Analysis of U.S. Flight Data from 2015

—

Database Management Final Project Report

Daniel (Dong) Lee, Madeline McKune, Tyler Quast, Galen Vincent

May 1, 2019

Contents

1 Data Description

The dataset we worked with for this project is entitled *2015 Flight Delays and Cancellations*. It contains information about every flight by major US air carriers in 2015. Specifically, it contains information about delays: scheduled take off and arrival time, actual take off and arrival time, etc. This dataset was interesting to us because being delayed for a flight is an experience that most of us have dealt with before. It isn't fun and nobody likes it. We figured it would be interesting to look at some of the factors that contribute to flights being delayed, which we can do using this dataset.

The dataset was obtained from Kaggle (https://www.kaggle.com/usdot/flight-delays), but the data itself came from from the US Department of Transportation's Air Travel Consumer Report from 2015. There are no licenses associated with the data; it is completely open to the public.

Figure 1 shows the entity-relationship diagram we used to store the dataset in the database. The three key entities are `flights`, `airport`, and `airline`. The most important of these is the `flights` entity, which contains over 5 million rows. Each of these rows corresponds to a single flight that took place in 2015. There are a total of 31 attributes associated with each flight. While each of these was stored in the database, we chose to only show the most interesting ones in the ERD to save space. These most interesting attributes are `departure_delay`, `arrival_delay`, date (split into `year`, `month`, `day`, and `day_of_week`), `flight_no`, `tail_no`, `distance`, and `id` as a serial primary key.

The `flights` entity is connected by the "offered on" relationship to the `airline` entity. This entity contains the abbreviated two-letter airline id as a primary key `id`, and also contains the full airline `name`. The "offered on" relationship is 1:N, meaning that one airline can offer many flights, but a specific flight can only be offered on a single airline. These two entities were connected by adding a foreign key constrained `airline_id` attribute to the `flights` entity. The "offered on" relationship is partial participation on the `slights` because a flight may be added to the database which is offered on a smaller airline not contained in the `airline` entity. Similarly, it is partial participation on the `airline` side because an airline may exist which had no recorded `flights` in 2015.

Lastly, the `flights` entity is connected to the `airport` entity through two relationships: "departed from" and "arrived at". The `airport` entity contains attributes for each airport including their three-letter abbreviated id as primary key `id`, their full `name`, and their location (split into `city`, `state`, `country`, `latitude`, and `longitude`). Both relationships are 1:N, meaning that each airport can send off and receive multiple flights, but each flight can only depart from one airport and arrive at one airport. The `airport` and `flights` entities were connected by placing foreign key constrained `destination_airport_id` and `origin_airport_id` attributes in the `flights` table. The connecting relationships are full participation on the `flights` side because a flight needs to have a defined starting and ending location. They are partial participation on the `airport` side because it is possible to have an airport in the database which may not have any recorded flights departing or arriving to it in 2015 (say because of renovations or some other temporary closure).

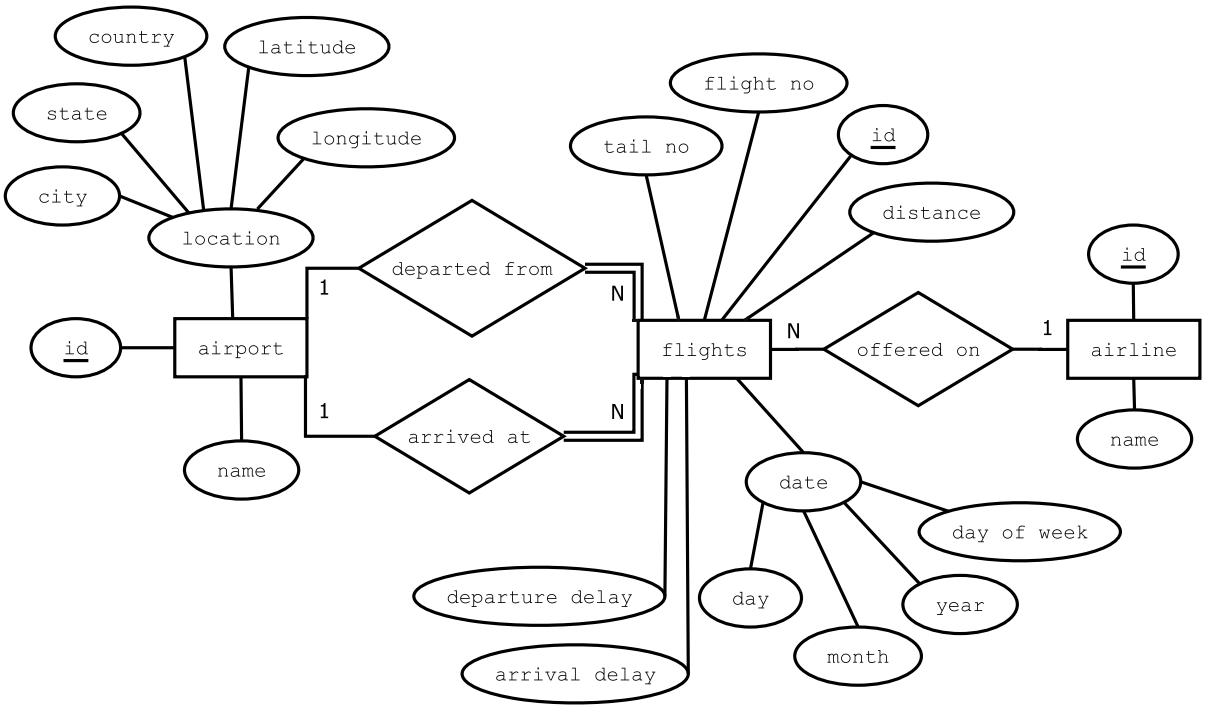

Figure 1: Entity relationship diagram for the organization of data used in this project.

2 Data Analysis

2.1 Delays vs. Location

It is possible that the busiest airports may be the ones with the largest departure delay times. We visualized the single worst delay times and the most popular flight routes to analyze the correlation between these two factors.

We first created a bubble plot to visualize the cities with the single worst delay times in 2015. This visualization is shown in Figure 2. The plot was created using the `ggplot2` library in R and a query that selected the maximum departure delay time for each airport in the database (this query can be found in the appendix). Delay time corresponds to bubble size and color so that the largest, most yellow bubble represents the longest delay. Also, we only selected the 200 worst airports because the plot became overcrowded when we mapped every airport. City names are shown for the top 15 delays of 2015.

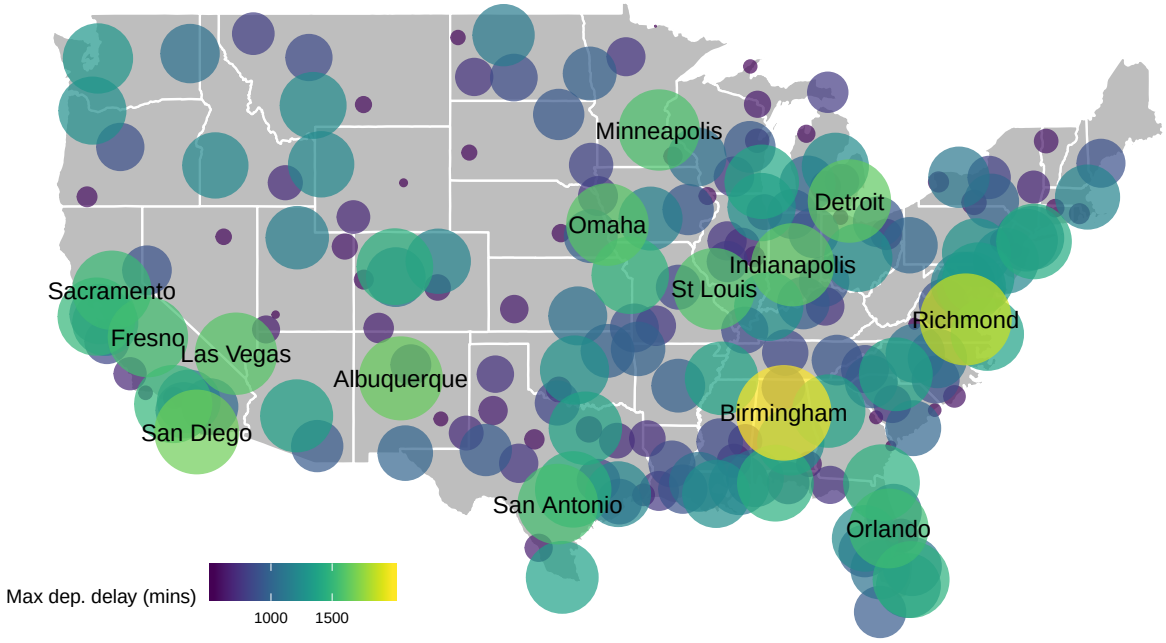

Figure 2: Map showing the cities with the worst single delay times in 2015.

As we can see from Figure 2, the majority of the bad delays were located on the Eastern half of the United States. Eastern airports likely have worse delays because of higher population density, or because of blizzards and other extreme weather.

California, however, had the most airports with the *worst* delays. California is a vertical state and is notorious for bad traffic, so it is likely that Californians prefer flying within the state to driving. Aside from being a popular tourist destination, California is also located on the West Coast, so it is likely people catch connecting flights from California to Canada, Asia, Mexico, or Hawaii, which are all factors that could contribute to the delays.

Next we plotted the most popular flight routes using a flight path map. This visualization is shown in Figure 3. The plot was created in Python using the `plotly` and `ScatterGeo` libraries. We first queried the database for the starting and ending airport's longitudes, latitudes, names, and cities. We then counted the number of flights from starting airport to ending airport and selected the top 100 most popular flight paths. In Figure 3, number of flights corresponds to path thickness and opacity so that the most popular flights are the thickest and darkest. We displayed city names on the top 10 most popular departure cities.

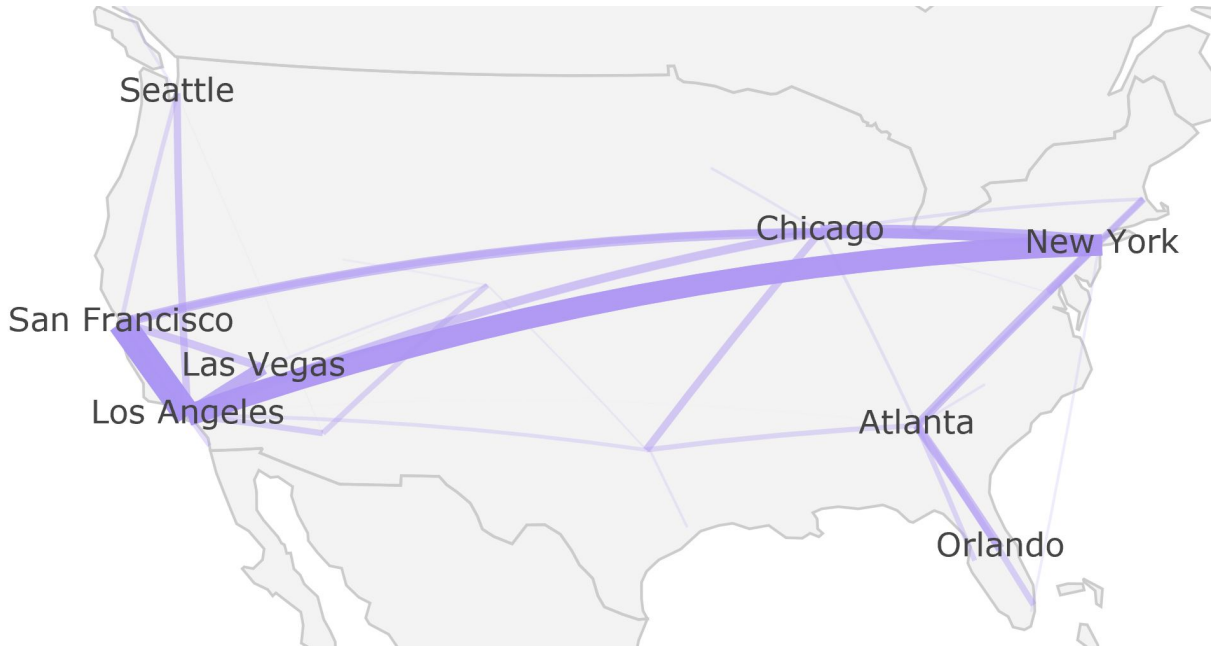

Figure 3: Flight path map of the most popular routes in 2015.

The two most popular routes in 2015 were Los Angeles to San Francisco and Los Angeles to New York. As mentioned earlier, it is likely that people frequently fly between California cities, so that would explain why an in-state flight was the most popular route. Also, Los Angeles and New York are both business hubs, so this flight is likely popular for work reasons.

There was not a direct correlation between the worst delay times and the most popular cities to fly to and from. We found that Birmingham and Richmond had the single worst delays but were not included in the most flown routes. Also, Seattle, Atlanta, Chicago, and New York were some of the most popular airports but had shorter worst departure delays than less frequently traveled airports. This could potentially be a result of the fact that more traveled airports are better equipped to handle heavy traffic and problems without having to delay flights.

2.2 Delays on Key Dates

The number of departing flights increases on important dates of the year due to high demand. Delays on these key dates are especially stressful for passengers because they are likely headed home to family or to important events. The five key dates we decided to analyze are New Years Day (1/1/2015), a day in mid may to represent when college students might fly home (5/15/2015), the Fourth of July (7/4/2015), a day in mid august to represent when college students might fly back to school (8/15/2015), and Christmas (12/25/2015).

Figure 4 shows the average delay *of all flights that were delayed* on these dates. So, these are averages of only flights which took off after when they were scheduled to take off. August 15th had the longest average delay of around 50 minutes per flight. The other dates averaged delays of around 30 minutes.

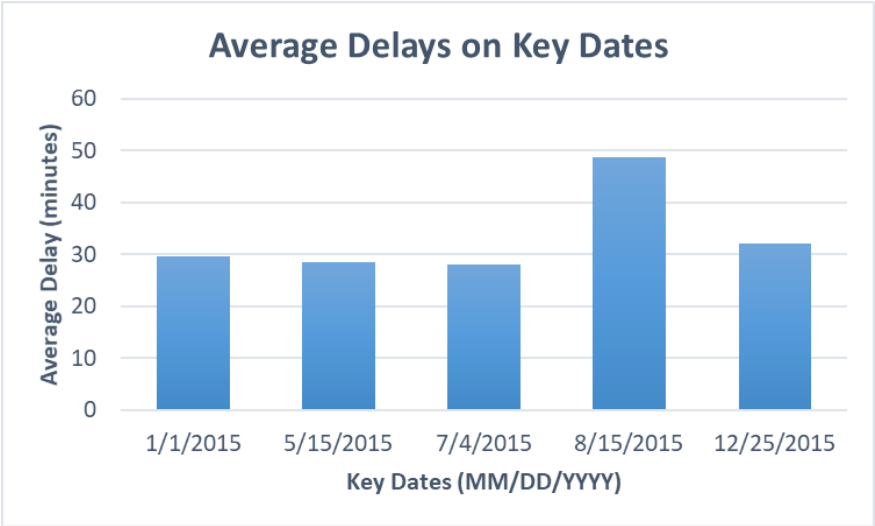

Figure 4: Average delay duration (minutes) of late flights on key dates.

We also looked at the average early departure time for each date. Once again, these are averages only of flights which took off before their scheduled take off time. These averages are shown in Figure 5. Overall, the average early departures were all around the same length. When comparing the delays to the early departures, it is clear that delays are on average much longer than early departures. Either of these situations is undesirable, as early departing flights could leave people behind, and delayed flights make people late to things.

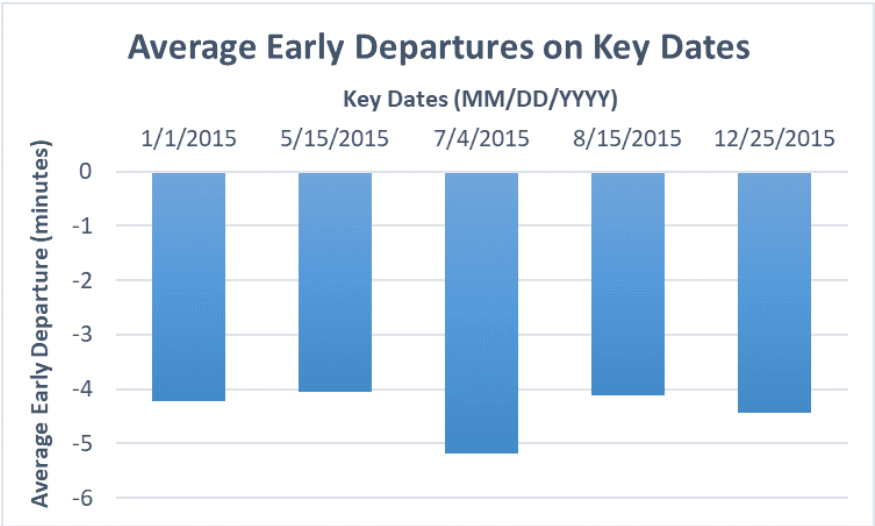

Figure 5: Average length (minutes) of early flight departures on key dates.

Figures 6 and 7 show the longest delays and earliest departures, respectively, on each of the key dates. Figure 6 shows the longest delay of these dates took place on New Years Day, when a flight was delayed 1150 minutes (20ish hours). Following close in second place was Christmas, with a worst delay of 1150 minutes (19ish hours). These delays were likely

due to harsh weather or large crowds from the holidays. The shortest delay of these dates occurred on 5/15, and was around 800 minutes (13ish hours).

Transitioning to Figure 7, we see that the earliest departure occurred on Christmas, where a flight left 55 minutes early. From this analysis, we can confirm that Christmas is a strange day for traveling, with long delays and even early departures.

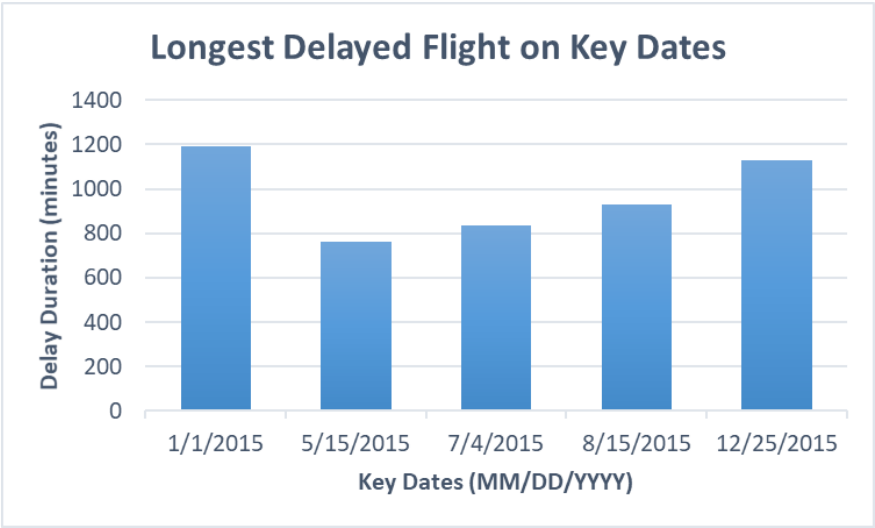

Figure 6: Longest delay duration (minutes) on key dates.

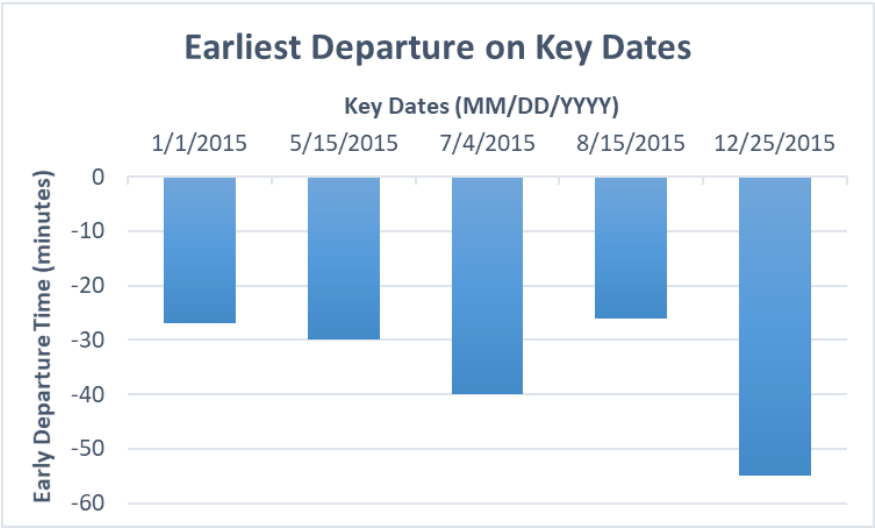

Figure 7: Earliest time before intended departure (minutes).

Lastly, we looked at the number of flights that left on time. Figure 8 shows a bar chart indicating the number of such flights on each key date. From this figure, we can see that 5/15 was the date with the most on time flights. The day with the fewest number of on time flights was the Fourth of July.

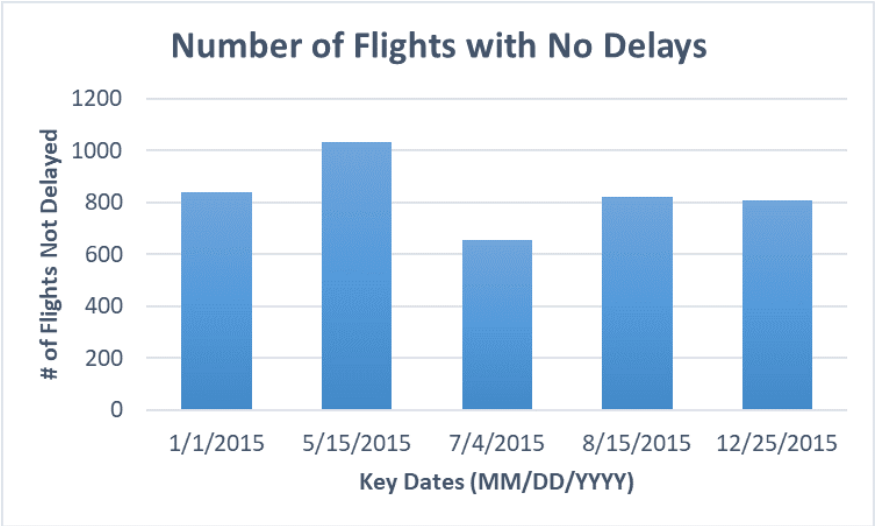

Figure 8: Number of flights without a delay on key dates.

2.3 What Attributes Cause Delays?

There are many factors that airlines and flight planners have to consider when organizing air travel. These factors uniquely effect the potential to add delay to travelers flights. In this section, we will take a look at various factors that impact delay time. All of this data was analyzed in python by querying the database and parsing the data to produce the following plots.

We first analyzed how different months affect delay times. The ideology behind this is that months with bad weather might have more delays. Figure 9 shows a plot of the average delay time for each months of the year. It is evident that there is no large correlation between month and delay time. The largest average delay is just under 14 minutes. Another interesting fact is that the data set does not include data from the month of October. Why the data set did not include this information is unknown, but we believe it is likely an error on the part of the person who published the data to Kaggle.

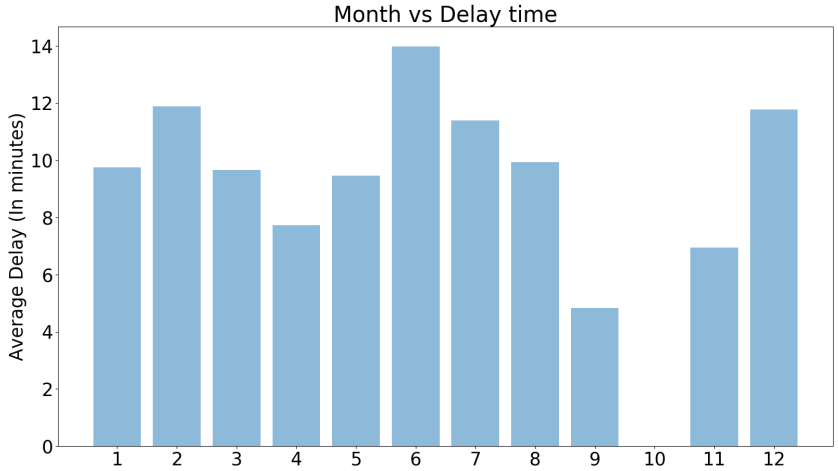

Figure 9: Bar graph depicting month vs delay time.

Next we analyzed airline versus delay time. Figure 10 shows more indication of correlation to delay time than the months did. There is a much larger spread of delay times, with some airlines sitting very close to 0, and some airlines sitting far from it. Hawaiian Airlines shows the lowest average delay time of under one minute, proving this company cares about its ability to get fliers air born on time. Companies like Spirit and Frontier, which are budget airlines, seem to have the largest delay times. These companies tend to sacrifice reliability, comfort, and quality in exchange for lower ticket prices. The rest of the companies reside between these two extremes, with the average delay time of all airlines being roughly 10 minutes.

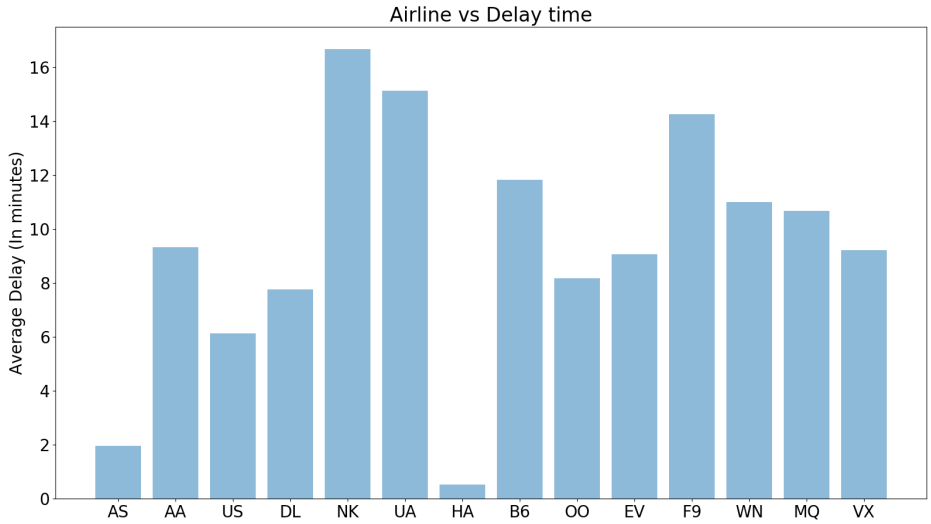

Figure 10: Bar graph depicting airline vs delay time.

The final relationship we looked at in this section was between airports and delay time. Airports could impact delays for a variety of reasons including their location, climate, or popularity. The data set provided dozens of airports, so for simplicity we only looked at some of the outliers.

To begin, Figure 11 shows average delay time from some airports including the one with the worst average delay of all: the Jack Brooks Regional Airport, located in Texas. We do not know why this airport had the highest delay time in 2015, but its clearly much worse than other airports in the same class. The other airports averaged around 9 minutes for their average delays.

Both Figure 11 and Figure 12 contain some interesting data points where airports have an average delay less than zero. This implies that on average these airports get fliers out earlier than scheduled. The EKO or Elko Regional Airport has the earliest average departure of about 5 minutes ahead of schedule. While this seem like a good thing, it could prove to be confusing to fliers who plan to leave at a certain time, show up, and realize the flight has already left. However, for the average traveler, leaving early sounds like a great alternative to leaving late.

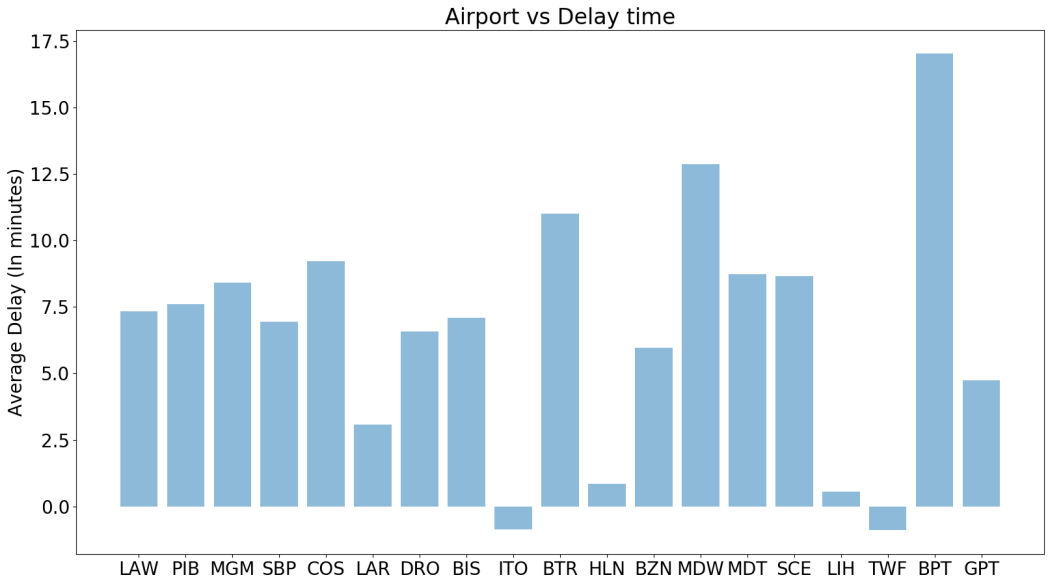

Figure 11: Bar graph depicting airport vs delay time of all the data, including the largest value.

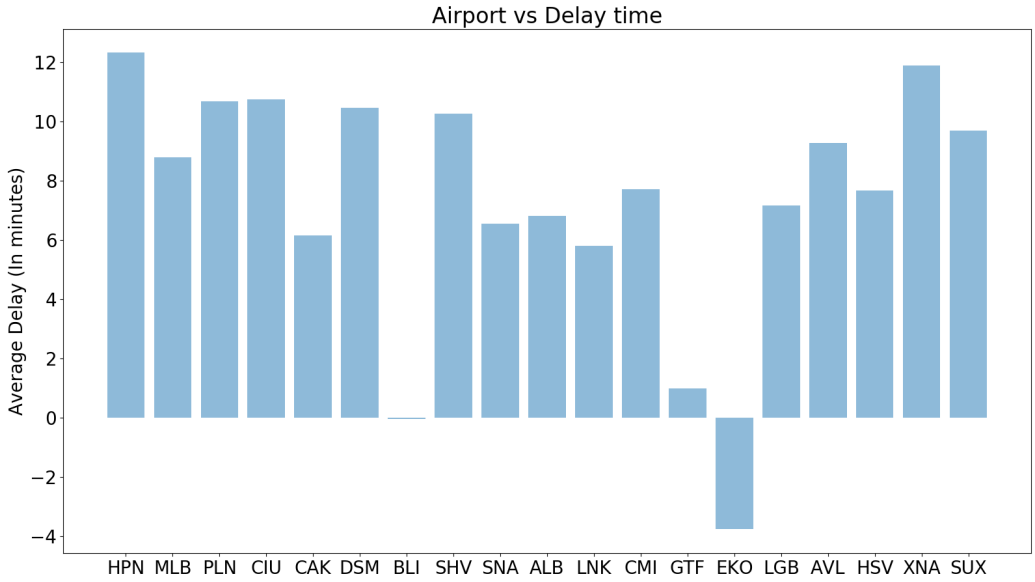

Figure 12: Bar graph depicting airport vs delay of all the data, including the smallest value.

2.4 Denver Airport Analysis

Living in and around Golden, most people at Mines are probably the most impacted by flights associated with Denver International Airport (DIA). Because of this, we thought it would be interesting to look at some of the flight data from DIA.

First, we looked at the most popular days of the week to fly out of DIA. Figure 13 shows a bar graph of the number of departures from DIA on each day of the week. It appears that there is a significant drop in number of departing flights on Saturdays. This could be due to the fact that there are less business travelers on Saturday, or that most people who leave on weekend trips fly out on Friday and return on Sunday. Either way, if you are looking to fly out of DIA on the least crowded day of the week, consider doing it on a Saturday.

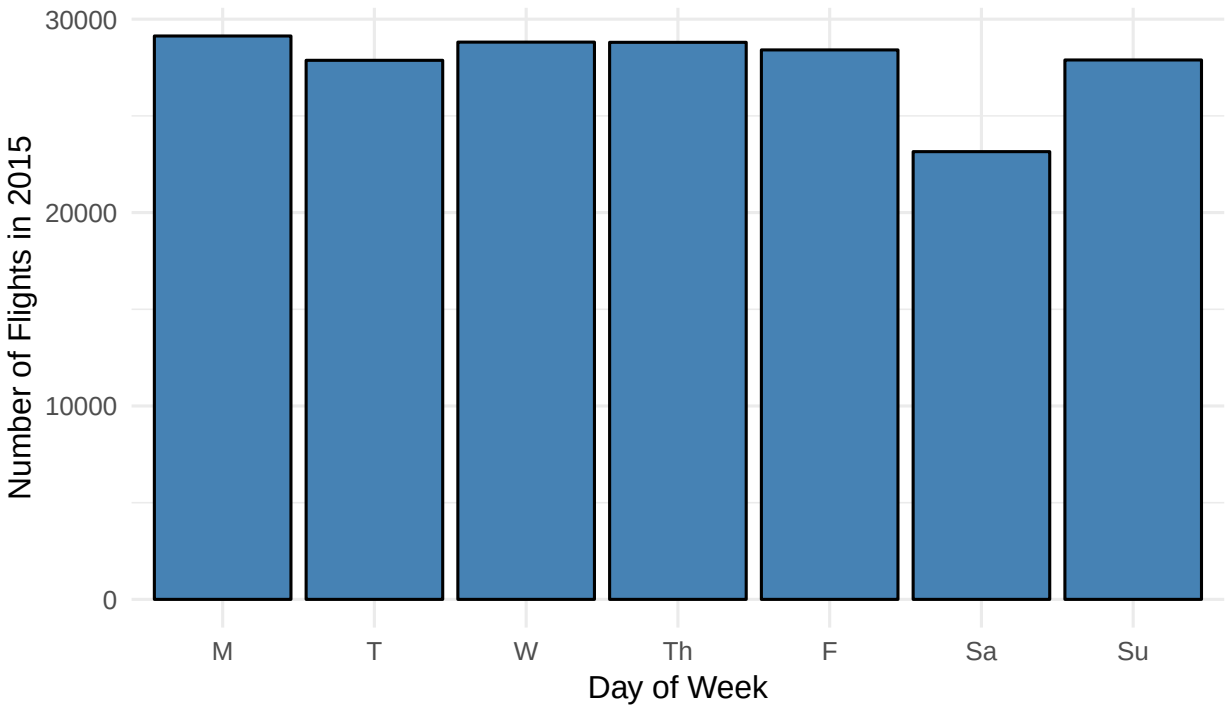

Figure 13: Bar chart of flight departures for each day of the week for Denver International Airport in 2015.

Next, we decided to look at what time of day is most popular for departures from DIA. Figure 14 shows a histogram of departure times for all of 2015, split up into hour increments. From this Figure, it seems that 10AM - 12PM, 3PM, and 7PM are the most popular times to fly. If you are looking to avoid crowds at the airport, consider scheduling flights late at night, or in one of the other apparent lulls throughout the day.

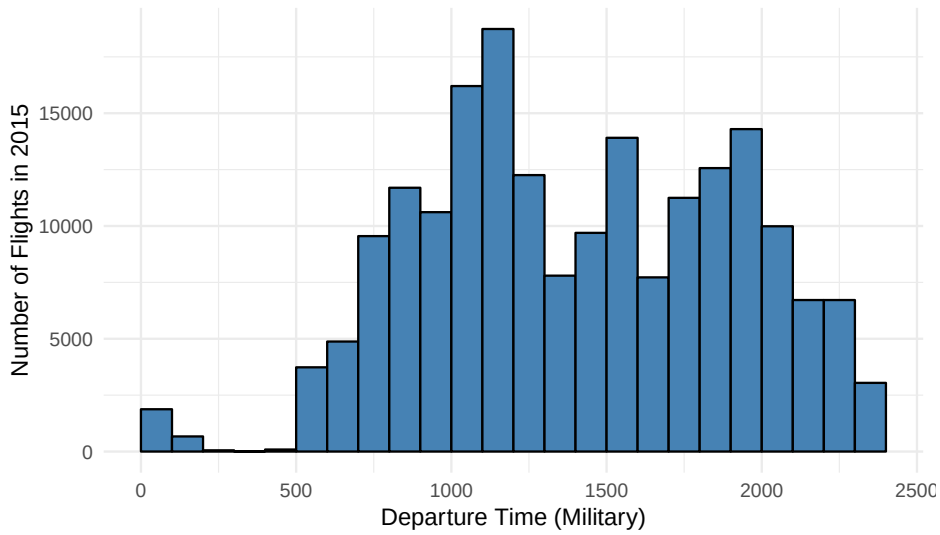

Figure 14: Histogram of flight departure times for Denver International Airport in 2015. Note that the x-axis is in military time. Bins are split up by hour.

Lastly, we decided to look at the most popular cities to fly to from Denver, based on total number of flights departing from DIA in 2015. These results are presented in Figure 15, which shows a bubble map of the most popular destination cities. In this plot, the number of outgoing flights from DIA is represented by both bubble size and color of the bubble. The top 15 most popular cities have their bubbles labeled with their name. Chicago, Phoenix, Las Vegas, Houston, and LA are at the top of the list.

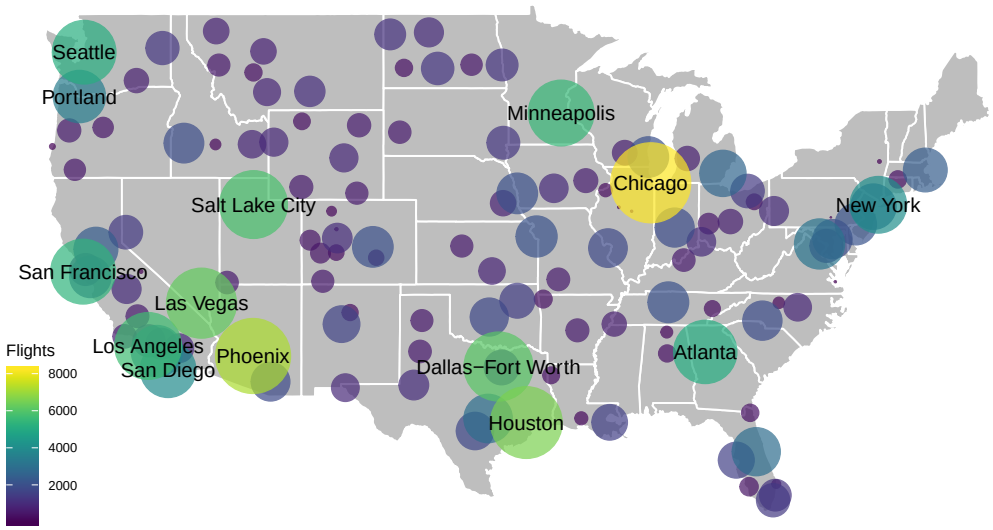

Figure 15: Most popular flight destinations from Denver International Airport. Color and size indicate number of outgoing flights in 2015.

3 Challenges

One challenge we faced had to do with getting our data into the database. The CSV containing the flight data has over 5 million rows, so uploading to the database was a non-trivial task which was taking a significant amount of time. One problem was that we were originally trying to upload the data from off-campus using a VPN, and the network latency really slowed down the data import (it took about 15 minutes). To fix this problem, we figured out a few things. First, we used the psql `\COPY` command to get the data from CSV to the flowers database. Next, we figured out that applying the foreign key constraints to the imported tables after uploading all of the data was faster than creating the tables with them and making the database check those constraints while importing the data. Lastly, we figured out that being on campus while running the code made things significantly faster. In the end, we got the total upload time down to 2 or 3 minutes.

A second challenge that we faced was deciding how we would pull the data from the database to analyze. Some of the group decided to use python. Specifically, we used the libraries `psycopg` to connect to the database and `matplotlib` to visualize the data. Other people in the group decided to export their queries to csv files and import them into excel to visualize. Lastly, some used `R` with the `RPostgreSQL` and `ggplot2` libraries. Deciding on languages and libraries to use can sometimes be a problem, but in our case, the splitting up of the group into different languages based on our individual skills worked well.

4 Contributions

Section	Author
Data Description	Galen Vincent
Delays vs. Location	Madeline McKune
Delays on Key Dates	Daniel Lee
What Attributes Cause Delays?	Tyler Quast
Denver Airport Analysis	Galen Vincent
Challenges	Everyone

5 Appendix - Queries

Included below are the queries that we used to both import data into the database with the desired structure and pull data from the database to do the analysis in section 2.

5.1 Importing Data to Database

This is the code we used to import the data into the database from the three csv files containing the rows for the flight entity, the airport entity, and the airline entity. After creating the tables with the correct columns and data types in the database (this code is long so we won't include it), these are the commands:

```
\COPY airline FROM 'airlines.csv' WITH CSV HEADER;

\COPY airport FROM 'airports.csv' WITH CSV HEADER;

\COPY flights (year, month, day, day_of_week, airline_id, flight_no,
tail_no, origin_airport_id, dest_airport_id, scheduled_dep_time,
dep_time, dep_delay, taxi_out_time, wheels_off_time, scheduled_time,
elapsed_time, air_time, distance, wheels_on_time, taxi_in_time,
scheduled_arr_time, arr_time, arr_delay, diverted, cancelled,
cancel_reason, air_system_delay, security_delay, airline_delay,
late_aircraft_delay, weather_delay) FROM 'flights.csv'
WITH CSV HEADER;

-- get rid of origin airports with numbers as IDs instead of letters
DELETE from flights where length(origin_airport_id) != 3;

-- add foreign constraints
ALTER TABLE flights ADD CONSTRAINT constraint_fkey1
FOREIGN KEY (airline_id) REFERENCES airline (id);

ALTER TABLE flights ADD CONSTRAINT constraint_fkey2
FOREIGN KEY (origin_airport_id) REFERENCES airport (id);

ALTER TABLE flights ADD CONSTRAINT constraint_fkey3
FOREIGN KEY (dest_airport_id) REFERENCES airport (id);
```

5.2 Delays Versus Location

For this section, we first created a separate table to store the worst delays from each airport, and populated this table with the associated data from the database. This was done with the following SQL commands:

```
CREATE TABLE worst_delays(
id TEXT ,
dep_delay INTEGER,
PRIMARY KEY (id));

INSERT INTO worst_delays (id, dep_delay)
SELECT origin_airport_id, MAX(dep_delay)
FROM flights GROUP BY origin_airport_id;
```

We then extracted the city names, worst delay, and city longitude and latitude from the databse by joining our new table with the airport table. Here is the query:

```
SELECT port.city, w.dep_delay, port.longitude, port.latitude
FROM worst_delays AS w, airport AS port WHERE w.id = port.id
ORDER BY w.dep_delay DESC LIMIT 200;
```

5.3 Delays on Key Dates

To make the charts for delays and early departures on key dates, we queried data from the database into csv files and then imported them into excel to visualize. Below, the queries for only one date are shown, but these queries were repeated to gather data for each of the other dates.

The first query was to get all the flights where there was a delay, along with the length of these delays. The second query was to get all the flights where there was an early departure, along with the magnitude of this early departure. The third query was to count the number of on time flights.

```
\COPY (SELECT month, day, year, dep_delay
FROM flights WHERE dep_delay > 1 AND month = 12 AND day = 25)
TO 'christmas_delay.csv' DELIMITER ',' CSV HEADER;

\COPY (SELECT month, day, year, dep_delay
FROM flights  WHERE dep_delay < 1 AND month = 12 AND day = 25)
TO 'christmas_early.csv' DELIMITER ',' CSV HEADER;

SELECT COUNT(dep_delay)
FROM flights WHERE dep_delay = 0 AND month = 12 AND day = 25;
```

5.4 What Attributes Cause Delays?

Here are the queries we used to extract data about various attributes and their connection to delays from the database:

```
CREATE TABLE freq_flights(
arr_id TEXT,
dep_id TEXT,
```

```
freq INTEGER,
PRIMARY KEY (arr_id, dep_id));

INSERT INTO freq_flights (arr_id, dep_id, freq)
SELECT origin_airport_id, dest_airport_id, COUNT(*)
FROM flights  GROUP BY origin_airport_id, dest_airport_id;

CREATE TABLE flight_csv(
dep_lon numeric,
dep_lat numeric,
arr_lon numeric,
arr_lat numeric,
airport TEXT,
nb_flights INTEGER,
PRIMARY KEY (arr_lon, dep_lon, nb_flights));

INSERT INTO flight_csv
SELECT dep.latitude, dep.longitude, arr.latitude, arr.longitude,
dep.city, f.freq
FROM freq_flights AS f, airport AS arr, airport as dep
WHERE f.arr_id = arr.id AND f.dep_id = dep.id
AND dep.longitude IS NOT NULL AND arr.longitude IS NOT NULL
AND dep.latitude IS NOT NULL AND arr.latitude IS NOT NULL
ORDER BY f.freq DESC;
```

5.5 Denver Airport Analysis

The query for getting number of flights for each day of the week leaving DIA in 2015 was:

```
SELECT f.day_of_week, COUNT(*) FROM flights f, airport a
WHERE f.origin_airport_id = a.id AND a.city = 'Denver'
GROUP BY day_of_week;
```

The query used to get the data for the histogram of departure times from DIA in 2015 was:

```
SELECT f.dep_time FROM flights f, airport a
WHERE f.origin_airport_id = a.id AND a.city = 'Denver';
```

Lastly, the query used to get the data for most popular flight destinations from DIA was:

```
SELECT a.name, a.city, f.freq, a.state, a.longitude, a.latitude FROM
(SELECT a2.id dest_city_id, COUNT(*) freq
FROM flights f, airport a1, airport a2
WHERE f.origin_airport_id = a1.id AND a1.city = 'Denver'
AND f.dest_airport_id = a2.id GROUP BY dest_city_id) f, airport a,
WHERE f.dest_city_id = a.id
AND a.state NOT IN ('HI', 'AK') AND country = 'USA';
```