# Mini Project 1

Xiaofeng Cai (5804513)

April 22, 2024

## Table of contents

## Section 1: Data Cleaning and EDA

```
  YEAR MONTH DAY_OF_MONTH OP_UNIQUE_CARRIER ORIGIN DEST CRS_DEP_TIME DEP_TIME
1 2023     1            1                AA    ATL  LAX         1230     1226
2 2023     1            1                AA    AUS  LAX          900      856
3 2023     1            1                AA    AUS  LAX         1335     1331
4 2023     1            1                AA    AUS  LAX         2100     2056
5 2023     1            1                AA    AUS  SNA         1251     1303
  DEP_DELAY CRS_ARR_TIME ARR_TIME ARR_DELAY CRS_ELAPSED_TIME
1        -4         1434     1426        -8              304
2        -4         1023     1038        15              203
3        -4         1502     1440       -22              207
4        -4         2227     2200       -27              207
5        12         1401     1424        23              190
  ACTUAL_ELAPSED_TIME
1                 300
2                 222
3                 189
4                 184
5                 201
```

The table above shows the first 5 rows of the combined data frame. Each row in the dataset represents a single flight, which is the unit of observation. There are a total of 1,267,353 observations, encompassing 14 variables including year, month, day of month, airline carrier, origin, destination, scheduled departure time, actual departure time, the amount of delat in departure, scheduled arrival time, actual arrival time, the amount of delay in arrival, the scheduled flight duration in minutes and actual flight duration in minutes.

```
      YEAR          MONTH         DAY_OF_MONTH   OP_UNIQUE_CARRIER
 Min.   :2023   Min.   : 1.000   Min.   : 1.00   Length:1267353
 1st Qu.:2023   1st Qu.: 4.000   1st Qu.: 8.00   Class :character
 Median :2023   Median : 7.000   Median :16.00   Mode  :character
 Mean   :2023   Mean   : 6.603   Mean   :15.75
 3rd Qu.:2023   3rd Qu.:10.000   3rd Qu.:23.00
 Max.   :2023   Max.   :12.000   Max.   :31.00

   ORIGIN             DEST            CRS_DEP_TIME     DEP_TIME
 Length:1267353     Length:1267353     Min.   :   4   Min.   :   1
 Class :character   Class :character   1st Qu.: 907   1st Qu.: 908
 Mode  :character   Mode  :character   Median :1315   Median :1320
                                       Mean   :1342   Mean   :1341
                                       3rd Qu.:1755   3rd Qu.:1759
                                       Max.   :2359   Max.   :2400
                                                      NA's   :11748
  DEP_DELAY        CRS_ARR_TIME     ARR_TIME       ARR_DELAY
 Min.   : -52.00   Min.   :   1   Min.   :   1    Min.   : -97.00
 1st Qu.:  -5.00   1st Qu.:1106   1st Qu.:1053    1st Qu.: -14.00
```

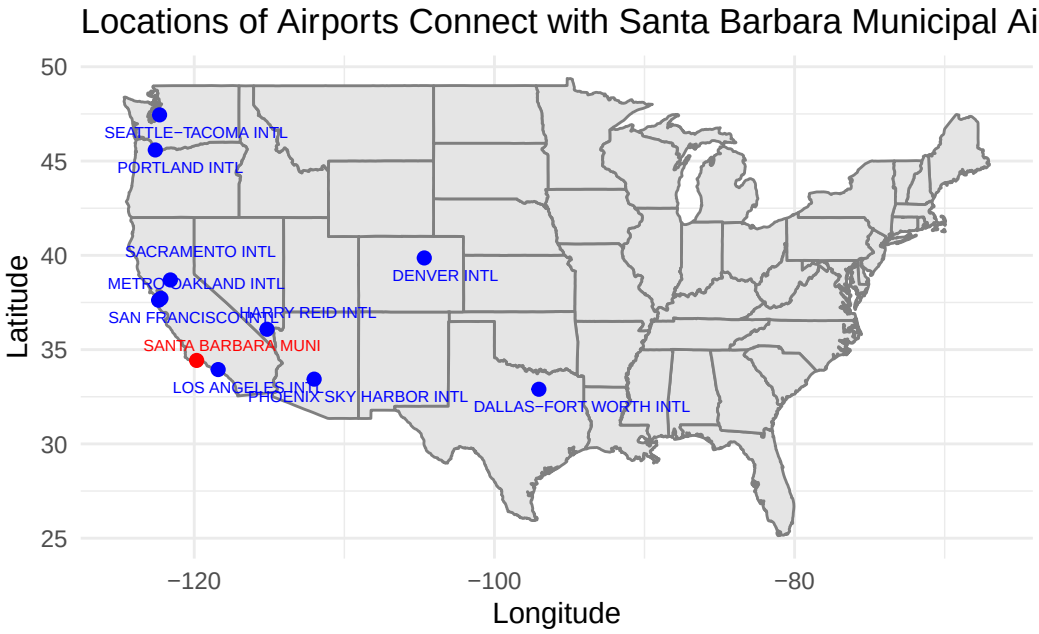
```
Median :  -1.00   Median :1525   Median :1515    Median :  -5.00
Mean   :  11.36   Mean   :1493   Mean   :1471    Mean   :   5.33
3rd Qu.:  10.00   3rd Qu.:1941   3rd Qu.:1936    3rd Qu.:   9.00
Max.   :2895.00   Max.   :2400   Max.   :2400    Max.   :2900.00
NA's   :11749                    NA's   :12591   NA's   :15015
CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
Min.   :  17.0   Min.   : 25.0
1st Qu.:  90.0   1st Qu.: 88.0
Median : 145.0   Median :140.0
Mean   : 177.3   Mean   :171.2
3rd Qu.: 255.0   3rd Qu.:246.0
Max.   :1003.0   Max.   :595.0
NA's   :3        NA's   :15015
```

And I used the summary function to find out that there are many missing values, which are filled in with 'NA.' Missing values are present in the departure time, the amount of delay in departure, arrival time, the amount of delay in arrival, the scheduled flight duration, and actual flight duration. All of these missing values will be removed before further data analysis.

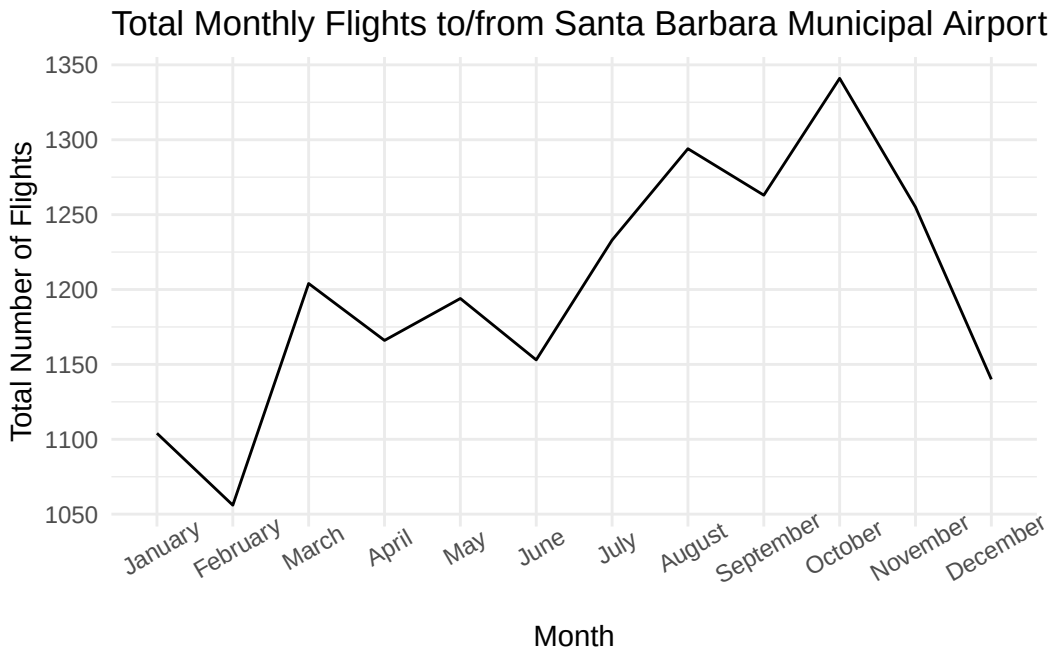
I have also joined the airport info with the original dataset. To improve readability, I have changed all months represented by numbers such as 1, 2, 3 to corresponding month names, such as January, February, etc.

## Section 2: Santa Barbara Airport

Ten airports have flights connecting with Santa Barbara Municipal Airport: Los Angeles International Airport, San Francisco International Airport, Denver International Airport, Harry Reid International Airport, Metropolitan Oakland International Airport, Sacramento International Airport, Portland International Airport, Dallas-Fort Worth International Airport, Phoenix Sky Harbor International Airport, and Seattle-Tacoma International Airport.

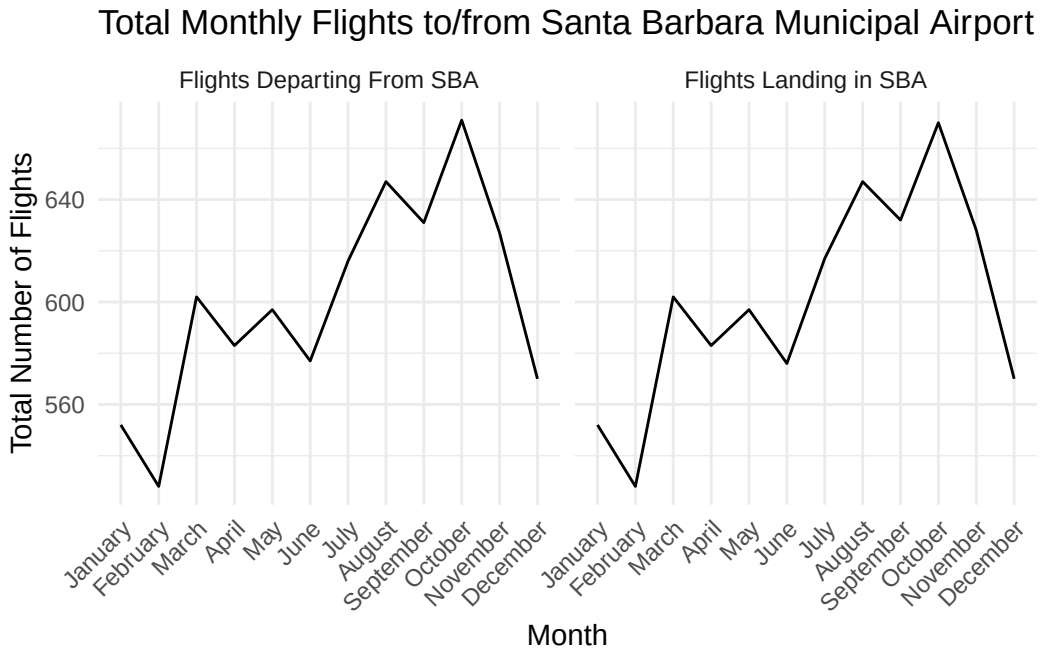



```
Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```




From the graph above, the highest seasons for flights travel to and from Santa Barbara is October and the lowest seasons for flights travel to and from Santa Barbara is February. In general, flights to or from Santa Barbara are high from July to November and low from December to February.

```
`summarise()` has grouped output by 'MONTH'. You can override using the
`.groups` argument.
```

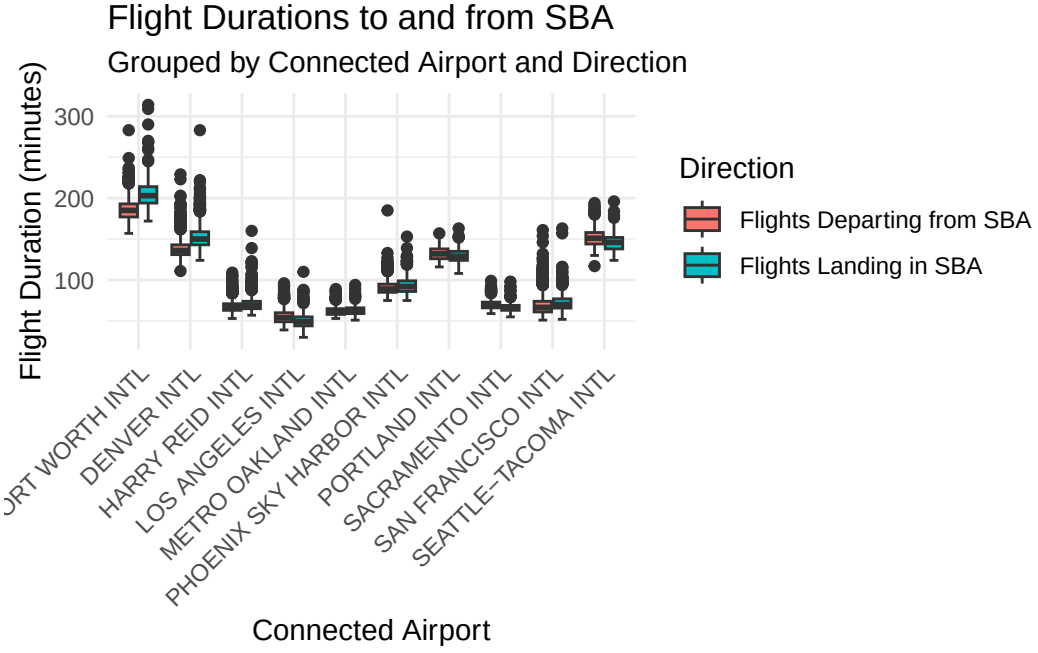



From the above graph, I did not observe any significant differences between the peaks and troughs of flights traveling to and from Santa Barbara. Whether departing from or landing in SBA, the high season appears to be October, while the low season seems to be February.

```
# A tibble: 12 x 4
# Groups:   MONTH [12]
   MONTH     `Flights Departing From SBA` `Flights Landing in SBA` Difference
   <ord>                            <int>                    <int>      <int>
 1 January                            552                      552          0
 2 February                           528                      528          0
 3 March                              602                      602          0
 4 April                              583                      583          0
 5 May                                597                      597          0
 6 June                               577                      576          1
 7 July                               616                      617         -1
 8 August                             647                      647          0
 9 September                          631                      632         -1
10 October                            671                      670          1
11 November                           627                      628         -1
12 December                           570                      570          0
```

From the table above, in the months of June, July, September, October, and November, the number of flights departing from and landing in SBA is different. In June and October, the number of flights departing from SBA is one more than the number of flights landing in SBA. And in July, September, and November, the number of flights departing from SBA is one less than the number of flights landing in SBA.

```
Warning: Removed 258 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



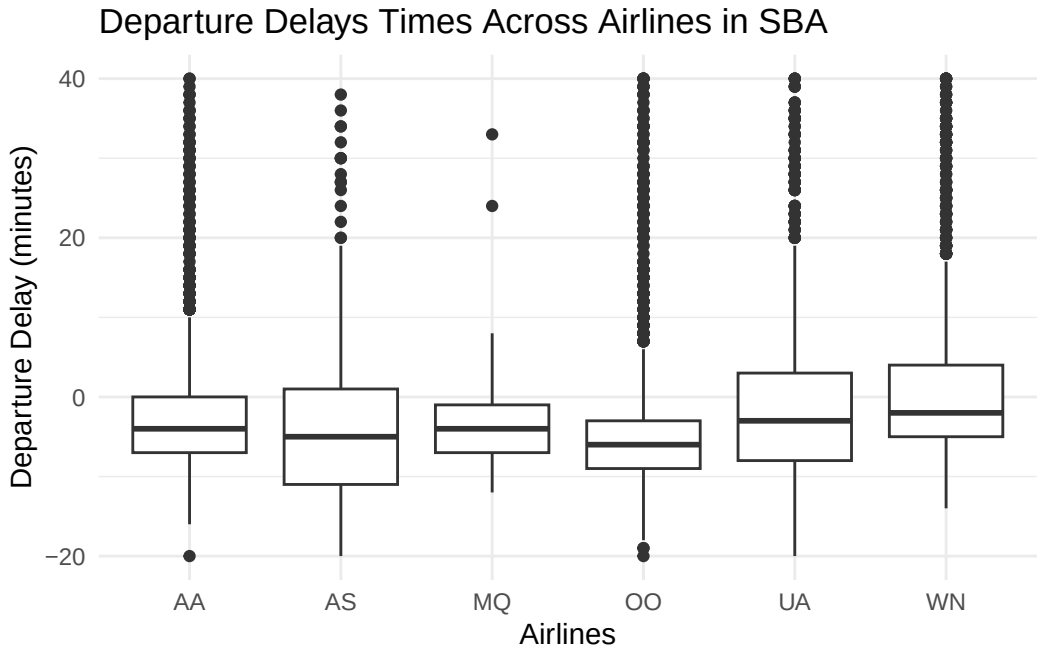

Dallas-Fort Worth International Airport has the highest flight durations for flights landing in SBA. The difference in flight duration between landing in SBA and departing from SBA is noticeable at

Dallas-Fort Worth International Airport, Denver International Airport. And the difference is not significant in the rest of airports.

```
[1] -29
```

```
[1] 1584
```

```
Warning: Removed 754 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




```
[1] "Median Departure Delays: -3"
```

```
[1] "Average Departure Delays: 10.65265"
```

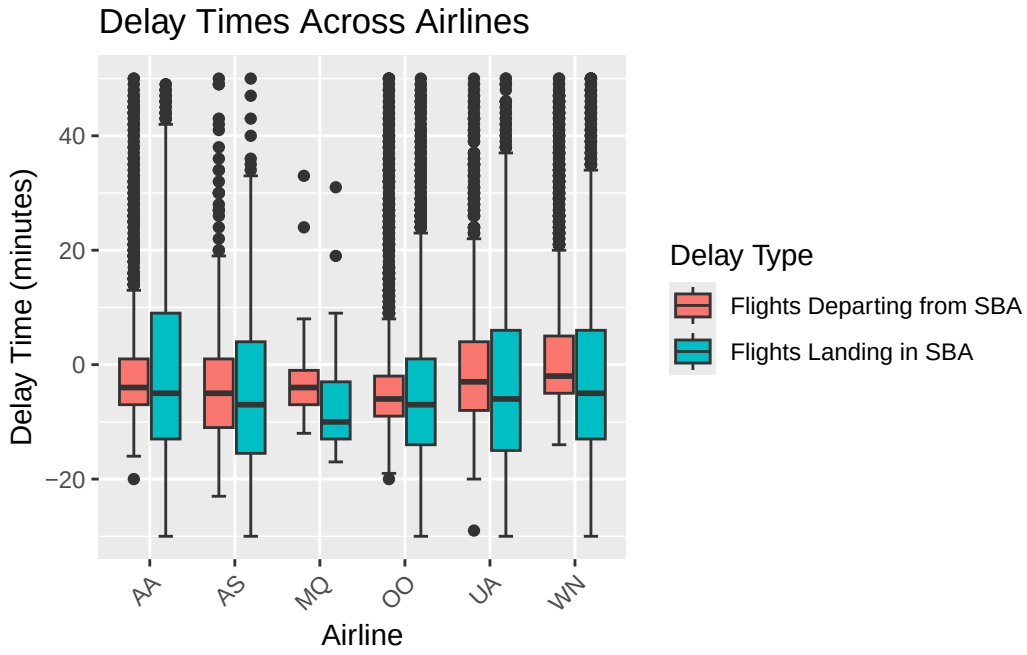
```
# A tibble: 6 x 2
  OP_UNIQUE_CARRIER median_delay
  <chr>                     <dbl>
1 AA                           -3
2 AS                           -5
3 MQ                           -4
4 OO                           -6
5 UA                           -2
6 WN                           -2
```

From the boxplot, it is evident that the middle number is below 0 for all airlines, indicating that airlines had flights departing before their scheduled departure time across all airlines. However, the presence of positive outliers in the box plot suggests that departure delays still occur.

In addition, I calculated the median, which is -3, and the average, which is 10.65265, further supporting my conclusion drawn from the boxplot. The median suggests that overall flights depart before their scheduled departure times, but there are some instances of significant departure delays that elevate the mean number substantially.

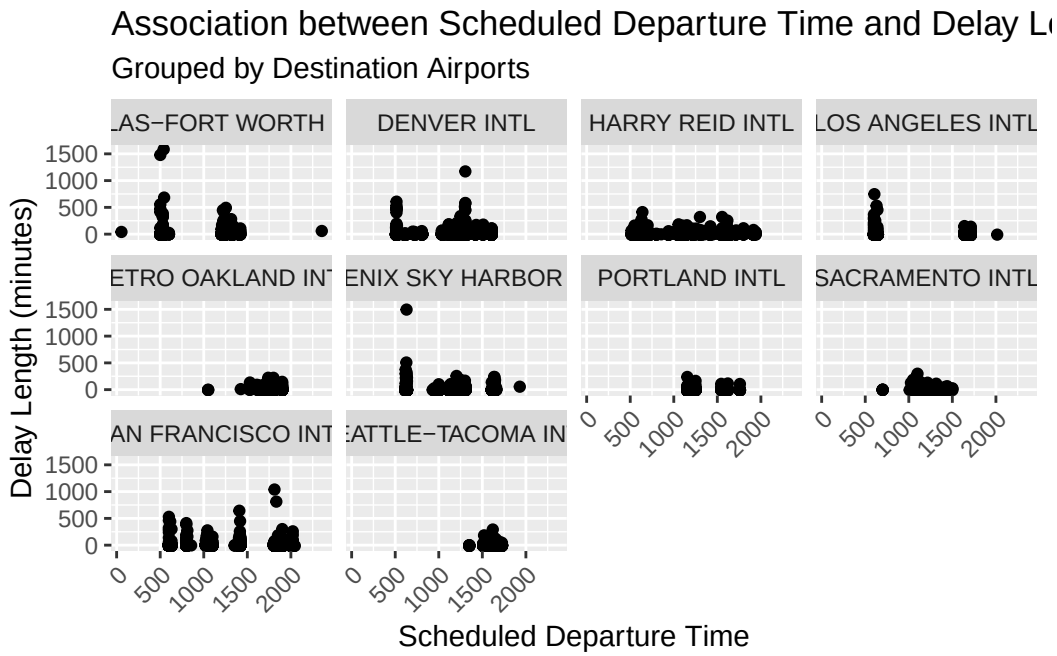
Furthermore, based on the boxplot and table, I concluded that the difference in the median departure delay across airlines is not significant.

```
Warning: Removed 1328 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




From the box plot above, it is evident that flights generally land and depart earlier in SBA, as indicated by all the medians being below 0. When comparing each airline, all show a median indicating earlier landing in SBA than departure. Additionally, although MQ airlines generally arrive and depart earlier, it has the largest difference between its departure and arrival times.

```
Warning: Removed 113 rows containing missing values or values outside the scale range
(`geom_point()`).
```



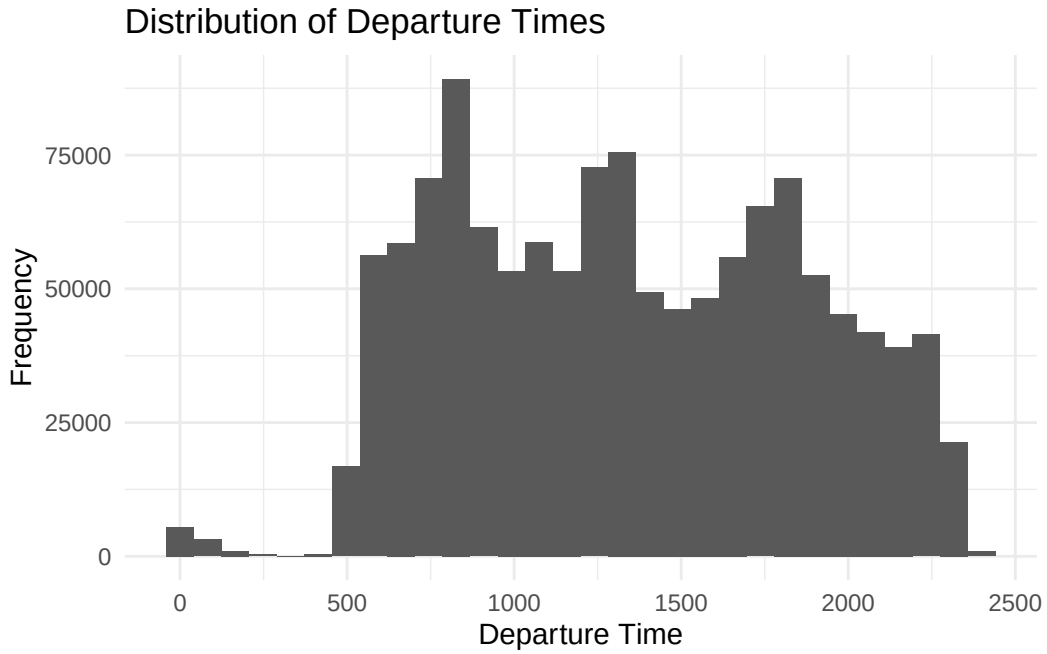

From the scatter plots above, flights tend to have longer delays from 6 am - 8 pm, particularly noticeable for Dallas-Fort Worth International Airport, Los Angeles International Airport, and Phoenix Sky Harbor Airport, which have very long delays around 6 am. Thus, there appears to be an association between the scheduled departure time and the length of the delay that flights with scheduled departure time during the day time are more likely to have longer delay times. However, this association does not seem to be influenced by the airport's location.

## Section 3: Branching Out

Now, broaden our scope and stop focusing solely on flights routing through SBA.

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

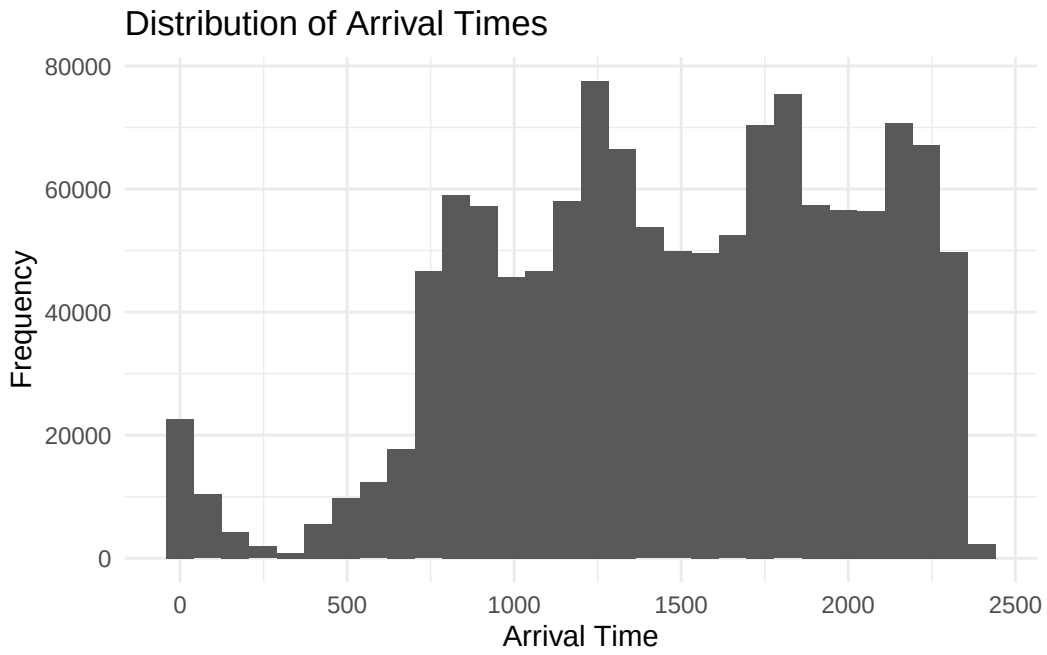
```
Warning: Removed 11748 rows containing non-finite outside the scale range
(`stat_bin()`).
```




The peaks of departure time are around 8 am, while the troughs of departure time throughout the day are between 12 am - 5 am.

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 12591 rows containing non-finite outside the scale range
(`stat_bin()`).
```




The troughs of arrival time occur between 1 am - 5 am. The peaks of arrival time occur around 12 pm, 4 pm and 10 pm.

The troughs of arrival time corresponds with departure time around 1 am - 5 am, indicating fewer flights during the middle of the night. And the peaks of arrival time corresponds with departure time in general from 9 am to 12 am (midnight).

```
# A tibble: 12 x 2
   MONTH     Average_Departure_Delay
   <ord>                       <dbl>
 1 January                      12.9
 2 February                     11.2
 3 March                        14.4
 4 April                        11.6
 5 May                          10.1
 6 June                         15.5
 7 July                         16.6
 8 August                       12.8
 9 September                     9.19
10 October                       8.04
11 November                      6.03
12 December                      7.96
```

By calulating the mean of departure delays, November has lowest average departure delays, July has highest average depature delays.

```
# A tibble: 12 x 2
   MONTH     Average_Arrival_Delay
   <ord>                     <dbl>
```

```
 1 January                7.53
 2 February               5.59
 3 March                 10.5
 4 April                  6.48
 5 May                    4.14
 6 June                  10.1
 7 July                  10.3
 8 August                 6.71
 9 September              3.02
10 October                1.32
11 November              -1.55
12 December              -0.0776
```

By calculating the mean of arrival delays, November has lowest average arrival delays, March has highest average arrival delays. June and July also have quite high average arrival delays.

To consider the information in the dataset, filtering solely based on arrival or departure airport will not provide us with all the information about flights routing through those airports. According to the Bureau of Transportation Statistics, the destination is defined as "the farthest point of travel from the point of origin of a trip of 75 miles or more one-way." Thus, not every single flight is recorded. Additionally, the dataset only contains flights from 2023 that routed through California, either having a California airport as either their point of origin or their final destination. Moreover, the dataset can also have missing information.