DESCRIPTIVE STATISTICS USING R ASSIGNMENT REPORT

Krishnan Chathadi_12420068

AMPBA Foundation Term

Data Overview

This report will study the "*nycflights13*" dataset, containing flight details for all departures in New York airports in 2013. The programming language R will be used to summarize and visualize the data.

Since most of the analysis is based around the departure delays, let us start with checking the summary of the *dep_delay* column in the *flights* dataset.

Table 1 Summary of Departure Delay (Values in minutes)

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Number of Missing values
-43.00	-5.00	-2.00	12.64	11.00	1301.00	8255

The above table shows that there are 8255 missing values. For the missing values in the *dep_delay* column, it was also found that the departure time was missing. The missing values will be ignored during computation.

Another important aspect about the departure delay data is the presence of both positive and negative values. Here, a positive value of 'x' indicates that the flight was delayed by 'x' minutes. Negative values mean that the flight departed ahead of time. There is a case for ignoring negative values or converting them to zero so that only the delays are captured. However, the negative values have been retained in the calculations so that there is no unnecessary penalty for early departures.

A box plot of the *dep_delay* column is also shown below:

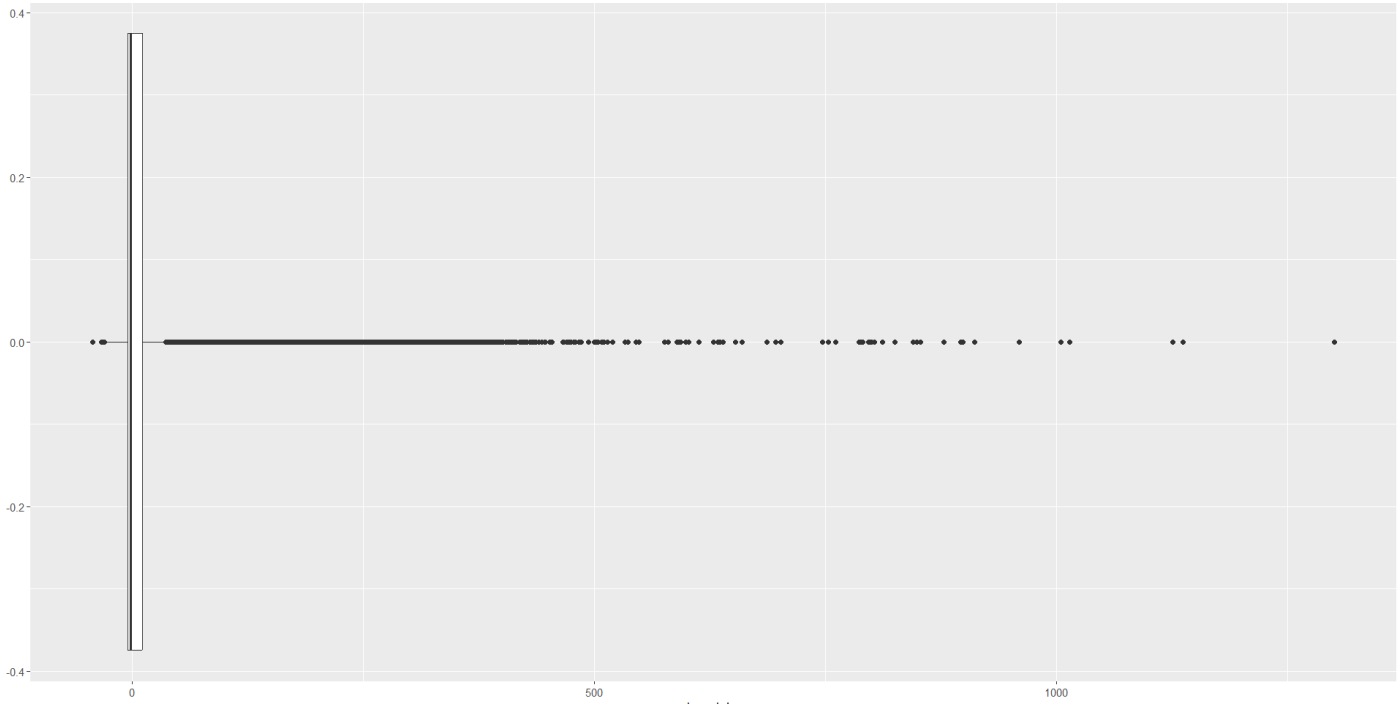

Figure 1 Box plot of departure delay

This indicates the presence of a lot of outliers. It is also hard to visualize the quartile and median from the above plot. To improve the visualization, one of the two methods would be used:

- Removal of the outliers from the plot.
- A log transformation of the departure delay column.

Question 1

To determine which airlines have the most and least delays on average, the mean and median of the *dep_delay* column will be computed, after grouping the data by the *carrier* column.

a. The results of mean and median departure delays are shown below:

carrier	name	Mean_Delay	Median_Delay
US	US Airways Inc.	3.782418	-4.0
HA	Hawaiian Airlines Inc.	4.900585	-4.0
AS	Alaska Airlines Inc.	5.804775	-3.0
AA	American Airlines Inc.	8.586016	-3.0
DL	Delta Air Lines Inc.	9.264505	-2.0
MQ	Envoy Air	10.552041	-3.0
UA	United Air Lines Inc.	12.106073	0.0
OO	SkyWest Airlines Inc.	12.586207	-6.0
VX	Virgin America	12.869421	0.0
B6	JetBlue Airways	13.022522	-1.0
9E	Endeavor Air Inc.	16.725769	-2.0
WN	Southwest Airlines Co.	17.711744	1.0
FL	AirTran Airways Corporation	18.726075	1.0
YV	Mesa Airlines Inc.	18.996330	-2.0
EV	ExpressJet Airlines Inc.	19.955390	-1.0
F9	Frontier Airlines Inc.	20.215543	0.5

Figure 2 Mean and Median Departure Delay by Airline

b. The table shown in the above figure has been sorted in the increasing order of mean. This order will be used as the ranking for the airlines. In terms of departure delay, the **best airlines is US Airways Inc and the worst is Frontier Airlines Inc.**

c. The median is the value around the mid-point of the data. It would give us a picture of how much the delay is 50% of the time. Most of the airlines have a negative median value, which indicates that the flights depart ahead of time in more than 50% of the cases.

   The more useful metric in this context, is the mean since it includes the effect of all the values. When the mean departure delay is large, it indicates the presence of large individual values. Customers in general, would not mind smaller delays of 5-10 minutes, but the longer delays are more impactful.

   Since the mean captures the effect of the longer delays, the **mean departure delay is more useful in this context**.

Question 2

To study the departure delays by airport, the mean is computed for the dep_delay column, after grouping the data by origin (airport).

a. The average departure delay for flights from JFK, LGA and EWR is shown in the figure below.

origin	Num_Flights	Mean_Delay
EWR	120835	15.10795
JFK	111279	12.11216
LGA	104662	10.34688

Figure 3 Mean Departure Delay by Airport

b. Box plot for departure delay for each airport. The outliers have been removed for better visualization.

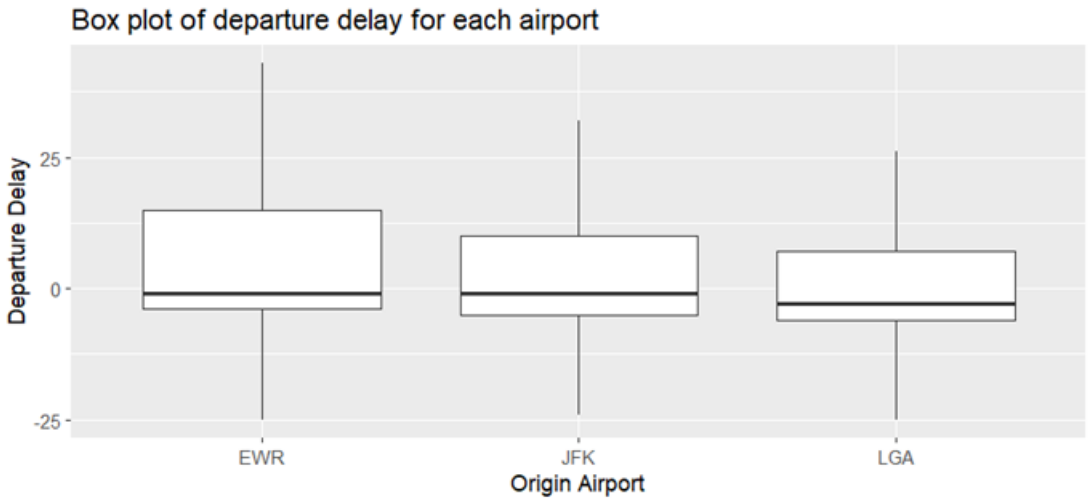

Figure 4 Box plot of departure delay for each airport (no outliers)

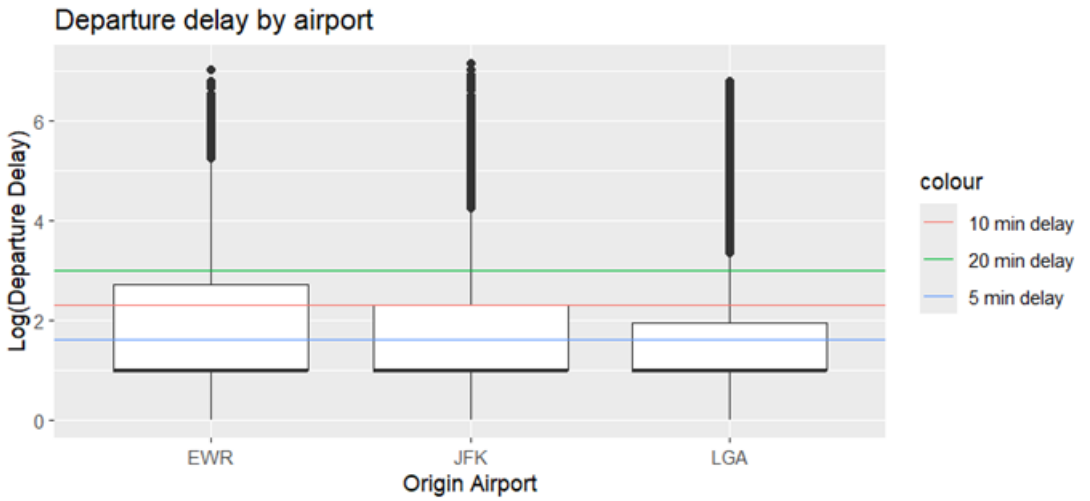

Figure 5 Box plot of log (departure delay) for each airport

c. The data shows that LGA has the lowest mean and lowest median departure delay. Airlines who have their presence in multiple airports could consider shifting a few flights from EWR or JFK to LGA so that they can reduce the departure delays.

Question 3

a. The figure shows the average departure delay by month:

month	Mean_Delay
1	10.036665
2	10.816843
3	13.227076
4	13.938038
5	12.986859
6	20.846332
7	21.727787
8	12.611040
9	6.722476
10	6.243988
11	5.435362
12	16.576688

Figure 6 Mean departure delay by month

b. Line plot of mean departure delay by month:

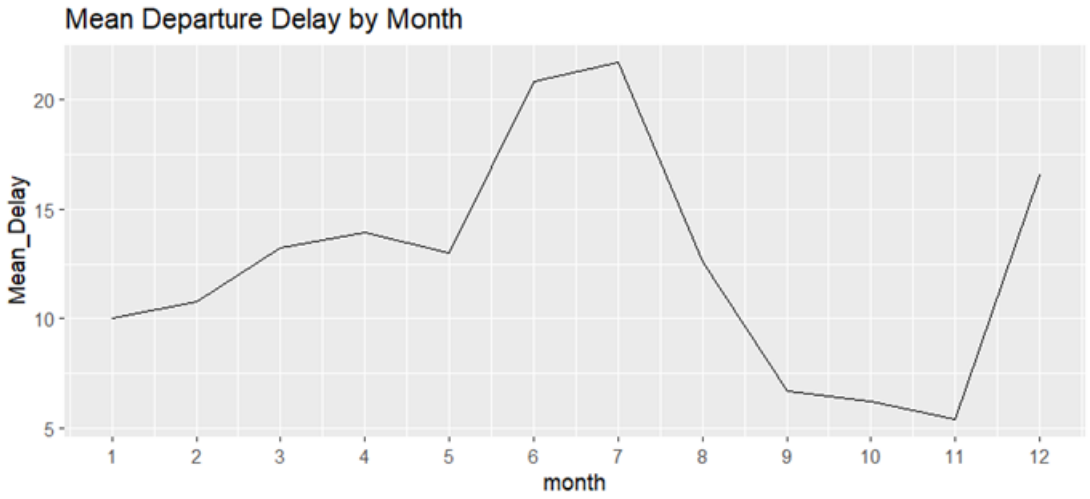

Figure 7 Line plot of mean departure delay by month

c. The average departure delay is higher for the months of June, July and December. These months coincide with the summer break and the winter break in the US. Holidays could be a possible reason for departure delay. Holidays could impact the departure delay in two ways. The first is the increase in the number of passengers flying out of the city. The second reason is the airport staff going on vacations. A deeper analysis with more data is required to confirm these statements.

The impact of weather-based delay can be checked by studying the correlation between departure delay and weather parameters like temperature, wind, precipitation etc. The correlation value between dep_delay and the rest of the columns are very low. Based on this chart, there is no evidence to suggest that departure delay is caused by weather conditions. However, this analysis is not sufficient because the delays could be caused by extreme weather conditions. This study would be done as a future exercise.

	dep_delay	temp	precip	visib	wind_speed	wind_gust	pressure
dep_delay	1.00000000	0.06194966	0.049991278	-0.09601502	0.020367140	0.023778820	-0.08218498
temp	0.06194966	1.00000000	-0.016725625	0.04327649	-0.316431079	-0.339796146	-0.22317726
precip	0.04999128	-0.01672562	1.000000000	-0.47043870	0.003211876	0.004520581	-0.10038595
visib	-0.09601502	0.04327649	-0.470438700	1.00000000	-0.053756950	-0.055634250	0.09434299
wind_speed	0.02036714	-0.31643108	0.003211876	-0.05375695	1.000000000	0.873813201	-0.22621085
wind_gust	0.02377882	-0.33979615	0.004520581	-0.05563425	0.873813201	1.000000000	-0.23948901
pressure	-0.08218498	-0.22317726	-0.100385949	0.09434299	-0.226210855	-0.239489008	1.00000000

Figure 8 Correlation between departure delay and weather parameters

Question 4

a. The flights are grouped by departure hour and the average delay is computed for each.

hour	numFlights	Mean_Dep_Delay
1	1	NaN
5	1953	0.6877572
6	25951	1.6427956
7	22821	1.9140778
8	27242	4.1279478
9	20312	4.5837378
10	16708	6.4982946
11	16033	7.1916503
12	18181	8.6148485
13	19956	11.4376504
14	21706	13.8188742
15	23888	16.8945646
16	23002	18.7570165
17	24426	21.1006059
18	21783	21.1100818
19	21441	24.7847911
20	16739	24.3041048
21	10933	24.1957431
22	2639	18.7910972
23	1061	14.0171756

Figure 9 Mean departure delay at different hours during the day

b. A line plot has also been created.

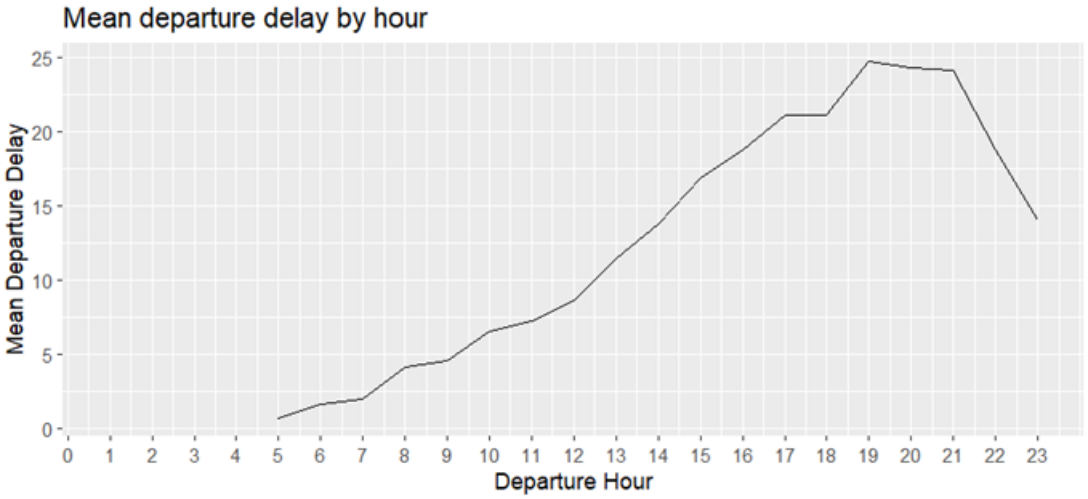

Figure 10 Mean departure delay by hour

c. The line plot of departure hour vs mean departure delay indicates that the departure delays increase throughout the day. Between 5 pm and 9 pm, it is more than 20 minutes on an average.

   If I were an airline, I would schedule more flights in the morning hours to minimize delays. Before rescheduling flights, there are other factors to be considered – the demand for flights is one of them. The table also indicates that the number of departures is much more in the morning hours, compared to evenings. Running empty flights cannot really be profitable!

Question 5

a. Correlation between distance and departure delay is -0.02167

Dist_Delay_Correlation
-0.02167079

Figure 11 Correlation between departure delay and flight distance

b. The scatter plot between distance and departure delay is shown below:

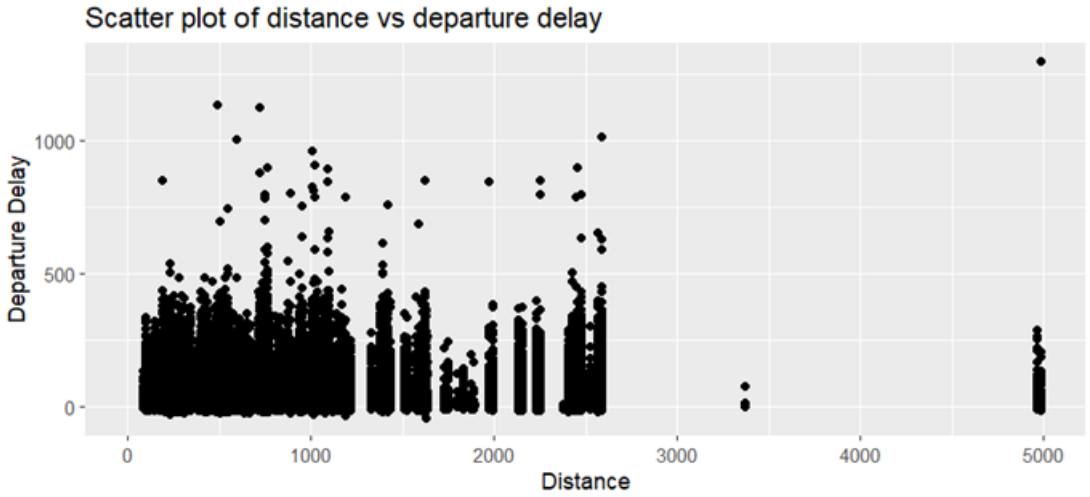

Figure 12 Scatter plot of flight distance vs departure delay

c. In the scatter plot, there is no observable trend which suggests a relationship between departure delay and distance. The correlation value is also low. Thus, there is no data to support the idea that longer flights experience more departure delays. There is no strong correlation between distance and arrival delay as well.

Question 6

The columns of interest here are *air_time* and *distance*. A summary of *air_time* shows the presence of missing values. These values will not be considered in the calculations.

Table 2 Summary of air time

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Number of missing values
20.0	82.0	129.0	150.7	192.0	695.0	9430

a. The ratio of air time to distance for each airline is shown below. It has been reordered based on effective travel time.

carrier	name	Avg_Air_Time	Avg_Distance	Eff_Travel_Time
HA	Hawaiian Airlines Inc.	623.08772	4983.0000	0.1250427
VX	Virgin America	337.00235	2499.4326	0.1348315
AS	Alaska Airlines Inc.	325.61777	2402.0000	0.1355611
UA	United Air Lines Inc.	211.79135	1531.3214	0.1383063
DL	Delta Air Lines Inc.	173.68880	1237.9791	0.1403003
AA	American Airlines Inc.	188.82230	1343.2799	0.1405681
B6	JetBlue Airways	151.17717	1069.6896	0.1413281
F9	Frontier Airlines Inc.	229.59912	1620.0000	0.1417279
WN	Southwest Airlines Co.	147.82481	996.9714	0.1482739
FL	AirTran Airways Corporation	101.14394	664.7874	0.1521448
US	US Airways Inc.	88.57380	560.8259	0.1579346
MQ	Envoy Air	91.18025	570.3746	0.1598603
EV	ExpressJet Airlines Inc.	90.07619	562.8650	0.1600316
9E	Endeavor Air Inc.	86.78160	529.8896	0.1637730
OO	SkyWest Airlines Inc.	83.48276	509.2759	0.1639244
YV	Mesa Airlines Inc.	65.74081	376.4375	0.1746394

Figure 13 Ranking of airlines based on effective travel time

b. The above figure shows the effective travel time for different airlines in ascending order. It can be concluded that Hawaiian Airlines Inc has the shortest time per mile.

The table also shows the average time and the average distance covered on each flight. Carriers which are ranked higher in terms of effective travel time, cover a longer distance on an average. This gives them the opportunity to cruise at higher speeds for longer periods of time. The carriers who focus on short distance flights do not have that luxury.

Question 7

For a business traveller who values punctuality, the most important consideration would be arriving at the destination on time. The arrival delay will be given top priority.

Table 3 Summary of arrival delay

Min.	1st Quartile	Median	Mean	3rd Qu.	Max.	Number of missing values
-86.000	-17.000	-5.000	6.895	14.000	1272.000	9430

The arrival delay will be evaluated based on 2 methods – mean arrival delay and percentage of on-time arrival. To calculate the percentage of on-time arrival, a threshold of 10 minutes will be set on the arrival time. In other words, an arrival delay of 10 minutes or less, will be marked as on-time arrival.

$$\text{Percentage of on-time arrival} = \frac{\text{Number of on-time arrivals}}{\text{Number of arrivals}} * 100$$

carrier	name	num_flights_daily	num_destinations	Mean_Arr_Delay	Percentage_OnTime_Arr
AS	Alaska Airlines Inc.	1.95616438	1	-9.9308886	80.95238
HA	Hawaiian Airlines Inc.	0.93698630	1	-6.9152047	82.74854
AA	American Airlines Inc.	89.66849315	19	0.3642909	75.71573
DL	Delta Air Lines Inc.	131.80821918	40	1.6443409	77.03180
VX	Virgin America	14.14246575	5	1.7644644	77.35374
US	US Airways Inc.	56.26301370	6	2.1295951	74.90261
UA	United Air Lines Inc.	160.72602740	47	3.5580111	72.82196
9E	Endeavor Air Inc.	50.57534247	49	7.3796692	67.25352
B6	JetBlue Airways	149.68493151	42	9.4579733	68.67576
WN	Southwest Airlines Co.	33.63013699	11	9.6491199	69.17312
MQ	Envoy Air	72.32054795	20	10.7747334	64.23836
OO	SkyWest Airlines Inc.	0.08767123	5	11.9310345	68.75000
YV	Mesa Airlines Inc.	1.64657534	3	15.5569853	57.57072
EV	ExpressJet Airlines Inc.	148.41917808	61	15.7964311	60.74244
FL	AirTran Airways Corporation	8.93150685	3	20.1159055	57.73006
F9	Frontier Airlines Inc.	1.87671233	1	21.9207048	56.49635

Figure 14 Airlines with number of daily flights, destination, and arrival delays

From Figure 13, the top two airlines, based on minimum arrival delay and minimum percentage of late arrival, are Alaska Airlines Inc and Hawaiian Airlines Inc. However, these two choices cannot be recommended to a business traveller, unless they are flying to the one destination, that carrier offers. As a result, the top recommendation for a business traveller who values punctuality is American Airlines Inc, who operates around 90 flights daily to 19 destinations. Delta Air Lines Inc, with more than 130 flights to 40 different destinations, is an excellent second choice. Both these airlines have a mean arrival delay of less than 2 minutes and an on-time arrival of more than 75%.

Question 8

The departure delays have been studied by considering a lot of different parameters. The different carriers were first ranked in the order of mean departure delays. Then, the delays were checked based on airports and it was found that LGA has the lowest mean departure delay and EWR has the highest. The monthly and hourly delays were also studied. Based on these initial assessments, a couple of factors could be considered deeper.

Actionable recommendations:

1. Reduce the number of evening flights, if possible.

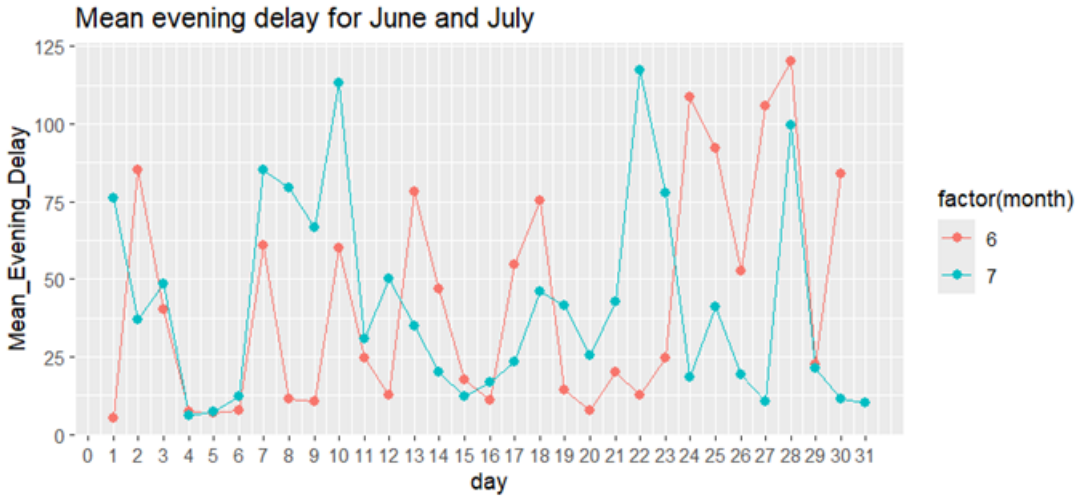

Figure 15 Daily evening delays for the months of June and July

When checking the data pertaining to a time period of 1700 to 2159 hours for the months of June and July, it was found that certain evenings had huge departure delays. In this filtered dataset, there was a moderate correlation of 0.3959 between the number of flights and the mean departure delay. The passenger data needs to be studied before considering the adjustments.

2. Ensure that the airlines, which are responsible for higher mean departure delays, function better.

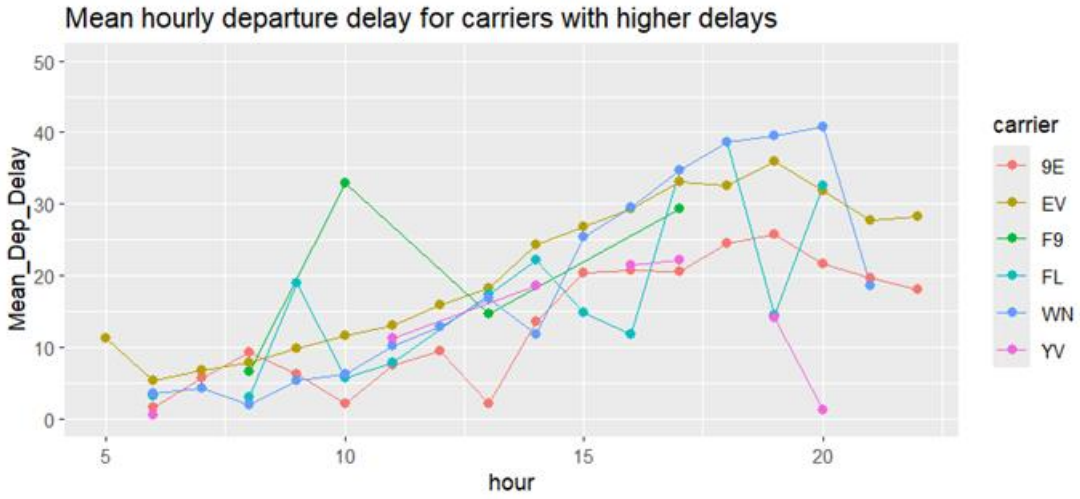

Figure 16 Hourly delay for carriers with higher mean departure delay

To elaborate this point, let us consider the hourly departure delay trends of the bottom 6 carriers from the Question 1 section. For each of these carriers, we could identify the problematic time periods, where the departure delay is much larger than the overall mean departure delay. For example, 9E has a higher-than-average departure delay in the mornings, around 7 and 8 am, but their delays at other times are not significantly worse. WN has very high departure delays from 3 pm to 8 pm. With this insight, the concerned airlines could reschedule flights or allocate more staff at certain times to reduce the delays.