ST2195 Programming for Data Science
Academic Year 2021-22

Student Number: 190399205

Table of Contents

Gathering Data

We will be exploring a domestic flights dataset that contains information from the year 1987 but here we will be focusing on the years 2003-2004. The following are the *CSV* files used as dataset to answer the questions which we will be addressing later:

- airports
- carriers
- plane-data
- 2003
- 2004

We will be providing both explanations and results based on R and Python programming languages.

To begin, we will be connecting to an SQLite database as we will retrieve certain data directly from the tables in our database. Upon connecting to SQLite, we create a database called 'airline' where the aforementioned CSV files are loaded into it.

CSV file	**Table in 'airline' database**
airports	airports
carriers	carriers
plane-data	planes
2003 2004	yearsdf

Data Preparation

We assume the delays mentioned in the questions are referring to Arrival Delays as it would be more significant to travellers hence, we start by removing rows with any empty fields or blank spaces values in the columns ArrDelay and DepDelay from the 'yearsdf' table. Afterwards, we create three copies of the 'yearsdf' table, renaming it to: 'yearsdf3', 'yearsdf4', and 'Q5flights', to be used for question 3, 4, and 5 respectively. Any subsequent data wrangling will be mentioned before we answer each question.

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

We approach this question by dividing it into three parts:

I. When is the *best time of day* to fly to minimise delays?
II. When is the *best day of the week* to fly to minimise delays?
III. When is the *best time of year* to fly to minimise delays?

We will be using the 'yearsdf' table, further removing rows where there are Cancelled or Diverted flights as we assume that these do not contribute to flight arrival delays. In R, we convert DepTime's data type from an integer to datetime. In Python, DepTime's data type remains as an integer. Bar charts will be used to represent the data which will assist in answering this question.

When is the best time of day to fly to minimise delays?

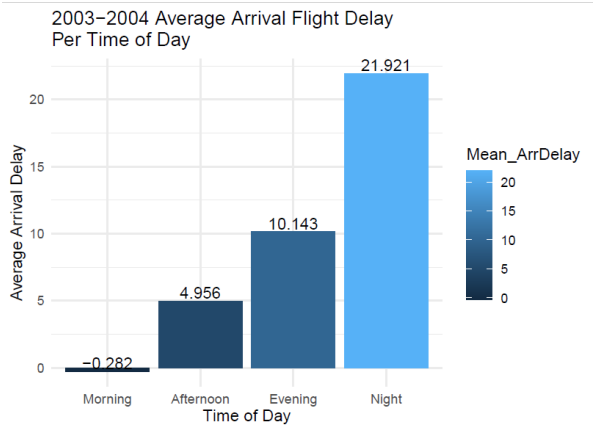
A data frame 'Q1flights' is created and will return a query that includes Year, Month, DayofMonth, DayofWeek, DepTime, ArrDelay and DepDelay from the 'yearsdf' table. Then, we check and handle any missing values.

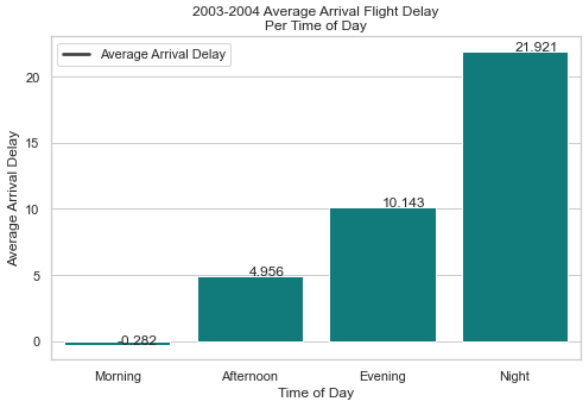
We group DepTime into irregular intervals and separate them into 4 categories:

- Morning (6 hours, between 05:00 and 11:59)
- Afternoon (4 hours, between 12:00 and 16:59)
- Evening (3 hours, between 17:00 and 20:59)
- Night (7 hours, between 21:00 and 04:59)

This data will be stored in a new column called 'Time_of_Day'/'category'. 'Q1flights' table is then loaded into our 'airline' database for easy querying.

A data frame 'timeofday' is created by returning a query that involves selecting the columns category/Time_of_Day and the averaged ArrDelay renamed as 'Mean_ArrDelay' from 'Q1flights' table, grouped by category/Time_of_Day. In R, we convert Time_of_Day's data type into ordered factors.

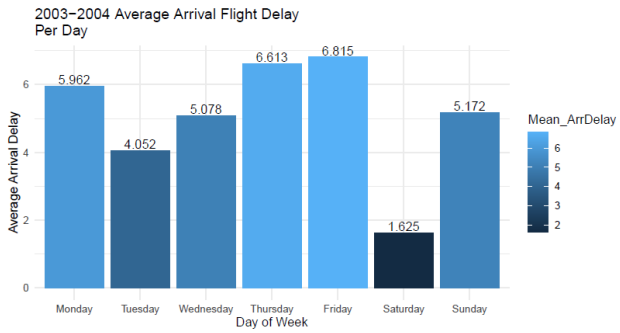


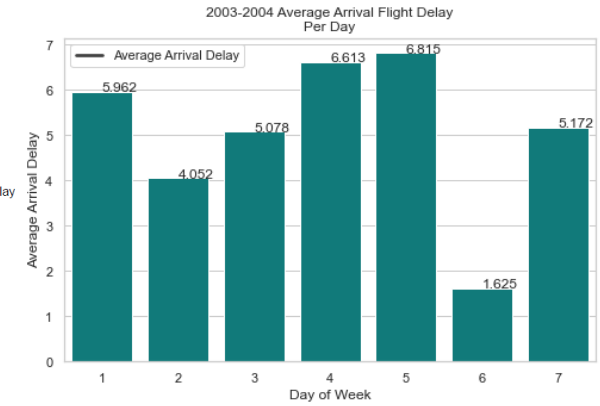


We can observe that Morning has the lowest average Arrival Delay of -0.28176 minutes. A negative delay would mean that the flight arrives earlier than its scheduled time. Meanwhile, Night has the highest average Arrival Delay of 21.92110 minutes. Hence, Morning is the best time of day to fly to minimise delays. This is not surprising as research shows it is more likely that early flights will stay on schedule, compared to nighttime as it gets busier throughout the day.

When is the best day of the week to fly to minimise delays?

Data frames ‘Q1.2’ and ’dayofweek’ are created by returning a query that involves selecting the columns DayofWeek and the averaged ArrDelay renamed as ‘Mean_ArrDelay’. The integer values in DayofWeek are recoded and renamed to each day of the week instead.

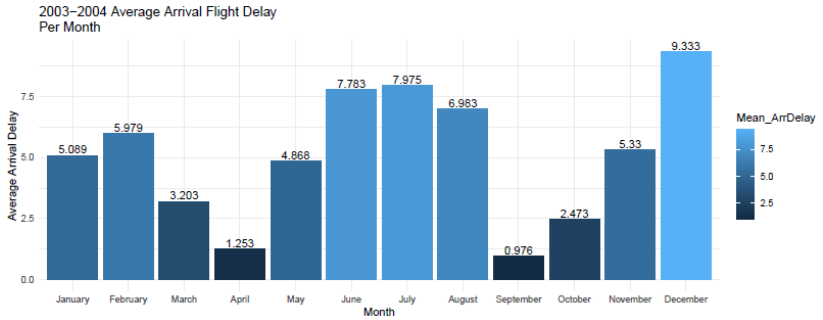


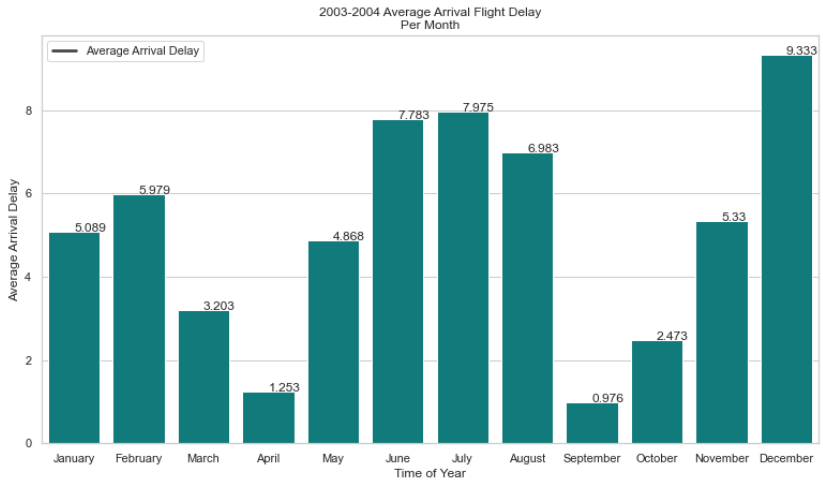


We can observe that Saturday has the lowest average arrival delay of 1.62479 minutes. Meanwhile, Friday has the highest average arrival delay of 6.81455. Hence, Saturday is the best day of week to fly to minimise delays. Business travellers tend to fly back home on a Friday which leads to heavier customer traffic and hence more flights and eventually higher flight delays. Meanwhile, people on vacation tend to return on Sunday instead of Saturday therefore making Saturday the optimal day to fly for less delay.

When is the best time of year to fly to minimise delays?

Data frames ‘Q1.3’ and ’timeofyear’ are created by returning a query that involves selecting the columns Month and the averaged ArrDelay renamed as ‘Mean_ArrDelay’. The integer values in Month are recoded and renamed to each month of a year. Here, the average arrival day is rounded up to 2 decimal places.





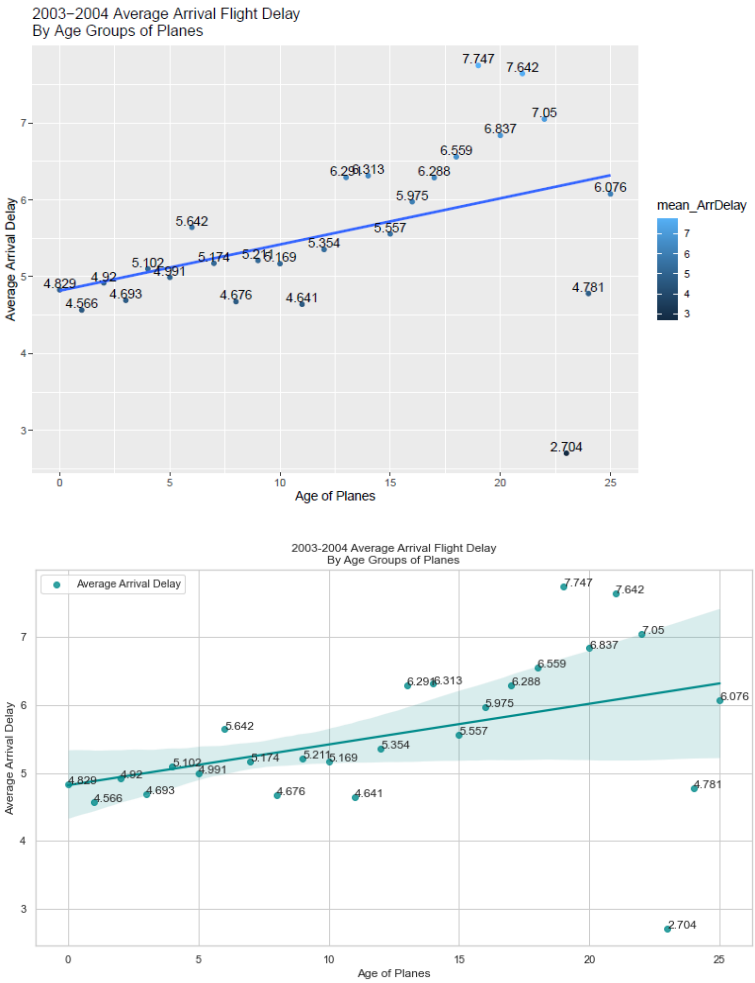
We can observe that September has the lowest average arrival delay of 0.98 minutes while December has the highest average arrival delay of 9.33 minutes. Hence, September is the best time of year to fly to minimise delays. This is foreseeable as people often fly home for the holidays such as Christmas and New Year’s. In conclusion, from analysing this data set, we assume that the overall best time to travel to minimise arrival delays would be on a Saturday morning in September.

Question 2: Do older planes suffer more delays?

Data frame ‘planes_age’ is created by joining tables ‘yearsdf’ and ‘planes’ with TailNum as its matching value and returning a query that involves selecting the columns Year, ArrDelay, and TailNum from ‘yearsdf’ table and Year from ‘planes’ table. Year from ‘yearsdf’ is renamed as flight_year while Year from ‘planes’ is renamed as plane_year.

We noticed that there are a few odd values in the plane_year column that does not contribute to the analysis such as blank spaces, 0000, 2007, and ‘None’. These data are removed. The data type for plane_year is converted from string to integer and a new ‘Age’ column is created in ‘planes_age’ data frame. The values contain the difference between flight year and plane year, representing the age of

the plane. For this analysis, we cap the age limit of the plane at 25. A scatter plot is generated to indicate the relationship between age of planes and average arrival delays.



We observe that there is a strong, positive linear association between the variables age of planes and average arrival delay. There are however some outliers present where if removed, the plot could provide us with a more accurate analysis. However, it is safe to assume that older planes indeed suffer more delays.

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

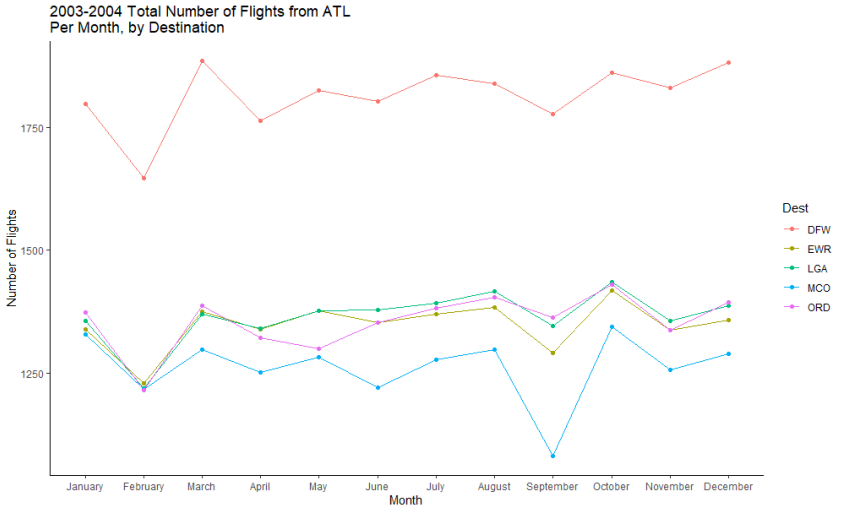
We will be using the 'yearsdf3' table, removing rows where flights are cancelled as we assume that this does not affect the number of people flying. We will also use the number of flights to represent the number of people flying. As the data set is huge, we will only concentrate on flights from a single origin airport to other destinations. We start with finding which origin airport has the highest number of flights. This turns out to be ATL (Atlanta), with a total of 77,3319 flights. Next, we find out which destination airport has the highest number of flights, where its origin is ATL. Then, we plot a line graph

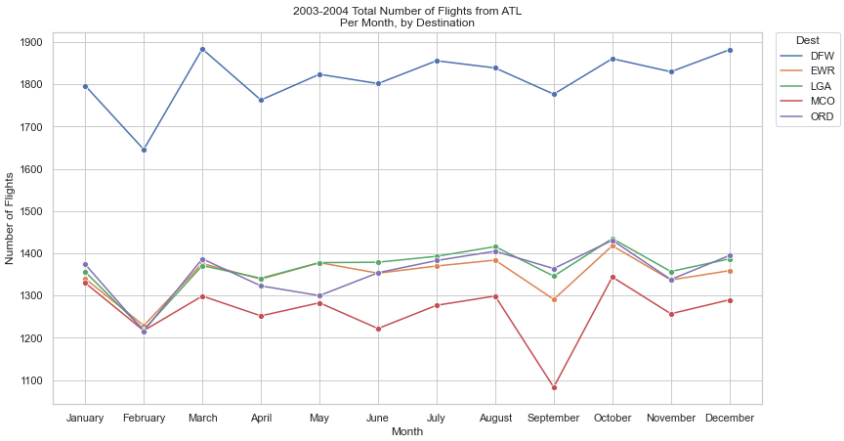
152	15153	MCO
153	16174	EWR
154	16270	ORD
155	16378	LGA
156	21761	DFW

	TotalFlights	Dest
151	15153	MCO
152	16174	EWR
153	16270	ORD
154	16378	LGA
155	21761	DFW

to show the changes of average number of flights between different locations over a time span of a year.

We will use the top 5 destination airports with the highest number of flights, and these are: DFW (Dallas/Fort Worth), LGA (Queens), ORD (Chicago), EWR (Newark), and MCO (Orlando).





These plots show us the total average number of flights from ATL to different destinations over a year. We observe that flights from ATL to DFW are the highest compared to the rest of the destinations, with March having the highest average number of flights of 1,884. DFW is one of the top 10 busiest airport in the world by passenger traffic and is American Airlines' largest hub hence it is a possible reason why there are a lot of flights flying into DFW from ATL. The rest of the destination airports have similar average number of flights across the year however flights to MCO is slightly lower than others, and we can see that the number of flights took a plunge in September, with a total of 1,083 flights. This could be due to Hurricane Charley which impacted Florida that lasted over 6 weeks.

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

We will be using the 'yearsdf4' table, removing rows where flights are cancelled as we assume that this does not affect the number of people flying. Initially, data frame 'delays' is created by returning a query that includes Year, Month, DayofMonth, DepTime, CRSDepTime, ArrTime, CRSArrtime, FlightNum, TailNum, ArrDelay, DepDelay, Origin, Dest, LateAircraftDelay from 'yearsdf4' table, on a

condition that ArrDelay and LateAircraftDelay are greater than 0. Data frame ‘latedep’ is created by further filtering out data and keeping values where DepDelay is greater than 0.

	Dest	n
1	ORD	53942
2	ATL	38915
3	DFW	33815
4	LAS	25617
5	LAX	24027
6	PHX	22171

ORD	53942
ATL	38915
DFW	33815
LAS	25617
LAX	24027
	...
LWB	14
IYK	11
GUC	11

We approach this question by finding out which destination has the highest number of flights, which is ‘ORD’. Then, we create two new data frames and retrieve values where Destination and Origin is ‘ORD’ respectively and merge these two data frames together. As the data set is quite huge, we narrow down this observation by showing data from the 12th of December 2003.

DepTime	CRSDepTime	ArrTime	CRSArrTime	FlightNum	TailNum	ArrDelay	DepDelay	Origin	Dest
1501	1437	1801	1723	746	N11641	38	24	IAH	ORD
1843	1800	2142	2124	1174	N11641	18	43	ORD	EWR
1555	1355	1811	1607	2399	N12552	124	120	ORD	CLE
1555	1547	1922	1830	246	N17309	52	8	IAH	ORD
2001	1915	2238	2154	1647	N17309	44	46	ORD	IAH
2156	2020	2207	2045	2227	N17928	82	96	CLE	ORD

Index	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	FlightNum	TailNum	ArrDelay	DepDelay	Origin	Dest	LateAircraftDelay
33171	2003	6	12	1501	1437	1801	1723	746	N11641	38	24	IAH	ORD	20
33523	2003	6	12	1843	1800	2142	2124	1174	N11641	18	43	ORD	EWR	15
16256	2003	6	12	1555	1355	1811	1607	2399	N12552	124	120	ORD	CLE	115
33159	2003	6	12	1555	1547	1922	1830	246	N17309	52	8	IAH	ORD	8
33540	2003	6	12	2001	1915	2238	2154	1647	N17309	44	46	ORD	IAH	44
15086	2003	6	12	2156	2020	2207	2045	2227	N17928	82	96	CLE	ORD	81

We observe that TailNum ‘N11641’ was scheduled to depart at 1437 (2:37PM) but it departed at 1501 (3:01PM) instead, resulting in 24 minutes of departure delay. It flew from IAH (George Bush Intercontinental Airport) to ORD and was scheduled to arrive at 1723 (5:23PM) but arrived at 1801 (6:01PM) instead, resulting in 38 minutes of arrival delay. Observing the same tail number, it was scheduled to depart from ORD at 1800 (6:00PM) but departed at 1843 (6:43PM) instead, resulting in 43 minutes of departure delay. It was scheduled to arrive in EWR at 2124 (9:24PM) but arrived at 2142 (9:42PM) instead, resulting in 18 minutes of arrival delay.

DepTime	CRSDepTime	ArrTime	CRSArrTime	FlightNum	TailNum	ArrDelay	DepDelay	Origin	Dest
1200	1137	1312	1225	4120	N623MQ	47	23	MSN	ORD
1709	1638	1803	1743	4371	N623MQ	20	31	MDT	ORD
1339	1328	1633	1608	4070	N623MQ	25	11	ORD	MDT
1846	1822	1942	1923	4209	N623MQ	19	24	ORD	IND
1232	1159	1237	1219	4316	N624MQ	18	33	CLE	ORD
1251	1242	1359	1258	4190	N626MQ	61	9	CMH	ORD

Index	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	FlightNum	TailNum	ArrDelay	DepDelay	Origin	Dest	LateAircraftDelay
25583	2003	6	12	1200	1137	1312	1225	4120	N623MQ	47	23	MSN	ORD	7
26095	2003	6	12	1709	1638	1803	1743	4371	N623MQ	20	31	MDT	ORD	16
25496	2003	6	12	1339	1328	1633	1608	4070	N623MQ	25	11	ORD	MDT	11
25749	2003	6	12	1846	1822	1942	1923	4209	N623MQ	19	24	ORD	IND	9
25982	2003	6	12	1232	1159	1237	1219	4316	N624MQ	18	33	CLE	ORD	2
25703	2003	6	12	1251	1242	1359	1258	4190	N626MQ	61	9	CMH	ORD	9

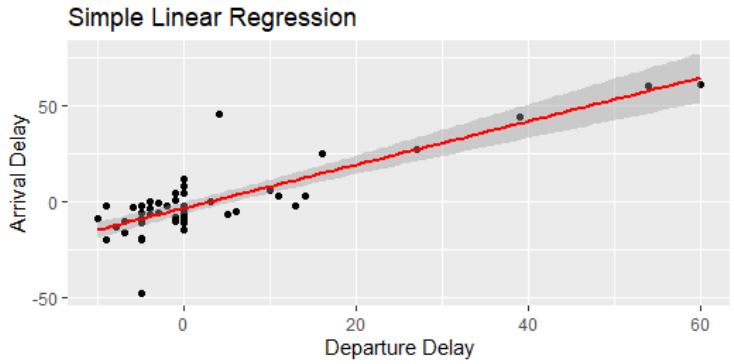
From this extracted data, we observe that TailNum ‘N623MQ’ was scheduled to depart MSN (Dane County Regional Airport) at 1137 (11:37AM) but departed at 1200 (12:00PM) instead, resulting in 23 minutes of departure delay. It was scheduled to arrive in ORD at 1225 (12:25PM) but arrived at 1312 (1:12PM) instead, resulting in 47 minutes of arrival delay. It was then scheduled to depart from ORD at 1328 (1:28PM) but departed at 1339 (1:39PM) instead, resulting in 11 minutes of departure delay and was scheduled to arrive in MDT (Harrisburg International Airport) at 1608 (4:08PM) but arrived at 1633 (4:33PM) instead, resulting in 25 minutes of arrival delay. From MDT to ORD, it had a departure and arrival delay of 31 and 20 minutes respectively. Finally, from ORD to IND, it had a departure and arrival delay of 24 and 19 minutes respectively.

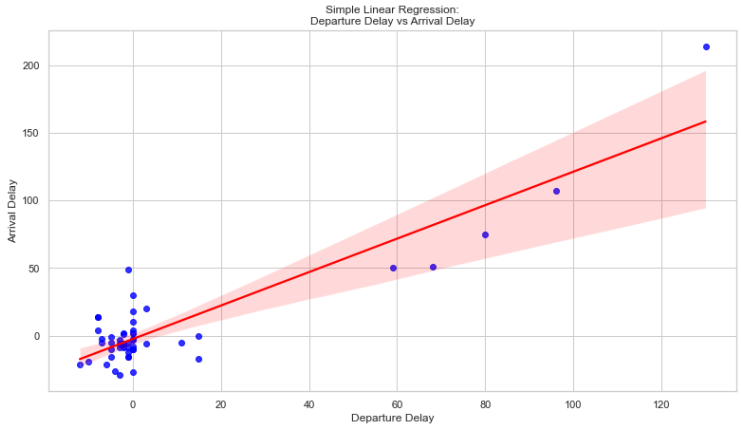
A late departure in one airport does not necessarily result in a longer departure delay in another but it is safe to assume that we can detect cascading failures as delays in one airport create delays in others.

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

Data frame ‘Q5flights’ is created by returning a query that includes ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay from Q5flights’ table, on a condition that flights are not cancelled and diverted as we assume that cancelled and diverted flights will affect the predictability of flight delays. We remove missing data in the data frame, apply the function set.seed to reproduce a particular sequence of 'random' numbers and extract a random sample of 50 to create a simple linear regression model.

We fit a linear regression model where the independent variable is departure delay and the dependent variable is arrival delay and plot a scatter plot that includes a best fitting line.





term	estimate	std.error	statistic	p.value
(Intercept)	-3.412078	1.5783157	-2.161848	0.0356457
DepDelay	1.126557	0.1094207	10.295648	0.0000000

There is a strong positive linear association between the Arrival and Departure delay variables. The equation of the line is defined as $\hat{y} = b_0 + b_1 x$ where b_0 and b_1 represents the y-intercept and slope coefficient respectively. In R, our output for our equation of line is $\widehat{Arr\_Delay} = -3.412 + 1.127 Dep\_Delay$ while it is $\widehat{Arr\_Delay} = -2.464 + 1.238 Dep\_Delay$ in Python.

When a plane has a 0-minute departure delay, the intercept $b_0 = -3.412$ can be understood as the average associated arrival delay. On average, flights that leave on time arrive 3.412 minutes early (a negative delay). For our slope, $b_1 = 1.127$ means that for every one-minute increase in departure delay, an average increase in arrival delay of 1.127 minutes occurs. In R and Python, this model has an R-Squared of 0.6883 and 0.53492, respectively, which shows that this model might not be the best fit for this data set and that we could possibly include other variables as well for a more accurate predictive model.

References

University of London. (2022). *ST2195 Programming for Data Science*. [online] Available at: https://emfss.elearning.london.ac.uk/course/view.php?id=382 [Accessed: 13 Mar. 2022]

Koh, CH and University of London. (2022). *ST2195 Programming for Data Science*. [online] SIMConnect. Available at: https://simconnect.simge.edu.sg [Accessed: 13 Mar. 2022]

Angelova, M. (2021). *Which flights are least likely to be delayed?* [online] 203 Travel Challenges. Available at: https://www.203challenges.com/which-flights-are-least-likely-to-be-delayed/ [Accessed: 14 Mar. 2022]

Wikipedia. (2022). *List of busiest airports by passenger traffic (2000–2009)*. [online] Available at: https://en.wikipedia.org/wiki/List_of_busiest_airports_by_passenger_traffic_(2000%E2%80%932009) [Accessed: 14 Mar. 2022]

Pearso, J. (2021). *Dallas Decoded: A Look At American Airlines' Largest Hub*. [online] Available at: https://simpleflying.com/dallas-american-airlines-largest-hub/ [Accessed: 14 Mar. 2022]

Feltgen, D. (2019). *4 hurricanes in 6 weeks? It happened to one state in 2004.* [online] Available at: https://www.noaa.gov/stories/4-hurricanes-in-6-weeks-it-happened-to-one-state-in-2004 [Accessed: 14 Mar. 2022]

Edureka! (2019). [online] Available at: https://www.edureka.co/community/51489/what-is-set-seed-in-r [Accessed: 14 Mar. 2022]

Ismay et al. (2017). *An Introduction to Statistical and Data Sciences via R*. [online] Available at: https://bookdown.org/fjmcgrade/ismaykim/6-regression.html [Accessed: 14 Mar. 2022]

Sawant, P. (2021). *Airline on-time performance*. [online] Medium. Available at: https://medium.com/@pranaysawant22/airline-on-time-performance-9520d9f2d72b [Accessed: 14 Mar. 2022]