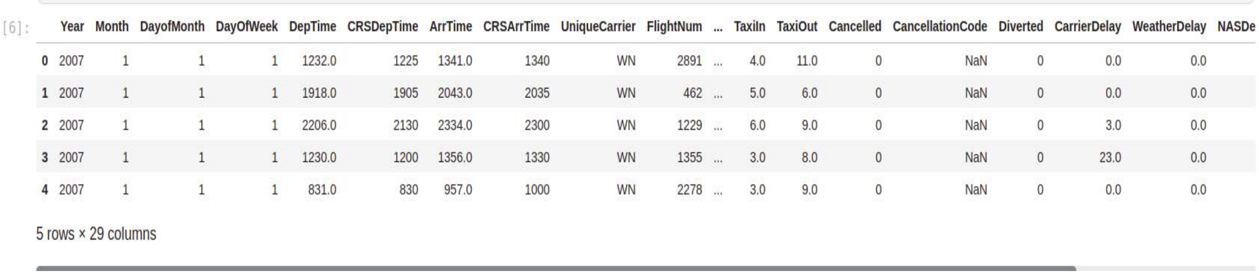
Wang Yi Ming
200548371

Introduction

In this project, I will be using flight dataset of year 2007 and 2008, analyze it and find out the answer of different questions. The analysis will include numerical and graphical results. Also, the analysis will include predictive modeling to predict the delay. Annually, there are a lot of factors that cause delays in the flights. The purpose of this project is to find out different reasons and factor to improve flight system.

Dataset



[6]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	TaxiIn	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay	NASDe
0	2007	1	1	1	1232.0	1225	1341.0	1340	WN	2891	...	4.0	11.0	0	NaN	0	0.0	0.0	
1	2007	1	1	1	1918.0	1905	2043.0	2035	WN	462	...	5.0	6.0	0	NaN	0	0.0	0.0	
2	2007	1	1	1	2206.0	2130	2334.0	2300	WN	1229	...	6.0	9.0	0	NaN	0	3.0	0.0	
3	2007	1	1	1	1230.0	1200	1356.0	1330	WN	1355	...	3.0	8.0	0	NaN	0	23.0	0.0	
4	2007	1	1	1	831.0	830	957.0	1000	WN	2278	...	3.0	9.0	0	NaN	0	0.0	0.0	

5 rows × 29 columns

Above image, shows the head of the dataset with subset of features. The dataset has 29 variables in total including variables related to time and delays.

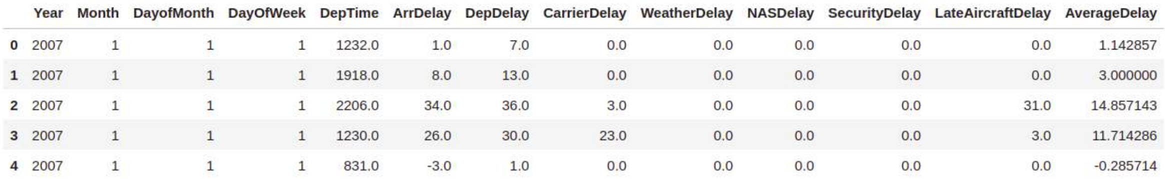
1. When is the best time of day, day of the week, and time of year to fly to minimize delays?

In this section, I will try to find three main point of time as given below:

1. Best time of the day.
2. Best day of the week.
3. Best time of the year

The dataset will be analyzed to find out these points in time that have minimum delay in flights. The problem can be tackled in multiples ways with different analysis.

To find the year with minimum delay, First, we extracted all the necessary features including Month, DayofMonth, DayOfWeek, Year, DepTime and all types of delays. We took the mean by row of all the columns representing different types of delays and we added a new feature a AverageDelay.

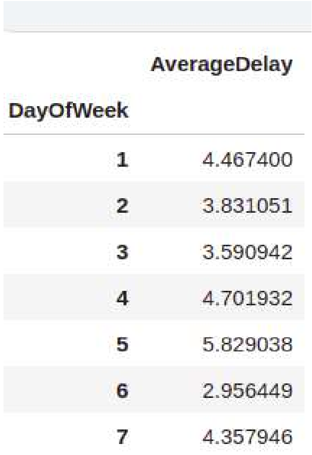


	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrDelay	DepDelay	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	AverageDelay
0	2007	1	1	1	1232.0	1.0	7.0	0.0	0.0	0.0	0.0	0.0	1.142857
1	2007	1	1	1	1918.0	8.0	13.0	0.0	0.0	0.0	0.0	0.0	3.000000
2	2007	1	1	1	2206.0	34.0	36.0	3.0	0.0	0.0	0.0	31.0	14.857143
3	2007	1	1	1	1230.0	26.0	30.0	23.0	0.0	0.0	0.0	3.0	11.714286
4	2007	1	1	1	831.0	-3.0	1.0	0.0	0.0	0.0	0.0	0.0	-0.285714

After that, we calculated the mean by year and found out that year 2008 and average delay of 3.59 mins which is less than overall average delay for 2007. Hence, best year with minimum delay is 2008.

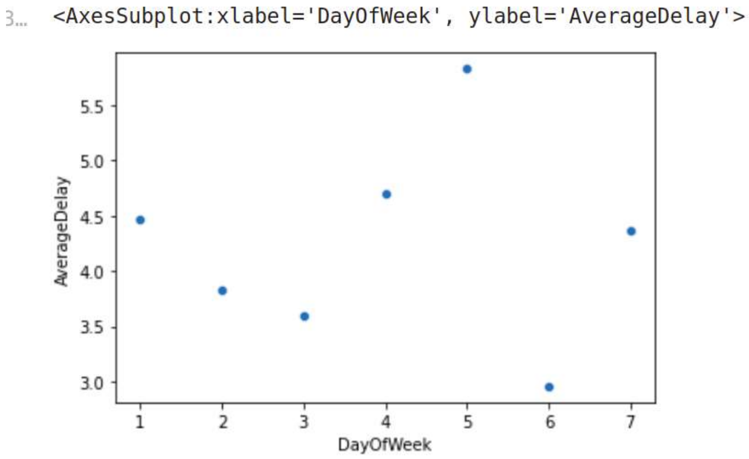
Year	Average Delay
2007	4.950638
2008	3.597905

Next, I will be finding best day of the week with minimum delay. For that, I calculated the average delay separately for each day of the year as given below:



	AverageDelay
DayOfWeek	
1	4.467400
2	3.831051
3	3.590942
4	4.701932
5	5.829038
6	2.956449
7	4.357946

Let’s plot the above data and we get the following scatter plot.

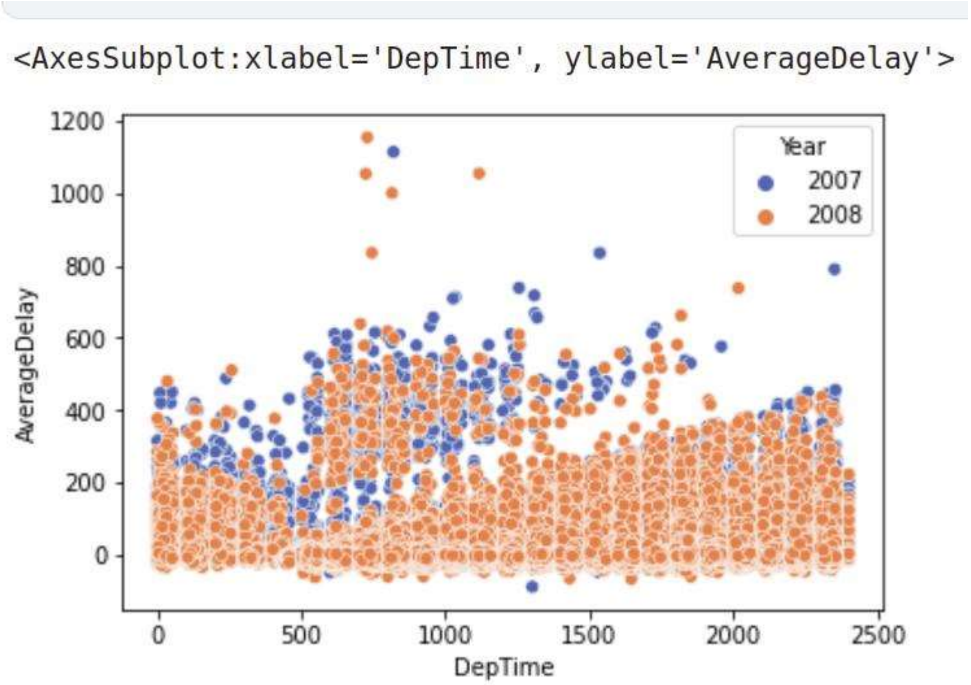


```
3...  <AxesSubplot:xlabel='DayOfWeek', ylabel='AverageDelay'>
```

We can clearly see that 6th day of the week has the minimum average delay and day 5 has the maximum or highest average delay. Hence, we can conclude that day 6 is the best day of the week for flights.

As, we have analyzed the dataset and found out the best year and best day of week. Let’s move on and understand the best time of the in year 2007 and 2008 with minimum delay.

Let’s plot a scatter plot to understand how delay behaves at different times through the year.



```
<AxesSubplot:xlabel='DepTime', ylabel='AverageDelay'>
```

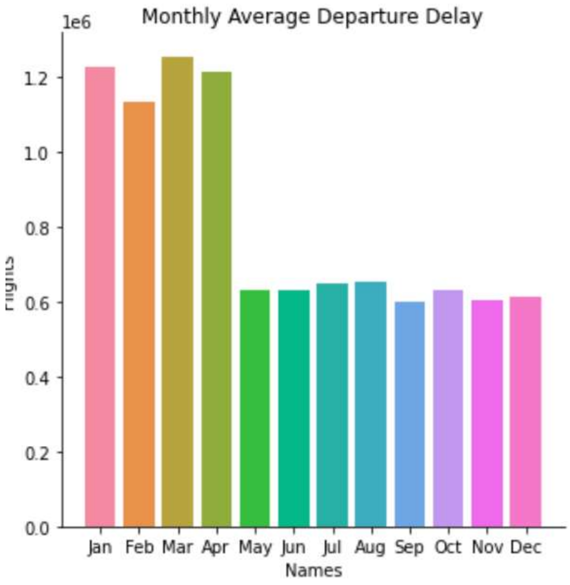
Above scatterplot defines Average Delay by Departure Time for year 2007-08. From 0th hour, the delay decreases gradually till 500 hour and then starts increasing gradually with few spikes to 1200. We can see lower delay around 500 hours as compared to other points in time. Hence, we can conclude that best time of the day for departure is nearly around 500 hours.

2. Do older planes suffer more delays?

Using this limited data, we cannot comment on this, because there is no specific column which speak about aircraft is OLDER or NEWER. But if we can get more data in this direction, we can definitely explore the same aspect.

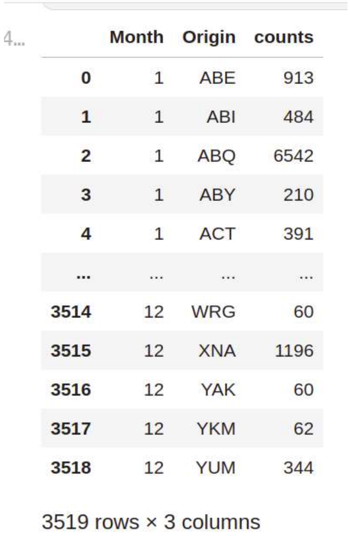
3. How does the number of people flying between different locations change over time?

In this section, I have manipulated the dataset to find the flights usage across different months and also to find which origin has more outgoing flights to estimate number of people.



Take a look at the above graph. This depicts total number of flights for each month for the year of 2007-2008. We see that more flights are book in the month of Jan, Feb, Mar and Apr in year of 2007-08. But we see almost equal number of flights in the rest of months for year 2007-08. Hence, we can say that the beginner 4 fours moths where more people travel, hence total number of people in these months are far greater than the rest.

But the above graph depicts for only months. Let take a look at another graph that might tell us which origin has the highest outgoing traffic.



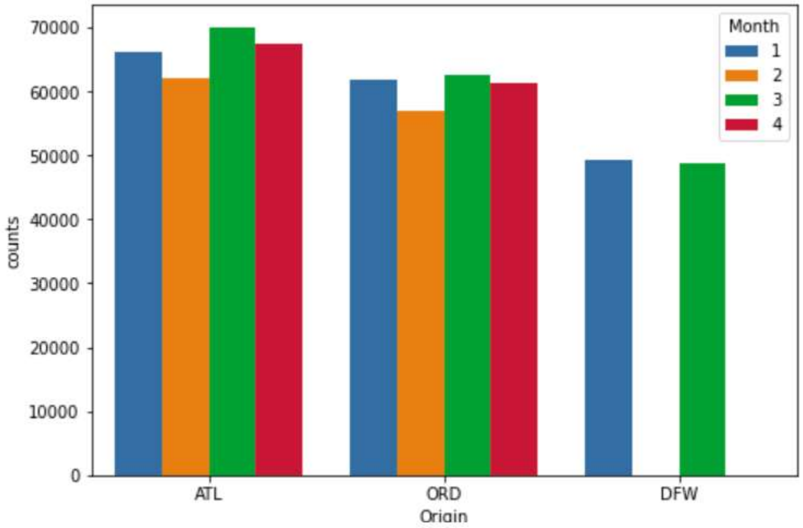
4...

	Month	Origin	counts
0	1	ABE	913
1	1	ABI	484
2	1	ABQ	6542
3	1	ABY	210
4	1	ACT	391
...	...	...	...
3514	12	WRG	60
3515	12	XNA	1196
3516	12	YAK	60
3517	12	YKM	62
3518	12	YUM	344

3519 rows × 3 columns

The above tables show total flights for every origin for all the months. Now, let plot but before that, I would sort this in non-increasing order to find the origins with highest traffic with high traffic months.

```
!1...  <Axes:xlabel='Origin', ylabel='counts'>
```

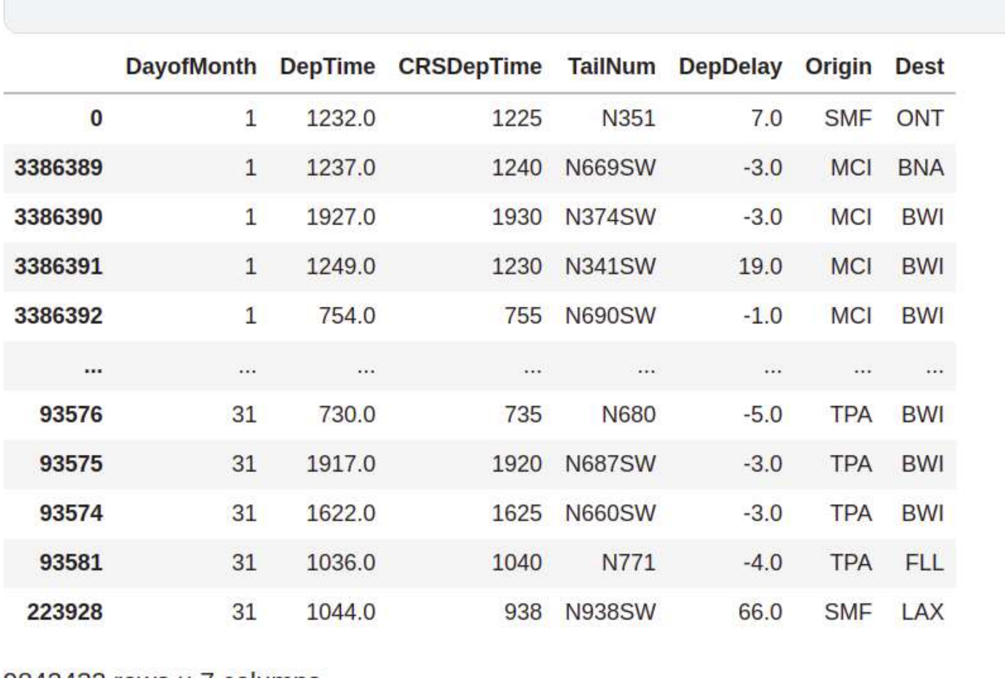


The above frequency plot shows top 10 origins that has the highest traffic and we see that ATL has the highest traffic, followed by ORD, then DFW and so on and so forth. Hence, from this analysis, we can conclude that more people fry from ATL and ORD followed by other origins.

4. Can you detect cascading failures as delays in one airport create delays in others?

In this section, I will be analyzing the dataset to find the cascading failures as delays in one airport affects other or not.

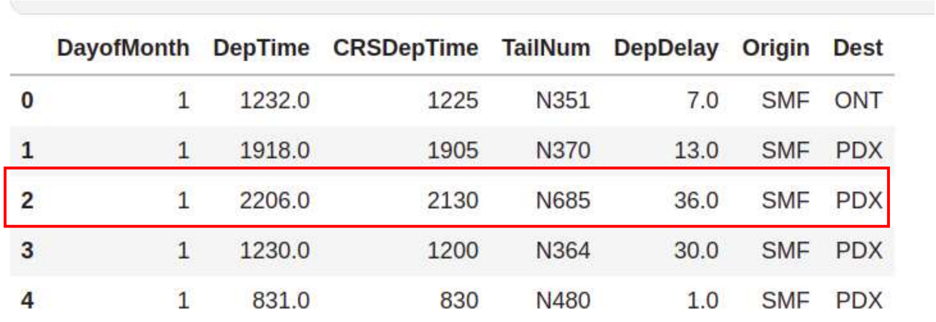
Let's fetch only important variable that will help us analyze the cascading failures.



	DayofMonth	DepTime	CRSDepTime	TailNum	DepDelay	Origin	Dest
0	1	1232.0	1225	N351	7.0	SMF	ONT
3386389	1	1237.0	1240	N669SW	-3.0	MCI	BNA
3386390	1	1927.0	1930	N374SW	-3.0	MCI	BWI
3386391	1	1249.0	1230	N341SW	19.0	MCI	BWI
3386392	1	754.0	755	N690SW	-1.0	MCI	BWI
...	...	...	...	...	...	...	...
93576	31	730.0	735	N680	-5.0	TPA	BWI
93575	31	1917.0	1920	N687SW	-3.0	TPA	BWI
93574	31	1622.0	1625	N660SW	-3.0	TPA	BWI
93581	31	1036.0	1040	N771	-4.0	TPA	FLL
223928	31	1044.0	938	N938SW	66.0	SMF	LAX

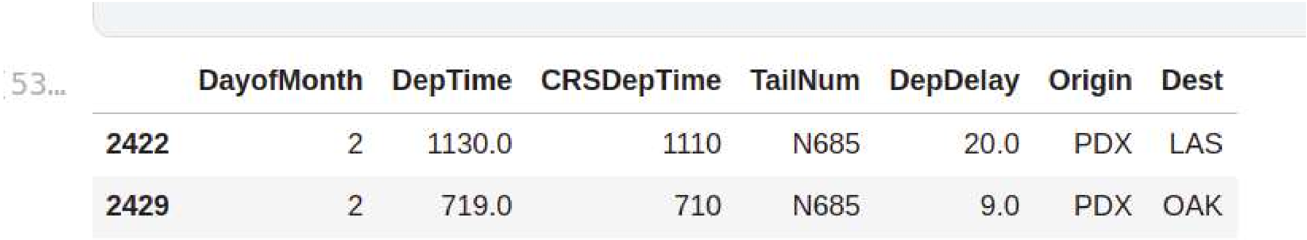
[illegible] rows × 7 columns

We will be using these 6 variables to find out the answer. There could not a lot of examples but let's take a look at the top 5 flights and see if we find something interesting.



	DayofMonth	DepTime	CRSDepTime	TailNum	DepDelay	Origin	Dest
0	1	1232.0	1225	N351	7.0	SMF	ONT
1	1	1918.0	1905	N370	13.0	SMF	PDX
2	1	2206.0	2130	N685	36.0	SMF	PDX
3	1	1230.0	1200	N364	30.0	SMF	PDX
4	1	831.0	830	N480	1.0	SMF	PDX

Let's take a look at Flight 2 with Tail Number as N685. It was supposed to depart at 2130 but departed at 2206 means it got delayed by 36 mins. Now, let's take a look at the PDX on next day flight.



	DayofMonth	DepTime	CRSDepTime	TailNum	DepDelay	Origin	Dest
2422	2	1130.0	1110	N685	20.0	PDX	LAS
2429	2	719.0	710	N685	9.0	PDX	OAK

Here, we can see that N685 reached PDF 36 mins late so it was supposed to depart as 1110 but departed at 1130 to LAS and got delayed by 20 mins. Hence, we can clearly see the cascading failure that is creating delays in other airports. The deeper we go and analyze different flights and airports we will find this pattern. So, delay of flight in one airport does affect other airports as well.

5. Use the available variables to construct a model that predicts delays.

In this section, I trained three different models including Random Forest, Linear Regression and Support Vector Machine to predict the delays. Before stepping into the modeling, I cleaned the dataset, normalized it for better performance.

As we know there are several variables either one can use PCA or best feature selection algorithm to extract important features. I have used Recursive Feature Elimination method to extract best features. Due to high number of observations and limited resources the models have been trained on 50,000 datapoints but it can be changed to whole dataset.

1. Random Forest is giving us MSE as 4.59.
2. Linear Regression is giving us MSE as 33.52
3. Support Vector Machine is giving us MSE as 24.04

Hence, we can say that Random Forest outperforms remaining two models.