ST2195 Programming for Data Science Coursework Project

Gerald Tan Yong Keng
22/03/2022

SIM Global Education University of London
SIM Student ID: 10199353

Table of Contents **Page**

Overview

Throughout the coursework, the dataset used has been stored on DB Browser. An RMarkdown and Jupyter notebook has been created for each question, another set of codes are being used for the initial setup of the database.[1]

The dataset selected is 2005 and 2006, the total observations are 14 282 518. We then note that there are 7 140 596 flights in 2005 and 7 141 922 in 2006. Upon deeper inspection, we confirm that the dataset contains only the flight data of the USA. So far, the data includes all types of flights including cancelled and diverted. After filtering out such observations, the number of flights is reduced to 6 982 428 and 6 997 085 in 2005 and 2006 respectively. Giving us a total of 13 979 513. The average arrival delay, avg_arr_delay, is 7.10 and 8.63 minutes for the respective years.

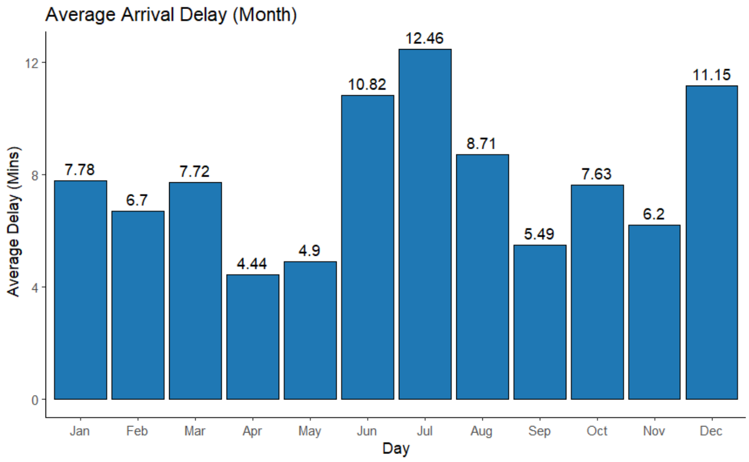
Year	flights	avg_arr_delay	max_arr_delay	min_arr_delay
2005	6,982,428	7.1041	1,925	-939
2006	6,997,085	8.6255	1,779	-592

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

We will work on finding the time of the year, followed by the day of the week and then the time of the day. In the process of deriving the average timings, we excluded unusual observations such as flight cancellations and flights that were diverted. A column showing the number of flights is denoted as No.of Flights.

Time of the Year

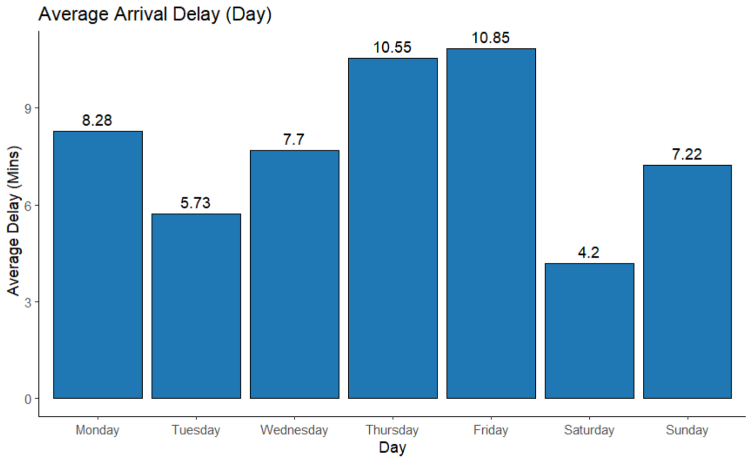
It is observed that the month of April has the lowest average arrival delay with 4.44 minutes, followed by May and September with a timing of 4.9 and 5.49 respectively. Thus, April is the best month to fly.



Month	No. of Flights	Average Arrival Delay (Mins)
Jan	1,136,706	7.78
Feb	1,052,193	6.7
Mar	1,201,414	7.72
Apr	1,162,500	4.44
May	1,201,410	4.9
Jun	1,182,382	10.82
Jul	1,218,516	12.46
Aug	1,230,862	8.71
Sep	1,134,742	5.49
Oct	1,179,656	7.63
Nov	1,134,330	6.2
Dec	1,144,802	11.15

Day of the Week

Based on our observations, Saturday is the best day to travel. With an average arrival delay of 4.2 minutes, it is the lowest compared to the other days. Followed by Tuesday, 5.73 minutes, and Sunday, 7.22 minutes.

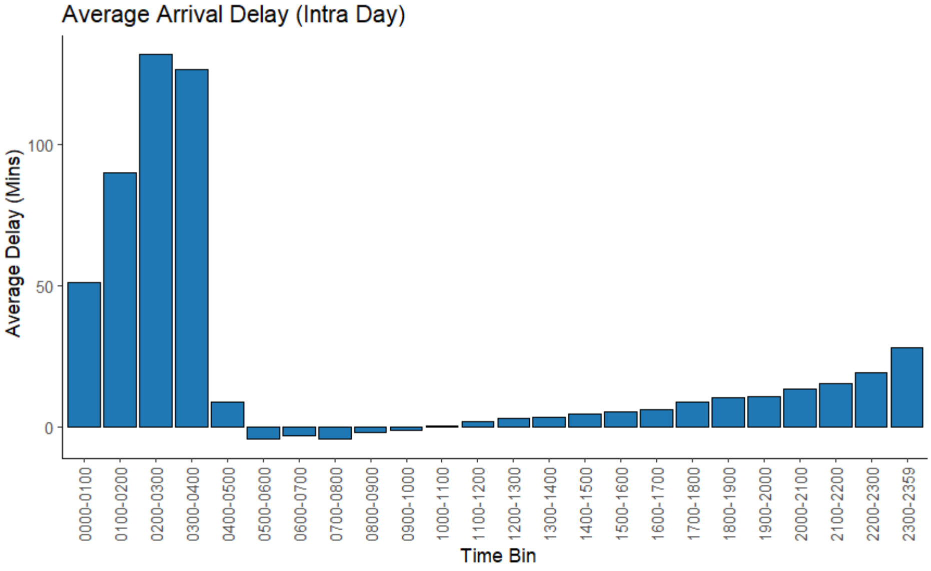


DayOfWeek	No. of Flights	Average Arrival Delay (Mins)
Monday	2,051,847	8.28
Tuesday	2,025,076	5.73
Wednesday	2,039,089	7.7
Thursday	2,051,644	10.55
Friday	2,061,542	10.85
Saturday	1,785,588	4.2
Sunday	1,964,727	7.22

[1] See Reference: (Segments & File names)

Time of the Day

To find the best time of the day, we create a bar plot of the Arrival Delay for each period. In a single day, we separated into 24 time bins (period), each time bin is 1 hour. Based on the diagram it is hard to observe which time bin has the lowest average arrival delay. According to the numerical data obtained, the best time to fly is between 0500-0600, with an estimated mean arrival delay of -4.23 minutes. This implies that flights arrive earlier than scheduled if their scheduled arrival time is between 0500-0600. This is followed by 0700-0800 and 0600-0700 with an average arrival delay of -4.11 and -2.98 minutes respectively.



Time_Bin	No._of_flights	Avg_Delay
0000-0100	163,504	51.28
0100-0200	54,661	90.20
0200-0300	17,301	131.86
0300-0400	6,797	126.62
0400-0500	16,668	8.73
0500-0600	74,371	-4.23
0600-0700	154,188	-2.98
0700-0800	461,290	-4.11
0800-0900	669,782	-2.03
0900-1000	805,849	-1.14
1000-1100	900,897	0.43
1100-1200	851,274	2.07
1200-1300	871,414	3.20
1300-1400	840,575	3.53
1400-1500	865,611	4.44
1500-1600	829,053	5.37
1600-1700	940,982	6.23
1700-1800	889,739	8.71
1800-1900	885,950	10.44
1900-2000	894,075	10.83
2000-2100	862,162	13.52
2100-2200	798,330	15.56
2200-2300	686,235	19.20
2300-2359	438,805	28.01

2. Do older planes suffer more delays?

Based on the data, the departure time is the time when the airplane leaves the city of origin, and the arrival time is the time when the plane arrives at its destination. We create a regression model to find out what factors affect the arrival delays of planes.

Consider the following regression model:

$$ArrDelay = \beta_0 + \beta_1 DepDelay + \beta_2 CRSArrTime + \beta_3 CRSDepTime + \beta_4 old$$

Where,

- ***ArrDelay*** is the arrival delay in minutes
- ***DepDelay*** is the departure delay in minutes
- ***CRSArrTime*** is the Scheduled Arrival Time
- ***CRSDepTime*** is the Scheduled Departure Time
- ***Old*** is a dummy variable to determine whether a plane is old or not. For Old = 1, the plane is considered old, and 0 otherwise.

We will be running a regression and testing the coefficient β_3 to see if it is significant or not. A significant result would imply that “Old” planes are statistically significant in causing arrival delays.

Defining Old

We first need to determine what is “Old”. After tidying the dataset, the years in which planes were manufactured have a range of 1956 to 2007. With a mean and median of 1996 and 1999 respectively.

Thus, we set the year as 1996 (include) for planes to be defined as old. Hence, by our definition, planes manufactured before 1996 are considered old. In terms of age, planes that are older than 10 years (include) are defined as old. The age can be obtained by taking:

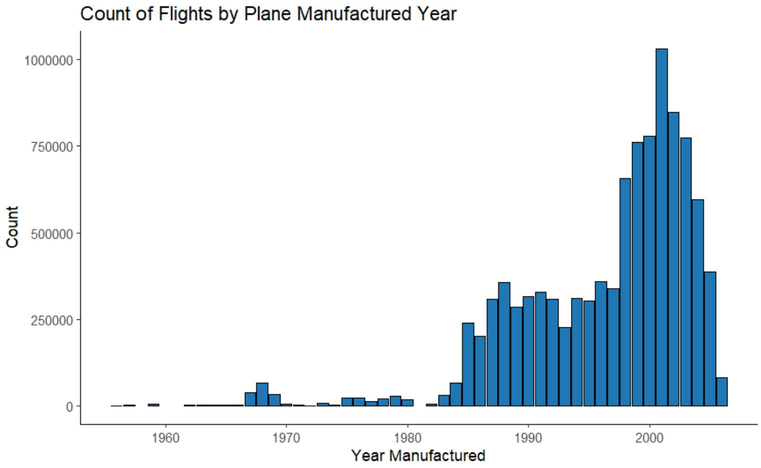
$$Age = Year - Year\_Manufactured$$

- ***Year*** is the Year of the observation. i.e either 2005 or 2006
- ***Year_Manufactured*** is the year which the plane is serviced

Based on our definition, of the 10 205 851 flights, a total of 3 792 824 flights are by old planes and 6 413 027 are by new planes.

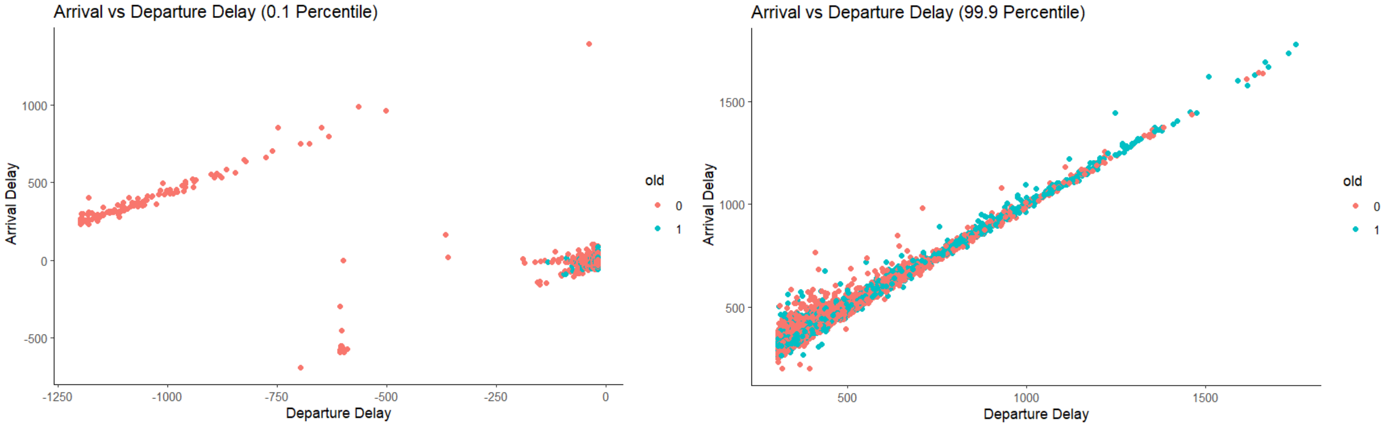
Count of Flights by Plane Manufactured Year

The diagram below shows the number of flights handled by planes according to their manufactured year. Overall, we observed that the number of flights for planes manufactured before the 1980s is much lower. By our definition, 1996 is considered Old, and graphically, we see a sharp increase around the mid-1980s, and again in the late 1990s followed by a significant drop in 2006.



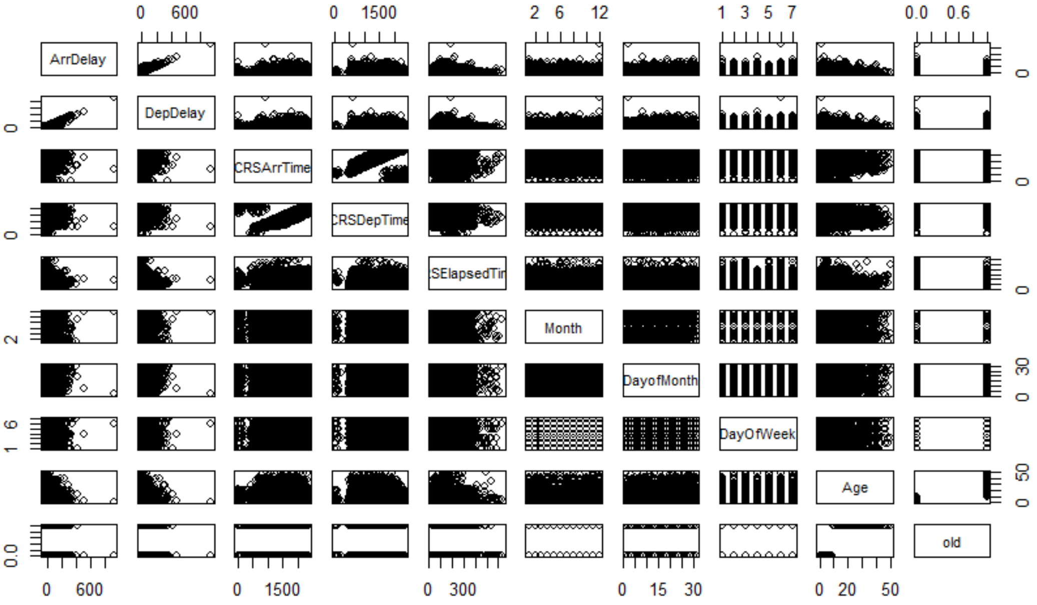
Graphical Observations

We plot Arrival Delay against Departure Delay for the 0.1 and 99.9 percentile to check for extreme values. We observe that in the 0.1 percentile chart, there are extreme values, consider the following situation where departure delay is less than 500 minutes while arrival delay is around 250 minutes and above. This implies that a plane that departs early by more than 8 hours (~500 mins), can still arrive late by about 4 hours (~250mins) or more. This does not make sense. Hence, we will remove data of planes with departure delays below -250. One thing we would like to point out is the apparent linear relationship between arrival delays and departure delays.



Correlation

The bottom plot shows the correlation of the variables, while some variables exhibit a linear correlation, others are harder to tell. We derive a correlation matrix to see the correlation between the variables, correlation falls within a range of -1 to 1. Numbers close to 1 indicate a strong positive correlation while numbers close to -1 indicate a strong negative correlation. Numbers close to zero indicate week correlation. Upon further inspection, we used variables that are correlated with arrival delay. While 'old' has no correlation, we are testing to see if it is significant to our model.



Correlation Matrix	ArrDelay	DepDelay	CRSArrTime	CRSDepTime	CRSElapsedTime	Age	Year_Manufactured	old
ArrDelay	1	0.92	0.13	0.14	0	0	0	0
DepDelay	0.92	1	0.14	0.15	0.01	0	0	0
CRSArrTime	0.13	0.14	1	0.77	0.05	0.02	-0.02	0.02
CRSDepTime	0.14	0.15	0.77	1	-0.01	0	0	0
CRSElapsedTime	0	0.01	0.05	-0.01	1	0.02	-0.02	0.04
Age	0	0	0.02	0	0.02	1	-1	0.81
Year_Manufactured	0	0	-0.02	0	-0.02	-1	1	-0.81
old	0	0	0.02	0	0.04	0.81	-0.81	1

Regression Results

With that, we can run a regression to obtain the following results:

term	estimate	std.error	statistic	p.value
(Intercept)	-1.1304	0.0148	-76.5311	0.0000
DepDelay	1.0196	0.0001	7596.1290	0.0000
CRSArrTime	0.0006	0.0000	41.6193	0.0000
CRSDepTime	-0.0007	0.0000	-48.5916	0.0000
old	-0.0040	0.0089	-0.4524	0.6510

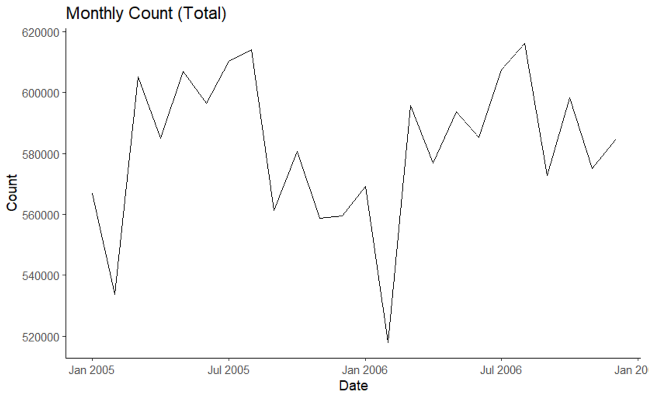
Residual standard error: 13.76 on 10205703 degrees of freedom
Multiple R-squared: 0.8527
Adjusted R-squared: 0.8527
F-statistic: 1.477e+07 on 4 and 10205703 DF
p-value: < 2.2e-16

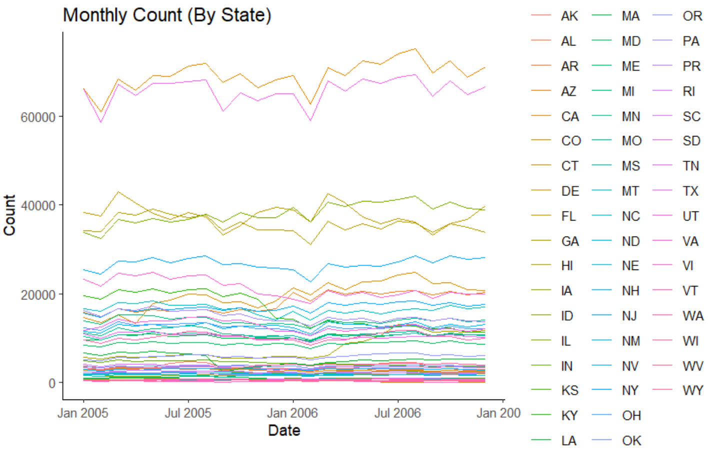
Based on our data set of 10 205 851 observations, our model has an R-squared of 0.853, which implies that the model has high explanatory power. The t-value of 'Old' is not significant. We determine that 'old' planes are not statistically significant to the model. This implies that there is no evidence of older planes causing more delays.

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

Currently, with 52 states[2], it gives us a total permutation of 2652 (52 * 51). Thus, we attempt to address the issue by monitoring the number of flights in each state over the period of 2005 – 2006. First, we look at the overall monthly count of flights, as seen below. While it seems that there is no apparent difference in the total flight count over the 2 years, we cannot conclude that there is no difference in the number of people flying between different locations. Likewise, when we look at the Monthly Count of each state, graphically it is hard to tell.

[2] (U.S. BUREAU OF LABOR STATISTICS, n.d.)



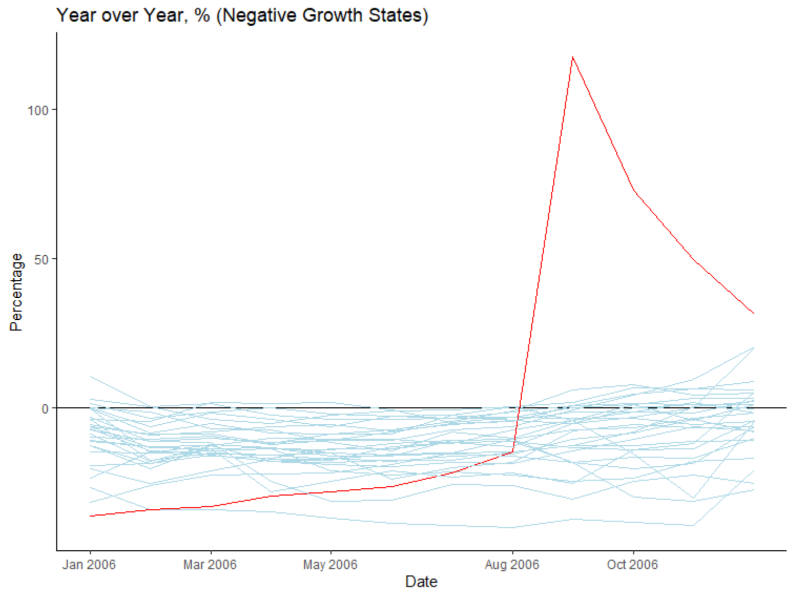


Based on our data set, we have a total of 13 970 740 observations after filtering out the missing values. The top three states with the highest number of flights over the two years are CA, California, TX, Texas, and IL Illinois. With California at the top. The states with the lowest flight counts are WV, West Virginia, VI, Virgin Islands, and DE, Delaware, with the lowest count of 336 flights.

	state	n
CA	California	1,661,513
TX	Texas	1,577,417
IL	Illinois	914,149
WV	West Virginia	11,045
VI	Virgin Islands	7,140
DE	Delaware	336

Negative Growth States [3]

After manipulating our dataset, we make the following observations. Of the 52 states in the dataset, 28 states reflect a decrease in flights. Of the 28 states, LA, Louisiana stands out from the rest, as indicated by the red line. Following the month of August 2006, Louisiana saw an increase in flights.



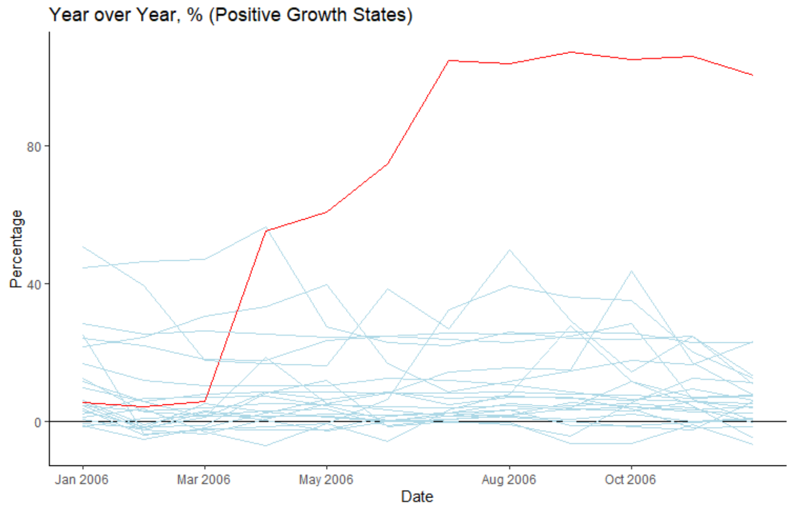
OH	Ohio	AK	Alaska
MN	Minnesota	FL	Florida
UT	Utah	CT	Connecticut
PA	Pennsylvania	VI	Virgin Islands
ID	Idaho	LA	Louisiana
NY	New York	RI	Rhode Island
GA	Georgia	MT	Montana
MO	Missouri	SC	South Carolina
VA	Virginia	TN	Tennessee
IN	Indiana	AL	Alabama
NH	New Hampshire	VT	Vermont
NC	North Carolina	ME	Maine
MI	Michigan	AR	Arkansas
KY	Kentucky	WV	West Virginia

Positive Growth States [4]

For the remaining 24 states, we see positive growth. One state stands out from the rest, which is HI, Hawaii, as indicated by the red line. For Hawaii, following March 2006, there is an increase in the year-over-year percentage of flights.

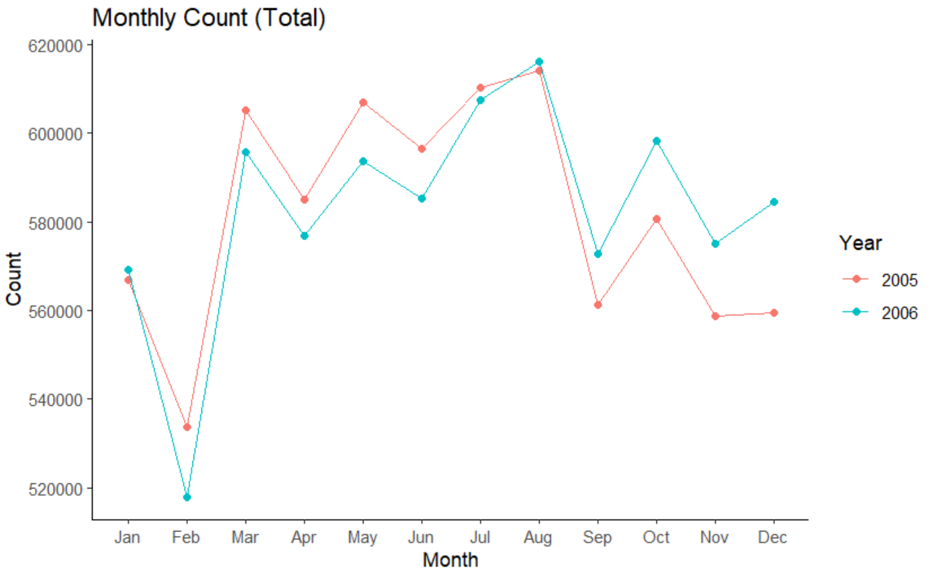
[3] See Table: Negative Growth States

[4] See Table: Positive Growth States



IL	Illinois	NV	Nevada
MA	Massachusetts	WY	Wyoming
TX	Texas	NM	New Mexico
CA	California	KS	Kansas
CO	Colorado	PR	Puerto Rico
OR	Oregon	HI	Hawaii
NE	Nebraska	AZ	Arizona
IA	Iowa	WI	Wisconsin
WA	Washington	MS	Mississippi
OK	Oklahoma	SD	South Dakota
NJ	New Jersey	ND	North Dakota
MD	Maryland	DE	Delaware

One exception is DE, Delaware, where flights were not recorded in 2005. Thus making it impossible to gauge the year-over-year growth for Delaware.



To conclude, we can say that over the period of 2005 and 2006. 23 states saw an increase in the number of flights while 28 states saw a decrease. With one exception, Delaware. While it has to tell if there is an increase graphically, based on the above figure, we see that there is some form of trend between the number of flights and the month.

The subsequent two tables show the year-over-year percent changes for negative growth states and positive growth states.

Negative Growth States, %

Date	AK	AL	AR	CT	FL	GA	ID	IN	KY	LA	ME	MI	MN	MO
Jan-06	-3.952	-12.781	-6.189	-4.129	1.324	-0.539	10.665	-11.163	-26.957	-36.203	-7.101	-6.289	-11.134	2.768
Feb-06	-4.725	-18.265	-14.192	-17.667	-3.613	-8.518	0.1	-12.029	-34.255	-33.993	-13.153	-8.362	-13.274	0.508
Mar-06	1.697	-13.461	-14.967	-15.343	-1.225	-5.293	-3.919	-13.608	-34.071	-32.803	-14.198	-8.566	-12.807	1.208
Apr-06	1.168	-13.068	-15.6	-15.931	-0.052	-8.626	-5.508	-13.561	-35.103	-29.734	-28.227	-11.702	-10.338	-2.357
May-06	1.743	-15.247	-14.907	-17.386	-2.262	-8.795	-2.606	-21.255	-36.93	-28.08	-25.881	-8.583	-10.757	-3.937
Jun-06	-0.624	-23.638	-15.382	-15.745	-2.711	-8.764	-1.149	-22.396	-38.624	-26.621	-21.196	-7.505	-10.715	-3.894
Jul-06	-5.177	-19.57	-14.839	-15.197	-3.501	-2.053	-0.976	-21.273	-39.27	-21.625	-23.463	-4.593	-11.924	-3.425
Aug-06	-3.111	-18.253	-15.679	-14.94	-3.148	-4.264	0.155	-22.316	-40.002	-14.566	-20.952	-1.073	-7.364	-1.611
Sep-06	-3.671	-7.944	-18.416	-12.928	-0.12	-1.537	-4.163	-24.754	-37.258	117.462	-25.802	5.727	-4.443	0.708
Oct-06	-3.55	-6.084	-16.594	-12.743	1.125	-1.545	-1.549	-23.62	-38.38	72.814	-14.55	7.664	1.023	1.24
Nov-06	-5.571	-6.442	-16.743	-11.434	-4.269	1.871	-1.787	-17.692	-39.51	50.209	-13.488	4.303	3.063	0.879
Dec-06	-4.723	-6.455	-10.164	-10.913	0.827	-1.835	2.368	-16.468	-20.917	31.303	-3.348	5.038	3.077	1.018

Date	MT	NC	NH	NY	OH	PA	RI	SC	TN	UT	VA	VI	VT	WV
Jan-06	-20.642	-3.329	-12.884	-0.236	-5.686	-14.718	-8.738	-21.895	-9.867	-0.486	-19.551	0	-7.389	-31.842
Feb-06	-25.333	-11.369	-20.545	-6.691	-9.314	-14.933	-18.004	-13.906	-10.67	-15.249	-18.713	-1.445	-11.481	-24.094
Mar-06	-21.197	-10.214	-14.023	-2.11	-8.042	-12.286	-12.644	-15.527	-9.566	-14.574	-16.141	-7.317	-11.429	-20.69
Apr-06	-17.408	-11.736	-18.982	-4.218	-6.299	-13.218	-18.009	-16.155	-12.465	-13.813	-18.346	-7.576	-24.499	-20.729
May-06	-17.663	-10.989	-17.574	-6.443	-5.393	-16.219	-19.065	-12.977	-11.228	-14.292	-18.375	-11.875	-31.534	-19.134
Jun-06	-13.078	-10.932	-19.039	-2.971	-7.617	-16.12	-19.787	-11.814	-8.236	-11.922	-17.493	-14.465	-30.943	-15.81
Jul-06	-11.545	-7.111	-17.74	-2.537	-5.736	-10.711	-18.627	-9.884	-5.15	-11.135	-17.643	-8.036	-26.066	-11.043
Aug-06	-11.567	-6.29	-16.536	-0.073	-4.354	-10.52	-18.587	-11.052	-4.485	-11.209	-14.972	-10.135	-26.397	-5.87
Sep-06	-18.536	-0.99	-10.631	1.379	-5.462	-7.579	-15.33	-5.042	0.359	-13.403	-12.873	-4.487	-30.818	-18.658
Oct-06	-20.493	4.014	-8.488	6.393	-0.125	-6.895	-11.728	-2.605	4.472	-14.035	-8.076	-15.429	-24.737	-30.172
Nov-06	-18.628	9.649	-3.735	6.307	0.993	-3.487	-6.502	1.785	6.516	-12.235	0.351	-30.42	-22.422	-30.755
Dec-06	-6.331	20.679	-7.308	8.869	-0.23	-1.694	-8.067	20.211	5.963	5.159	2.288	-4.601	-25.378	-26.978

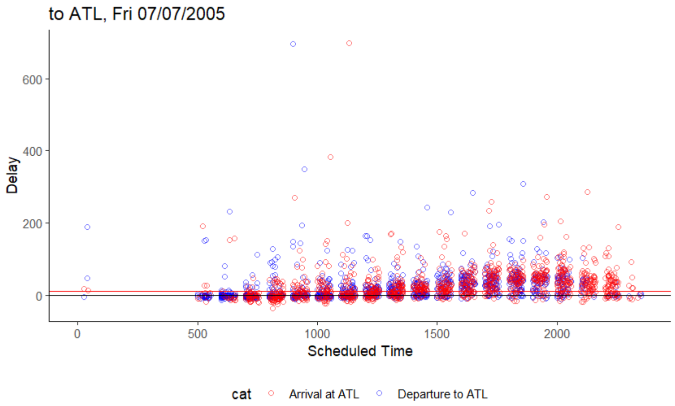
Positive Growth States, %

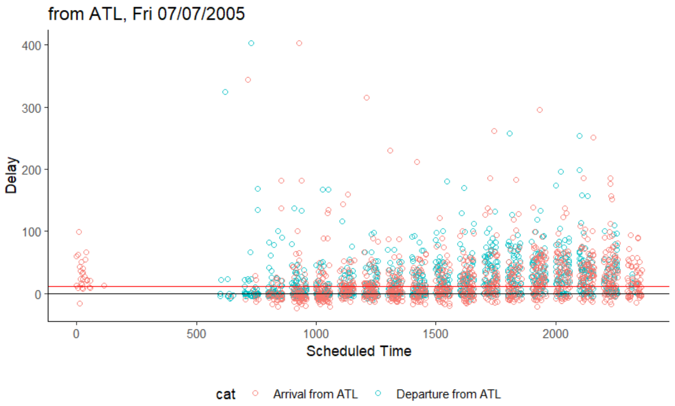
Date	AZ	CA	CO	HI	IA	IL	KS	MA	MD	MS	ND	NE
Jan-06	28.369	4.385	44.405	5.451	21.702	16.569	25.802	4.369	3.592	2.442	-0.857	11.711
Feb-06	25.306	2.899	46.342	4.197	24.328	11.711	-3.804	-2.669	-4.088	5.519	-2.083	2.02
Mar-06	26.15	3.833	46.979	5.558	30.356	10.325	-2.448	2.685	-2.403	-1.891	1.062	2.836
Apr-06	25.378	4.947	56.29	55.116	33.212	10.089	-1.749	0.176	-2.443	-6.805	18.278	4.108
May-06	24.248	4.706	27.265	60.852	39.515	10.363	-0.76	-2.908	-2.697	0	4.942	4.358
Jun-06	24.555	4.054	22.65	74.926	16.346	12.254	5.804	-0.058	1.325	-4.38	8.123	-0.047
Jul-06	25.542	3.82	21.753	104.84	8.665	11.679	32.034	2.399	2.276	6.055	14.168	1.802
Aug-06	25.229	4.523	25.914	103.76	11.143	10.586	38.261	5.158	3.308	10.208	15.746	2.421
Sep-06	25.821	3.167	23.921	107.266	14.474	8.311	35.286	3.842	4.425	27.129	15.051	3.479
Oct-06	25.497	4.274	23.623	105.238	17.668	6.326	34.831	11.543	5.173	11.269	43.478	3.01
Nov-06	22.966	3.316	24.552	105.921	16.676	5.682	20.07	6.724	3.853	4.551	17.113	6.432
Dec-06	22.832	4.065	13.035	100.355	23.377	5.024	12.258	7.552	-0.33	-4.807	7.528	4.62

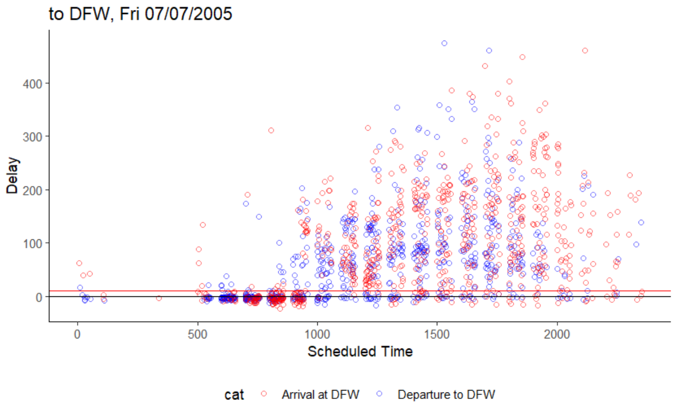
Date	NJ	NM	NV	OK	OR	PR	SD	TX	WA	WI	WY
Jan-06	5.368	1.834	9.786	5.538	11.506	-1.541	50.649	-1.662	0.534	24.194	2.914
Feb-06	-1.979	6.74	5.787	-1.027	5.496	-5.42	39.347	0.633	-2.626	21.65	-1.743
Mar-06	4.924	6.925	7.902	2.888	0.999	-2.067	17.536	1.311	-3.742	17.797	-1.264
Apr-06	2.422	7.309	8.55	1.254	8.289	2.408	16.519	1.438	0.352	17.851	8.031
May-06	0.956	4.412	8.435	2.331	6.219	6.377	15.865	1.443	5.096	23.445	11.645
Jun-06	-0.127	0	7.964	1.987	8.514	8.522	38.289	0.052	3.312	24.729	-1.525
Jul-06	0.498	-0.461	8.001	3.255	6.545	4.777	26.623	1.426	1.353	23.496	-0.234
Aug-06	1.352	-0.244	8.511	1.863	7.231	6.872	49.649	1.66	3.351	22.2	-0.936
Sep-06	0.573	-6.549	7.911	0	6.545	6.713	29.134	5.408	-1.428	24.075	-4.371
Oct-06	2.094	-6.453	7.604	-1.421	6.033	3.44	14.198	3.988	-1.294	28.038	5.458
Nov-06	-0.221	-1.097	6.605	-1.45	9.399	-2.022	24.613	2.456	-0.366	6.329	12.305
Dec-06	0.893	-6.968	7.301	6.358	6.082	-1.75	10.386	2.17	0.333	0.051	11.174

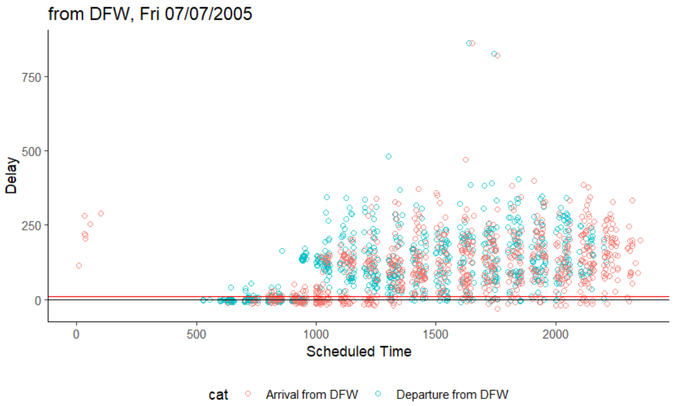
4. Can you detect cascading failures as delays in one airport create delays in others?

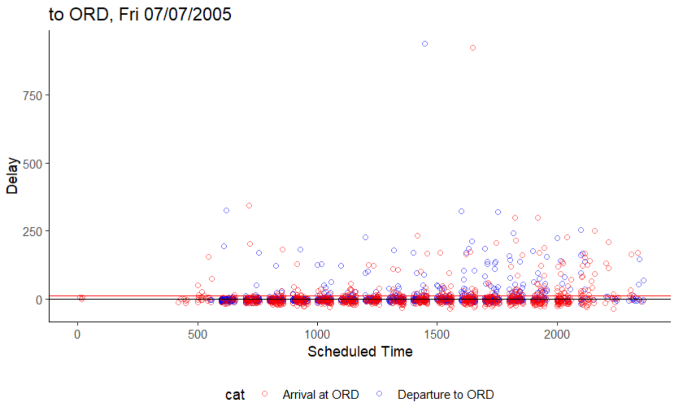
We will compare the arrival delays of flights to a target airport and the arrival delay of flights from the target airport, our goal is to see if arrival delays in one airport will create arrival delays in other airports. We focus on the 2005 data set, the top 3 airports with the highest count of flights are ATL (Atlanta), DFW (Dallas-Fort Worth), and ORD (Chicago). The number of flights is 413 342, 327 943, and 302 266 respectively. The figures on the left show flights to the target airport, while the right shows flights from the target airport. For example, we will be looking at the arrival delays of flights traveling to ATL and the arrival delays of flights arriving from ATL. Our target airports are ATL, DFW, and ORD.

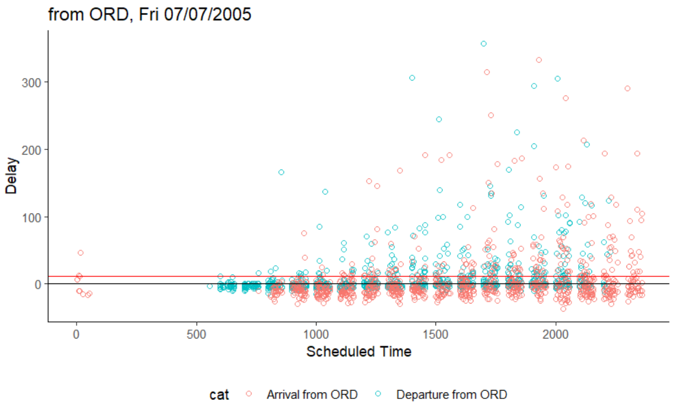












From Q1, we identified that Fridays and July have the highest arrival delay, hence we observe flights on 07/07/2005. The red horizontal line indicates the average arrival delay from Q1, which gives us a value of 10.85 minutes.

ATL, Atlanta

From 0500 to around 1400, we see that arrival delays are concentrated around the mean. As we approach 2000, we see that the arrival delays deviate away from the mean, increasing significantly. While on the right, likewise, we observe that arrival delays from the ATL are around the mean, and beyond 1400, we start to see larger deviations in the plot.

DFW, Dallas-Fort Worth

In DFW, beyond 0900 we start to see a larger deviation of the arrival delays, it is observed that there is a larger variation in delays. Conversely, we see that the arrival of planes from DFW to other airports is delayed. This is reflected in the large variation in the arrival delays.

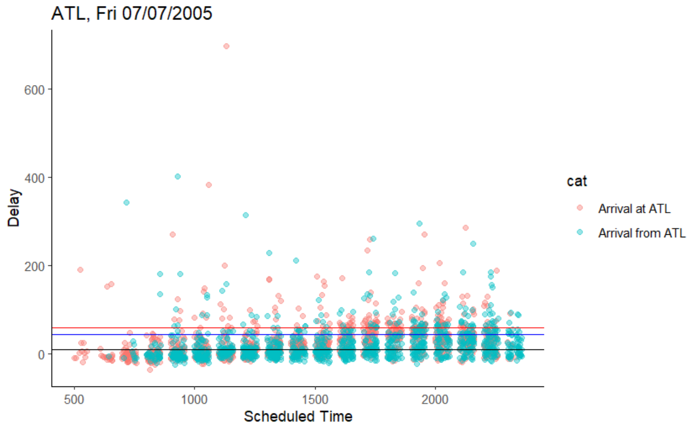
ORD, Chicago

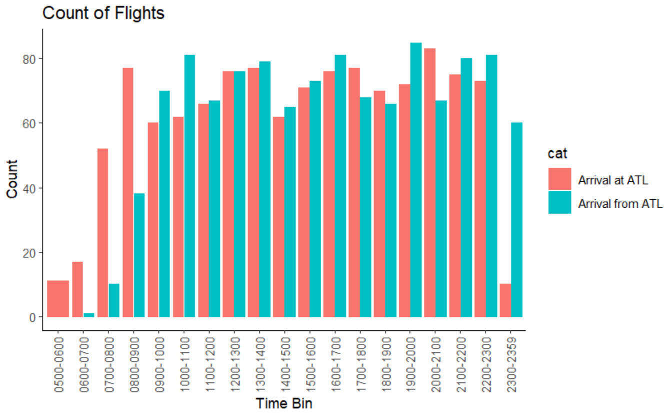
Compared to the previous two, arrival delays in ORD hover around the mean throughout. We start to see a slight increase in arrival delay beyond 1500. Compared to the arrival delays from ORD, we see that there is an increase in delay at 1500, and deviations are larger as well.

Graphically, it appears that as the spread of the arrival delay at a target airport increases, the spread of the arrival delays from the target airport increases as well. Now let us further investigate ATL, we want to see the number of flights and the mean delay in each period. The table below shows the consolidated list of flights to ATL and flights from ATL, with the mean arrival delays.

- ***count_toATL*** number of flights to ATL
- ***avg_toATL*** average arrival delay at ATL
- ***count_fromATL*** number of flights from ATL
- ***avg_fromATL*** average arrival delay from ATL

Timebin	count_toATL	avg_toATL	count_fromATL	avg_fromATL
0500-0600	11	20.18	NA	NA
0600-0700	17	12.71	1	-5
0700-0800	52	-1.71	10	35.7
0800-0900	77	1.31	38	7
0900-1000	60	17.08	70	12.4
1000-1100	62	22.95	81	7.81
1100-1200	66	25.26	67	14.51
1200-1300	76	13.5	76	15.01
1300-1400	77	22.48	79	12.68
1400-1500	62	22.1	65	14.38
1500-1600	71	30.03	73	13.6
1600-1700	76	31.33	81	16.65
1700-1800	77	48.94	68	28.53
1800-1900	70	40.73	66	23.73
1900-2000	72	53.54	85	32.86
2000-2100	83	46.27	67	31.97
2100-2200	75	39.44	80	36.61
2200-2300	73	31.86	81	37.72
2300-2359	10	17.8	60	21.62





The plot on the left shows the delays against the scheduled time, there are 3 horizontal lines. Black represents the average arrival delay on Fridays which is 10.85. The blue line represents arrival delays of 45 minutes, and the red line is the arrival delay at 60 minutes. On the right shows the number of flights in each time bin, we note that from 0900 onwards, the number of flights scheduled in each period is at least 60 and in some periods, flight numbers exceed 80. Based on the chart on the left we take the blue line as a reference point, as the Arrival Delay at ATL increases, the Arrival Delay from ATL increases as well. Between 1000-1500, the cluster of flights is below the blue line, beyond 1500 we see a shift in the cluster to around the blue line. Thus, we can conclude that arrival delays in one airport, create arrival delays in another.

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

In question 2, we utilised a regression model to show the effects of how departure delays, scheduled departure and arrival time, and old planes from one airport lead to arrival delays in another. Building on that model, we will use a regression model to determine if we can predict the delays, more specifically the arrival delay.

Seasons

In question 3, we ended off by assuming that there is some form of trend that creates the demand for flights and that the trends are associated with the Month of the Year. It is possible that in periods of with higher demand for flights, could result in a delay and the converse could be true. We decide to break down the twelve months according to their seasons and introduce new dummy variables according to the seasons. There are four seasons, we use Winter as the reference category, which leaves us with Summer, Fall, and Spring as the 3 dummy variables to include. We define

Winter as 12, 1, 2 or December, January, and February. Summer as 6, 7, 8 or June, July, and August; Fall as 9, 10, 11 or September, October, November, and Spring as 3, 4, 5 or March, April, and May.

Louisiana

In Question 3, we arranged the states in descending order, based on the number of flights. In the middle at the 26th position lies LA, Louisiana. With a total of 94,389 observations, after excluding missing values. We used 3 models for Louisiana, Regression, Ridge regression with autotuning, and random forest. We use the mean square error (MSE) as a measure to see the performance of our models. Ideally, the model with the lowest MSE is the optimal model.

Model

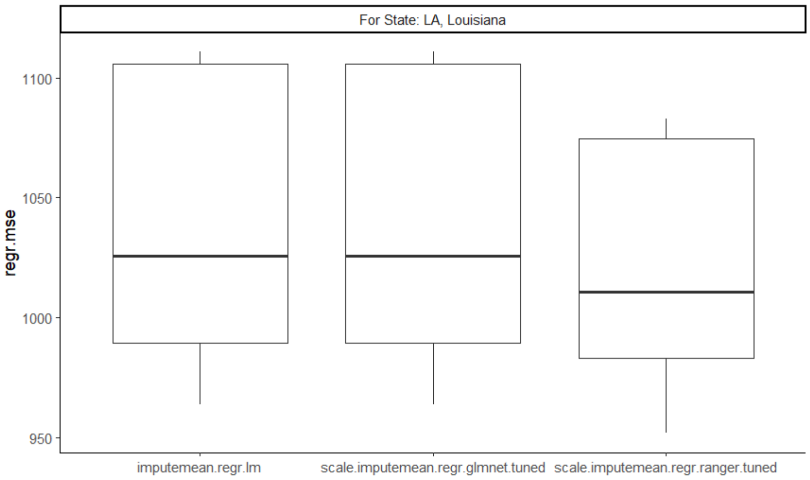
We build a machine learning model that involves predicting the Arrival Delays, based on the following variables. Month, DayofWeek, CRSDepTime, CRSArrTime, CRSElapsedTime, Age, Old, SummerDV[5], SpringDV, and FallDV. We use 70% of the Louisiana data to train our model and the remaining 30% to test.

After training the 3 models, we test our model on the test data set and the results of the top 30 rows are compiled into the table below on the left. **truth** indicates the actual value, **glrn_lm_response** indicates the predicted value by the linear regression model, **at_ridge_response** is the predicted value by the ridge regression model and **at_rf_response** is the predicted value by the random forest. After running our model, we obtain the following boxplots. It is observed that the **at_rf_response**, random forest, gives us the lowest MSE or mean squared error of 1017.031. MSE is the square of the difference between the actual and the predicted value. Ideally, we would prefer a model with the lowest MSE as it implies that the model will produce more accurate results. This suggests that we should use the random forest for forecasting arrival delays.

Conclusion

All three model allows us to predict the arrival delays but the random forest model is the optimal model among the three. In turn, this allows us to forecast the actual arrival time, simply by adding the arrival delays to the scheduled arrival time. We tried including departure delays into the model as well, it significantly reduces the MSE to below 200. However, as the flights have yet to take place, this would mean that we would need a model to estimate the departure delays.

row_ids	truth	glrn_lm_response	at_ridge_response	at_rf_response
2	36	19.50197686	19.53388885	64.89119811
3	-12	18.79706585	18.81602995	15.55767882
4	44	19.18225565	19.21476826	9.948878464
5	28	20.21589055	20.24851187	26.85845838
9	10	10.27539836	10.35959878	1.755428152
22	-16	19.96496914	19.96340707	-4.545783979
23	23	19.6686073	19.66621895	23.09633483
28	217	20.08615915	20.08365386	16.64130902
31	89	19.4379349	19.43538384	26.19090053
32	-3	19.33864137	19.33725471	7.380197079
34	-19	18.92108952	18.9198198	-1.503859914
36	-24	19.94307266	19.9412894	11.10478132
37	148	19.62481435	19.62198362	30.68994501
43	199	19.89927971	19.89705407	9.514330353
44	174	19.75619322	19.75468961	-0.89130547
46	-5	6.61685199	6.63705822	3.37344434
48	10	6.221196618	6.241740978	0.547081507
52	0	6.86942087	6.890011005	24.44362045
54	-7	6.517558454	6.538929092	5.402621932
55	-5	6.221196618	6.241740978	0.547081507
67	-6	6.429972543	6.450458432	-0.103174148
68	115	6.243093096	6.263858643	8.183401607
69	55	19.28317706	19.30206302	7.890367769
71	-5	18.86562521	18.88462811	17.3127393
73	42	19.81352068	19.83440206	20.55276347
75	45	19.4700565	19.48866281	20.67142954
76	5	20.6267279	20.64715421	25.50177677
77	56	19.07440113	19.09334557	14.98231405
78	86	18.84372873	18.86251045	24.98974186
87	181	19.81352068	19.83440206	20.55276347



[5] DV: Dummy Variable

References

(n.d.). Retrieved from U.S. BUREAU OF LABOR STATISTICS: https://www.bls.gov/respondents/mwr/electronic-data-interchange/appendix-d-usps-state-abbreviations-and-fips-codes.htm

Data Expo 2009: Airline on time data. (2008). Retrieved from Havard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

Ming, L. C. (2021). SIM UOL Lecture Notes.

Segments & File names. (n.d.).

Segments & File name	**RMarkdown**	**Jupyter Notebook**
Data Setup	Coursework_setupMD.rmd	pycoursework_setup
Question 1	Coursework_Q1MD.rmd	pycoursework_Q1
Question 2	Coursework_Q2MD.rmd	pycoursework_Q2
Question 3	Coursework_Q3MD.rmd	pycoursework_Q3
Question 4	Coursework_Q4MD.rmd	pycoursework_Q4
Question 5	Coursework_Q5MD.rmd	pycoursework_Q5