Exploratory Data Analysis & Cleaning In R

Billy Dang

April 29, 2024

Introduction

The **United States Bureau of Transportation Statistics** (BTS) is a part of the Department of Transportation designed to aid in the reporting and collection of transportation-related data. In this project, I will be exploring some of the aviation data that the BTS provides. Specifically, as a student of UC Santa Barbara that has taken flights to and from Santa Barbara Municipal airport, I am especially interested in unraveling local flight patterns and their implications on regional travel. More specifically I will examine only flights from 2023 that are routed through California (i.e. that have a California airport as either their point of origin or their final destination) and take a particular focus on those routed through Santa Barbara itself.

This project will mainly take the form of a visualization project, where the main aim is to better understand the provided data through exploration and analysis, rather than inference or modeling.

Background & Data Description

I will be using the following packages:

```
library(pander)
library(reshape2)
library(tidyverse)
library(dplyr)
library(maps)
library(ggthemes)
```

The data for this project is spread across several files;

- a series of 12 files containing flight informations for each of the 12 months in 2023 (these files all have the name CA_Flights_, where represents the month represented in the file)
- a file called Carrier_Codes.csv, which includes the full names for the various airline carriers included in the dataset
- a filed called Airport_Info.csv, which contains geographical information about major US airports.
- Each of the CA_flights_.csv files contain the following column names (and their description):

Variable Name	Description
year	the year of observation
month	the month of observation
day_of_month	the day of month of observation
op_unique_carrier	the airline carrier associated with the observation
origin	the airport code of the origin (i.e. point-of-departure) of the observation
dest	the airport code of the destination
crs_dep_time	the scheduled departure time
dep_time	the actual departure time
dep_delay	the amount of delay in departure; i.e. actual departure minus schedule departure (flights that departed early have a negative dep_delay value)
crs_arr_time	the scheduled arrival time
arr_time	the actual arrival time
arr_delay	the amount of delay in arrival; i.e. actual arrival minus schedule arrival (flights that arrived early have a negative dep_delay value)
crs_elapsed_time	the scheduled flight duration (in minutes)
actual_elapsed_time	the actual flight duration (in minutes)

Data Cleaning & Exploratory Data Analysis

As the data I will be taking a look at spans across multiple files, we'll begin this project with a bit of cleaning and merging before exploring the data set. This initial portion will combine the flight California flight logs for the 2023 calender year

```
files <- c("data/CA_Flights_Jan.csv",
           "data/CA_Flights_Feb.csv",
           "data/CA_Flights_Mar.csv",
           "data/CA_Flights_Apr.csv",
           "data/CA_Flights_May.csv",
           "data/CA_Flights_Jun.csv",
           "data/CA_Flights_Jul.csv",
           "data/CA_Flights_Aug.csv",
           "data/CA_Flights_Sept.csv",
           "data/CA_Flights_Oct.csv",
           "data/CA_Flights_Nov.csv",
           "data/CA_Flights_Dec.csv")

file_list <- lapply(files, read.csv)

flight_files <- bind_rows(file_list)
```

Notice

Taking a look at the data set after combining the aforementioned .CSV files reveals that there are over 1.2 million observational units with 14 variables when looking at 2023. It can also be seen, either through `view()` or `summary()`, that missing values are encoded as NA (with nearly 28,155 missing values in the "ACTUAL_ELAPSED_TIME" column!)

After joining the data frames, I'll also take this time to ensure that variables are encoded using the appropriate type (e.g. checking that ordinal variables are stored as ordered factors, numerical variables are stored with the data type numeric, etc.). Also, let's make sure that months have descriptive names (e.g. Not using 1 for January; instead, use Jan or January, etc.).

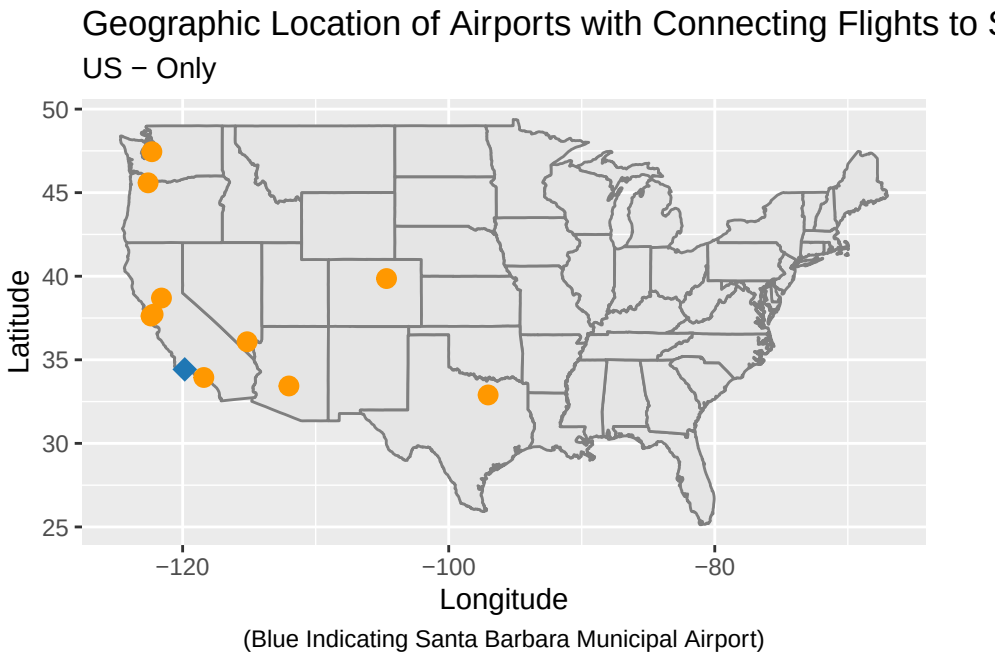
```
flight_files_final <- flight_files_updated %>%
  mutate(MONTH = case_when(
    MONTH == 1 ~ "Jan",
    MONTH == 2 ~ "Feb",
    MONTH == 3 ~ "Mar",
    MONTH == 4 ~ "Apr",
    MONTH == 5 ~ "May",
    MONTH == 6 ~ "Jun",
    MONTH == 7 ~ "Jul",
    MONTH == 8 ~ "Aug",
    MONTH == 9 ~ "Sep",
    MONTH == 10 ~ "Oct",
    MONTH == 11 ~ "Nov",
    MONTH == 12 ~ "Dec",
    TRUE ~ as.character(MONTH)
  ))
```

Looking at Santa Barbara Airport

We'll start off by restricting our considerations to flights routing through Santa Barbara Airport (SBA). Then, let's plot all these airports onto a map and take a look at SBA's reach and connectivity. Keep in mind that this refers to flights both originating from and arriving at SBA.

Notice

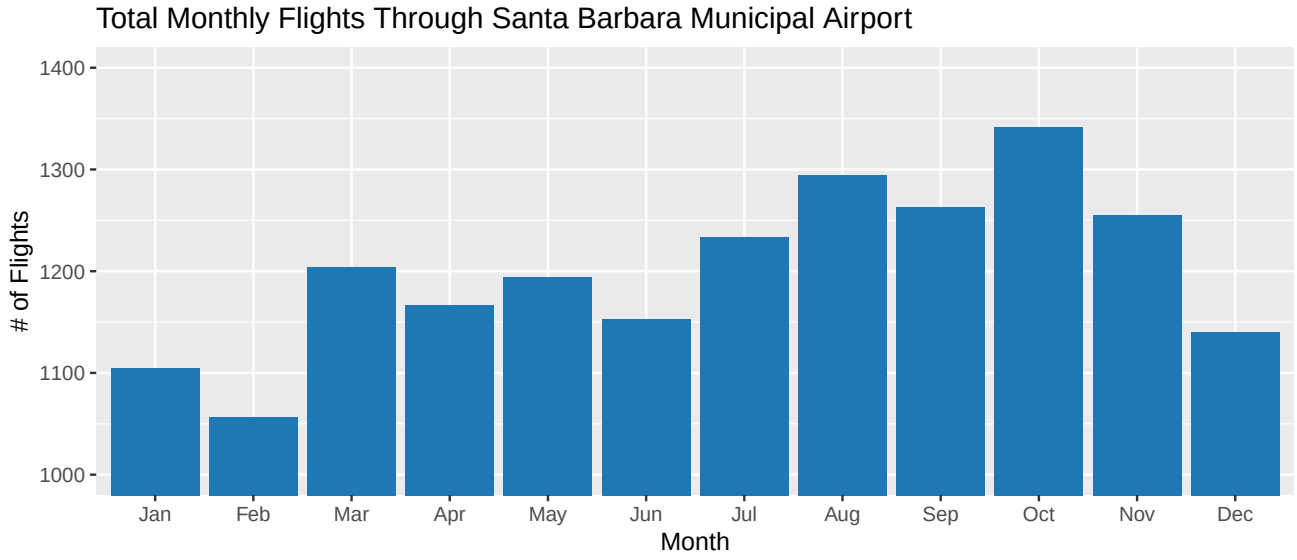
After these restrictions, one can see that there are 10 total airports that have flights to and from Santa Barbara: these include Dallas-Fort Worth INTL, Phoenix Sky Harbor INTL, Seattle-Tacoma INTL, Los Angeles INTL, San Francisco INTL, Denver INTL, Harry Reid INTL, Metro Oakland ITNL, Sacramento INTL, and Portland ITNL. Notably, the map also reveals that there is a lack of major east coast hubs served such as those in New York, Atlanta, Boston, or Miami. Wait, so how would travelers (i.e. students) on the East Coast get to Santa Barbara? We'll answer this later when evaluating the scope of inference of this report.



(Blue Indicating Santa Barbara Municipal Airport)

Exploring Flights

Now, let's delve a little deeper into the data to identify some more details, trends and aviation patterns associated with SBA and it's routed flights.



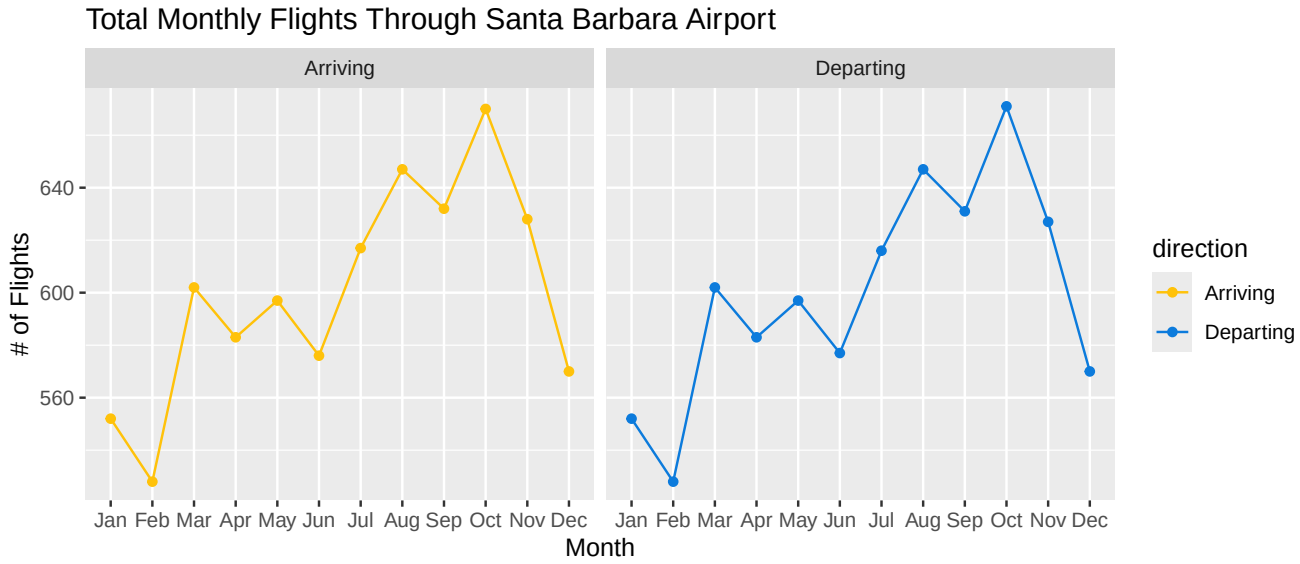
Notice

Here, we've made a graph to visualize the total # of monthly flights that route through SBA. Looking at this reveals the seasons that travelers most frequently fly from or to SBA. We see the lowest number of flights in Spring (with the least in February) and the highest frequency during Fall (with the most in October).

Why might this be?

There are a couple factors that may play into the distribution of flights across the months displayed above. One may be the tourism industry, as Santa Barbara is a popular tourist destination known for its mild climate and coastline, the jump in total flights during warmer months (March-July) may be from visitors seeking to bask in the summer Santa Barbara sun. As for the later months, holiday travel may be a significant contributing factor as it is widely known that the Sunday after Thanksgiving is the busiest day in the country's aviation history. This would account for the large spike around September & October.

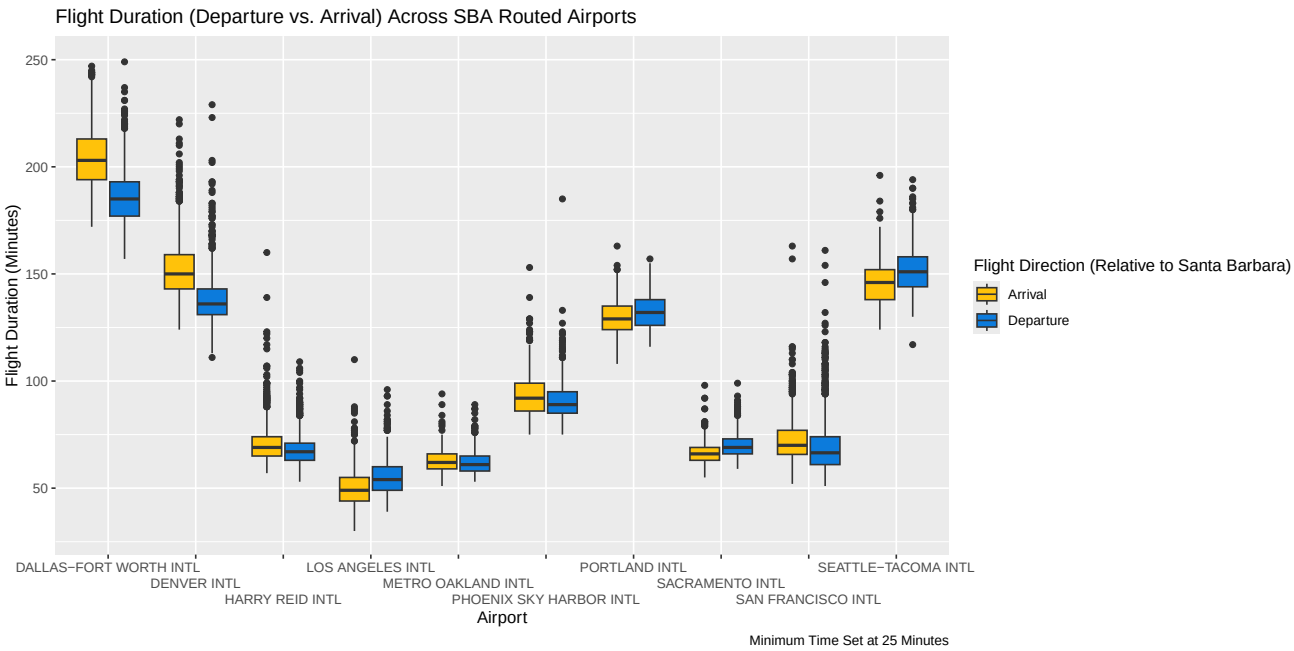
As mentioned earlier this only gives us an idea of aggregate flights, without taking into account whether travelers are leaving or entering. Let's take a look at the graph again, this time taking into account departure or arrival.



Notice

After reproducing the graph above and taking into account aviation traffic going in and out of Santa Barbara, the overall trends seem to hold with arrivals and departures having identical values across all calender months.

Alright, that's enough for examining flight volume and frequency. Let's go ahead and look at flight times and everyone's nightmare: flight delays. How might these flight times vary from airport to airport and based on arrival or departure?



Notice

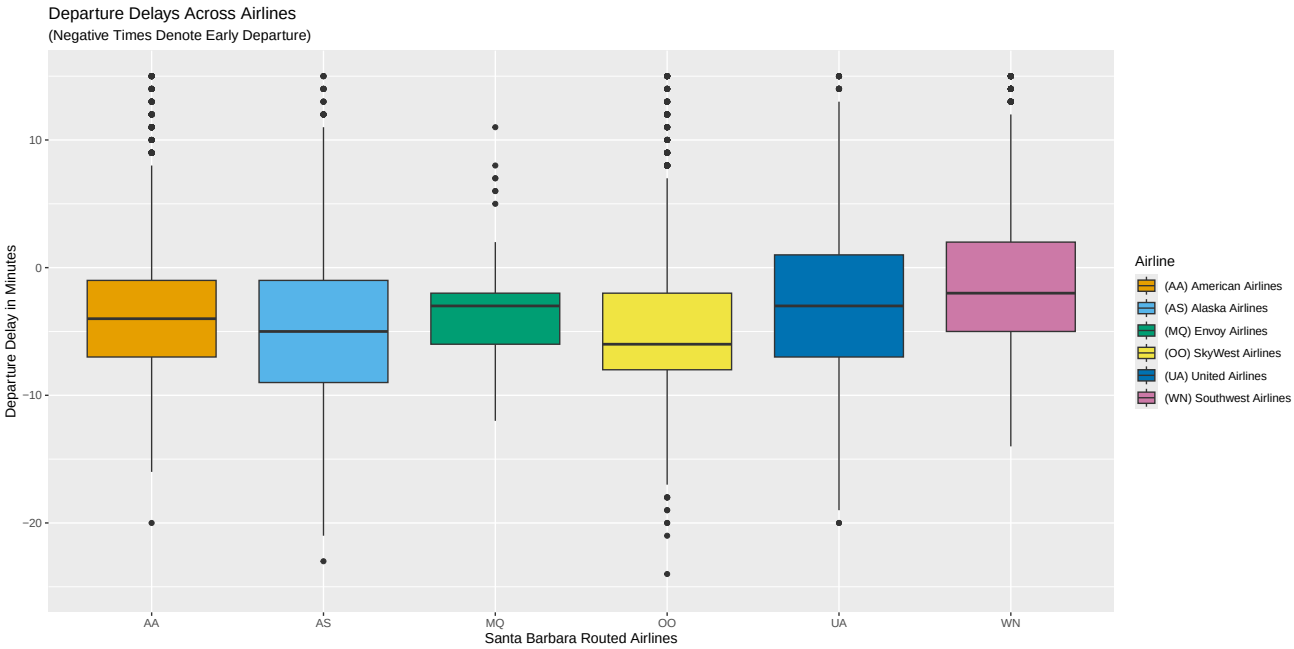
Looking at the graph produced, there are notable differences between certain airports and their respective flight times. Dallas, Seattle, and Denver hold some of the highest travel times which can be explained by their distance from Santa Barbara (especially when compared to that of Oakland or Los Angeles.) Additionally, for these longer flights their tends to be a notable difference between the distribution of flight times when it comes to whether or not these flights are arriving or departing from SBA, with flights arriving in SBA having longer flight durations (except for Seattle).

Why might this be?

The increased flight duration when arriving to Santa Barbara may be a result of having to take connecting flights in order to reach SBA as it is a smaller airport.

Investigating Delays

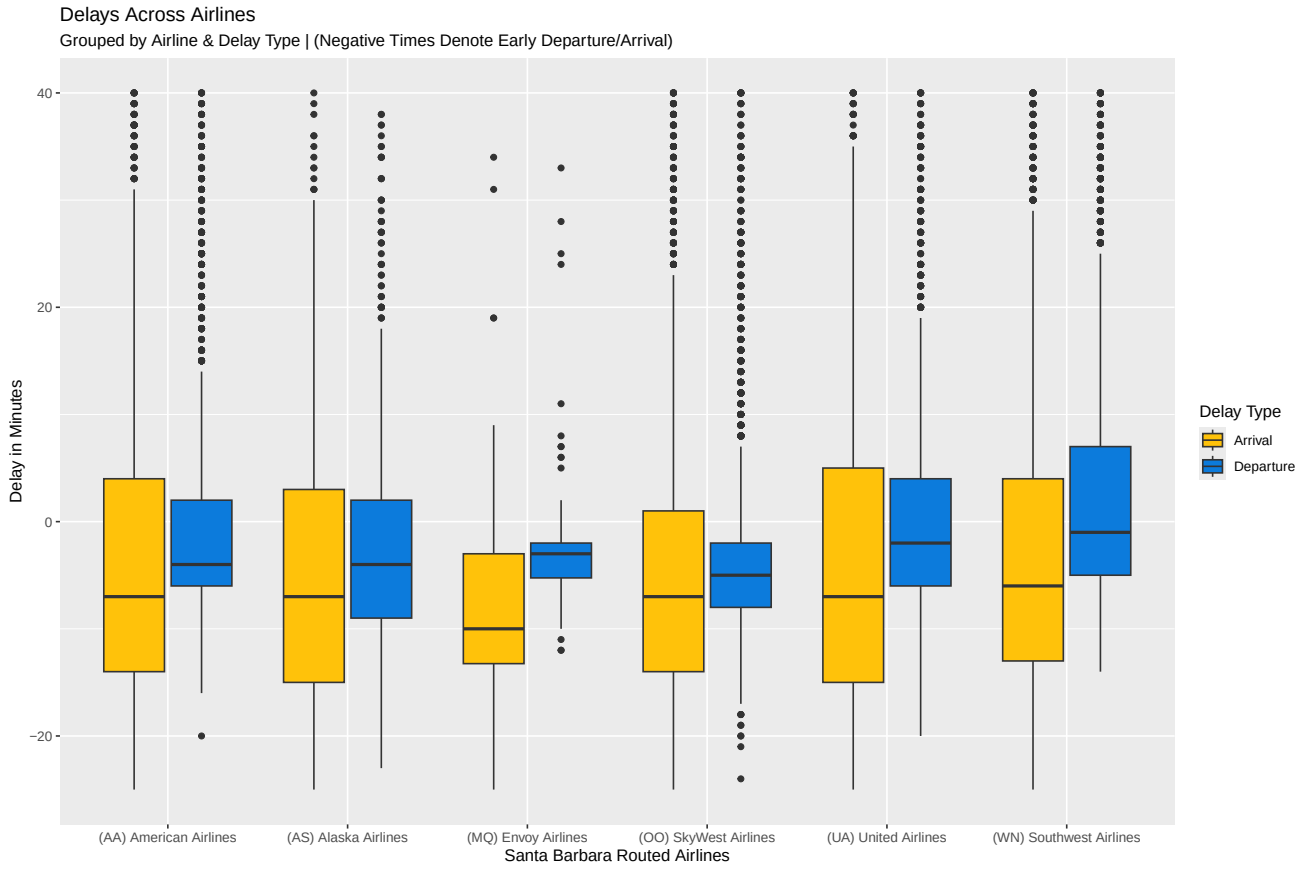
Now that we have a decent feel for the airports and flight durations included in the dataset, let's start to investigate the delays. Again, let's only consider flights that route through SBA (i.e. have SBA as either their point-of-origin or their destination) and look at flight delays across Airlines instead of Airports.

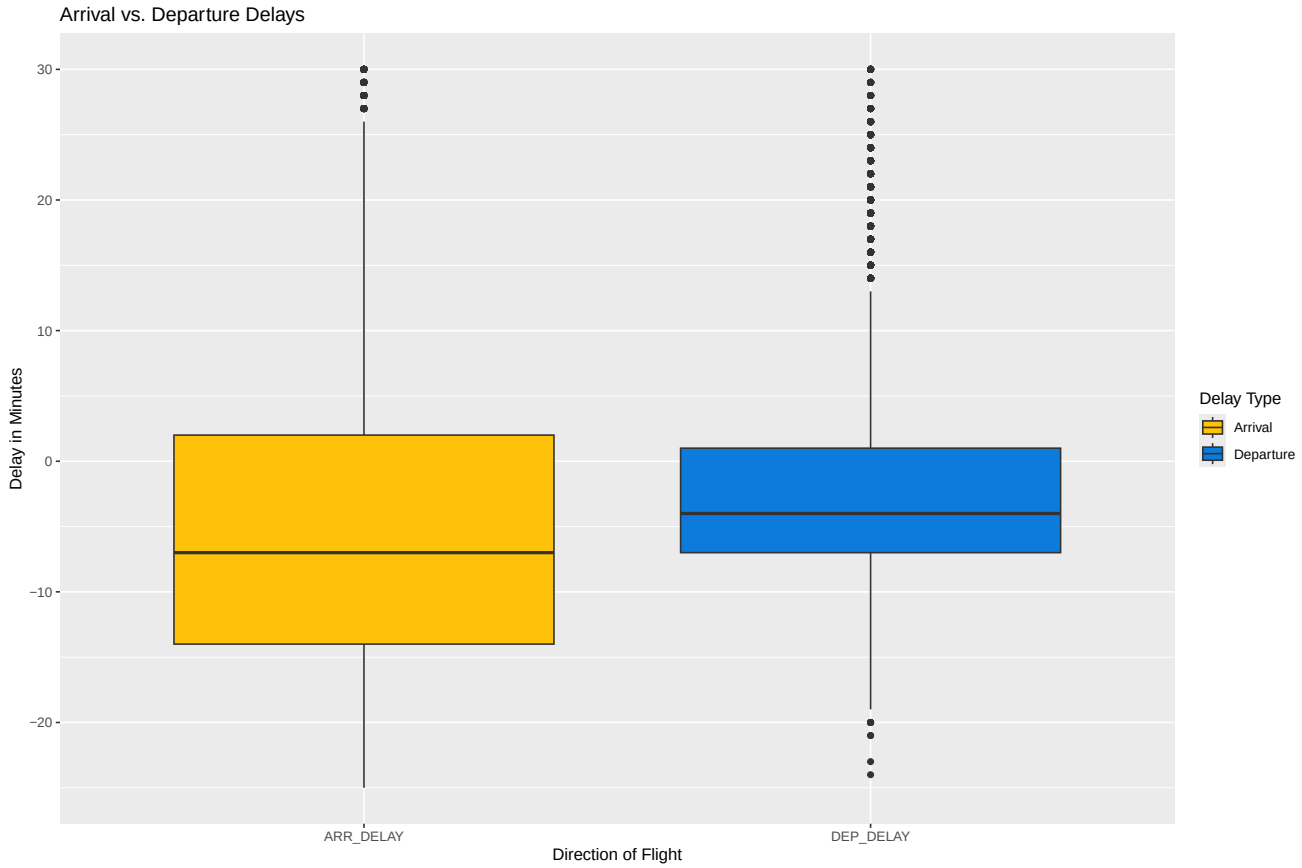


Notice

Using the bar plot pictured above, we can see that all airlines on average (median) depart earlier than their schedule time; with airlines like America, Alaska, Envoy, and Skywest having ~75% or more of their flights departing early! The airport with the lowest average departure delay time would be held by SkyWest Airlines, while Southwest Airlines would have the highest average (median) departure delay time.

Once again, this graph only looks at **departure** delays across Santa Barbara routed airlines. Let's take a look at how these delay times may vary when we look at both **departure** AND **arrival** across these airlines and in aggregate:


Delays Across Airlines
Grouped by Airline & Delay Type | (Negative Times Denote Early Departure/Arrival)
Delay in Minutes
40
20
0
−20
(AA) American Airlines
(AS) Alaska Airlines
(MQ) Envoy Airlines
(OO) SkyWest Airlines
(UA) United Airlines
(WN) Southwest Airlines
Santa Barbara Routed Airlines
Delay Type
Arrival
Departure


Arrival vs. Departure Delays
Delay in Minutes
30
20
10
0
−10
−20
ARR_DELAY
DEP_DELAY
Direction of Flight
Delay Type
Arrival
Departure

Notice

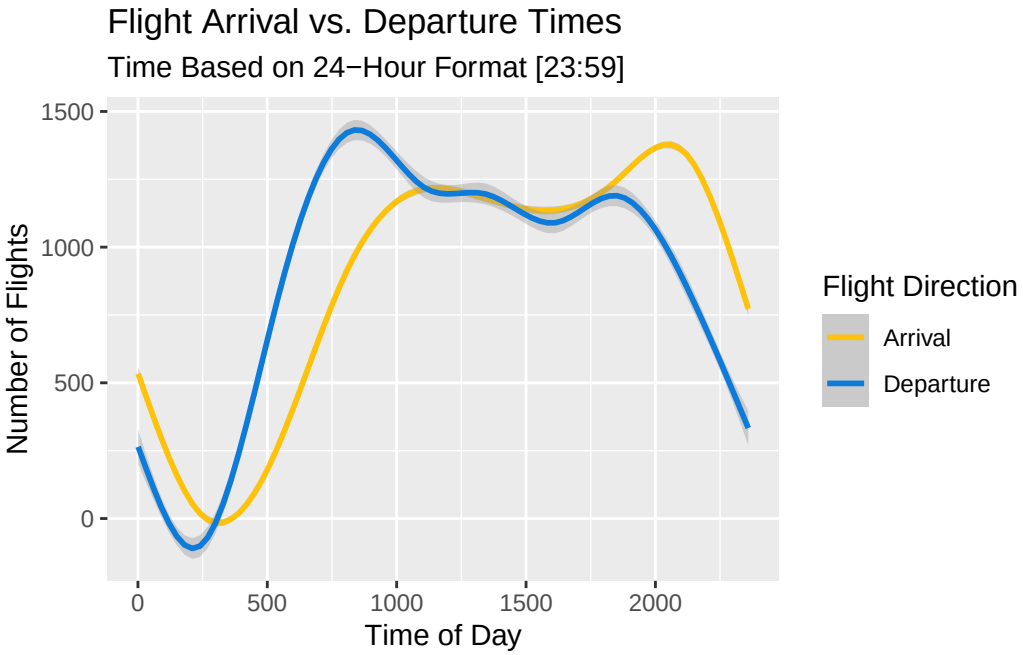
After taking into account the different delay types, let's start by identifying overall trends within the graph. First, it seems all airlines on average (median) run ahead of their schedule time. Another notable trend is that for all airlines, plane rides tend to have smaller delay times for arrival when compared to their departure counterparts. (Look at American Eagles Airline 'MQ' where 75% of departures & arrivals are run earlier than scheduled). While a bit unnecessary, the 2nd graph that has grouped all the airlines and separated only by delay type reinforces our findings above.

Why might this be the case?

A possible explanation could be that airlines often build buffer time into their schedules to account for potential delays, especially for arrival times. Departure times may be less padded with buffer time while having more variables that may lead to delay such as passenger boarding, baggage loading, and taxiing to the runway.

Branching Out

Alright, that's enough focus on Santa Barbara Airport. Next, we're going to broaden our scope & examine the aviation data for 2023 more holistically. Let's start by looking at *actual* departure & arrival times across all airports:

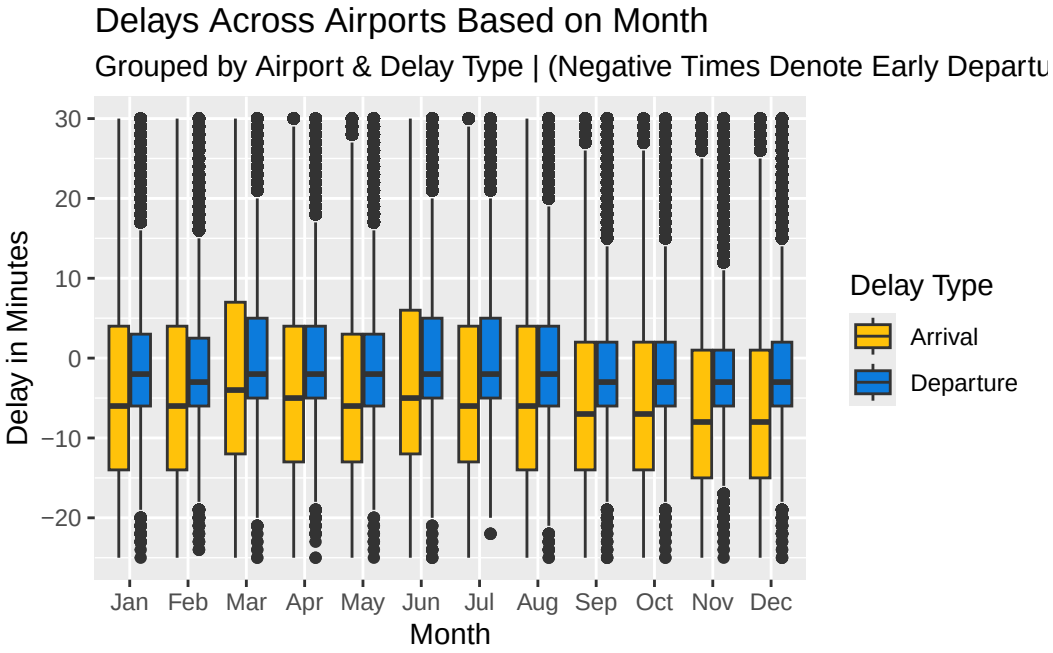


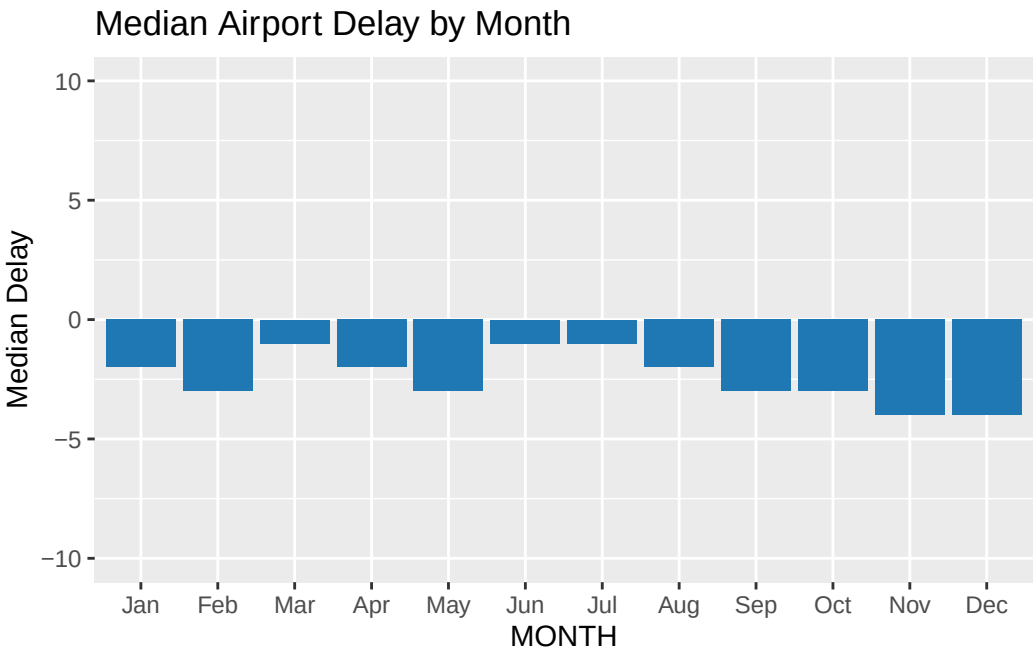
Notice

Examining the graph above shows us a relatively similar distribution for arrival and departure times with the majority of both occuring somewhere between ~7:00 AM and ~8:00PM. However, there is a notable difference at the ends of the graph as it is evident that more departures take place in the morning at ~8:00 AM, while more arrivals take place in the evening ~8:00 PM.

Why might this be?
There may be a multitude of factors that affect the scheduling of flight times, both for departure and arrival. A strong influence could come from passenger and worker preferences. Most people would not like a flight to be scheduled for 3:00 AM in the morning and airports are most likely short-staffed during this time frame as well, which may explain the limited flights during this time for both flight types.

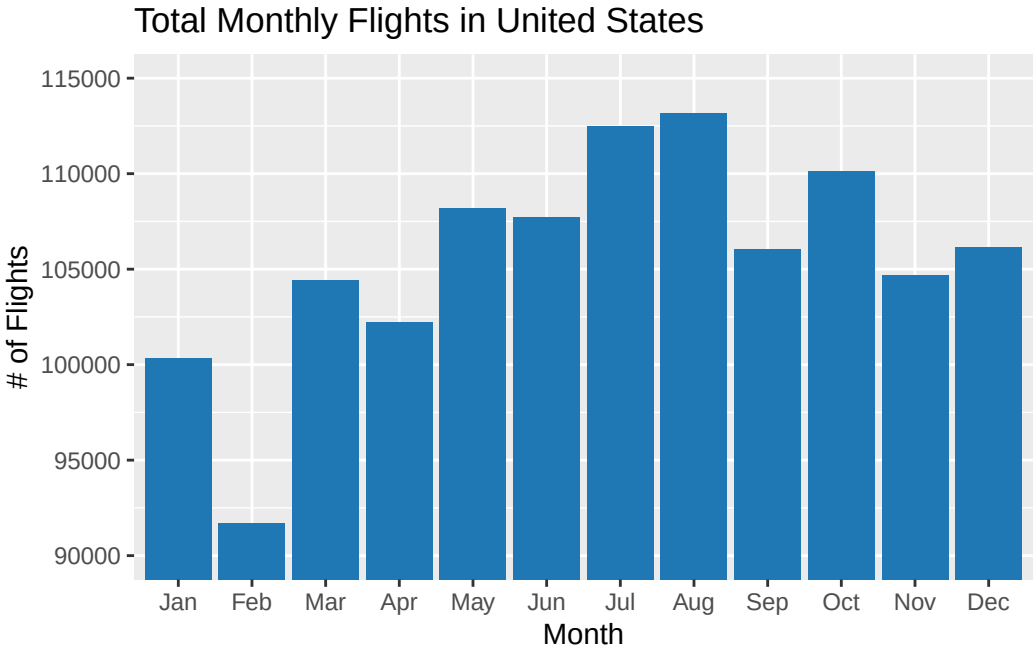
When exploring aviation data for Santa Barbara Routed airports earlier, we mainly focused on seeing if there was any correlation between delay times between airports and depending on flight type: **arrival vs. departure**. We did, however, look at aviation traffic and frequency across seasons so let's see if we'll discover any patterns in airport delays based on season as well. Could it be possible that increased aviation traffic during peak months (Aug-Oct) will in turn lead to more delays as well?





Notice

After plotting the grapsh above that measures delay time at all US-airports across each month, there seems to no noticeable trend in terms of longer or shorter delay times. In fact, the graph seems to show no assocation between month and median delay times. Our prediction earlier did not seem to pan out as October actually has a similar median delay time to January, a month that saw significantly less travel. Meanwhile, months like June & July have a slightly higher value in delay time than their Fall counterparts. (Though these differences can certainly be deemd negligible as they are within 1-2 minutes of eachother) On that note, before we make any inferences on why this may be, lets take a look at travel frequency by month once again, this time for all US-flights.



Notice

When taking all US-Flights into consideration, not just those routed through Santa Barbara, we see noticeable differences in some months. December, in particular, jumps from being the 3rd least traveled month for Santa Barbara Airport to the 6th most over all airports. Other notable changes in July going from 4th least to 2nd most, and May from 7th least to 4th most. But despite these differences, the data does not seem to suggest any correlation between aviation traffic and flight delays.

Scope of Inference

Let's backtrack to our initial maps for flights routed to Santa Barbara Airport and where we posed the question in regards to how those on the East Coast would reach Santa Barbara. Turns out, upon filtering flights off their departure and arrival airport, we've filtered out flights that requiring transit / connecting flights. So while routes exist that can take passengers to and from the East Coast, a connecting flight is necessary.

! Acknowledgements

Data is courtesy of the Bureau of Transportation Statistics, and was accessed from https://www.transtats.bts.gov/.
The data contained in the Airport_Info.csv file was also provided by the Bureau of Transportation Statistics, at the following location: https://geodata.bts.gov/datasets/usdot::aviation-facilities/about.
Special thanks to David Nichols for his website on coloring for colorblindness https://davidmathlogic.com/colorblind/. Accessibility is key to any report and his guidance on the matter helped generate the graphs within this project

Supplementary Material

All course materials, as well as instructions that guided the following project can be found online at: https://ucsb-pstat100.github.io/. Additionally, a data scientist's favorite resource: Google (https://stackoverflow.com/) was crucial in the construction and development of this project.

Appendix

All of the code to replicate the following project and graphs can be found at: https://github.com/billydanggg/Projects