Mini Project 1

Albert Chui (albertchui) Joseph Zaki (josephzaki)

2024-04-21

Table of contents

1 Abstract

In this paper we explore data from the United States Bureau of Transportation Statistics detailing 1,267,353 domestic flights routed through California in 2023. First, we explored flights routed through Santa Barbara Airport (SBA). We identified seasonal highs in flight counts occurring between August and October while seasonal lows occurred between December and February. We also observed what may be the effects of jetstreams in the atmosphere which led to East-bound flights having shorter durations than West-bound flights. In exmainig the entire dataset, we identified the distributions of flight departure and arrival times in order to observe the most and least common flight times. Last, we compared departure and arrival delays across the months of the year finding that March and June had the highest median arrival and departure delays.

2 Combining Dataframes

Result of `dim(flights)`:

```
[1] 1267353      14
```

After combining the CA Flight Data for each month into a single dataframe called `flights`, we see from the output of `dim(flights)` that this dataset contains 14 variables on 1,267,353 observational units (flights) listed below. We also see that missing values are encoded as `NA`.

The table below describes each column of our `flights` dataframe:

VARIABLE	DESCRIPTION
YEAR	This is the year the flight took place, for this dataset, all values are 2023.
MONTH	This is the month the flight took place represented as numeric values 1-12 for January-December.
DAY_OF_MONTH	This is the day of the month the flight took place represented as numeric values from 1-31.
OP_UNIQUE_CARRIER	This is the airline carrier associated with the flight represented as a two character abbreviation.
ORIGIN	This is the airport of origin for the flight represented as a three character airport code.
DEST	This is the destination airport for the flight represented as a three character airport code.
CRS_DEP_TIME	This is the scheduled departure time of the flight represented in 24-hour time.
DEP_TIME	This is the actual departure time of the flight represented in 24-hour time.
DEP_DELAY	This is the delay in departure time in minutes (DEP_TIME - CRS_DEP_TIME).
CRS_ARR_TIME	This is the scheduled arrival time of the flight represented in 24-hour time.
ARR_TIME	This is the actual arrival time of the flight represented in 24-hour time.
ARR_DELAY	This is the delay in arrival time in minutes (ARR_TIME - CRS_ARR_TIME).
CRS_ELAPSED_TIME	This is the scehduled flight duration in minutes.
ACTUAL_ELAPSED_TIME	This is the actual flight duration in minutes.

3 Add Airport Information

In addition to the data found in `flights`, we have information about each airport stored in `Airport_Info.csv`. By merging these data sets, we can easily reference airport information and flight information from one place. Merging these data sets leads to the following columns being added to `flights`:

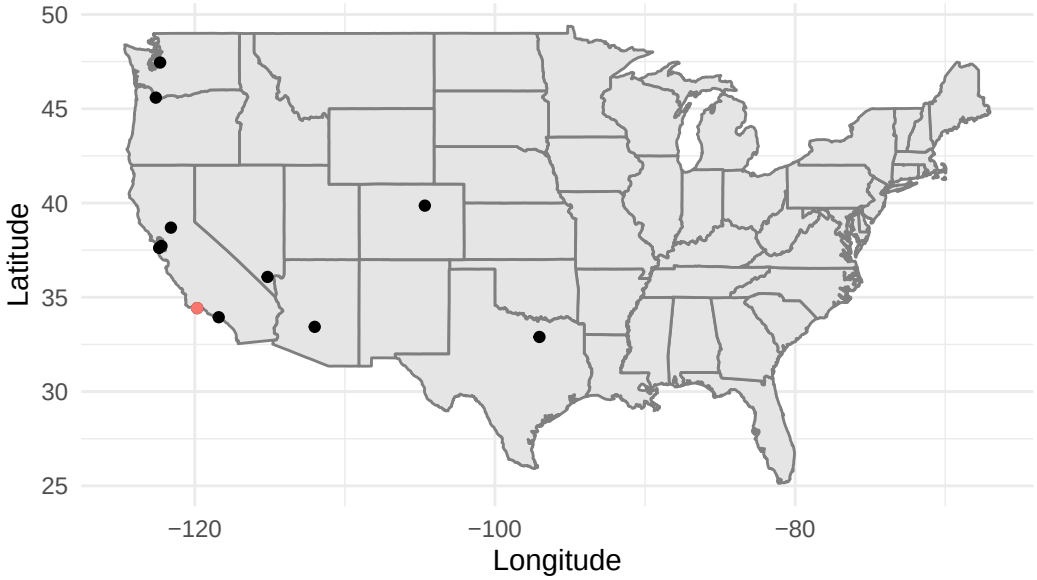
VARIABLE	DESCRIPTION
ORIGIN_ARPT_NAME	This is the full name of the airport of origin for the flight.
lon_origin	This is the longitudinal coordinate for the flight's airport of origin.
lat_origin	This is the latitudinal coordinate for the flight's airport of origin.
DEST_ARPT_NAME	This is the full name of the destination airport for the flight.
lon_dest	This is the longitudinal coordinate for the flight's destination airport.
lat_dest	This is the latitudinal coordinate for the flight's destination airport.

We have also altered the `MONTH` column to represent months as their full names rather than integers (e.g. "January" instead of 1).

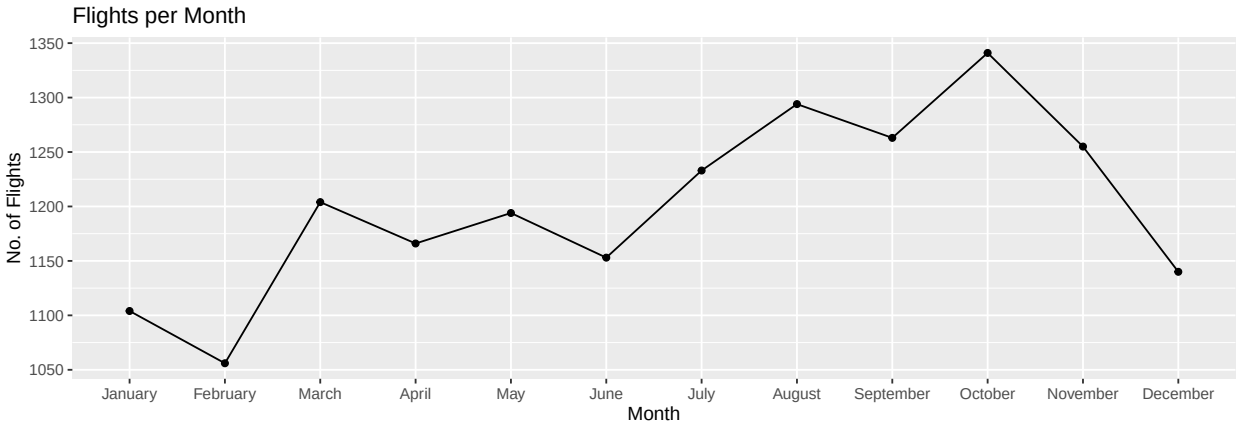
4 Explore Santa Barbara Airport

In this section, we explore flights routing through Santa Barabara Airport (SBA). Below is a list of the 10 airports that had flights routed through SBA in 2023 along with a map of these airports with SBA in red.

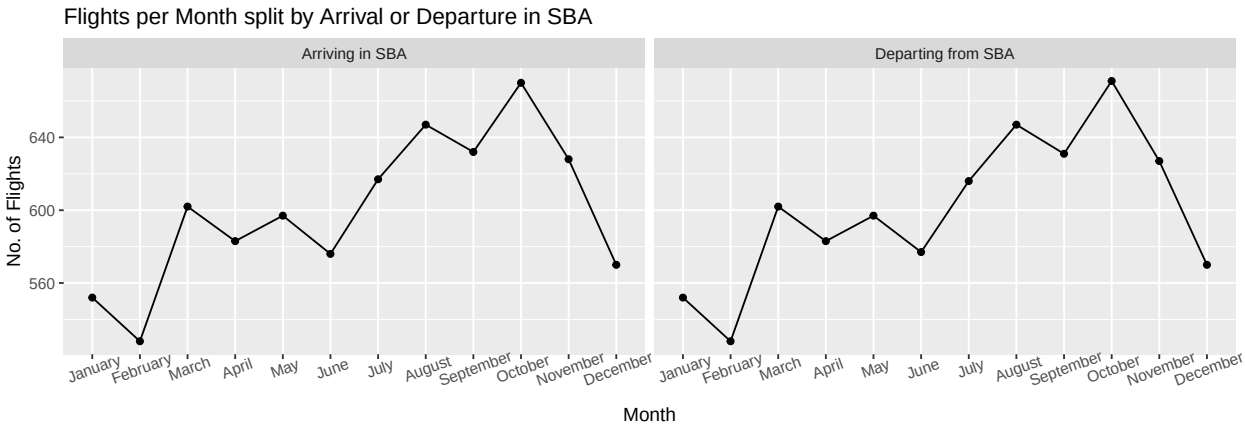
```
 [1] "DALLAS-FORT WORTH INTL"  "DENVER INTL"
 [3] "HARRY REID INTL"         "LOS ANGELES INTL"
 [5] "METRO OAKLAND INTL"      "PHOENIX SKY HARBOR INTL"
 [7] "PORTLAND INTL"           "SACRAMENTO INTL"
 [9] "SAN FRANCISCO INTL"      "SEATTLE-TACOMA INTL"
```



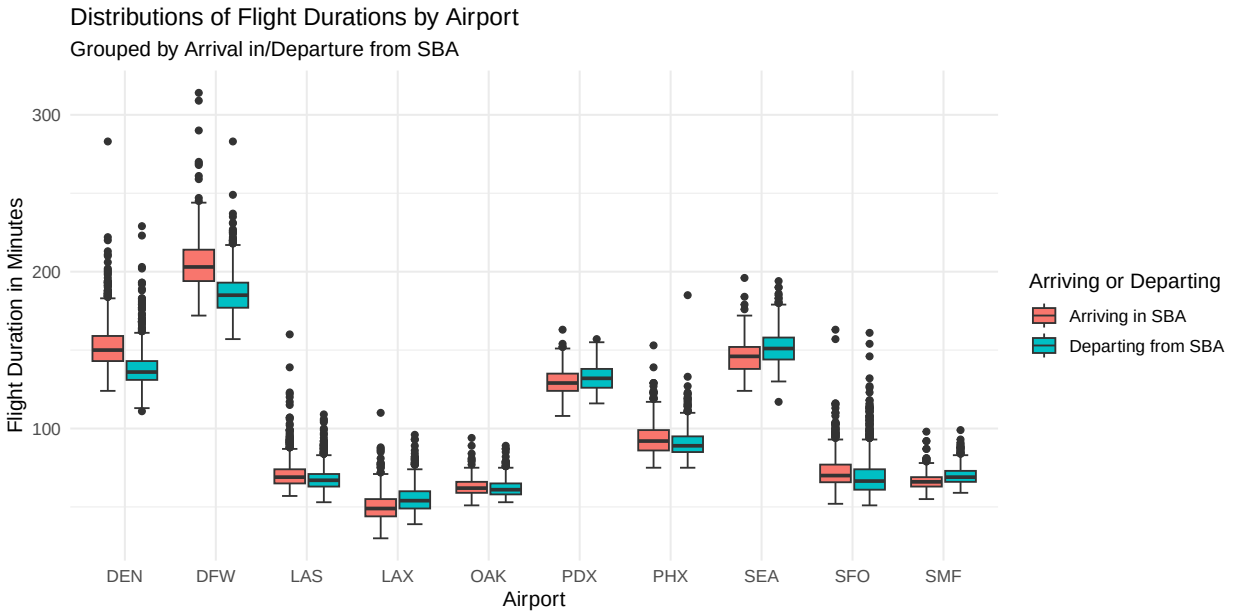
By plotting the number of flights routing through SBA each month, we can see that the winter months of December, January, and February are typically the slowest, while the late summer and early fall months of August, September and October are busiest.



Separating this information over two separate graphs, one for arrival and one for departure, reveals that there are the same amounts of flights arriving as departing in each month meaning the flights follow the same seasonal patterns as we identified when the arrivals and departures were combined.



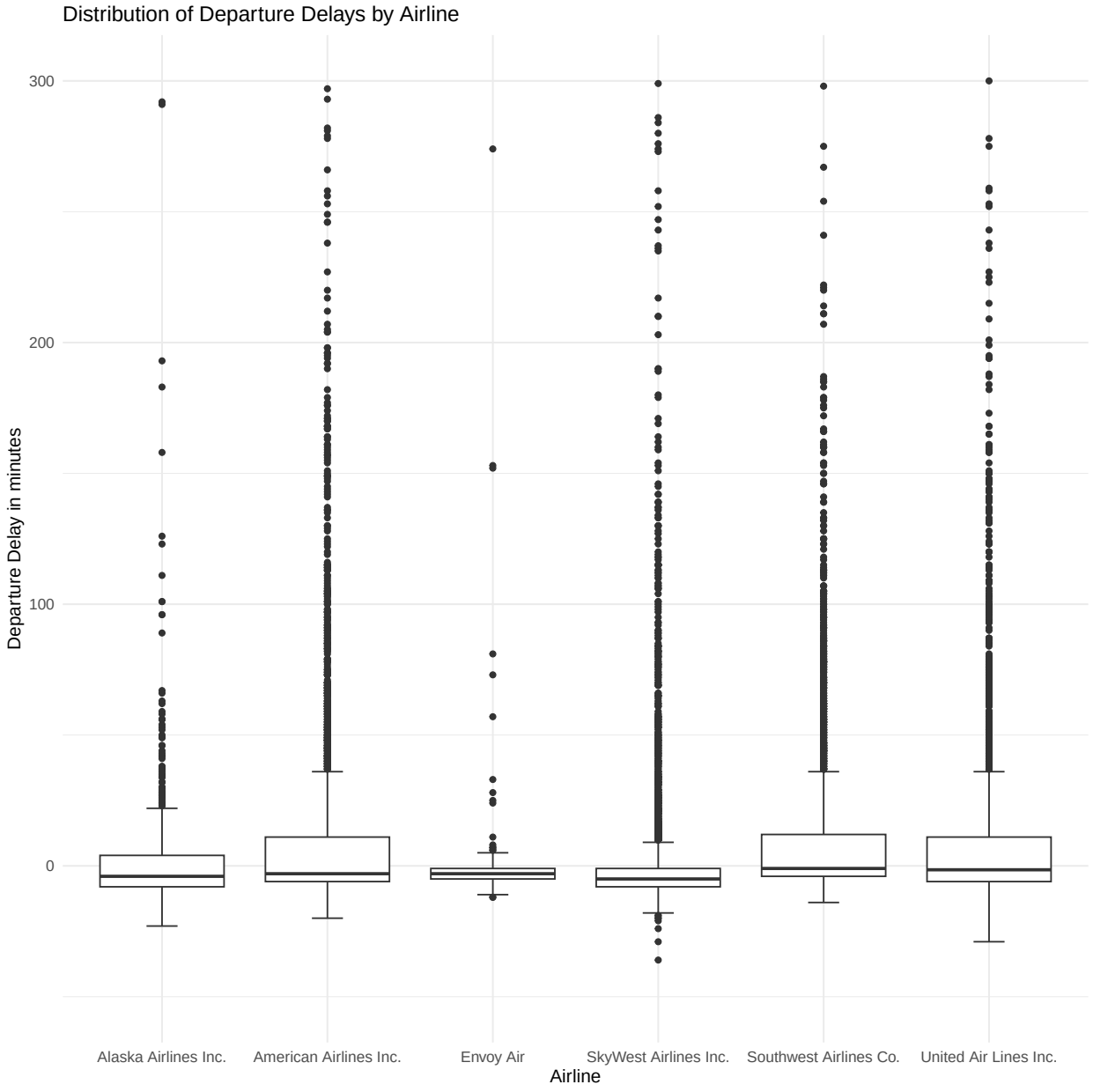
By graphing the distributions of flights routed through SBA in side by side box plots and separating them based on whether they are departing from or arriving in Santa Barbara, we observe that there is a difference in distributions for flights travelling to or from the Eastern United States. In this dataset, only two airports far to the East of Santa Barbara have flights that route through SBA. These airports are Denver International Airport (DEN) and Dallas Fort Worth International Airport (DFW). For both of these airports, flights departing from SBA and travelling East have much shorter flight durations than those travelling West to SBA. This is possibly due to the jetstreams[1] that travel Eastward which allow for higher airspeeds when travelling West to East.



[1]https://www.noaa.gov/jetstream/global/jet-stream

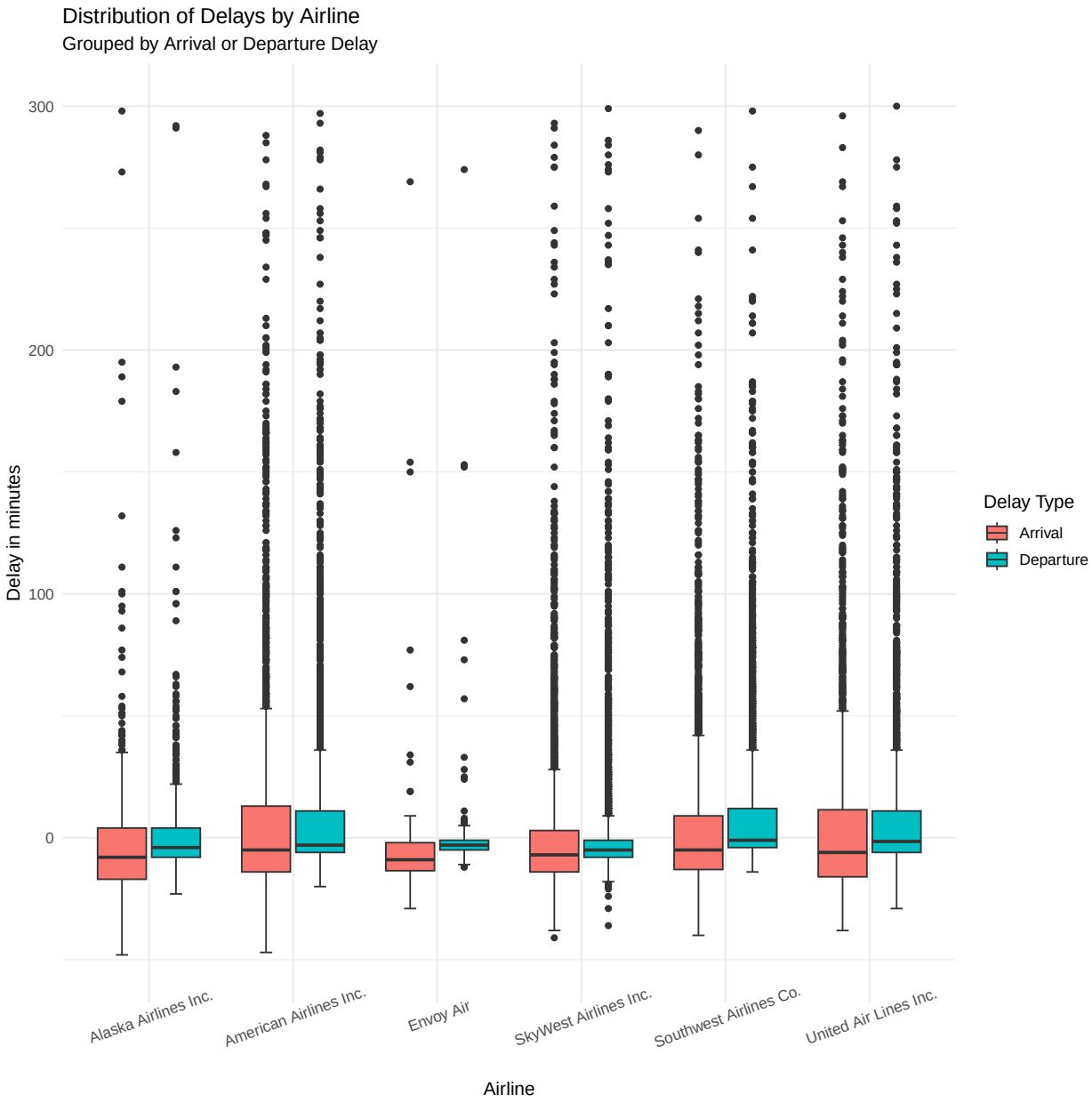
4.1 Investigate Delays

First, we examine the distributions of flight departure delays by airline. From the below box plots, we see that it is significantly more common for flights to depart after their scheduled departure time rather than before, which is expected. Additionally, we see that the median departure delays do not vary much from airline to airline, which we can confirm with the table below. The table also indicates that on average, flights across all airlines tend to depart before their scheduled time. Also from the box plot, we can see that Envoy Air was most likely to have flights depart on time, while American Airlines, Southwest Airlines, and United Airlines were most likely to have a flight depart late.

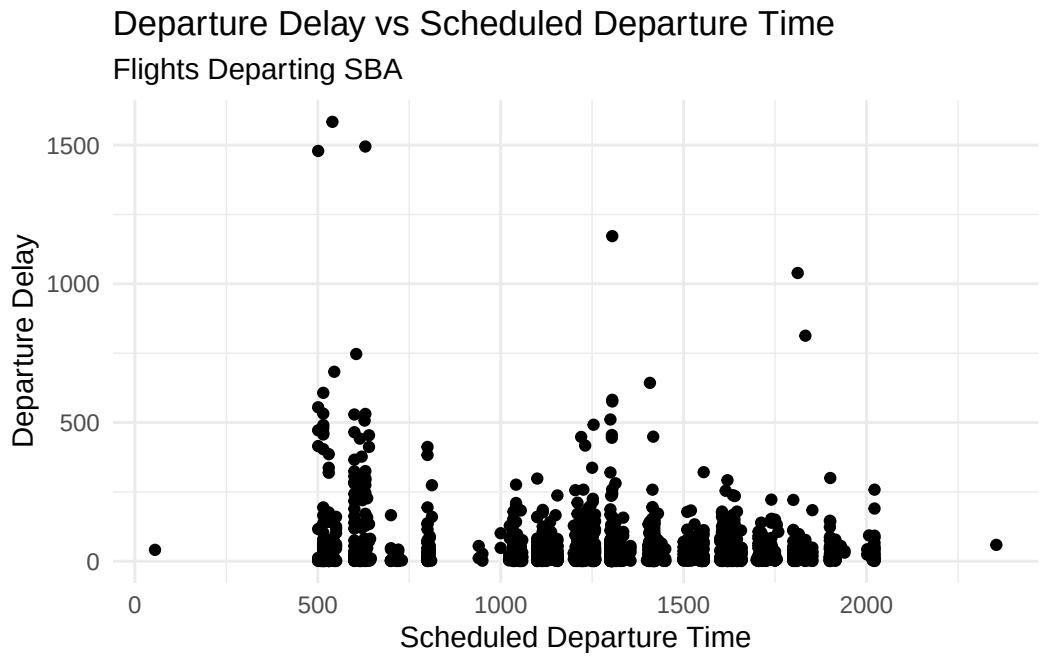


Airline	Median Departure Delay
Alaska Airlines Inc.	-4
American Airlines Inc.	-2.5
Envoy Air	-3
SkyWest Airlines Inc.	-5
Southwest Airlines Co.	-1
United Air Lines Inc.	-1

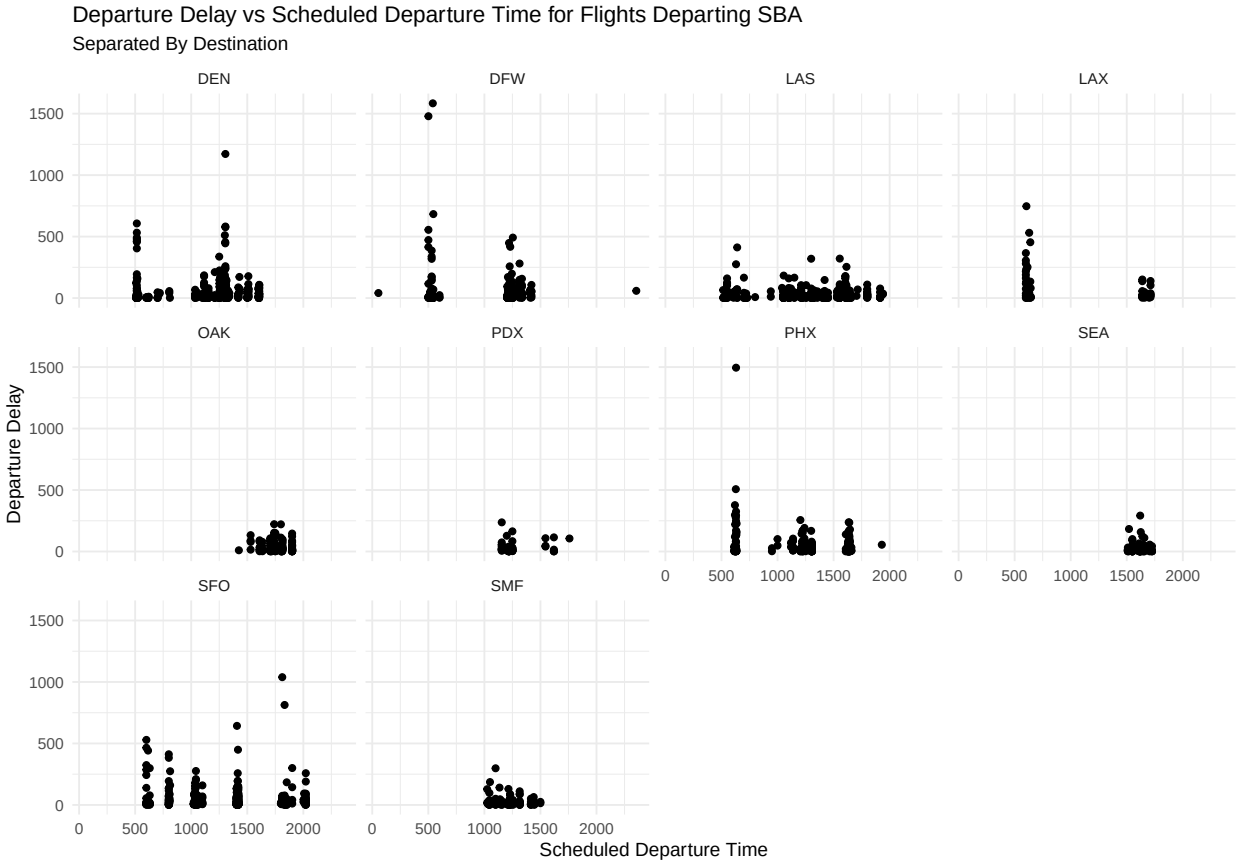
Next, we examine the relationship between Arrival Delays and Departure Delays. From this plot, we see that for all airlines, the median Departure Delay is greater than the median Arrival delay. This may indicate that most causes of delay occur before takeoff and could include late passengers or crew members, mechanical issues, weather restrictions on takeoffs, etc. Furthermore, the fact that all airlines show a longer box plot for arrival delays than departure delays indicates that although the average (median) flight will arrive ahead of schedule, there is greater variance in the punctuality of flights once they have taken off.



By plotting departure delays against scheduled departure time for flights we see that there isn't a very strong association between scheduled departure time and departure delay for flights leaving SBA. However, we see that many relatively short delays occur in the middle of the day between 11:00 AM and 7:00 PM. In contrast, there are fewer, but larger delays in the early morning hours between 5:00 AM and 7:00 AM. By splitting this scatter plot into a separate visual for each airport, we can see that his behavior arises largely due to flights with specific destinations consistently leaving at the same time tend to have delays.

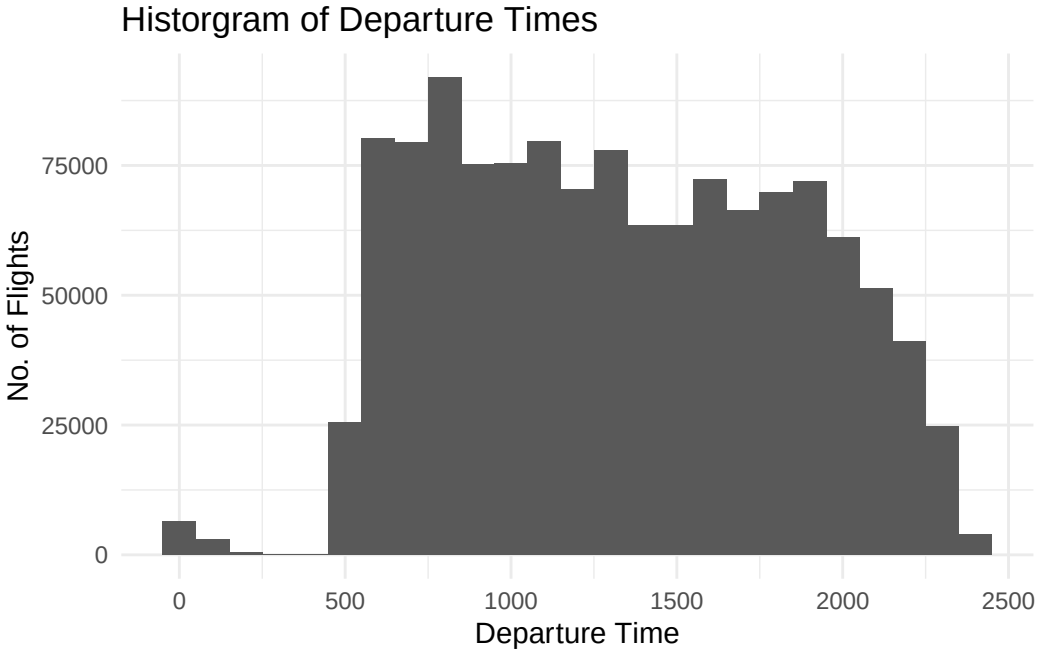


We can more clearly examine this by separating the flights by destination:

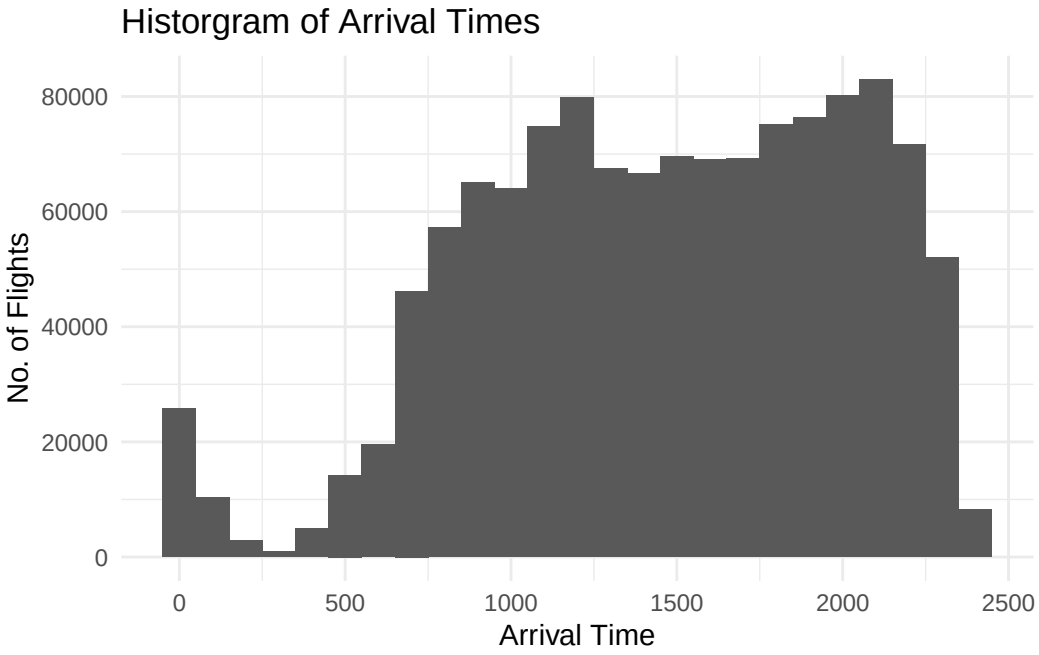


5 Explore All Flights

Next, we examine the distribution of departure times across all our flight data. From the histogram below, we can see that the number of departures remains high throughout the majority of the day starting from 5:00 AM until 8:00 PM. At this point, the number of departures begins to steadily decrease and remains very low through the late night and early morning hours.



Performing the same analysis on arrival times, we see that the most arrivals occur around 12:00 PM and 8:30 PM. Like departures, the number of arrivals remains high throughout most of the day, with a sudden drop after 10:00 PM, followed by a small spike in arrivals between 12:00 AM and 1:00 AM. The lowest number of arrivals occurs between 2:00 AM and 5:00 AM. Overall, the distribution of arrival times seems to be slightly shifted to the right when compared to departure times, which makes sense since a spike in departures is expected to cause a spike in arrival later in the day.



Last, we can take a look at the median arrival and departure delays each month. We can see that March and June have the most arrival delays as well as the most departure delays (along with July) with the median flight arriving only 2 minutes ahead of schedule and departing exactly as scheduled. This is possibly due to harsh weather conditions during these months, but also could be caused by an increase in travelers. According to United Airlines, Spring Break, which usually occurs in March is the busiest time of the year for air travel[2]. An increase in air travel may also explain the increase in arrival delays in June, as this is when most students begin their Summer Break and may coincide with family summer vacations.

In contrast, the late Fall and Winter months of September through December have the least delays for both arrivals and departures. The median flight in these months typically departs 2 minutes ahead of schedule and arrived 7 or 8 minutes ahead of schedule.

MONTH	Median Arrival Delay	Median Departure Delay
January	-4	-1
February	-5	-2
March	-2	0
April	-4	-1
May	-5	-1
June	-2	0
July	-4	0
August	-5	-1
September	-7	-2
October	-7	-2

[2]https://www.nbcnews.com/business/travel/spring-break-travel-forecast-airlines-and-vacations-how-much-money-rcna141729

MONTH	Median Arrival Delay	Median Departure Delay
November	-8	-2
December	-8	-2

6 Scope of Inference

Our data is collected from the The United States Bureau of Transportation Statistics (BTS), and specifically only examines flights from 2023 that routed through California. For example, if there was a flight directly from Seattle to Newark, this flight would not be included in our dataset as it did not route through California. This means that if we try to use our dataset to examine flights from a non-California airport, we will only see flights going to California.