






## UCCC3073 DATA SCIENCE ASSIGNMENT

<b>Programme(s)</b>	:	Bachelor of Computer Science (Honours)
<b>Trimester</b>	:	June 2022
<b>Course Leader</b>	:	Dr. Tong Dong Ling ( <a href="mailto:tongdl@utar.edu.my">tongdl@utar.edu.my</a> )
<b>Submission Date</b>	:	<b>Monday, 22 August 2022, 5 p.m.</b>
<b>Submission Platform</b>	:	<b>WBLE</b>

### Group Information

<b>Group number</b>	1			
<b>Student Name</b>	<b>Student ID</b>	<b>Individual Contribution %</b>	<b>Signature</b>	<b>Final mark</b>
1) Cheong Kin Fai	20ACB03897	40		
2) Lee Jian Zhen	19ACB02281	30		
3) Thiang Jian ru	19ACB01142	30		

### Course Learning Outcomes Assessed

CLO2	Assess and manipulate data sources from various online sources.
------	---

## Assignment mark sheet

Items	Poor	Marginal	Satisfactory	Good	Excellent	Marks
<b>Task 1: Data Summary [30 marks]</b>						
Data understanding [10m]	0 2	4	6	8	10	
Data acquisition [5m]	0 1	2	3	4	5	
Data description [10m]	0 2	4	6	8	10	
Outlined questions [5m]	0 1	2	3	4	5	
<b>Task 2: Data Analysis [30 marks]</b>						
Lead sentence[10m]	0 1	2	3	4	5	
Statistic summary [10m]	0 2	4	6	8	10	
Clarity of the visualisation [10m]	0 2	4	6	8	10	
<b>Task 3: Data Memo [30 marks]</b>						
Identification of audience [5m]	0 1	2	3	4	5	
Insights and recommendations [10m]	0 2	4	6	8	10	
Clarity of visualisation [10m]	0 2	4	6	8	10	
Reference sources [5m]	0 1	2	3	4	5	
<b>Overall report writing [10m]</b>	0 2	4	6	8	10	
<b>Total mark (100)</b>						

Assessment category	Criteria
Poor	Insufficient content, “rush work”, ambiguity, fatal errors
Marginal	Insufficient content, fairly written, some major fatal errors
Satisfactory	Fair content, reasonably written, minor problems with formatting and coding
Good	Good proportion of content, clearly written and coded
Excellent	Concise, high density, value creation to organization

Marking item	Description
Data understanding	Clear understanding on the chosen data set evident with a relatively detailed explanation on the data set.
Data acquisition	The origin of data
Data description	Data attributes are clearly explained
Outlined questions	questions are meaningful and are answerable from the data
Lead sentence	Most interesting thing in the data that would Interest readers
Statistic summary	Clear, concise summary to show lead and key findings of the data
Clarity of the visualisation to support the lead	The charts clearly illustrated lead sentence, key findings and/or insights of the data. The charts are appropriately sized to be easily read within the report
Identification of audience	Appropriateness of audience
Insights and recommendations	Clear, compelling points derived from the data and appropriate recommendations given
Reference sources	Sources have been well-summarised and referenced in the report. The sources are meaningful and greatly improve the readability and content of the report
Overall report writing	Well-structured report with proper use of grammar, punctuation and spelling

## Table of Contents

Table of Contents .....	3
Task 1: Data Summary.....	4
1.1 Problem(s) or Challenge(s) .....	4
1.2 Nature of Data Set .....	5
1.3 Question Outlines .....	7
Task 2: Initial Analysis .....	8
Task 3: Data Memo.....	13
Brief description of the intended audience .....	13
Finding 1: Distance .....	14
Suggestion.....	17
Finding 2: Flight Delay .....	18
Suggestion.....	22
Finding 3: Cancelled Flight .....	23
Suggestion 3.....	23
References.....	27

## Task 1: Data Summary

### **1.1 Problem(s) or Challenge(s)**

After a discussion, the problem and challenge were about to predict the likeliness of flight delay from the dataset. The number of flights that fail to take off on time likewise rises as more consumers prefer to travel by air. This expansion makes airports even more crowded and harms the airline industry's bottom line. Flight delays are a sign of the aviation system's inefficiency. It comes at a high price for both customers and airline corporations. As a result, forecasting delays can benefit airline operations and passenger pleasure, which will benefit the economy.

The next challenge was reducing the amount of flight cancelled by identifying the factor. The reason was that flight cancellation are the most frustrating things for traveler or others such as mess up their well scheduled time. However, there were always a reason for the cancellation of flights such as inclement weather, security, mechanical issues, and so on. Thus, identifying the factor that cause flight cancellation is a must so that the company can enhance or improve their security, system, and so on by using their carrier's name, date, time, and the location of departure and destination to search for the factor of flight cancelled.

The third challenge that is targeted to solve is to identify the rate of flight delay from origin airport to destination airport. This is because some of the airport flight delay might be caused by the airport to have overcrowded phenomenon. From here, if the flight delay is to be found caused by the congestion of airport, then the problem of flight delay can be prevented by limiting the number of flights that can be allowed in the airport. Likewise, even if the airport congestion is already under monitoring, it could be because of the location of the airport is not strategic. Thus, when flights are travelling to the destination airport from the origin, the chances of accident like having many clouds, rain, or even thunder during the flight can slow the flight down. Hence, causing the delay to happen.

## **1.2 Nature of Data Set**

The data is collected from the Bureau of Transportation Statistics, Govt. of the USA. This file includes all flights beginning on January 1 and lasting until January 31, 2019. This file contains approximately 400 000 rows and 21 feature columns that describe the elements of the flight, such as the origin airport, destination airport, information on the aircraft, departure time, and arrival time. The data was split into four theme which is Airport, Flight, Date, and Time.

### **Airport**

<b>Attribute Name</b>	<b>Attribute Description</b>	<b>Attribute Type</b>
ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.	Nominal
ORIGIN_AIRPORT_SEQ_ID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time.	Nominal
ORIGIN	Original Airport.	Nominal
DEST_AIRPORT_ID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.	Nominal
DEST_AIRPORT_SEQ_ID	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time.	Nominal
DEST	Destination Airport.	Nominal
DISTANCE	Distance between airports (miles).	Ratio

### **Flight**

<b>Attribute Name</b>	<b>Attribute Description</b>	<b>Attribute Type</b>
OP_UNIQUE_CARRIER	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2).	Nominal
OP_CARRIER_AIRLINE_ID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation	Nominal

OP_CARRIER	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique.	Nominal
TAIL_NUM	A unique identification of Tail Number for airplane.	Nominal
OP_CARRIER_FL_NUM	An identification number for Flight Number.	Nominal

**Date**

Attribute Name	Attribute Description	Attribute Type
DAY_OF_MONTH	Day of Month (1~31).	Interval
DAY_OF_WEEK	Day of Week starting from Monday (1~7).	Interval
CANCELLED	Cancelled Flight Indicator (1=Yes, 0=No).	Binary, Nominal
DIVERTED	Diverted Flight Indicator (1=Yes, 0=No).	Binary, Nominal

**Time**

Attribute Name	Attribute Description	Attribute Type
DEP_TIME	Actual Departure Time (local time: hhmm).	Ratio
DEP_DEL15	Departure Delay Indicator, 15 Minutes or More (1=Yes, 0=No).	Binary, Nominal
DEP_TIME_BLK	Departure Time Block, Hourly Intervals.	Ratio
ARR_TIME	Actual Arrival Time (local time: hhmm).	Ratio
ARR_DEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes, 0=No).	Binary, Nominal

### **1.3 Question Outlines**

The purpose of the first question was to identify the airline with the most flight delays. The results will also list the flight delays of other carriers after eliminating the carrier with the most delays. With this outcome, the challenge that mentioned earlier can be solved where the likelihood of a flight delay may be predicted, and the original outcome can then be contrasted.

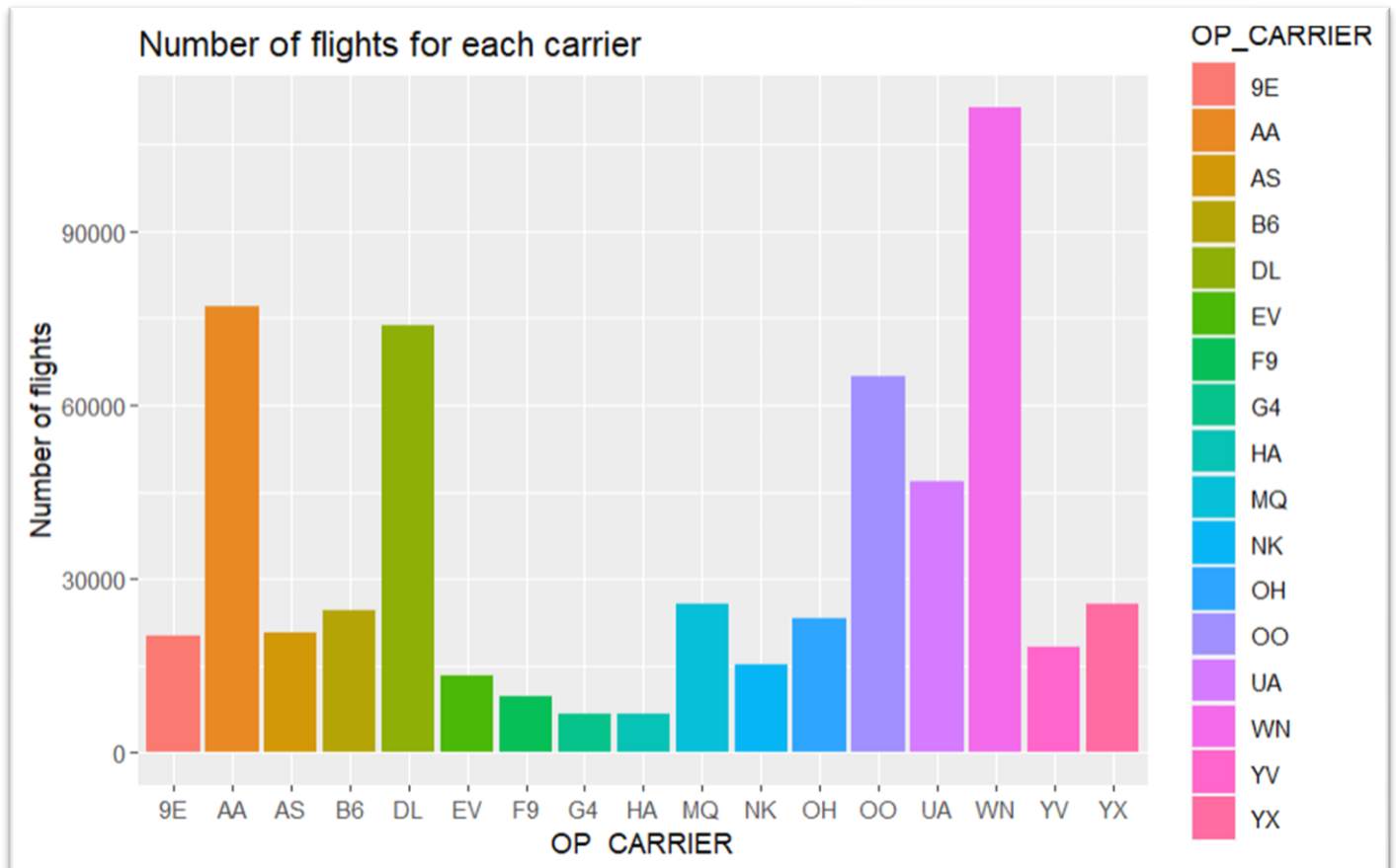
Next, the question outline was about to find out the coefficient of number of airplane and the number of flight delay. From here, the factor of cancelation of flight may also be able to detect by finding the relationship between flight delay and flight cancelled.

The third question outline was to search for which origin airport to destination airport has the highest rate of delay to identify the rate of flight delay from origin airport to destination airport. Furthermore, the result will be used and determine either what cause the flight to be delayed by looking for this date, time, and the airplane tail number.

## Task 2: Initial Analysis

### LOWER FLIGHT NUMBER BUT HIGHER FLIGHT DELAY!?

First, both G4 and HA are a flight carrier code, respectively. It is a very interesting findings as in the dataset, it was thought that the higher number of flights will contribute to higher rate of delay.



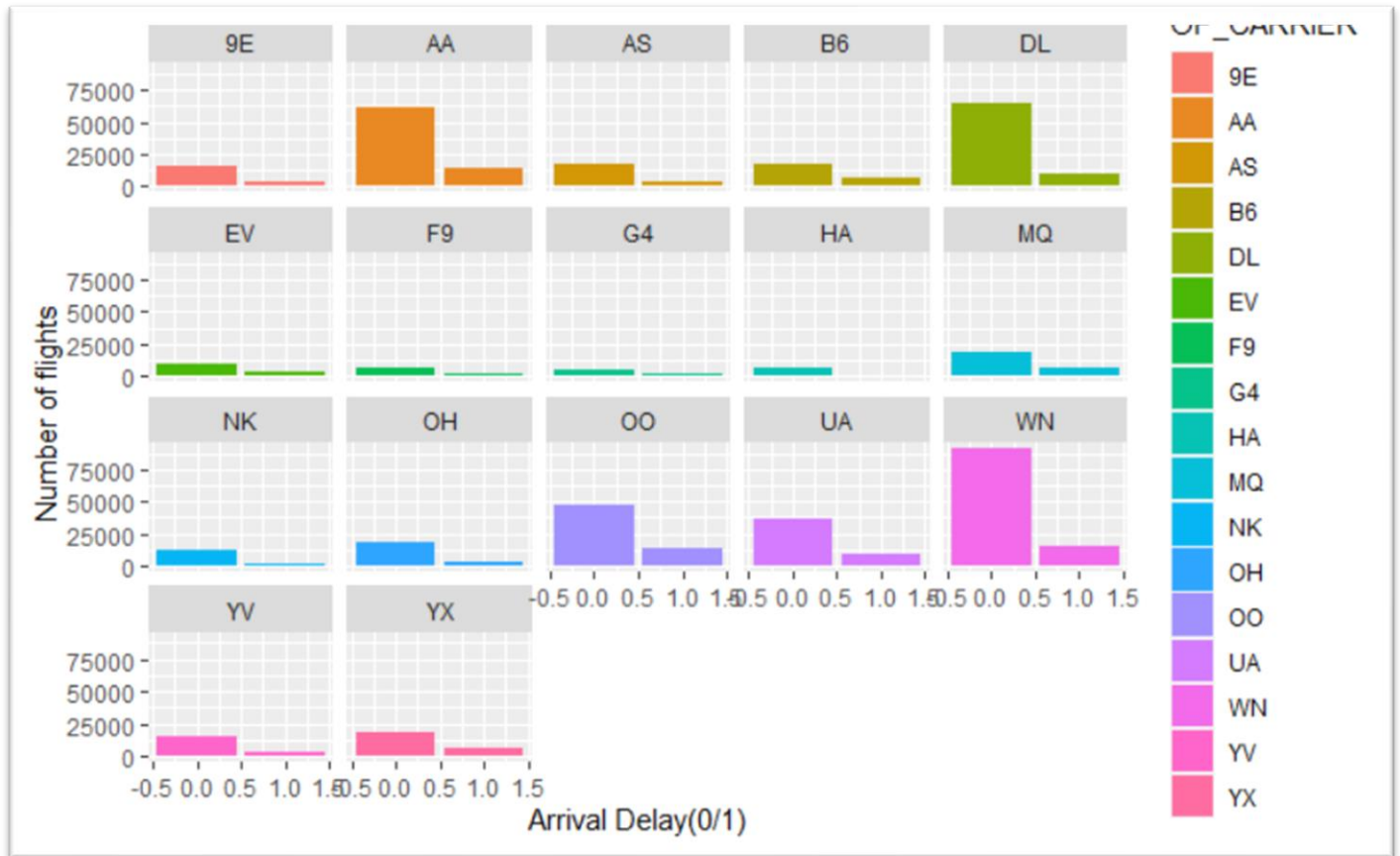
*Figure 2.1. Number of flights of each flight carrier*

A tibble: 2 × 2	
OP_CARRIER <chr>	Total_flights <int>
G4	6763
HA	6798
2 rows	

*Table 2.1. Total number of OP\_CARRIER of G4 and HA*



Based on Figure 2.1, the highest number of flights is WN carrier. Since the number of flights for G4 and HA is very close as shown in the Figure 2.1. It is hard to see that whether which one of them has the lowest of number of flights. Hence, table 2.1 is generated to show more closely on the number of flights. The lowest number of flights is G4 carrier has total flight of 6763 and HA has total flight of 6798.



**Figure 2.2.** Bar graph of Number of flights vs Arrival Delay

OP_CARRIER <chr>	NumberOfDelay <int>
G4	1728
HA	851
2 rows	

**Table 2.2.** Table for the number of delays of G4 and HA

Based on Figure 2.2, it is shown that the WN carrier which has the highest number of flights has the highest number of arrival delay which is 16111 as the 0 for arrival delay indicates FALSE ( no delay ) and the 1 for

arrival delay indicates TRUE ( flight delay ). It is initially thought that the higher the number of flights, the higher the number of delays. However, this is proven not the case as when analyzing the graph, it is found that the number of flights between G4 and HA is just 35 in difference as G4 has total flight of 6763 and HA has total flight of 6798. But the number of delay that G4 had is 1728 and HA only had 851 which is twice as worse flight delay compared to HA as shown in Table 2.2. Hence, what factors caused flight delay? It is interesting as when the dataset is analyzed, it is found that the DEP\_DEL15 which indicates that the departure delay of a flight contributes the most to arrival delay. In other words, if a flight departs late, it will arrive late. It has a noticeably high correlation compared to other columns in the dataset which is 0.71943 as shown in Table 2.3.

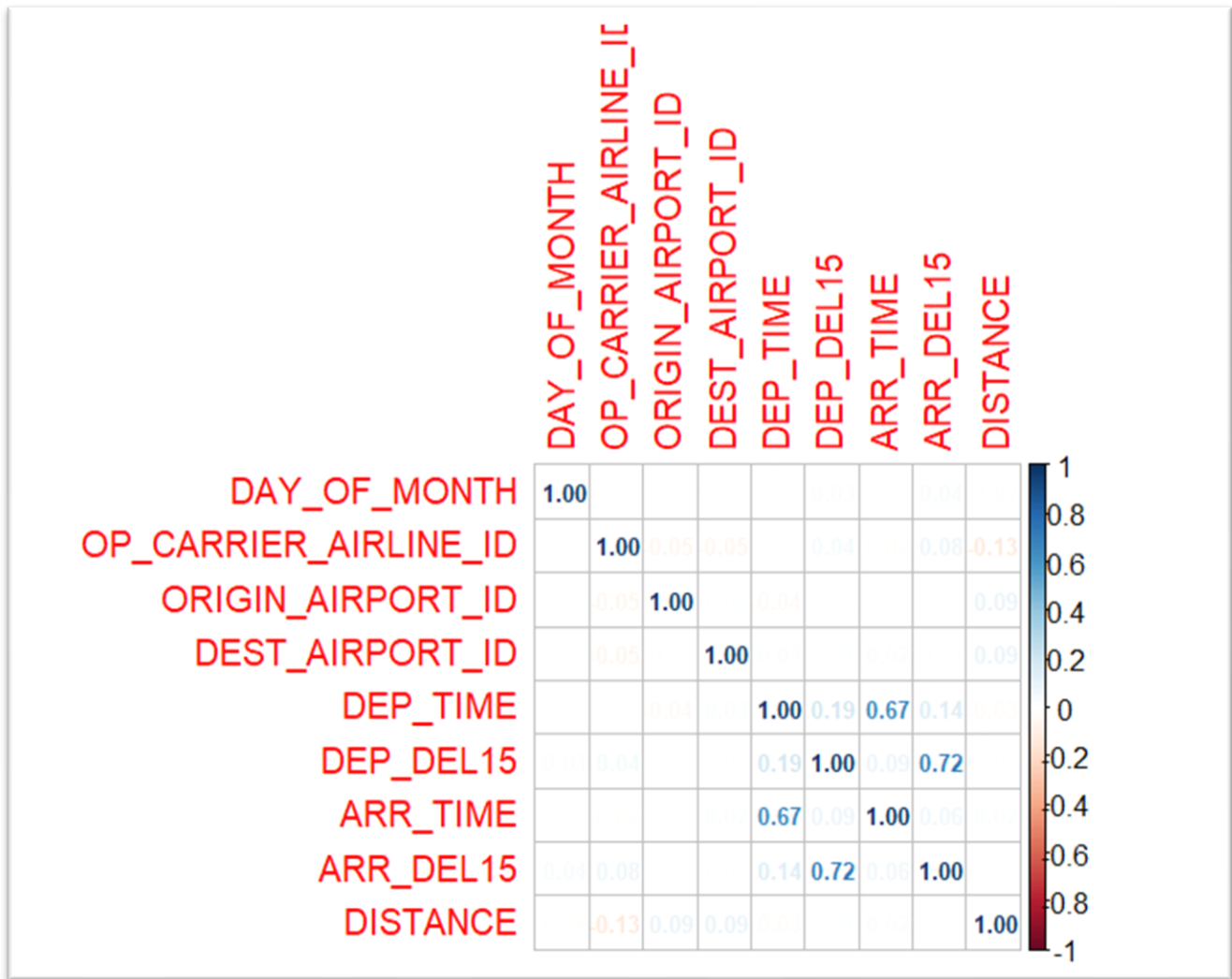
```

{r}
cor(dfdrop2,dfdrop2$ARR_DEL15, method = c("pearson", "kendall", "spearman"),
    use = "complete.obs") #the correlation between both

```

	[,1]
DAY_OF_MONTH	0.0386385307
DAY_OF_WEEK	-0.0006548303
OP_CARRIER_AIRLINE_ID	0.0792381287
OP_CARRIER_FL_NUM	0.0386400816
ORIGIN_AIRPORT_ID	0.0084819618
ORIGIN_AIRPORT_SEQ_ID	0.0084821482
DEST_AIRPORT_ID	0.0155397710
DEST_AIRPORT_SEQ_ID	0.0155398668
DEP_TIME	0.1432828501
DEP_DEL15	0.7194299212
ARR_TIME	0.0620308756
ARR_DEL15	1.0000000000
DISTANCE	0.0034065405

**Table 2.3.** Correlation table against all numeric variables.



**Figure 2.3** Correlation matrix graph against useful columns

The correlation of 0.71943 also tell us that departure delay does not mean that the arrival of a flight must be late. Let us take the highest number of flight delay, WN carrier to explain further. In table 2.4,

OP_CARRIER <chr>	TotalDelay <int>
WN	16111
1 row	

**Table 2.4.** WN carrier number of flight delay

OP_CARRIER <chr>	LateDep_Arr <int>
WN	12926
OO	10845
AA	9455
UA	6834
DL	6708
B6	5257
YX	4149
MQ	4094
9E	3191
OH	2625
1-10 of 17 rows	

**Table 2.5.** *Number of departure and arrival delay of each flight*

The number of flight arrival delay from WN is 16111 and the number of late departure and arrival delay is only 12926. Putting it into percentage, it means that there is 80% of the flight that are late after late departure. In another perspective, there might be more factors contributing to the delay.

## Task 3: Data Memo

### **Brief description of the intended audience**

The audience that this data analysis is for the researchers that would like to find out the reason of why the reason of flight delay. The objective of conducting this data analysis is to find out the factors of flight delay. This analysis of January Flight Delay Prediction is a dataset collected from Bureau of Transportation Statistic, USA. Hence, the data source is highly creditable, and the analysis performed on this dataset should be mirroring to the real-life scenario of everyday flights. However, in this EDA conducted by this group, this can only solely be used for researchers that like to investigate the situation of flights delay in US. This is because the dataset is collected from USA. Beside researchers, data scientists are also one of the intended audiences. This is because data scientists are always exploring variety of datasets depends on what they would like to explore. Thus, the data analysis conducted here may be referred by fellow data scientist in the future. Furthermore, flights company are also one of the intended audiences. This is because with the data analysis done by us here, the flights company will know the reason of why the flight is delaying. Thus, avoiding all the possible reasons and minimizing the delay even further after knowing the root of the causes.

**Finding 1: Distance**

This finding is about the relationship between total flight distance from each carrier and arrival delay as well as departure delay. To show the relationship, Tables 3.1.1 and 3.1.2 have been created. The correlation between arrival delay, departure delay, and distance are 0.011 and 0.0034. Both are very weak relationships. Based on tables 3.1.1 and 3.1.2, the carrier id with WN has the highest number of total distances, arrival delay, and departure delay as well as both delay percentages which are 26412706 miles, 16111 arrival delay, 18507 departure delay, and 15.31%, 18.71% of percentage delay while carrier id with HA has the lowest number of total distances, arrival delay, departure delay, and both percentages.

OP_UNIQUE_CARRIER <chr>	ARR_DEL15 <int>	arr_percentage <dbl>	DEP_DEL15 <int>	dep_percentage <dbl>	distance <int>
9E	4013	3.8138412	3746	3.7866305	3899059
AA	13741	13.0590561	12077	12.2079918	25694810
AS	3506	3.3320028	2789	2.8192506	8198358
B6	6429	6.1099390	6362	6.4310047	13841974
DL	9403	8.9363441	9462	9.5646285	17981020
EV	3022	2.8720230	2528	2.5554197	2556723
F9	2362	2.2447777	2338	2.3633588	5064900
G4	1728	1.6422421	1441	1.4566296	2920825
HA	851	0.8087662	585	0.5913451	2234598
MQ	5844	5.5539716	4737	4.7883793	4671378

*Table 3.1.1 Percentage of Delay and Total Distance for each Carrier*

OP_UNIQUE_CARRIER <chr>	ARR_DEL15 <int>	arr_percentage <dbl>	DEP_DEL15 <int>	dep_percentage <dbl>	distance <int>
NK	2383	2.2647355	2171	2.1945475	4533759
OH	3475	3.3025413	3328	3.3640968	2696601
OO	14024	13.3280113	12714	12.8519009	13838633
UA	9156	8.7016023	8401	8.4921205	21259832
WN	16111	15.3114368	18507	18.7077340	26412706
YV	3310	3.1457300	2914	2.9456064	3651670
YX	5864	5.5729790	4827	4.8793555	6299973

*Table 3.1.2 Percentage of Delay and Total Distance for each Carrier*

Its percentage of delay is 0.81 for arrival delays, and 0.59 for departure delays with a total number of 851 and 585. The total distance of HA has been flighted is 2234598 miles. Besides the highest and lowest values, there are some of the carrier ID have a similar percentage of delay among others which is MQ and

YX, the arrival delay is 5844, and 5864 and departure delay with 4737 and 4827. The percentage of arrival delays are 5.55% and 5.57% whereas departure delays are 4.79% and 4.88%.

Secondly, although the total number of arrival delays and departure delays in AS and OH are slightly different (3506, 3475) the percentages of arrival delays are similar (3.33% and 3.30%). Graphs 3.1.1 and graph 3.1.2 have been plotted for visualization of table1 and table 3.1.2 to show the finding. As the graph is plotted, it is easier to find out the highest and lowest values among the attributes. Based on the table and graph shown, a short conclusion can be stated is the longest the flight distance, the higher the delay rate. To examine the conclusion stated, Tables 3.1.3 and 3.1.4 are created. Before discussing in table 3.1.3 and table 3.1.4, how to define a flight distance in short or long. According to [1] [2], in the US, long-haul flights are those that are more than 3,000 miles (approx. 4800km), whereas short-haul flights are those that are fewer than 700 miles (approx. 1100km). According to table 3.1.3, there are a total of 55598 arrival delay cases on a short flight and, 51325 departure delays on short flights. Whereas long flights with arrival delays have 49624 and a total of 47602 departure delays. Table 3.1.4 show that the percentage of arrival delay and departure delay in short flight is lower than in long flights. 18.08% of arrival delays and 16.65% of departure delays on short flights whereas 21.51% arrival delays and 19.80% departure delays on long flights.

```
percentage = c(percentage_short_arr , percentage_short_dep, percentage_long_arr,percentage_long_dep)

percentage <- data.frame(percentage)
percentage <- percentage %>% rename("Short Flight (ARR_DEL)% " = n, "Short Flight (DEP_DEL)% " =n.1,
                                   "Long Flight (ARR_DEL)% " = n.2, "Long Flight (DEP_DEL)% " =n.3)
percentage
```

Short Flight (ARR_DEL)% <dbl>	Short Flight (DEP_DEL)% <dbl>	Long Flight (ARR_DEL)% <dbl>	Long Flight (DEP_DEL)% <dbl>
18.08283	16.64548	21.50793	19.79459

1 row

*Table 3.1.3 Total percentage of Delay from Short and Long Flight*



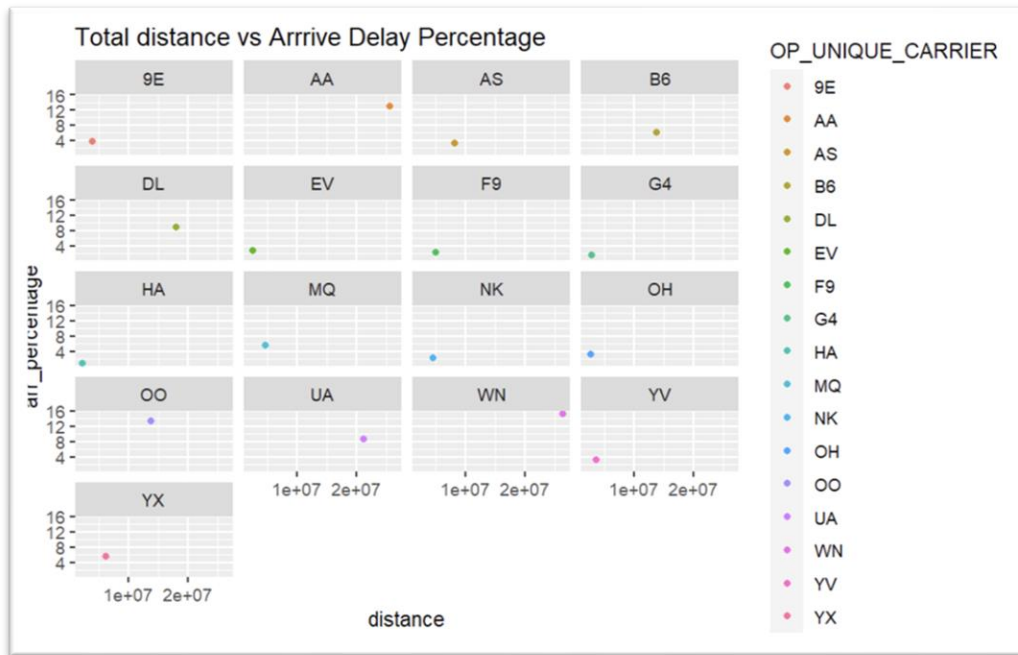
short_flight_arr_delay		
ARR_DEL15	n	
<int>	<int>	
1	55598	
1 row		
Hide		
short_flight_dep_delay		
DEP_DEL15	n	
<int>	<int>	
1	51325	
1 row		
Hide		
long_flight_arr_delay		
ARR_DEL15	n	
<int>	<int>	
1	49624	
1 row		
Hide		
long_flight_dep_delay		
DEP_DEL15	n	
<int>	<int>	
1	47602	
1 row		

*Table 3.1.4 Total number of Delay from Short and Long Flight*

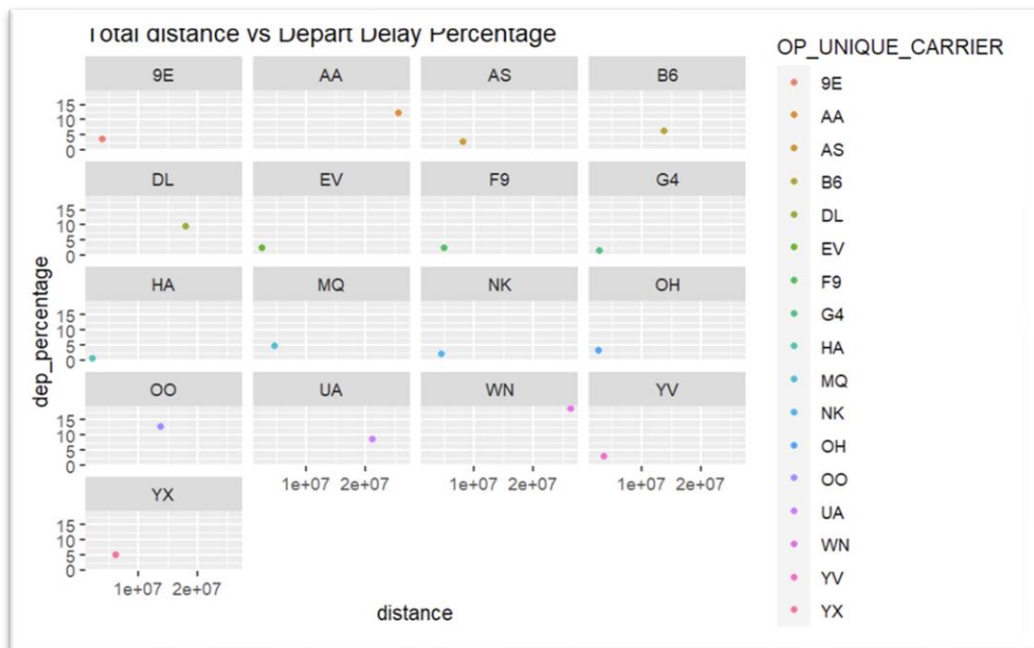
DAY_OF_MONTH	-0.01601396
DAY_OF_WEEK	0.01702414
OP_CARRIER_AIRLINE_ID	-0.12506184
OP_CARRIER_FL_NUM	-0.34254928
ORIGIN_AIRPORT_ID	0.09359686
ORIGIN_AIRPORT_SEQ_ID	0.09359724
DEST_AIRPORT_ID	0.09323709
DEST_AIRPORT_SEQ_ID	0.09323747
DEP_TIME	-0.02694862
DEP_DEL15	0.01139318
ARR_TIME	0.02090613
ARR_DEL15	0.00340654
CANCELLED	NA
DIVERTED	NA
DISTANCE	1.00000000

*Figure 3.1.1 Correlation Distance with another attributes*





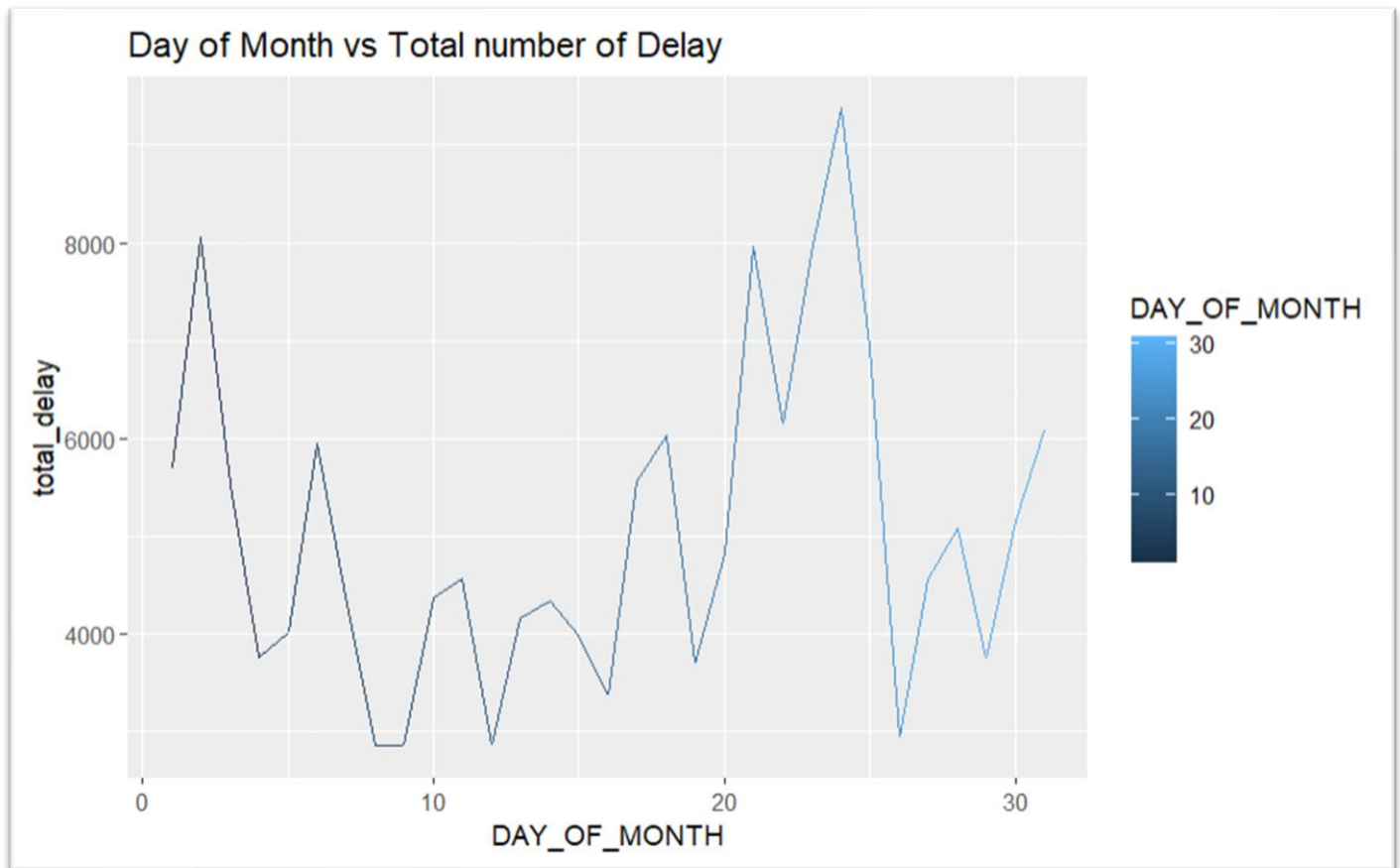
Graph 3.1.1 Total Distance vs Arrival Delay Percentage



Graph 3.1.2 Total Distance vs Departure Delay Percentage

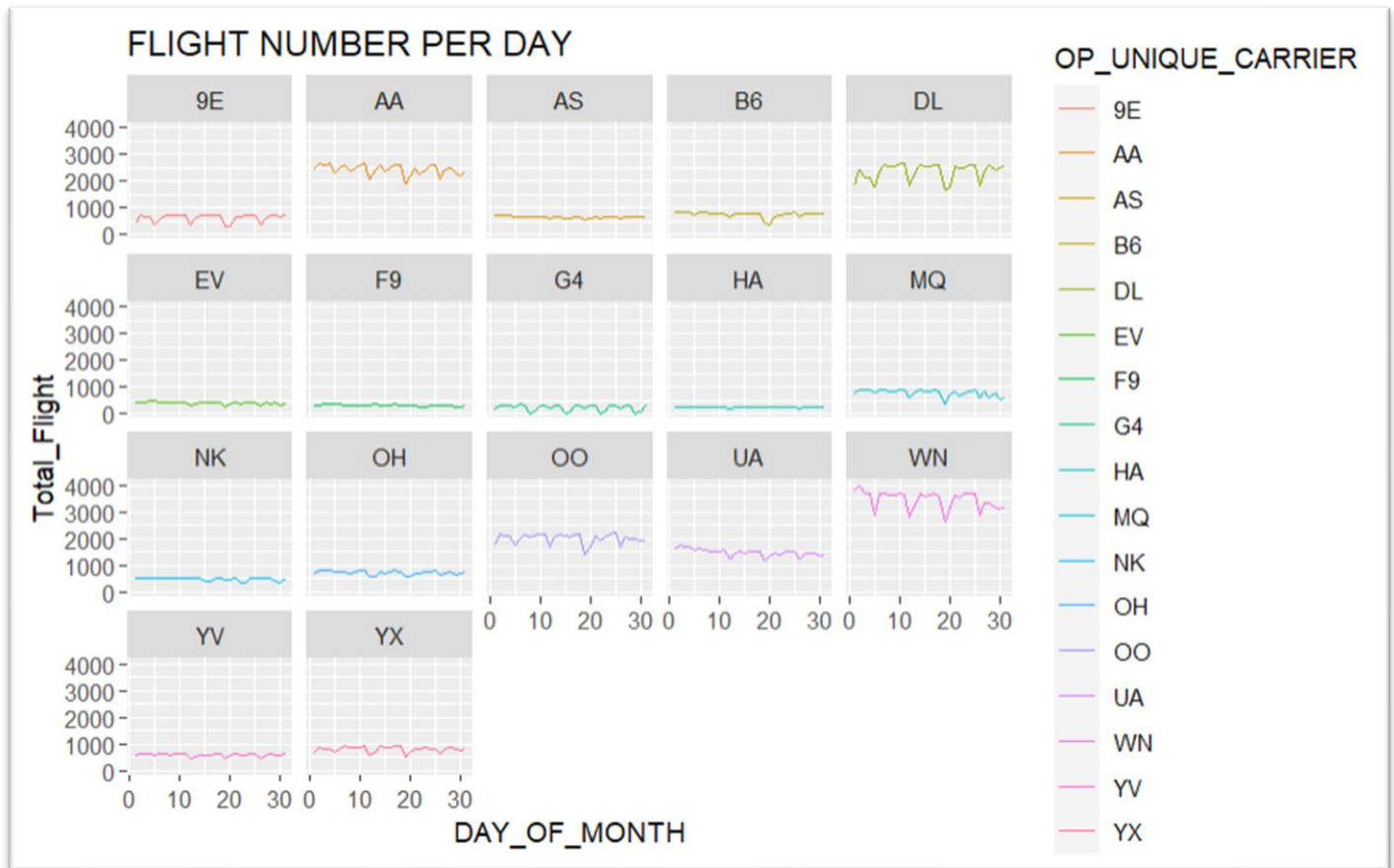
### Suggestion

No research explicitly states that distance is one of the factors that affect flight delays, even though the conclusion is based on the results in the table and graph. But according to their research [3], large passenger loads on any route at any time of year and at the busiest time of the year both increase a flight's revenue. This contrasts with delays in arrival and departure flights at the airport of origin and the distance between the two airports.

**Finding 2: Flight Delay**

**Figure 3.2.1** Graph of Day of Month against Total number of delays.

This finding is primarily focus on the reason of flight delay. In the previous discussion, it is mentioned that the reason of flight delay is caused by the departure delay of a flight. Thus, causing the arrival delay. However, since the correlation of the departure delay with arrival delay is only 0.71943 as shown in Table 3, it is suspected that there might be other factor contributing to flight delay. Hence, let us continue to look at other columns that might be contributing to it. So, based on this graph in Figure 3.2.1, it is shown that the peak of the delay is at Day of Month of 24. So, to investigate the relationship between the date of 24 to the flight delay. A graph of Figure 3.2.2 is plotted to show that during the 24<sup>th</sup> of the month, how many flights are flying to cause the flight delay to reach the peak amount? To show it clearly with numbers, Table 3.2.3 is created. In the Figure 3.2.2, almost every carrier had the highest amount of flight number during the date 24<sup>th</sup> of the month.



*Figure 3.2.2. Number of flights per day for each carrier*

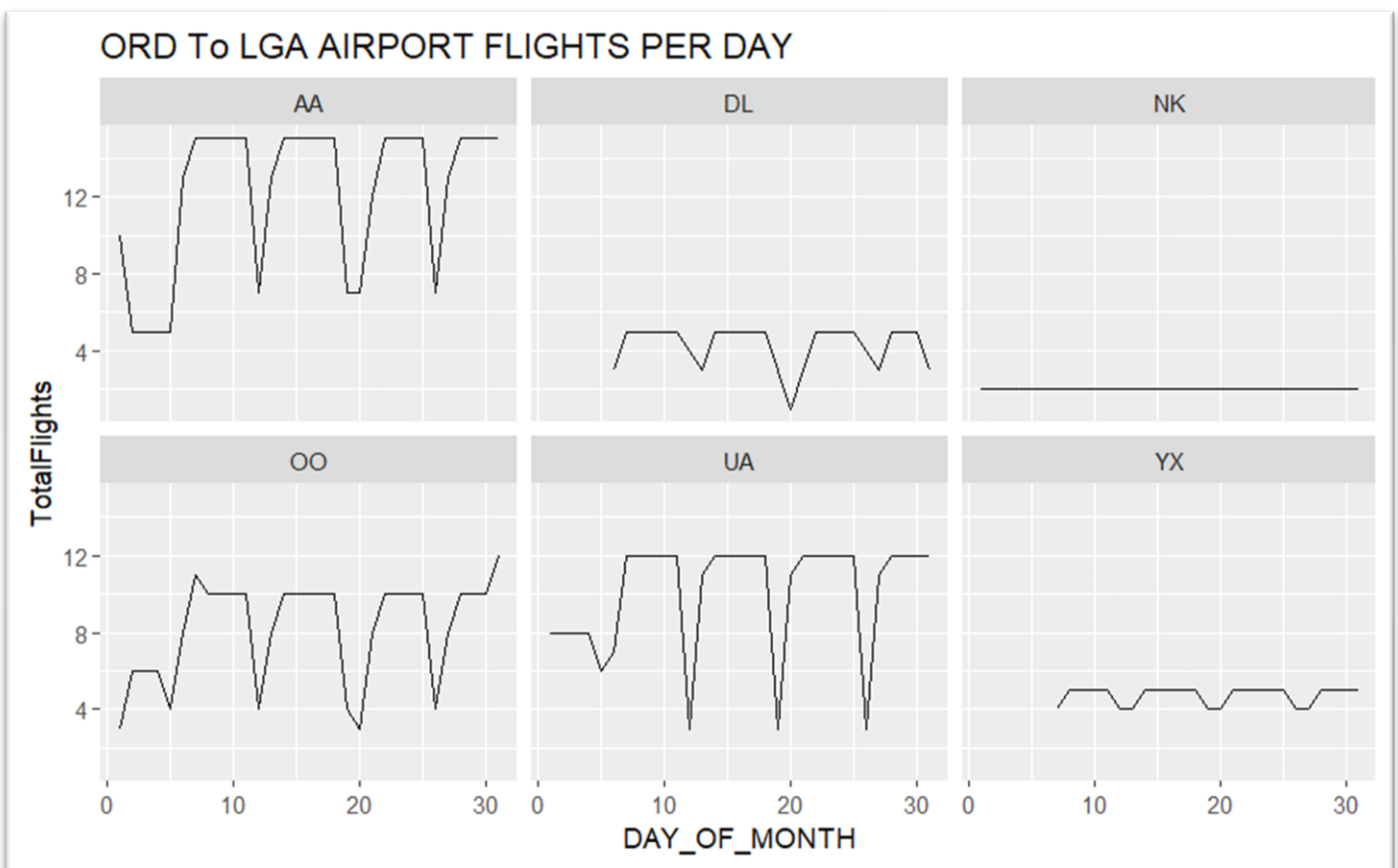
OP_CARRIER <chr>	DAY_OF_MONTH <int>	TotalFlights <int>
WN	24	3710
AA	24	2632
DL	24	2632
OO	24	2236
UA	24	1532
YX	24	907
MQ	24	892
B6	24	811
OH	24	796
9E	24	750
OP_CARRIER <chr>	DAY_OF_MONTH <int>	TotalFlights <int>
AS	24	682
YV	24	619
NK	24	508
EV	24	429
F9	24	329
G4	24	284
HA	24	214

**Table 3.2.3.** Flight number of day 24<sup>th</sup> in table form.

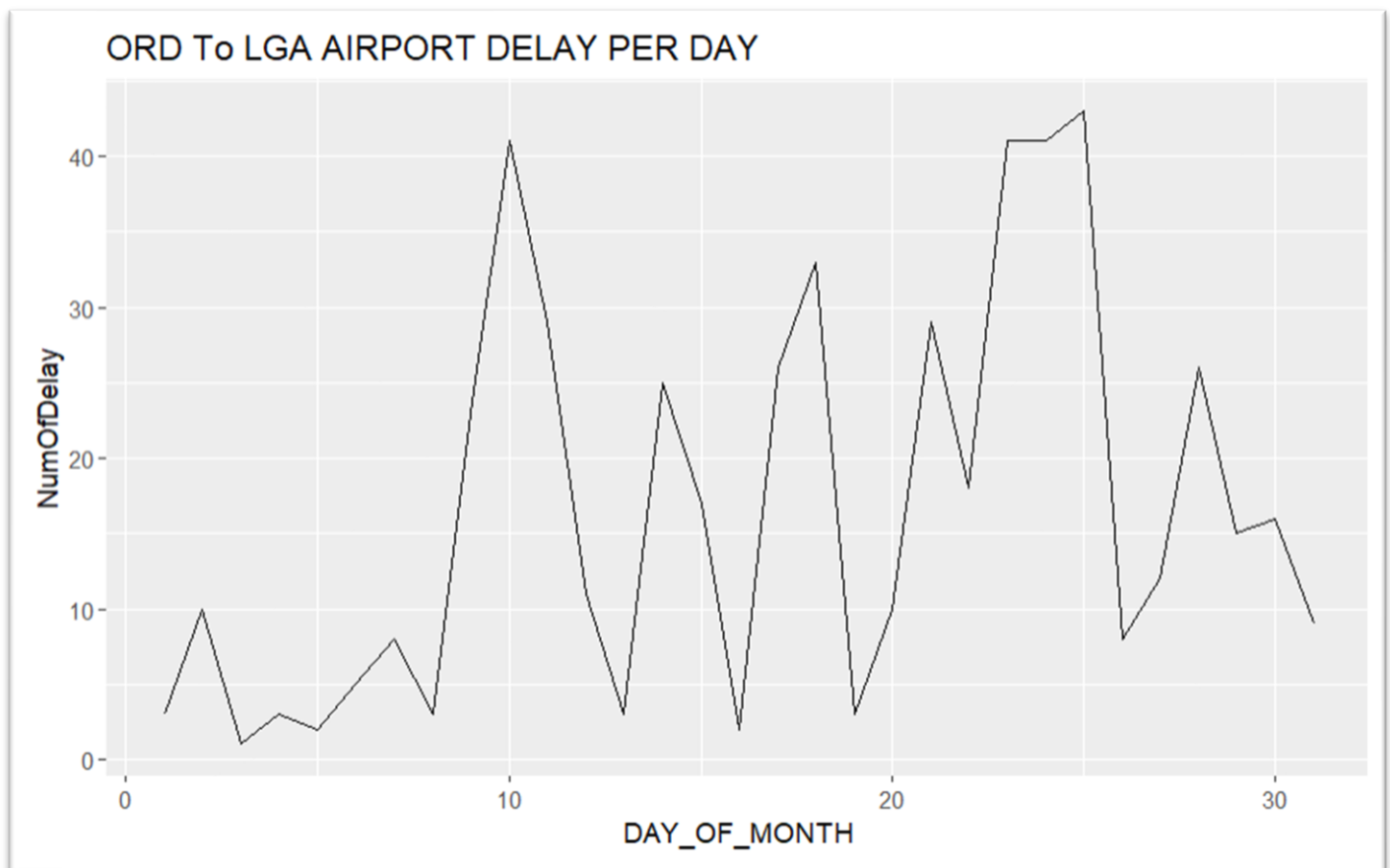
It is assumed that higher number of flights contributes to higher flight delay. However, in the previous discussion, it is already mentioned that the flight delay does not align with the higher number of flights. What else can be found from this information? To answer this, Table 3.2.4 is coded to show which airport has the highest number of delays. In the table, it is saw that airport origin from ORD to destination LGA has the highest flight delay of total 516. An assumption of airport origin from ORD to destination LGA has been made that during 24<sup>th</sup>, the number of flights going from ORD to LGA is the highest.

ORIGIN <chr>	DEST <chr>	TotalDelay <int>
ORD	LGA	516
LGA	ORD	431
LAX	SFO	370

3 rows

**Table 3.2.4.** Airport with the highest number of flight delay**Figure 3.2.5** Flights per day on the ORD to LGA airport.

Referring to the Figure 3.2.5, the flight per day of ORD airport to LGA airport during 24<sup>th</sup> is not peak. For OO carrier, the peak of flight from ORD to LGA is on 31<sup>st</sup>. For other carriers, during 24<sup>th</sup> the flight is peak, but the other days of the month are also the same number of flights. Hence, it cannot be concluded that the flight from ORD to destination LGA is peak which made the flight delay of 24<sup>th</sup> during the month is the highest. One more interesting finding to double confirm the relationship between flight number and flight delay can also be seen in Figure 3. This is because in Figure 3.2.5, the airport that has the highest number of flight delay does not have WN carrier. WN carrier is the largest carrier that has the highest flight number. Thus, it is concluded that higher flight number does not contribute to higher flight delay. In another perspective, we can look at the airport delay from ORD to LGA during 24<sup>th</sup> to answer the assumption of highest flight delay of 24<sup>th</sup> is caused by highest flight delay from ORD to LGA.



**Figure 3.2.6 Number of flight delay per day for ORD to LGA airport.**

Based on Figure 3.2.6, the flight delay of 24<sup>th</sup> is also not the maximum as on 25<sup>th</sup>, it reached the maximum for flight delays. From this finding, it can be said that the maximum number of flight delay during 24<sup>th</sup> is not caused by the airport that has the highest number of flight delay which is ORD to LGA. To elaborate even further, it does not cause by the number of flight and the flight delay of a specific airport as well. It is the overall flight delay contributed by every airport to achieve the highest amount of flight delay.

Flight delay needed to be minimized as in a research [4], flight delay will negatively be affecting the passengers to the flight carrier. To cover up the cost, flight company would have to raise the fares. Other factors that might affect flight delay can be found in research done by [5], human factor, which is the number of passenger and aircraft size also have relationship to cause flight delay. For this finding, the researchers must explore more into the data set.

### Suggestion

Based on the finding, it is found that the higher number of flights cannot be used as a factor to determine the number of flight delay. It is also found that when the day of number of flight delay reached the maximum, it does not mean that the airport with the highest flight delay is the maximum during the day. Hence, the related audience can skip the part to find the relationship between flight delay and number of flights, or the relationship of an airport contributing to flight delay. The most contributed factor is the departure delay. After the analysis above, it is advised that in the future, the direction to find the reason of flight delay should be finding what caused departure delay. To tackle the flight delay, what causing the departure delay is the key as departure delay contributes the most to arrival delay.

	A	B			
			14	2015	32.8
			15	2016	32.9
			16	2017	33.2
			17	2018	38.4
			18	2019	38.7
			19	2020	28.5
			20	2021	26.6
1	Weather's Percent Share of Total Delay Minutes (%)				
2	2003(Jun-Dec)	49.9			
3	2004	49.7			
4	2005	47.1			
5	2006	44.2			
6	2007	43.6			
7	2008	45.5			
8	2009	44.4			
9	2010	38.1			
10	2011	38.7			
11	2012	33.7			
12	2013	36.5			
13	2014	32.6			

**Figure 3.2.7.** Percentage of extreme weather condition contribute to the total delay in minutes. [6]

Secondly, the flight company should also take in account that since the departure delay will most probably causing flight delay. Thus, the flight company should always depart punctually to prevent any flight delay to be happened. However, this should not always be condemned on the flight company because sometimes flight delay is caused by the extreme weather conditions that made them arrived late instead of late departure. Based on a statistic, it is found that extreme weather conditions contribute a lot to delay as shown in Figure 3.2.7 [6]. Therefore, the flight will arrive late and caused late departure for the next flight.

**Finding 3: Cancelled Flight**

This finding is about the cancelled flight. First and foremost, Table 3.3.1 and Table 3.3.2 are created to visualize the total number of cancelled flights in each carrier. Table 3.3.3 shows that WN has the highest number of cancellation flights (3949) whereas HA has the lowest with a total of 7 cancelled flights among other. In previous findings, WN also has the highest values among others and HA has the lowest values among other carriers. Graph 3.3.5 is also plotted to visualize the highest and lowest number of cancelled flights. Is any relationship between cancelled flights and delays? Table 3.3.4 is created to show the result. As table 3.3.4 shows, canceled flights have no relation to arrival delay. This is because the flights have been canceled and will not arrive at another airport. Besides that, graph 3.3.7 is plotted to show when the highest number of total canceled flights is. On the 30<sup>th</sup>, WN had a total of 628 flights cancelled, among others. Based on the result stated in the table and graph, there is no relationship between flight delays and cancelled flights. This is because, previous flights might be delayed due to some reasons such as air traffic control and staff shortages, flight mechanical issues, ambitious airline schedules, and cause a chain reaction that affected another flight. Since the flight has not arrived, there is no value in the arrival delay.

**Suggestion 3**

The authors emphasized in [7] that the study does not attempt to establish a relationship between delay and cancellation rates and leakage. However, it illustrates that actual flight data might support customers' worries about reliability. While fares, flights, and travel time have all been shown to be important, flying into a hub airport on a nonstop flight may also be a good way to avoid issues in the case of delays or cancellations.

OP_UNIQUE_CARRIER <chr>	Total_Cancelled <int>
9E	341
AA	1511
AS	429
B6	980
DL	328
EV	964
F9	151
G4	50
HA	7
MQ	1939

*Table 3.3.1 Total Cancelled Flights for each Carrier*



OP_UNIQUE_CARRIER <chr>	Total_Cancelled <int>
NK	199
OH	751
OO	2821
UA	697
WN	3949
YV	477
YX	1132

Table 3.3.2 Total Cancelled Flights for each Carrier

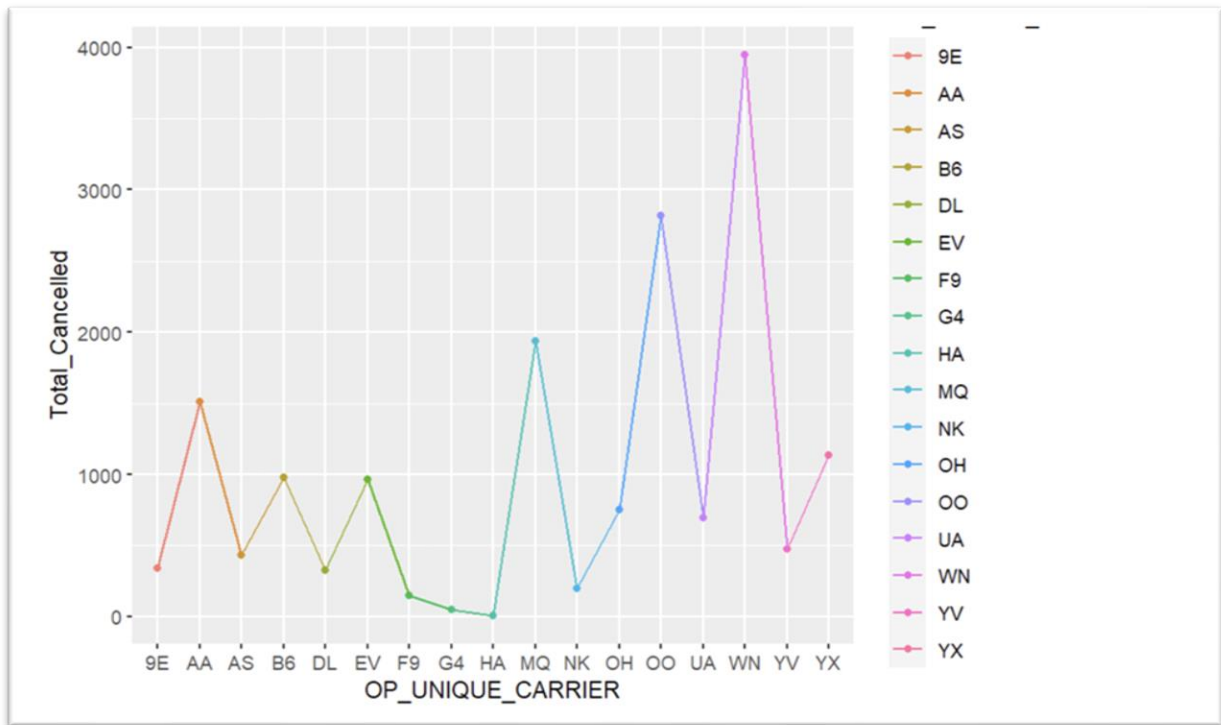
OP_UNIQUE_CARRIER <chr>	CANCELLED <int>
WN	3949
1 row	

Table 3.3.3 Highest Cancelled Flight

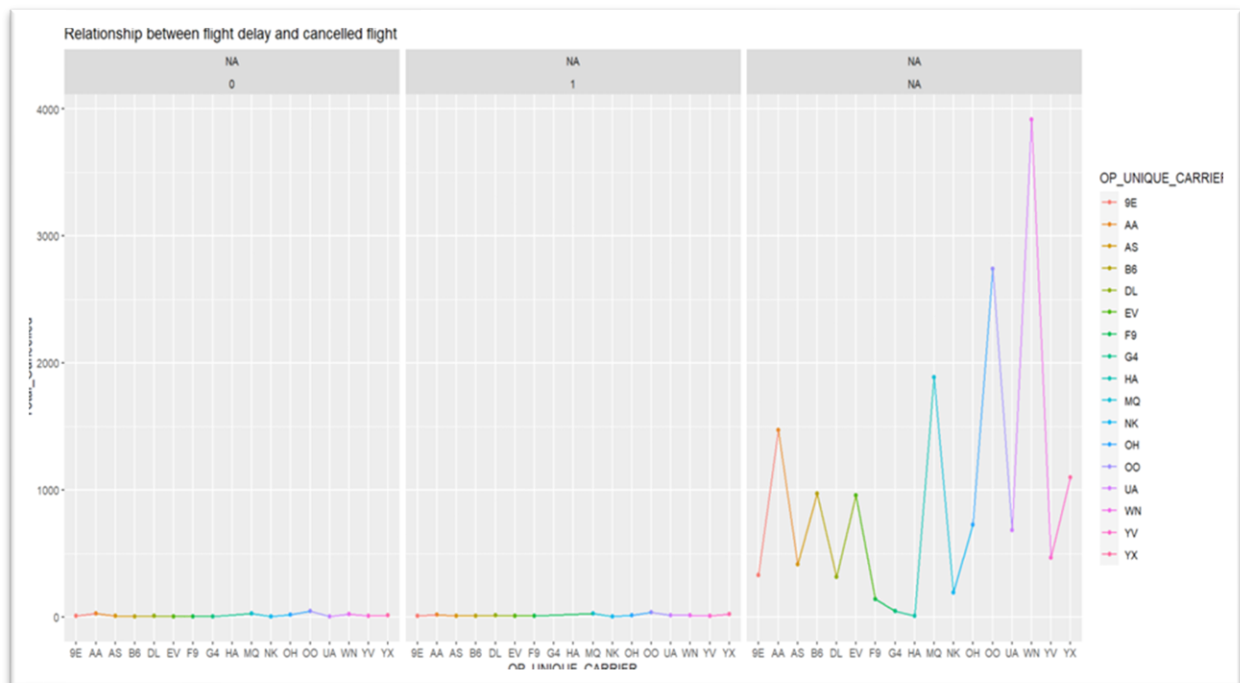
OP_UNIQUE_CARRIER <chr>	ARR_DEL15 <int>	DEP_DEL15 <int>	Total_Cancelled <int>
9E	NA	0	8
9E	NA	1	7
9E	NA	NA	326
AA	NA	0	24
AA	NA	1	16
AA	NA	NA	1471
AS	NA	0	6
AS	NA	1	8
AS	NA	NA	415
B6	NA	0	2

Table 3.3.4 Total Cancelled Flights with Arrival Delay &amp; Departure Delay

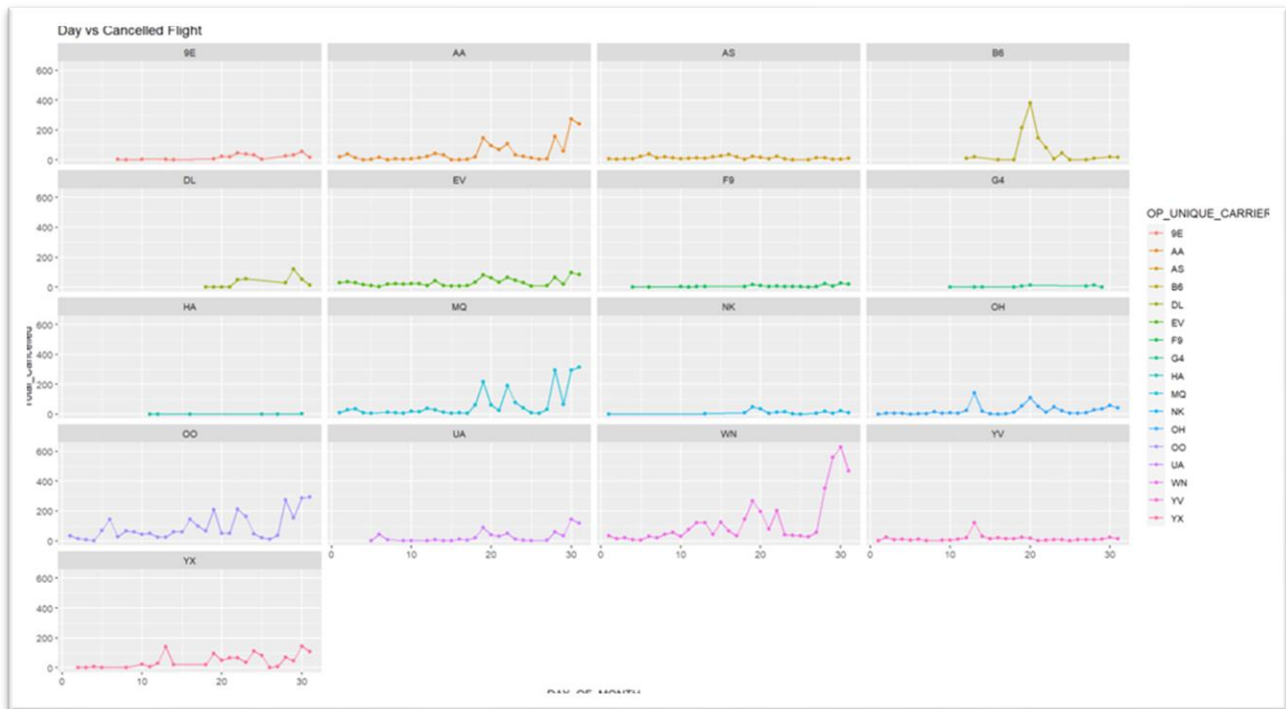




Graph 3.3.5 Total Cancelled flight of each Carrier



Graph 3.3.6 Total Cancelled Flights with Arrival Delay & Departure Delay



Graph 3.3.7: Relation between day of month with cancelled flight

## References

- [1] B. Cohen, "Sweet spots for redeeming United MileagePlus miles," USA TODAY, 22 April 2016. [Online]. Available: <https://www.usatoday.com/story/travel/roadwarriorvoices/2016/04/22/united-airlines-mileage-plus-points/83365166/>. [Accessed 15 August 2022].
- [2] UNITED, "United Airlines Offers Bonus-Miles Promotion for Premium-Cabin Travelers," UNITED, 9 January 2015. [Online]. Available: <https://united.mediaroom.com/2015-01-09-United-Airlines-Offers-Bonus-Miles-Promotion-for-Premium-Cabin-Travelers>. [Accessed 15 August 2022].
- [3] M. Guven, B. Cetinguc, E. Calik and B. Guloglu, "Kybernetes," *Assessing the effects of flight delays, distance, number of passengers and seasonality on revenue*, vol. 48, no. 9, p. 13, 2019.
- [4] R. Britto, M. Dresner and A. Voltes, "Transportation Research Part E," *The impact of flight delays on passenger demand and societal welfare*, vol. 48, no. 2, pp. 460-469, 2011.
- [5] F. J. Wang, J. Bi, D. F. Xie and X. M. Zhao, "IET Intelligent Transport Systems," *Flight delay forecasting and analysis of direct and indirect factors*, vol. 16, no. 7, pp. 890-907, 2022.
- [6] United States Department of Transportation, "Bureau of Transportation Statistic," 10 January 2022. [Online]. Available: <https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations>.
- [7] M. J. Stone, "Tourism and Hospitality Research," *Impact of delays and cancellations on travel from small community airports*, vol. 18, no. 2, pp. 214-228, 2018.