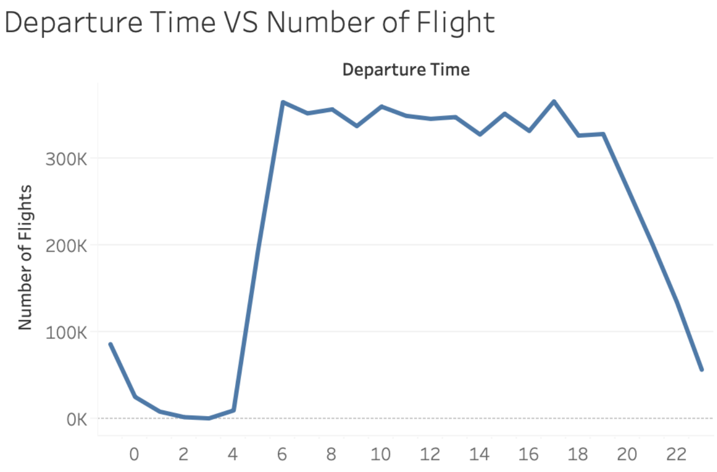


Fig9



First, we try to predict the departure delay over the busy hours during the day, from 5:00 to 18:00, as shown in the Fig Departure Time VS Number of Flight. This is because, during the active hours, which have more flights than other periods within a day, the data would be more representative and meaningful for both passengers and airlines.

We would achieve our goal by firstly dividing a month into representing hours(for example, 0 represent the first hour in the first day of a month, and 24 represents the first hour of the month's second day), grouping by the hours on the mean departure delay, then splitting the data into a training set and a test set, learning through the first three weeks, and lastly predicting the delay in the remaining week. We decided to choose the departure delays of Delta Airlines in JFK during April.

To better understand our prediction and an overview of the actual trend in the delay time, we constructed a plot of the average departure delay versus the representing hours in April of Delta Airlines in JFK shown below as Fig10.

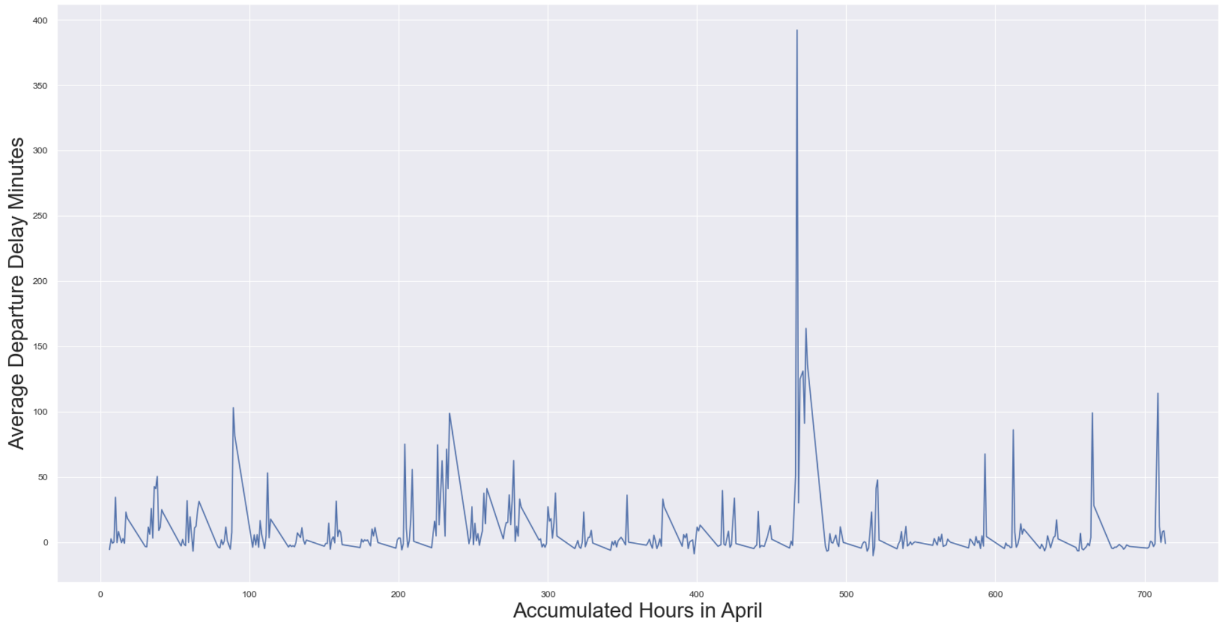


Fig10

From the plot shown above, we could easily see that there are always peaks during each day, and the pattern seems evident to human eyes.

After splitting the training and testing sets, we firstly fitted simple exponential smoothing to the training data, and used it to predict values on the test data. The original data was plotted in blue; the fitted values on the hours in the training data are in red, and the forecasted values on the hours in the test data were in black. As we can see from Fig11, the simple exponential smoothing isn't a good prediction for the black line is horizontal rather than periodical fluctuating in the original dataset. The blue line served as validation.

The Holt-Winters (ExponentialSmoothing) using additive seasonality and no trend might be a better model under our circumstance. Seasonality refers to the presence of variations that occur at certain regular intervals in time series data. Because of the peaks that we could see each day, we would take the seasonality factor into account and set the seasonal period to an appropriate level to lower the sum of squared errors. We chose not to add the trend factor because the original data we saw displayed no trend. After introducing the Holt-Winters model, we have seen that the forecasting line forecasts much better than the simple exponential smoothing model in Fig12.

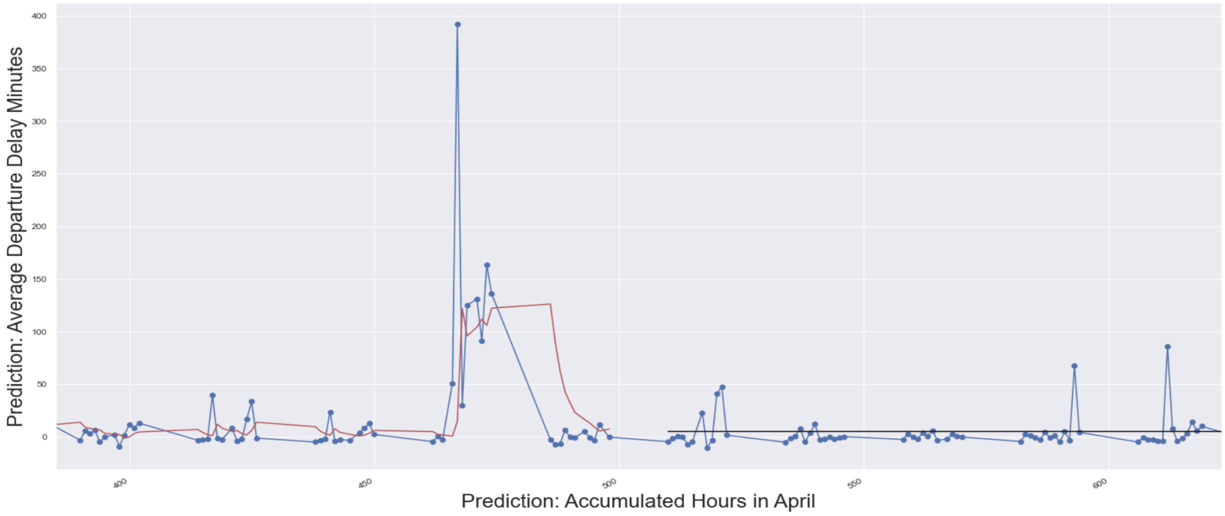


Fig11

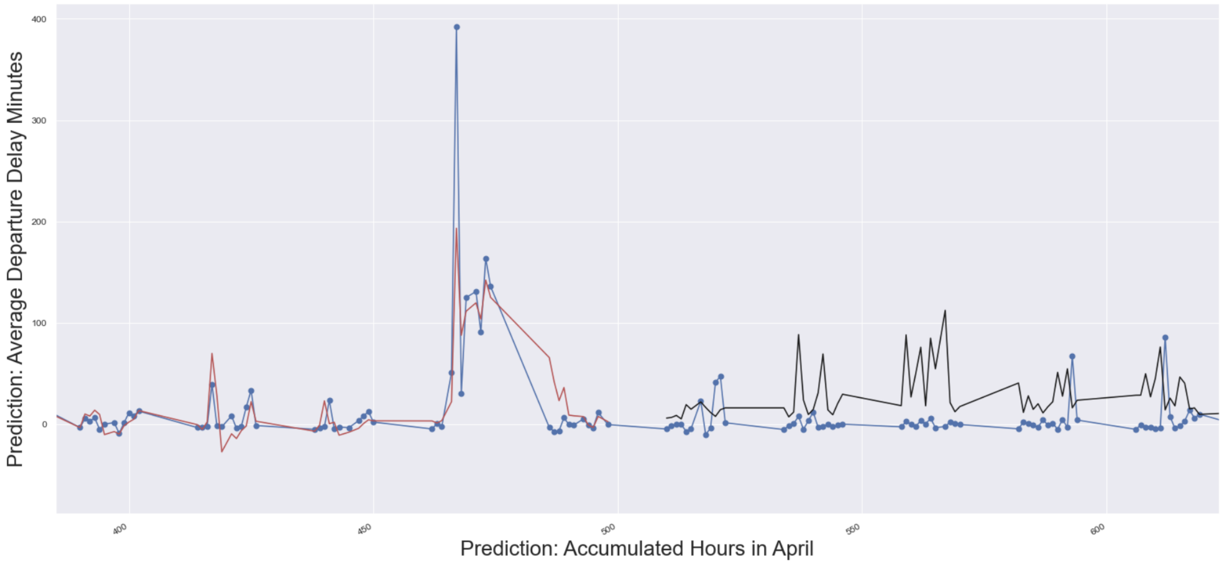


Fig12

We also tried to fit the above models to predict seasonal trends. However, the original data doesn't have an obvious pattern or a strong trend. The forecasting doesn't work well under this circumstance. We might need to work on a larger set of data or changing models for better forecasting.

Conclusion

In this report, we firstly identified that the arrival delay time has a high correlation to the departure delay, which means we only need to analyze the departure delay as the dependent variable. Then, we concluded that the departure delay depends on airlines, airports, date, and time by some visualizations. In particular, the departure delay in date and time has some period of change. Finally, we may use their periodicity to predict future delays. Even though there are unsolved trends and predictions, we believe that deeper investigation of the correlations between factors and analysis could reveal the mystery of this complex industry. For some advice to passengers, if we have the opportunity to choose our flight, we can choose the morning flight as much as possible to avoid delays, also choose the airlines that have less average delay time.