Report Analysis of flight details for all commercial flights within the USA from 2007-2008

Course ID: ST2195 Programming for Data Science

Student ID: 190536235

190536235_ST2195

Table of Contents

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

Due to its large file and constraints imposed on a lower-end computer system, the data set from 2007-2008 used for mining and analysis has been randomly subset 30% of the dataset. After a thorough analysis of sampled data, an assumption has been made on delays where delays exceeding 240minutes are deemed as cancelled and considered as outliers to be removed from the data set, as well as including only non-cancelled flights. This would prevent significant changes to any prediction model when outliers are present, affecting the outcome of the interest of prediction.

The data has been grouped and categorized into the following factors:

- Departure time – Morning (0500hrs to 1159hrs), Afternoon (1200hrs to 1659hrs), Evening (1700hrs to 2059hrs) and Night (2100hrs - 0459hrs).
- Month – 1st Quarter (1-3), 2nd Quarter (4-6), 3rd Quarter (7-9) and 4th Quarter (10-12).
- DayofWeek – Monday (1), Tuesday (2), Wednesday (3), Thursday (4), Friday (5), Saturday (6) and Sunday (7).

In figure 1.1. and 1.2. below shows the optimal period of the day, day of the week and quarterly period of the year to minimize flight delays generated from R and Python respectively. It can be observed that in figure 1.1. generated by R, the optimal schedule to travel to minimize departure delays based on average departure delays is on Sunday morning between 0500 to 1159hrs in 2nd quarter of the year between April-June as its average departure delay is 0.9 minutes which is lowest among others, indicating that majority of the flights are on time. Overall, the 2nd quarter of Sunday has the lowest average departure delays across periods compared to other Sundays in other quarters. Similarly in figure 1.2. generated by Python, it can be observed that morning tends to have the lowest average departure delays in Q2 where it has been highlighted as the lightest shades among other quarters, indicating close to 0 average departure delays.

This may be attributed to non-peak flight hours in the morning as generally, the morning tends to have the lowest average departure delays compared to other periods. Furthermore, morning flights are less prone to harsh weather conditions such as thunderstorms, and turbulence which occur frequently at noon as well as less congested air-traffic control (Rizzo, 2022). Thus, it is recommended to consider peak periods and amend the number of flights, departure, or arrival times accordingly to avoid further delays.

Figure 1.1. shows the heatmap of average departure delays (in minutes) of the period each day, day of the week and quarterly period of the year for the 2007-2008 data set generated by R.

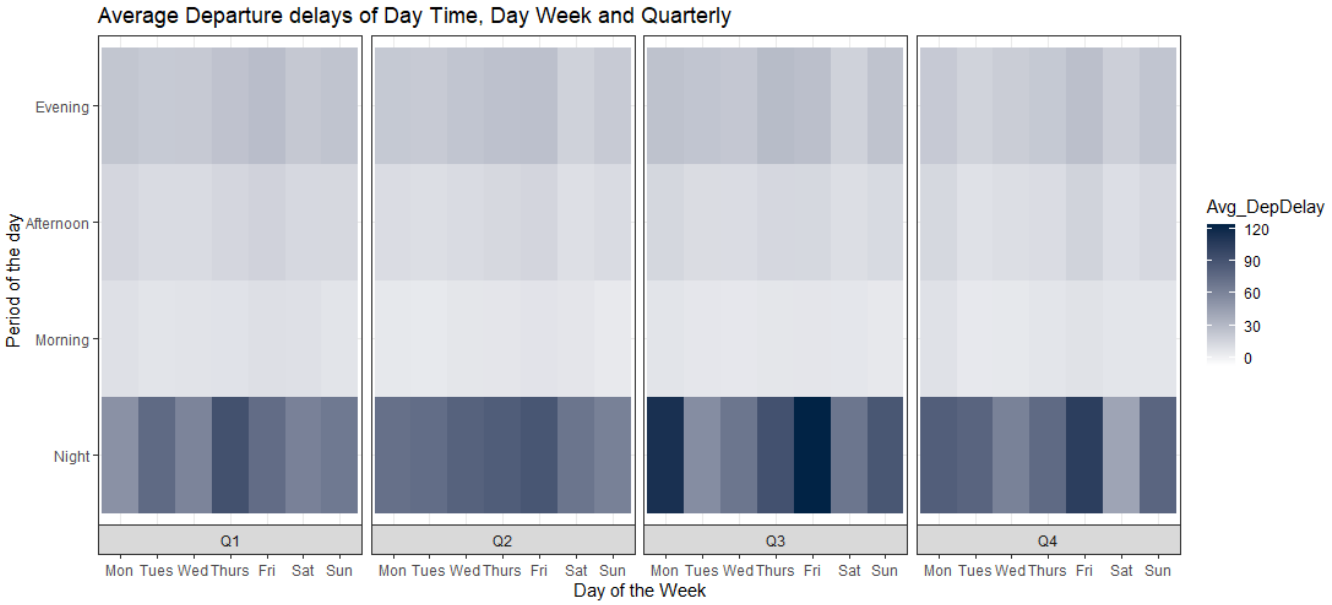


Figure 1.1. shows the heatmap of average departure delays (in minutes) of the period each day, day of the week and quarterly period of the year for the 2007-2008 data set generated by Python.

2. Do older planes suffer more delays?

The issue date has been used to determine the date when the planes are materialized and commercialized for use. To count the total delays, the flights have to take place and thus, only include non-cancelled flights. The data has been categorized into the following: Issue date – Before 2000 (<01-01-2000) and After 2000 (>= 01-01-2000).

In figure 2.1. below shows the arrival and departure delays of older and newer plane models generated by R and Python. It can be observed that from the left side of the plot, older plane models that were issued before 2000 tend to suffer slightly higher arrival and departure delays of 10.8 minutes and 9.8 minutes respectively than newer plane models that were issued after 2000. Furthermore, the fatness of the distribution indicates that there are more observations with lower departure delays for flights commercialized after 2000 than flights before 2000. This evidence is supported by looking at the right side of the plot generated by Python, where its average departure and arrival delays Before 2000 are higher than delays After 2000.

Possibly, the older plane model classified before 2000 does not have the access to cutting-edge technology that enhances its efficiency and minimizes delays as technologies after the 2000s, when the internet and other technological advancement breakthrough has boomed after 2000 due to the dotcom bubble burst (speculation in internet-based businesses) and hence, leading to a rise of internet access around the world. For example, a lack of real-time airport system and integration among air-traffic control, automation or poor infrastructure can contribute to flight delays. Thus, it is advisable for the aircraft industry to constantly upgrade and maintain technology to prevent future flight delays as integration provides maximum efficient and effective flight performance at the expense of investment.

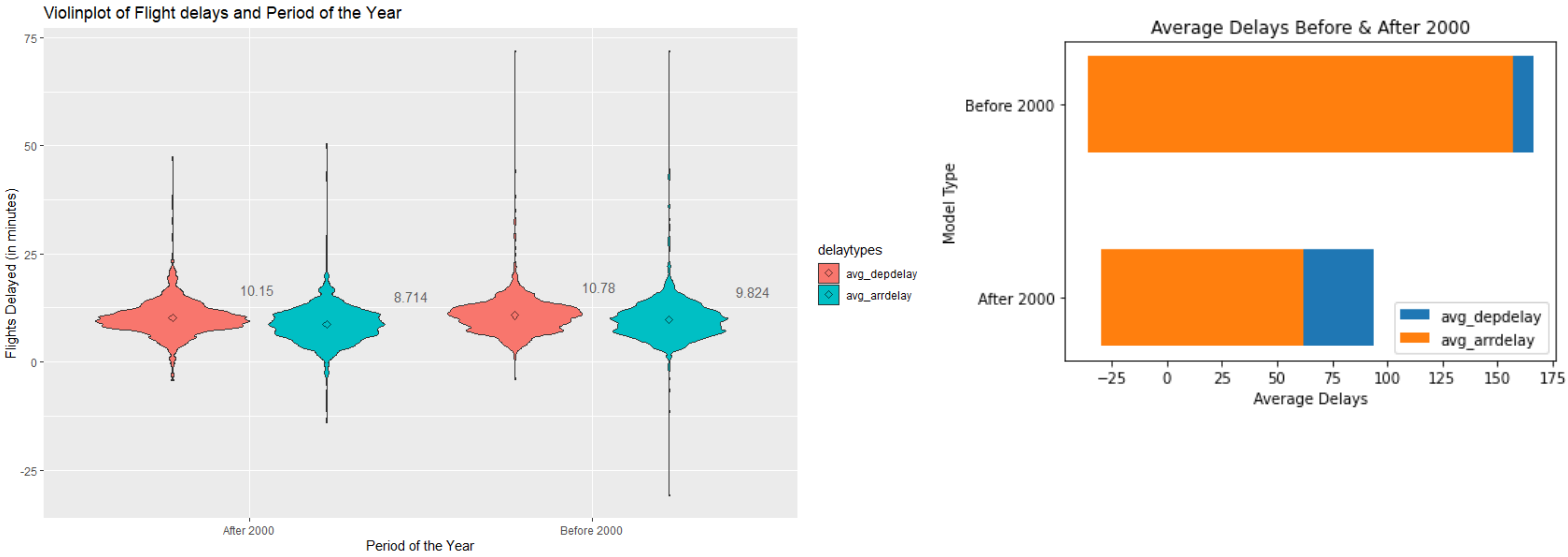


Figure 2.1. shows the comparison of average and departure flight delays (in minutes) among older planes that were issued before 2000 and after 2000. The circles at the centre of each plot represent the average flight delays generated from Left: R. The right side of the plot is generated by Python.

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

To analyze the number of people flying between different locations, an external source of the dataset from 'data.world' has been merged into the current dataset provided by 'Harvard Dataverse'. To count the number of passengers and flights, the flight has to take place and thus, only include non-cancelled flights.

Based on figure 3.1. generated by R below, it can be observed that the most popular travel destination is from Kahului (OGG) airport to Honolulu International airport (HNL), where it has the overall highest number of total passengers of 129,220,705 in Q1, 67,465,299 in Q2, 83,557,811 in Q3 and 41,897,935 in Q4. However, in figure 3.2. generated by Python, the highest overall number of total passengers travelled are from MCO to ATL airport of 727,646 in Q1, 709,027 in Q2, 661,659 in Q3 and 691,066 in Q4.

Both R and Python show similar results in the highest total number of flights recorded from Los Angeles International airport (LAX) to Las Vegas - McCarran International Airport (LAS) airport and vice versa, with total flights of 9,716 in Q1, 6,285 in Q2, 5,306 in Q3 and 7,001 in Q4 from LAX-LAS and 9,106 in Q1, 6,347 in Q2, 5,521 in Q3 and 5,431 in Q4 from LAS-LAX and total flights of generated from R. Total flights of 37,698 in Q1,25,001 in Q2, 21,387 in Q3 and 19,973 in Q4 from LAX-LAS and total flights of 38,060 in Q1, 23,162 in Q2, 18,965 in Q3 and 19,966 in Q4 from LAS-LAX generated from Python.

This indicates that travels between Kahului (OGG) and Honolulu (HNL) are a popular travel destination among travellers as it is situated in Hawaii where it offers diverse tourist attractions and food, as well as accessibility in travel stop-over due to direct connections to other continentals and countries, making one of the busiest airport traffic compared to other states. Similarly, MCO and ATL airport where offers tourist attraction such as the top largest aquarium, home to one of the most popular productions such as Coca Cola as well as the gateway to 80% of the United States area (Sood, 2022). Furthermore, it is observable that in 1st quarter of OGG-HNL, it has the highest amount of passengers and flights compared to 4th quarter and this may be attributed to cheaper air tickets and accommodation due to non-holiday or non-peak period and its weather condition which makes it pleasant to visit.

Similarly, flights between Los Angeles International airport (LAX) and McCarran International Airport (LAS) and vice-versa, are popular due to its major international gateway access to United States continent from other countries and provides a full range of services for private aircraft as well, making one of the top stop-over destinations. Furthermore, flights from LAX to LAS are highest in the 1st quarter has the compared to 4th quarter and vice versa, indicating peak flight hours. This allows aircraft carriers to estimate any congestion that exists among these airports to prevent flight delays, as well as targeting and promoting their plane tickets to attract more customers to travel to and fro from popular destinations.

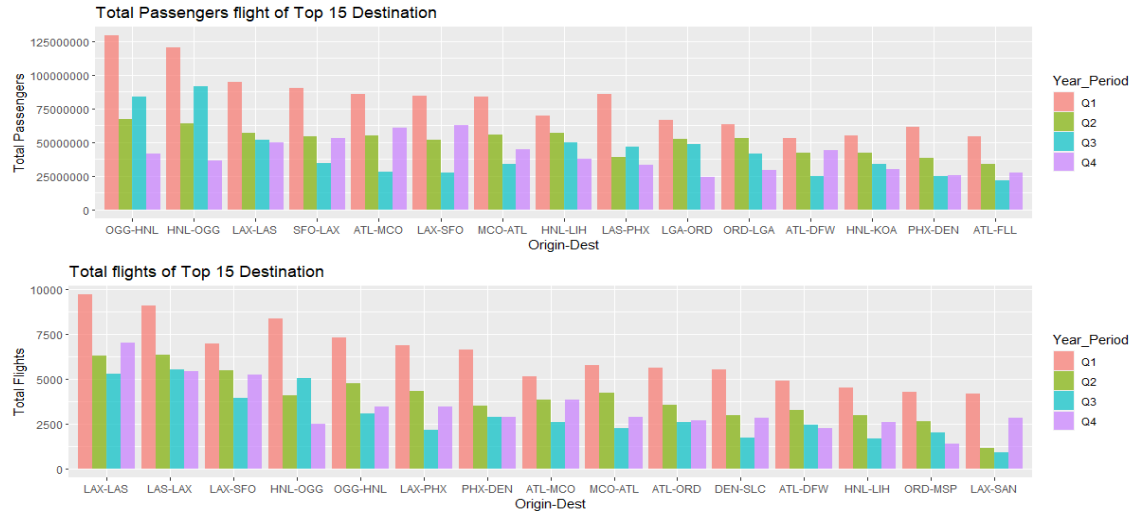
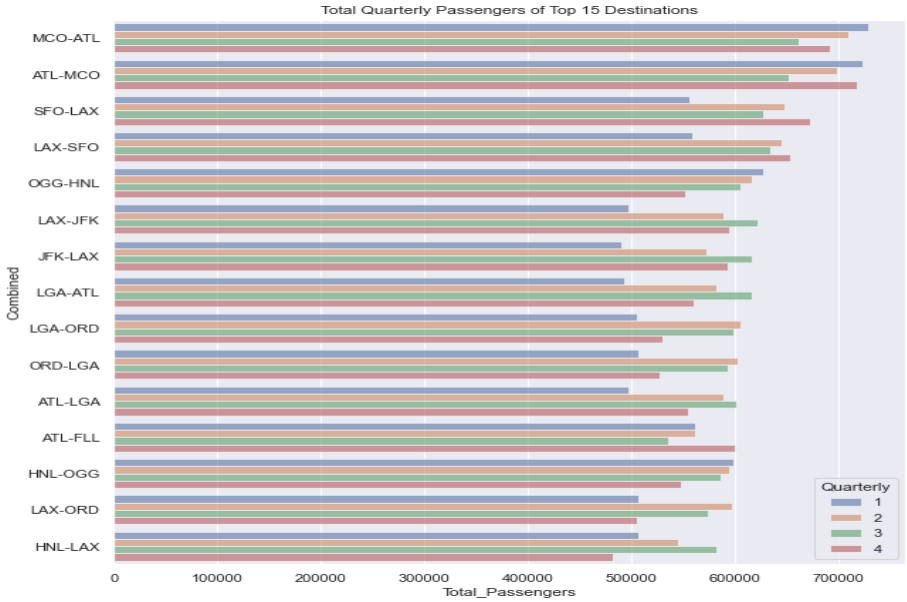


Figure 3.1. shows the quarterly of Total Passengers and Total Flights of top 15 Destinations over 2007-2008 from Origin to Destination airport generated by R.



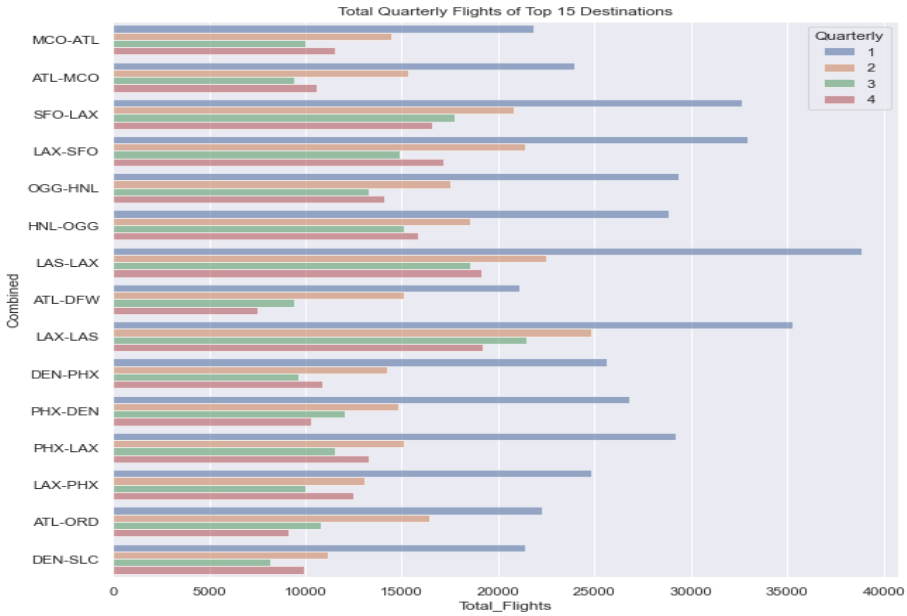


Figure 3.2. shows the quarterly Total Passengers and Total Flights of top 15 Destinations over 2007-2008 from Origin to Destination airport generated by Python.

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

To detect and visualize any cascading failures of departure delays in one airport causing delays in other airports, 2007 and 2008 data has been merged and clustered based on the days of each month from January to December and departure delays have been filtered to non-cancelled flights with more than 0 minutes, to show the cause-and-effect of cascading failures occurred on individual days of each month. According to figure 4.1. and 4.3 generated by Python, it can be seen that average flight departure delays of 147minutes on 1st of January 2007 where the flight departed at 0937hrs instead of scheduled time 0710hrs caused by ABQ airport that affected flights to DFW airport, which simultaneously, affects further departure flight delays of 164minutes and 160minutes to JAX and SFO airport where it departed on 2254hrs and 2305hrs instead of the scheduled time of 2010hrs and 2025hrs respectively, from DWF airport. Similarly, on 1st January 2008 where average flight departure delays of 149minutes where flight depart at 0829hrs instead of the scheduled time of 0600hrs caused by ROC airport affected flights to ORD airport which simultaneously, affects further average flight departure delays of 160 minutes where the flight departed at 1140hrs instead of 0900hrs to ROA airport.

In figure 4.2. and 4.4. generated by R, it can be observed that the average flight departure delays of 180minutes on 1st of January 2007 shown at the top, where the flight departed at 1037hrs instead of scheduled time 0737hrs caused by FAT airport that affected flights to DEN airport, which simultaneously, affects further departure flight delays of 115minutes and 31minutes to CPR and FSD airport where it departed on 1210hrs and 1559hrs instead of the scheduled time of 1528hrs and 1745hrs respectively, from DEN airport. Similarly, on 1st January 2008 shown at the bottom, where average flight departure delays of 186minutes where the flight departed at 1111hrs instead of the scheduled time of 0805hrs caused by BNA airport affected flights to DFW airport which simultaneously, affects further average flight departure delays of 198minutes where the flight departed at 2053hrs instead of 1735hrs to DTW airport.

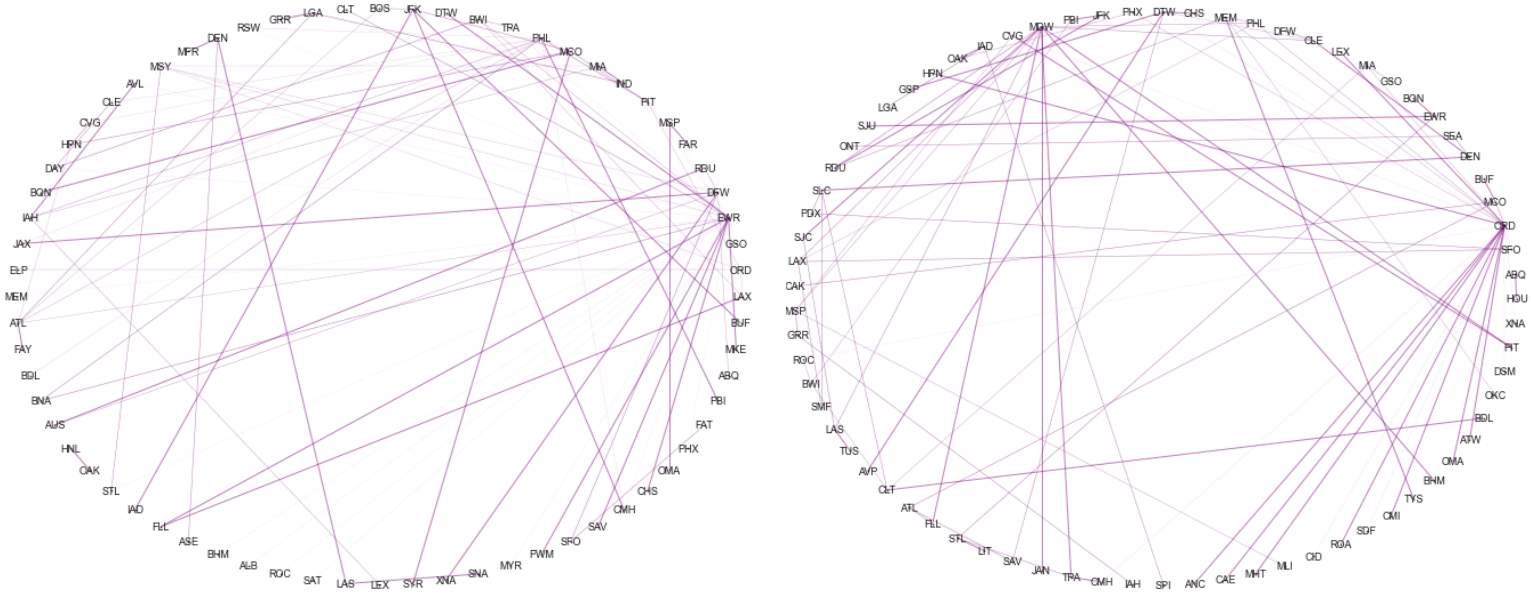
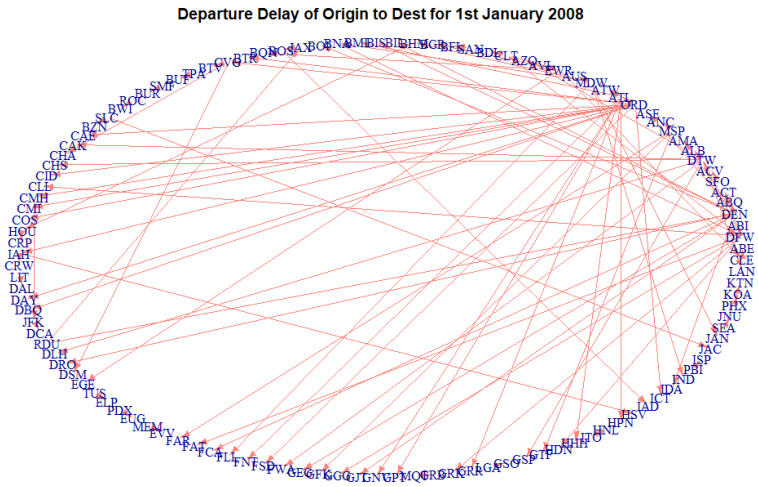


Figure 4.1. shows the cascading failures of departure delays of one airport to another from left: 1st of January 2007 and from right: 1st January 2008 generated by Python. Lighter shades of solid lines indicate lower average departure delays while dark shades indicate higher average departure delays.



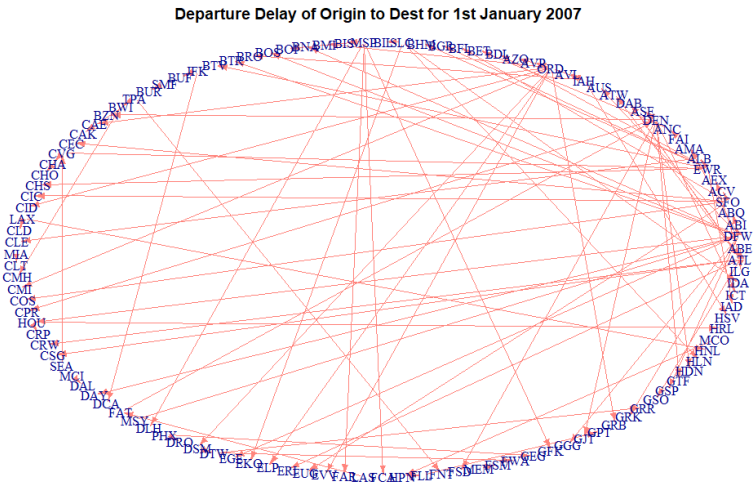


Figure 4.2. shows the cascading failures of departure delays of one airport to another from left: 1st of January 2007 and from right: 1st January 2008 generated by R.

avg_depdelay	dest	month	origin	dayofmonth	year	deptime	crsdeptime	pct_avg_depdelay
147	DFW	1	ABQ	1	2007	937	710	0.3

avg_depdelay	dest	month	origin	dayofmonth	year	deptime	crsdeptime	pct_avg_depdelay
164	JAX	1	DFW	1	2007	2254	2010	1
209	MSY	1	DFW	1	2007	1404	1035	0.228415
186	MSY	1	DFW	1	2007	1551	1245	0.203279
230	RDU	1	DFW	1	2007	120	2130	1
160	SFO	1	DFW	1	2007	2305	2025	0.517799
163	XNA	1	DFW	1	2007	1838	1555	1

avg_depdelay	dest	month	origin	ayofmont	year	deptime	crsdeptime	pct_avg_depdelay
149	ORD	1	ROC	1	2008	829	600	0.0686636

avg_depdelay	dest	month	origin	ayofmont	year	deptime	crsdeptime	pct_avg_depdelay
205	PHX	1	ORD	1	2008	825	500	0.380334
160	ROA	1	ORD	1	2008	1140	900	1
155	PHX	1	ORD	1	2008	1430	1155	0.28757
151	MIA	1	ORD	1	2008	1531	1300	0.288168
162	MHT	1	ORD	1	2008	1602	1320	1
199	HPN	1	ORD	1	2008	1643	1324	1
152	ATW	1	ORD	1	2008	1717	1445	1
187	DTW	1	ORD	1	2008	1852	1545	0.34375
162	CAE	1	ORD	1	2008	1857	1615	1
153	MEM	1	ORD	1	2008	1938	1705	0.298246

Figure 4.3. shows the scheduled and actual departure time, as well as departure delays computed from top: ABQ to DWF airport and to subsequent airports on 1st January 2007, from bottom: ROC to ORD airport and to subsequent airports on 1st January 2008, generated by Python.

	dest	origin	month	year	dayofmonth	deptime	crsdeptime	mean_depdelay
		All	All	All	All	All	All	All
15	DEN	FAT	1	2007	1	1037	737	180

	dest	origin	month	year	dayofmonth	deptime	crsdeptime	mean_depdelay
	All	DE	All	All	All	All	All	All
1	GJT	DEN	1	2007	1	1003	858	65
2	ICT	DEN	1	2007	1	1113	1020	53
3	ASE	DEN	1	2007	1	1148	1039	69
4	CPR	DEN	1	2007	1	1210	1015	115
5	FSD	DEN	1	2007	1	1559	1528	31
6	GTF	DEN	1	2007	1	1918	1745	93
7	BZN	DEN	1	2007	1	2118	2040	38
8	HDN	DEN	1	2007	1	2133	2048	45

	dest	origin	month	year	dayofmonth	deptime	crsdeptime	mean_depdelay
	All		All	All	All	All	All	All
10	DFW	BNA	1	2008	1	1111	805	186

	dest	origin	month	year	dayofmonth	deptime	crsdeptime	mean_depdelay
	All	DF	All	All	All	All	All	All
1	GGG	DFW	1	2008	1	1035	905	90
2	HDN	DFW	1	2008	1	1139	1130	9
3	CLT	DFW	1	2008	1	1152	1015	97
4	FAT	DFW	1	2008	1	1230	1140	50
5	CLL	DFW	1	2008	1	1311	1300	11
6	ABI	DFW	1	2008	1	1630	1605	25
7	GRK	DFW	1	2008	1	1817	1730	47
8	DTW	DFW	1	2008	1	2053	1735	198
9	ACT	DFW	1	2008	1	2136	2110	26

Figure 4.4. shows the scheduled and actual departure time, as well as departure delays computed from top: FAT to DEN airport and to subsequent airports on 1st January 2007. From bottom: BNA to DFW airport and to subsequent airports on 1st January 2008, generated by R.

Overall, based on the data set, the highest cascading delays that occurred on 1st January 2007 and 2008 are EWR airport and ORD airport respectively generated from Python. Similarly, the highest cascading delays occurred on 1st January 2007 and 2008 are MSY airport and BHM airport respectively generated from R. Furthermore, it is apparent that based on the data set given, these departure flight delays are attributed to the carrier, weather, security, nasdelays and security delays and thus, generated highest departure delays among these stated airports above. Therefore, these airport needs to consider reducing their departure delays by incorporating a better-integrated system and technology for better coordination and prediction to prevent the waste of resources on aircraft or airports that increases the cost of energy and fuel consumption, further dampening the greenhouse gases (SHAHEEN and LIPMAN, 2007).

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

Due to computational constraints on a lower-memory computer sy, the data has been randomly subset further 10% for R and 20,000 for Python of the previously sampled 30% data set for efficiency performance on machine learning algorithms. To facilitate computational performance, data has been pre-processed by removing NA values to avoid encountering programming errors that would affect the result of the analysis. Before model selection and assessment, a correlation plot has been drawn to visualize variables that have a relation in predicting changes to departure delays. In figure 5.1. below generated from R and Python, it can be observed that CRSDepTime, CRSArrTime, CarrierDelay, WeatherDelay, NasDelay and LateAircraftDelay are positively correlated to departure delay, indicating that these variables are useful in predicting changes to departure delays. Furthermore, these data can be obtained through internal or external forecasted data which are relevant in predicting departure delay and thus, would be included in the final prediction model. The average arrival delay will be excluded from the prediction model due to its perfect correlation with the average departure delay that may affect the outcome of the prediction model if dominant variables are included, making other significant independent (signal) variables less important in contributing to flight delays thus, lesser precision in predicting flight departure delays. Additionally, non-relevant data such as actual departure and arrival time, and taxiout were excluded in predicting future departure delays as flights have yet to depart or materialize.

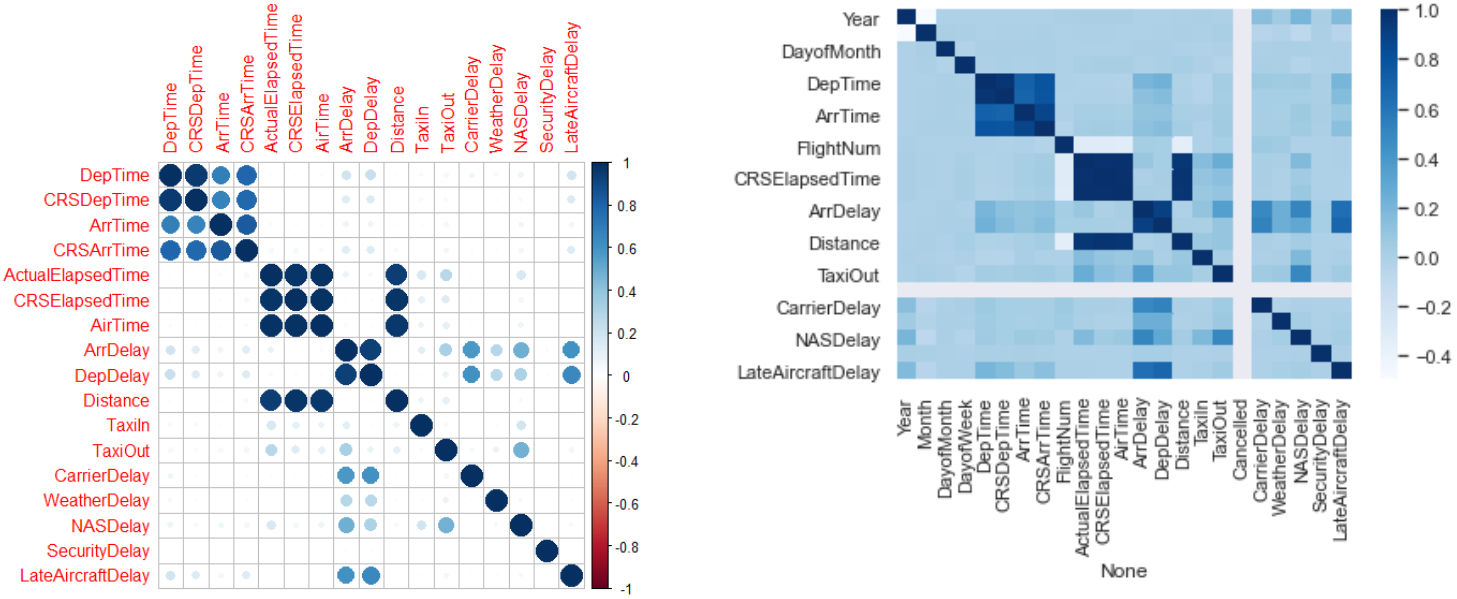


Figure 5.1. shows the correlation between individual variables from the 2007-2008 dataset generated from left: R and from right: Python.

After selection, assessment and evaluation using train-test split and 10-fold cross-validation, the test-set Root Mean Square Error (RMSE) has been generated in figure 5.2. and 5.3. for comparison between outcomes generated from R and Python across different machine learning methods. It can be observed that the lowest aggregated Root Mean Square Error (RMSE) is 9.77 generated from Support Vector Regressor via benchmark function but it is the highest RMSE in Python. The lowest RMSE is 10.51 from Ridge and Lasso Regression generated from Python. This is attributed to tuning hyperparameter that was set in R but not in Python and thus, the RMSE differs when default settings are used compared to tuned hypermeters. Thus, the prediction model generated from R is more accurate in predicting departure delay than Python. Possibly for the lowest RMSE, due to its less intensive

computation and complex algorithm structure as it is computing at a higher dimension or polynomial to better segregate the best fitting (predicted) line that has the maximum number of actual observations that are close to the predicted model, resulting in producing higher accuracy, deeper insights and credibility compared to other models (Rawat, 2022). Thus, it makes a better model to predict future departure flight delays as this would allow the aircraft or airports to optimize their current resources and contingency plans to accommodate for future flight delays.

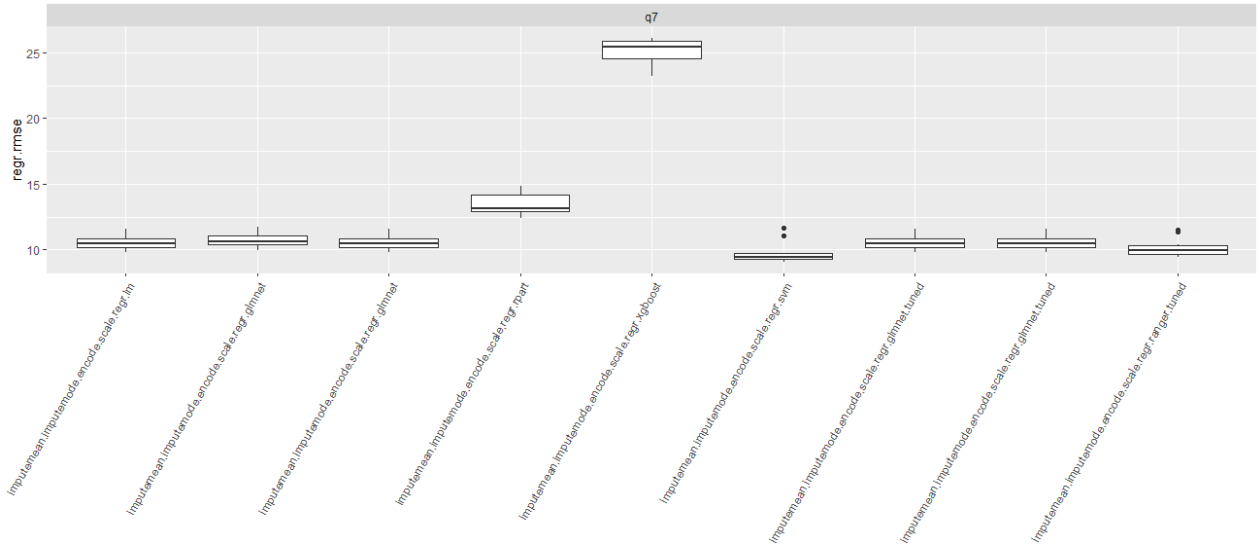


Figure 5.2. shows the Root Mean Square Error (RMSE) of individual machine learning algorithms from left to right: Linear Regression, Ridge Regression, Lasso Regression, Classification Regression, Extreme Gradient Boosting Regression, Support Vector Regression, Tuned Ridge Regression, Tuned Lasso Regression and Tuned Random Forest Regression generated by R.

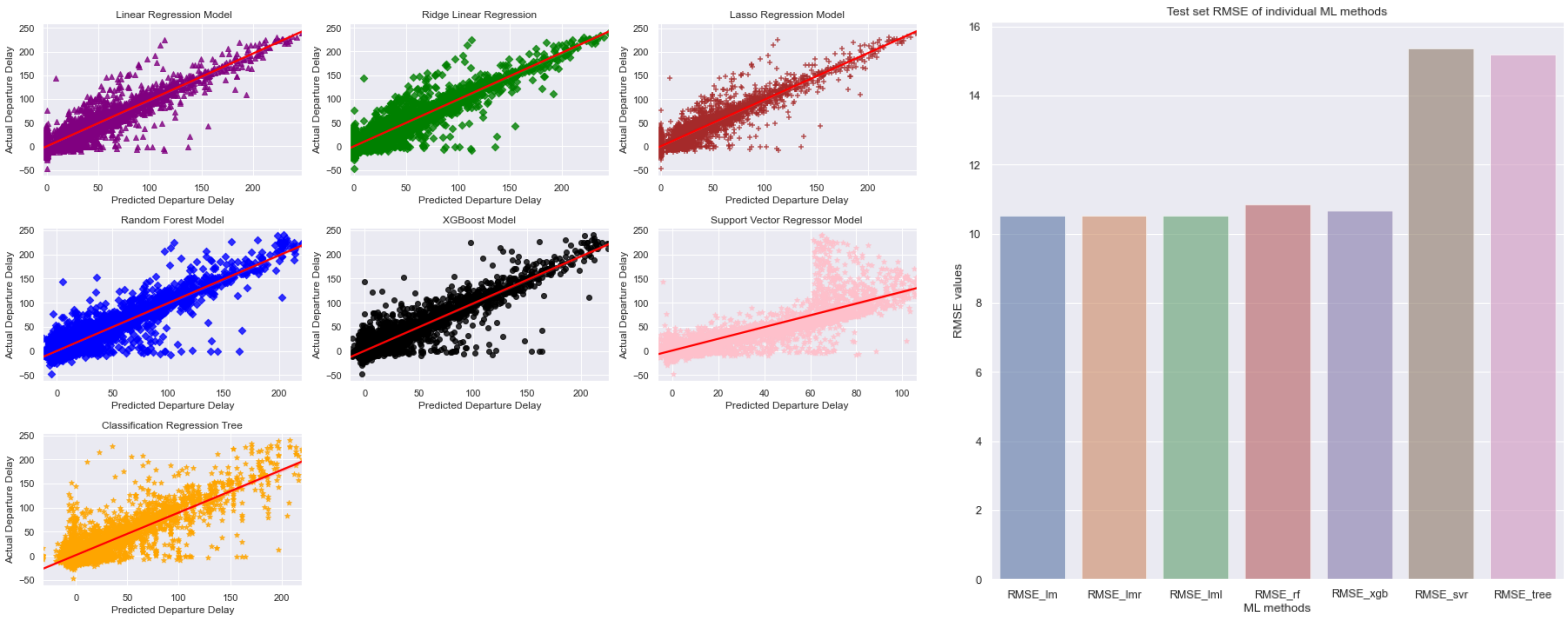


Figure 5.3. Left: shows the Regression plot of prediction against actual observations of individual machine learning algorithms and Right: RMSE values of individual machine learning methods in order from left to right: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Extreme Gradient Boosting Regression, Support Vector Regression, Classification Regression, generated by Python.

6. References:

Citizensadvice.org.uk. 2022. *Claim compensation if your flight's delayed or cancelled*. [online] Available at:<https://www.citizensadvice.org.uk/consumer/holiday-cancellations-and-compensation/if-your-flights-delayed-or-cancelled/> [Accessed 20 March 2022].

Rizzo, C., 2022. *Why You Should Only Ever Book an Early Morning Flight*. [online] Travel + Leisure. Available at: <https://www.travelandleisure.com/travel-tips/early-morning-best-time-to-fly> [Accessed 28 March 2022].

Hillyer, M., 2022. *Here's how technology has changed the world since 2000*. [online] World Economic Forum. Available at: <https://www.weforum.org/agenda/2020/11/heres-how-technology-has-changed-and-changed-us-over-the-past-20-years/> [Accessed 20 March 2022].

Corporate Finance Institute. 2022. *Dotcom Bubble*. [online] Available at: <https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/dotcom-bubble/> [Accessed 20 March 2022].

SHAHEEN, S. and LIPMAN, T., 2007. REDUCING GREENHOUSE EMISSIONS AND FUEL CONSUMPTION. *IATSS Research*, 31(1), pp.6-20.

Rawat, A., 2022. *What is Support Vector Regression?* | *Analytics Steps*. [online] Analyticssteps.com. Available at: <https://www.analyticssteps.com/blogs/what-support-vector-regression> [Accessed 20 March 2022].

Tripster Travel Guide. 2022. *10 Irresistible Reasons to Visit Hawaii*. [online] Available at: <https://www.tripster.com/travelguide/top-5-reasons-to-visit-hawaii/#:~:text=World%2Dclass%20beaches%2C%20pristine%20rainforests,matter%20which%20way%20you%20turn> [Accessed 20 March 2022].

Collections of Waikiki. 2022. *Hawaii in January: 5 Reasons Why This is a Good Time to Visit*. [online] Available at: <https://collectionsofwaikiki.com/hawaii-in-january/> [Accessed 20 March 2022].

En.wikipedia.org. 2022. *Los Angeles International Airport - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Los_Angeles_International_Airport> [Accessed 20 March 2022].

Cber.unlv.edu. 2022. *Airports* | *Center for Business and Economic Research* | *University of Nevada, Las Vegas*. [online] Available at: <https://cber.unlv.edu/SNBDI/airports.html> [Accessed 20 March 2022].

Sood, S., 2022. *Why is Atlanta the world's busiest airport?*. [online] Bbc.com. Available at: <https://www.bbc.com/travel/article/20130207-why-is-atlanta-the-worlds-busiest-airport> [Accessed 30 March 2022].