Mini Project 1

Dylan Thomas (dylan_thomas)

April 22, 2024

The observational unit in these datasets would consist of the flights taken in each month. There are 1,267,353 flights total in the datasets. Variables for this set include the year, month, and day of the flight, it includes the airline that ran the flight, it includes the origin and destination airport of the flight. It also has time information, such as the scheduled departure and arrival, the actual departure and arrival times, how much delay there was in those times, and the scheduled and real elapsed time of the flight. Any missing datapoints are stored as NA.

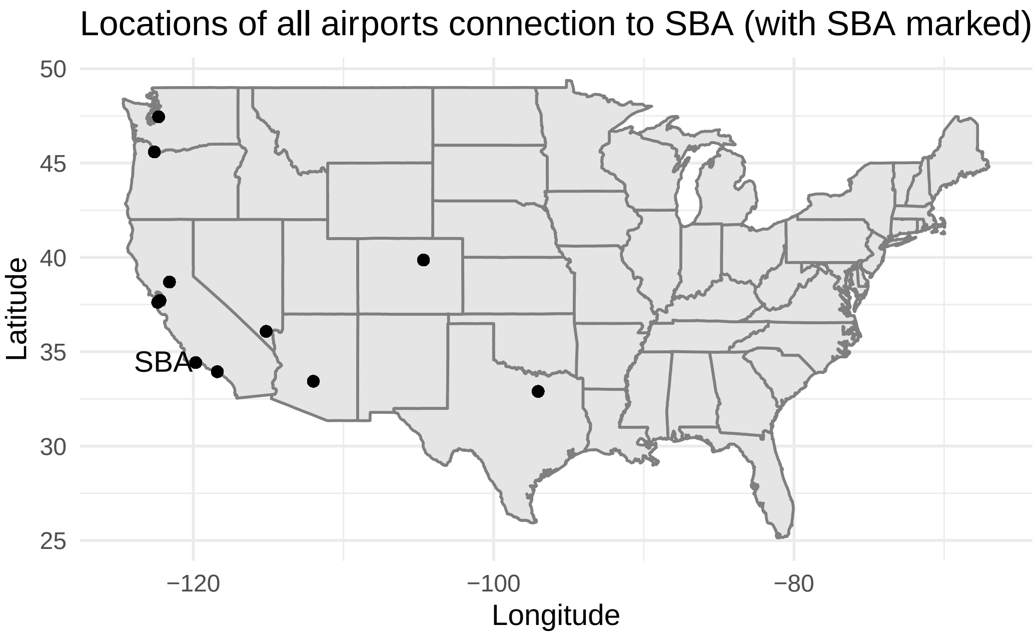
Now that we have this data in a dataset, we can combine it with the airport information and reformat/rename some of the variables.

Now that the data has been organized and combined properly, we can continue on to drawing conclusions and manipulating the data more.

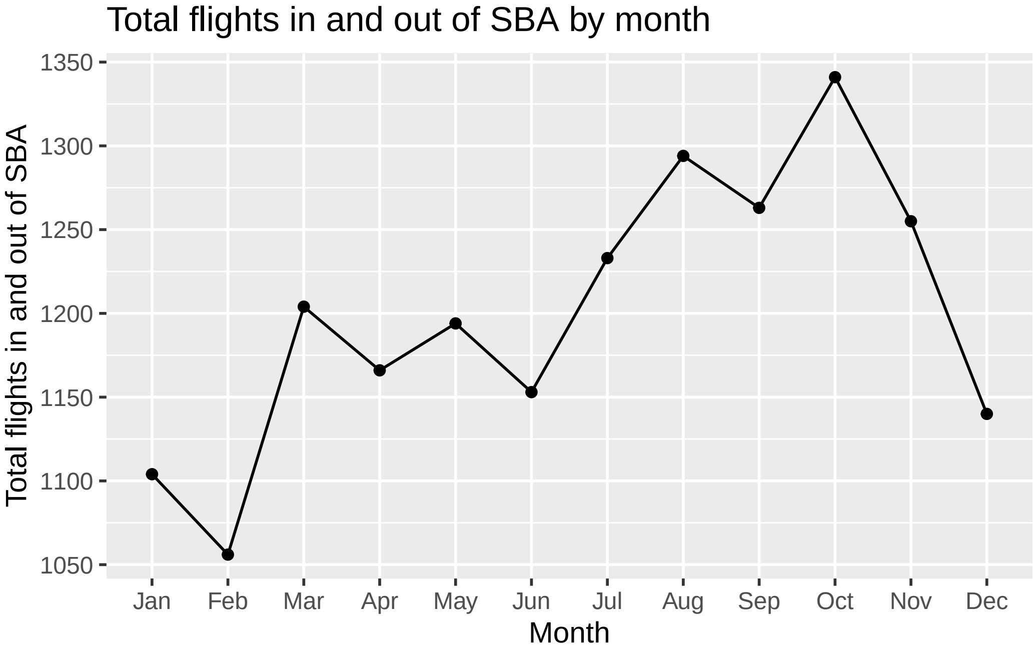
```
 [1] "DALLAS-FORT WORTH INTL"  "PHOENIX SKY HARBOR INTL"
 [3] "SANTA BARBARA MUNI"      "SEATTLE-TACOMA INTL"
 [5] "LOS ANGELES INTL"        "SAN FRANCISCO INTL"
 [7] "DENVER INTL"             "HARRY REID INTL"
 [9] "METRO OAKLAND INTL"      "SACRAMENTO INTL"
[11] "PORTLAND INTL"
```

```
 [1] "SANTA BARBARA MUNI"      "DALLAS-FORT WORTH INTL"
 [3] "PHOENIX SKY HARBOR INTL" "SEATTLE-TACOMA INTL"
 [5] "LOS ANGELES INTL"        "SAN FRANCISCO INTL"
 [7] "DENVER INTL"             "HARRY REID INTL"
 [9] "METRO OAKLAND INTL"      "SACRAMENTO INTL"
[11] "PORTLAND INTL"
```

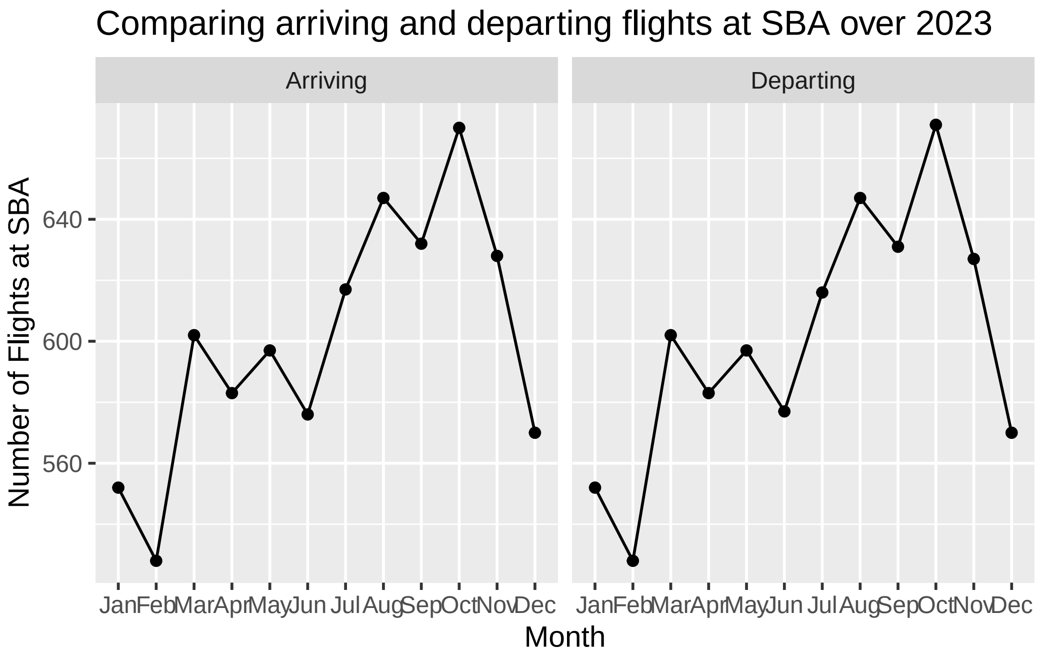
Using the unique function, we can find that the airports that service flights to or from SBA include Dallas-Fort Worth, Phoenix, Seattle-Tacoma, LAX, San Fransisco, Denver, Harry Reid (las vegas), Oakland, Sacramento, and Portland. Now we can use the coordinate data we have to map these locations.


Locations of all airports connection to SBA (with SBA marked)
50
45
40
35
30
25
Latitude
SBA
−120
−100
−80
Longitude

This map shows the location of all airports who have sent or recieved a flight from Santa Barbara in 2023. Santa Barbara Airport is also marked and labelled as SBA. Now we can also use this data to produce a line graph to display how busy (how many flights are coming in or going out) SBA is depending on the month.



This line graph shows us that every month falls between 1050 to 1350 flights, with february being the minumum and october being the maximum. The trend that pops out is that summer and fall months are clearly higher traffic than the winter and spring months. One possible explanation (that could be further explored) for this is increased tourism during those months.

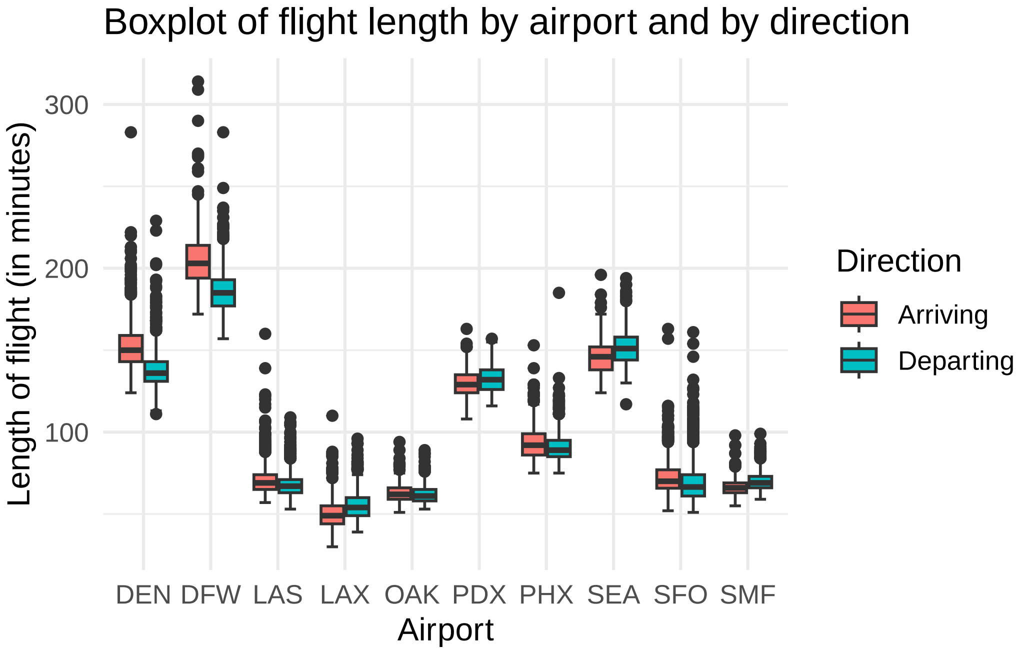


```
# A tibble: 12 x 2
   MONTH count
   <ord> <int>
 1 Jan     552
 2 Feb     528
 3 Mar     602
 4 Apr     583
 5 May     597
 6 Jun     577
 7 Jul     616
 8 Aug     647
 9 Sep     631
10 Oct     671
11 Nov     627
12 Dec     570
```

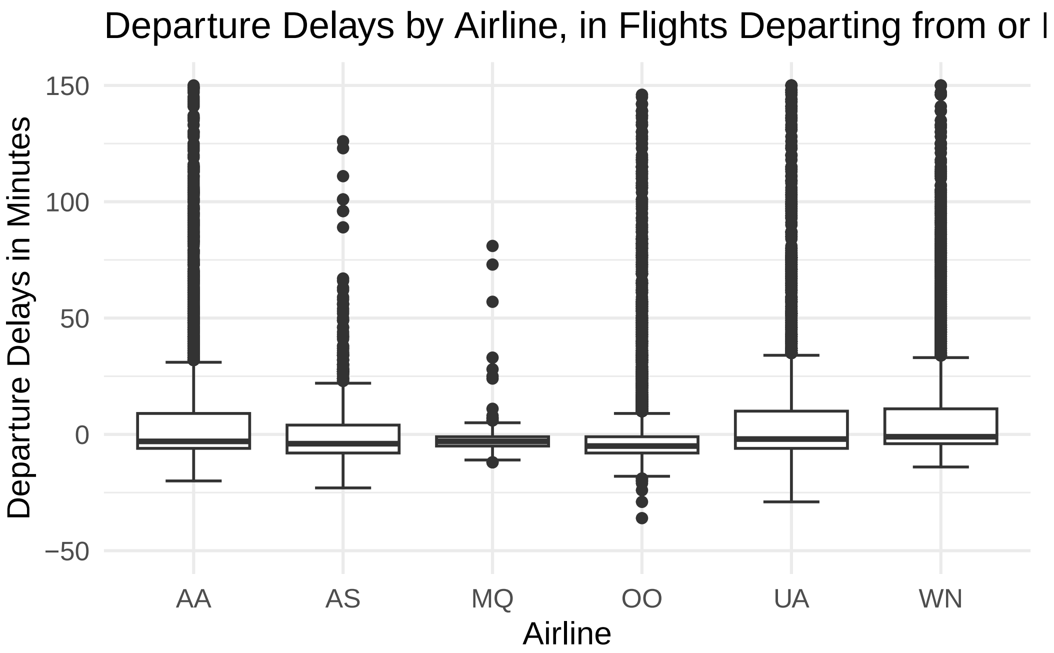
```
# A tibble: 12 x 2
   MONTH count
   <ord> <int>
 1 Jan     552
 2 Feb     528
 3 Mar     602
 4 Apr     583
 5 May     597
 6 Jun     576
 7 Jul     617
 8 Aug     647
 9 Sep     632
10 Oct     670
```

```
11 Nov     628
12 Dec     570
```

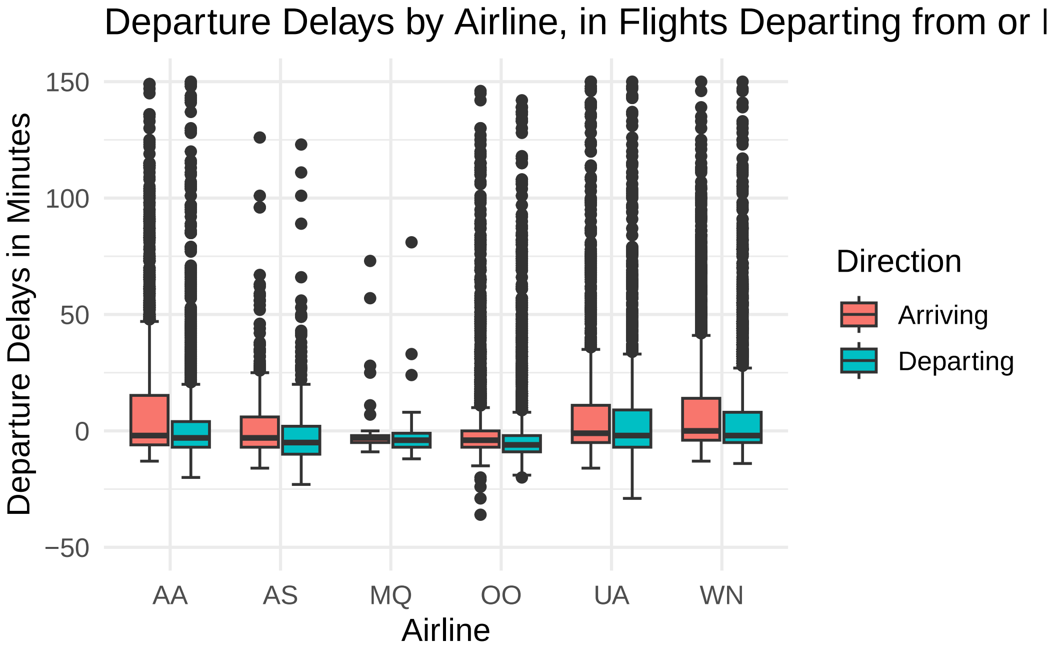
As we can see from this graph and table, the two numbers are generally the same (as would be expected, the plane has to take off if it lands), but some exceptions can be seen in June and October (where there are more arrivals than departures) and in July and November (where there are more departures than arrivals). One possible hypothesis for this phenomenom is that if a plane arrives on the last day of the month, and then doesn't take off until the next day, it would cause the first month to have extra arrivals, and the second month to have extra departures. This tracks with our observation that june/july and october/november are the months with this discrepancy. Overall, there is no noticeable difference between the arriving and departing flights.



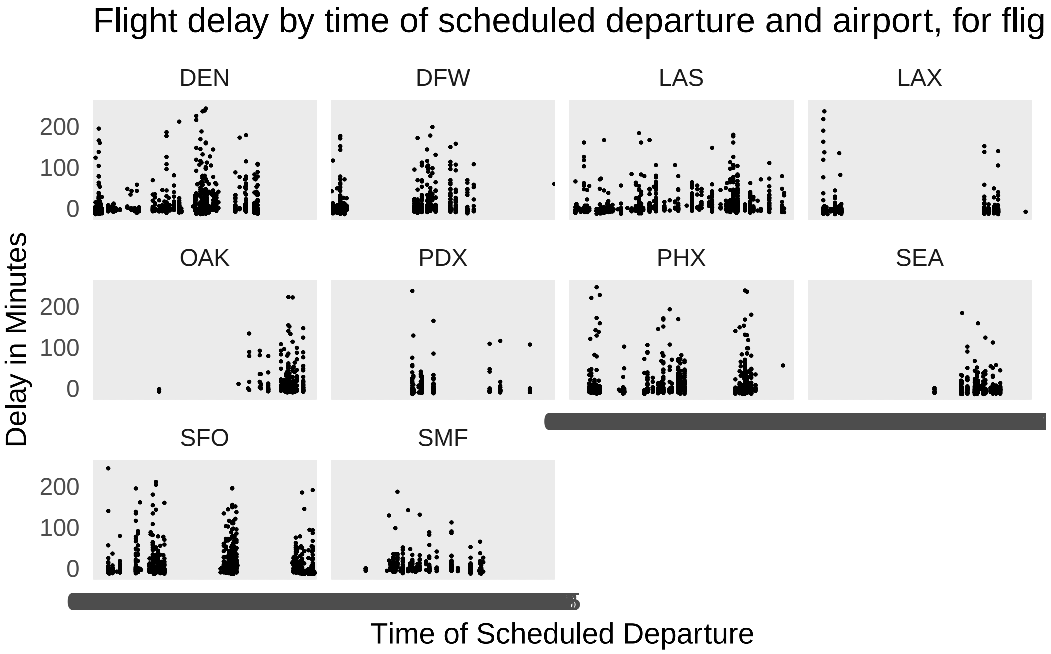
As this boxplot shows, while generally the time of a flight inbound or outbound is pretty similar, there are some noteable airports where there is a visual difference, specifically Denver and Dallas, both of which are longer coming to Santa Barbara compared to leaving. These are also the flights that travel the most distance east/west, so I would be interested to investigate if that has something to do with this difference.



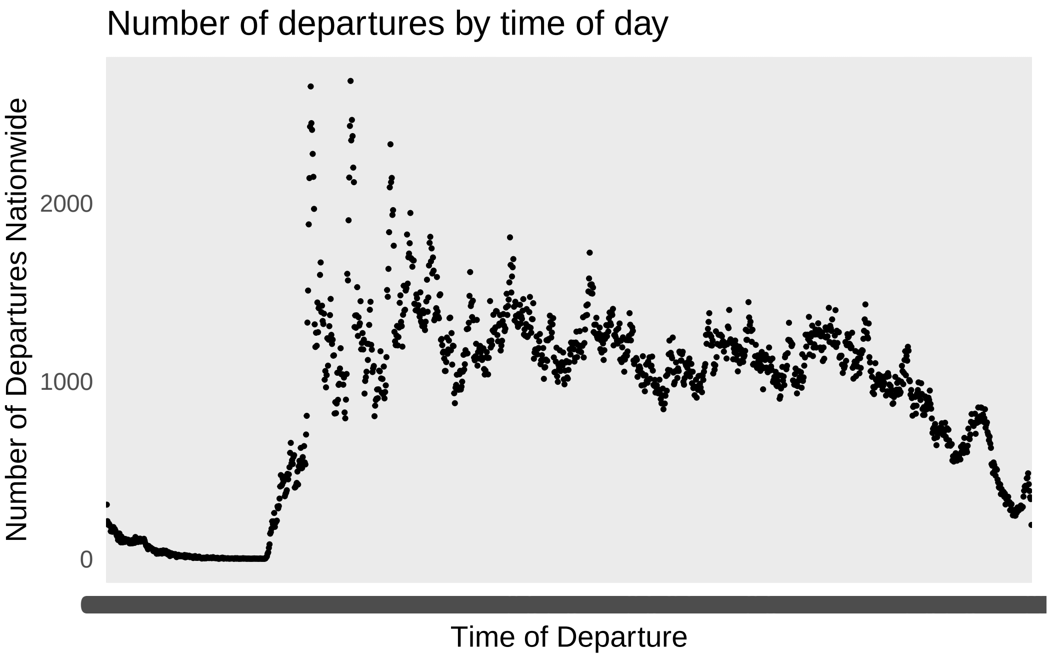
As can be seen in this box and whisker plot, flights generally departed before their scheduled departure time, no matter the airline. We can also see however, that it is very rare for a plane to leave more than 20-30 minutes before scheduled departure time, but the range of delays can go for hours in the other direction. The box plot also tells us that airlines MQ and OO (my research says that those are envoy air and skywest respectively) have the densest interquartile range, meaning that they are usually the closest to the scheduled departure time. AA and UA (American and United Airlines) on the other hand have the largest IQR, meaning their times fall on a bigger range, most likely due to more delays, given that they lie higher on the y-axis as well.



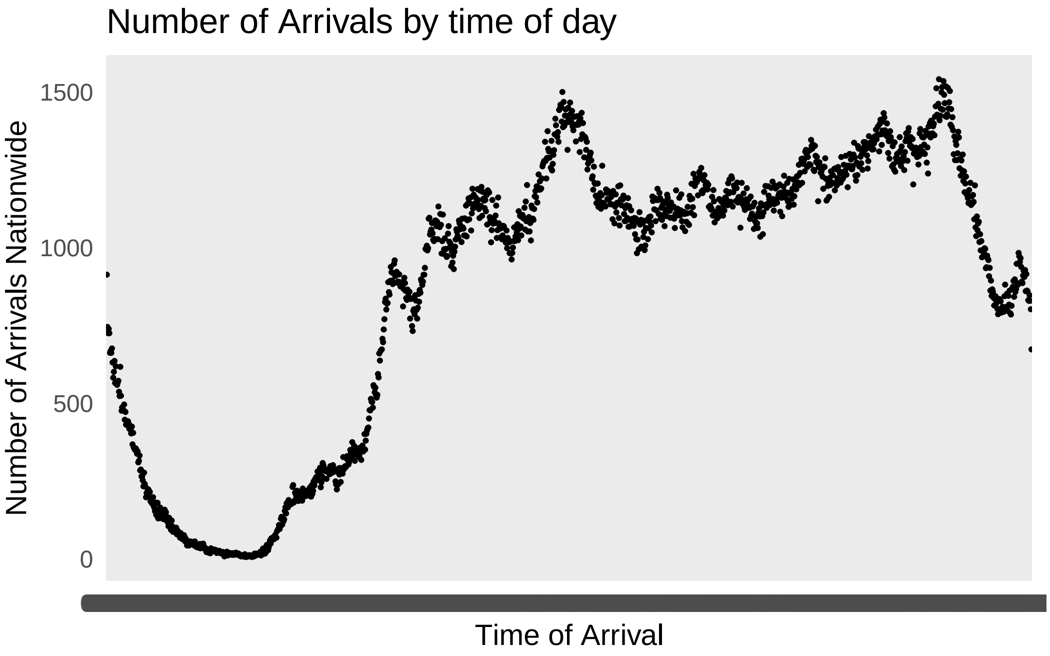
This plot is the same as the last one, but split to represent whether the flight was arriving at SBA or departing. Some interesting quirks we can see here is that American is much more accurate with their scheduled time when leaving SBA rather than other airports. There also seems to be a trend across all airlines that flights at SBA leave earlier than flights heading to SBA.



By splitting this graph by destination airport, we can see this interesting timeline of when certain flight routes are flown most often, and we can see which ones get delayed the most. Flights to Phoenix, Oakland, and Denver seem to always get delayed no matter what time of day they are scheduled for, but looking at a route like SMF (Sacramento), and it seems that morning/midday flights get delayed much more than afternoon/evening ones do. Las vegas also has an interesting gap in its schedule, flights in the middle of the day seem to rarely get delayed more than about an hour or so, but morning/evening flights get delayed for much longer times. Overall, however, there doesn't seem to be much of a relationship over all the routes and time of departure, each route has a very different relationship.



Here we have the number of departures at each minute over the course of 2023. What immediately sticks out is that there is almost no departures early in the morning (about 1-6 AM). There are also massive spikes throughout the morning, around the hour/half hour marks, which would line up with how flights are normally scheduled to depart on round numbers.There are 2 more dips at night, along side some short peaks representing red-eye flights leaving very late at night.



This graph makes sens as a transformation of the departures graph, with the biggest peaks shifted 4-8 hours later, which would line up with the length of a domestic flight. There is still a major trough in the early morning, and the spikes are a bit more compact, with less sharp outliers at specific minute marks.I also notice that in the first morning spike, the datapoints are in an almost perfect (very steep) upwards slope, with rarely one minute having less arrivals than the one before it.

```
# A tibble: 12 x 3
   MONTH departure_delay_median arrival_delay_median
   <ord>                  <dbl>                <dbl>
 1 Jan                       -1                   -4
 2 Feb                       -2                   -5
 3 Mar                        0                   -2
 4 Apr                       -1                   -4
 5 May                       -1                   -5
 6 Jun                        0                   -2
 7 Jul                        0                   -4
 8 Aug                       -1                   -5
 9 Sep                       -2                   -7
10 Oct                       -2                   -7
11 Nov                       -2                   -8
12 Dec                       -2                   -8
```

Keeping this in the table format makes the information easiest to compare, as the numbers are relatively similar over time, so the better comparision to make is between departures and arrivals. We can see that departures always have a median of leaving on time or early, there is never a month in which arrivals are not earlier than departures. In fact, in every single month, arrivals are 2-6 minutes earlier than departures. I would like to further explore how this may have changed over time (from previous years) as I would assume predictive technology has gotten better at knowing how long a flight will take. Something that stands out to me in this table is that the end of the year (sep-dec) is the earliest months in both departures and arrivals. It doesn't seem to be a seasonal phenomenon, at least from January and February.

For the Newark-Seattle direct flight, according to the parameters of the data that we started with, we only have flights that either left from or went to California based airports. This means that while a Newark-Seattle flight exists, it is not within the scope of this dataset.