Operation On-Time: The Quest to Enhance Airline Efficiency

Fletcher Newman | fletcht13@tamu.edu | 332000276

Intro

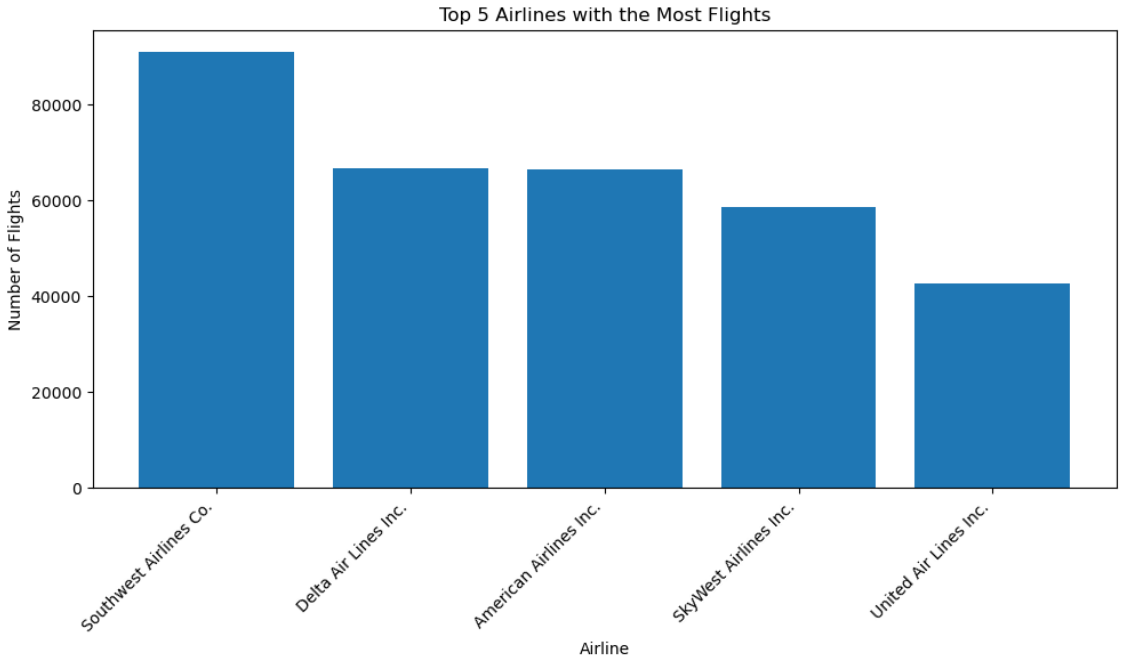
This report discusses observations made from analyzing the Reporting Carrier On-Time Performance (1987-present) data table of the "On-Time" database from the TranStats data library, specifically in January 2022. The dataset was uploaded to a pyspark data frame, analyzed using different queries in pyspark SQL, and visualized by using matplotlib to create graphs for better understanding of the results of the queries.

Results

Most Popular Airlines

First, a query was taken to find the top 5 most popular airlines in terms of flight density. Below is the resulting bar graph.

```
+-----------------+-----------+
|Reporting_Airline|FlightCount|
+-----------------+-----------+
|               WN|      90859|
|               DL|      66573|
|               AA|      66370|
|               OO|      58651|
|               UA|      42744|
+-----------------+-----------+
```



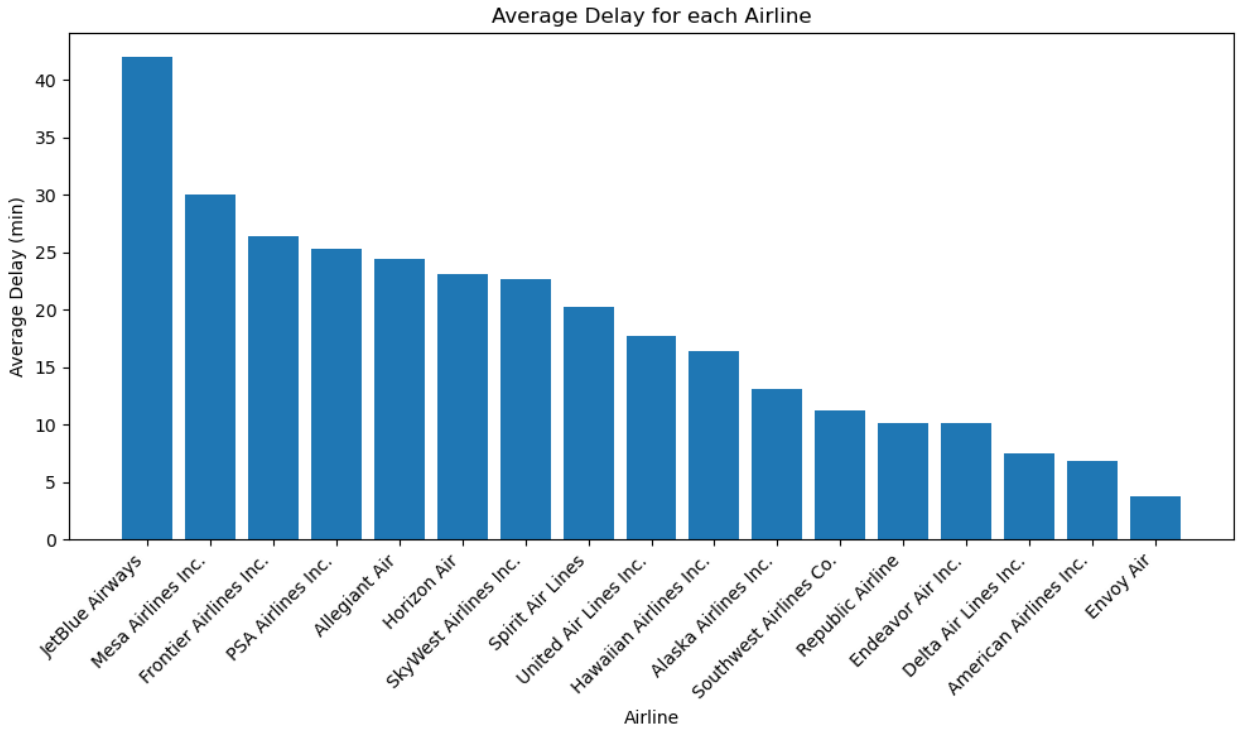
It appears that Southwest Airlines is by far the most popular airline in terms of flight density by a relatively large margin. Delta Air Lines and American Airlines are practically equivalent as well, with SkyWest Airlines close behind and another relatively large drop off for United Airlines.

Average Flight Delays

Next, the average flight delay for each airline will be visualized through a query. Every airline in the dataset will be taken into consideration (17 in total) to get a full comparison of the range of airlines and the average delays. The bar graph from this query can be seen below.

```
+-----------------+-------------------+
|Reporting_Airline| Average_Delay_min|
+-----------------+-------------------+
|               B6| 41.97723009587328|
|               YV|30.085887096774194|
|               F9|26.455685181960646|
|               OH|25.293724674186837|
|               G4|24.393228840125392|
|               QX| 23.05726238724016|
|               OO|22.685120458304205|
|               NK|20.248272603791413|
|               UA|17.761112670784204|
|               HA|16.389893617021276|
|               AS|13.122665088134701|
|               WN|11.267579436269385|
|               YX|10.176639934196997|
|               9E| 10.09348931339158|
```

```
|                DL| 7.539107445961576|
|                AA|  6.81136055446738|
|                MQ|3.7098199748802156|
+------------------+------------------+
```



In this graph, it can be seen why most people fly with the above most popular airlines. Four out of the five top airlines fall below the 2nd quartile in terms of average flight delay, with American and Delta having the second and third shortest average flight delay. Customers would be more satisfied with one of these airlines that has less delays and thus more likely to fly with them again. On the other hand, it can be stated that since these airlines have more flights than the others, their average flight delay will tend towards the true mean than the others due to the central limit theorem. This is a valid argument, but could be countered by stating that if this were true, Delta and American would not be so far to one side of the graph relative to the other airlines and would favor the middle of the pack in terms of average flight delay (unless all other airlines contain heavy outliers that the top airlines do not, which is very unlikely).

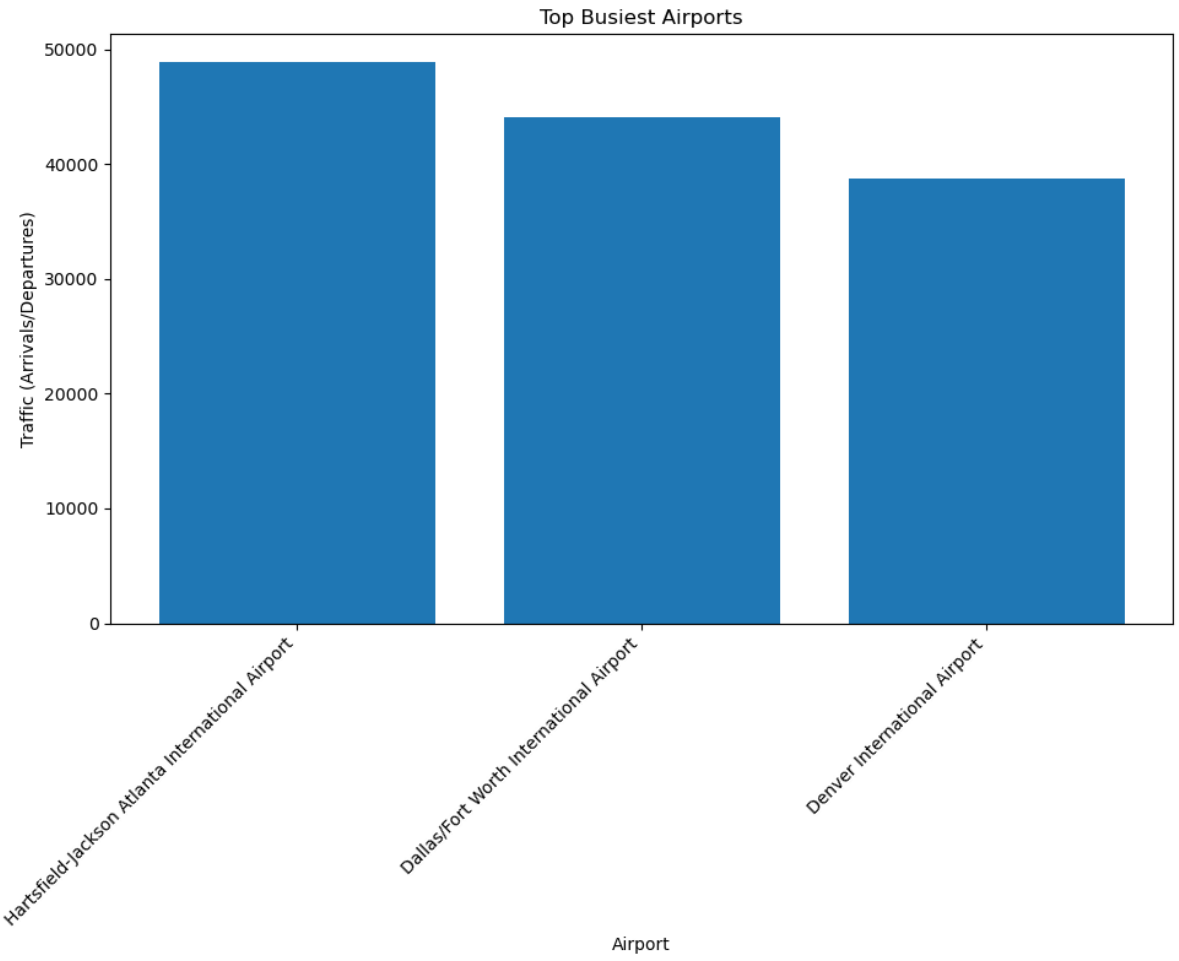
Another observation that can be made from this graph is the very high flight delay for JetBlue. JetBlue seems to have the only flight delay that does not fit into the linear trend of all the other airlines, with the difference between JetBlue and the next highest average delay airline being much higher than any other airline. In this dataset, JetBlue has 19,192 listed flights, which is more than enough flights to smooth out any outlier flight delays if there are any. From this data, it can be concluded that significant improvements to JetBlue's flight efficiency would have

a large positive effect on the total average flight delays across all airlines, improving airline efficiency.

Busiest Airports and Cities

Next, the top three busiest airports will be analyzed. This was determined by combining the number of departures and arrivals to each airport as total traffic and taking the three highest ranking airports. Below is the table and graph produced from this query.

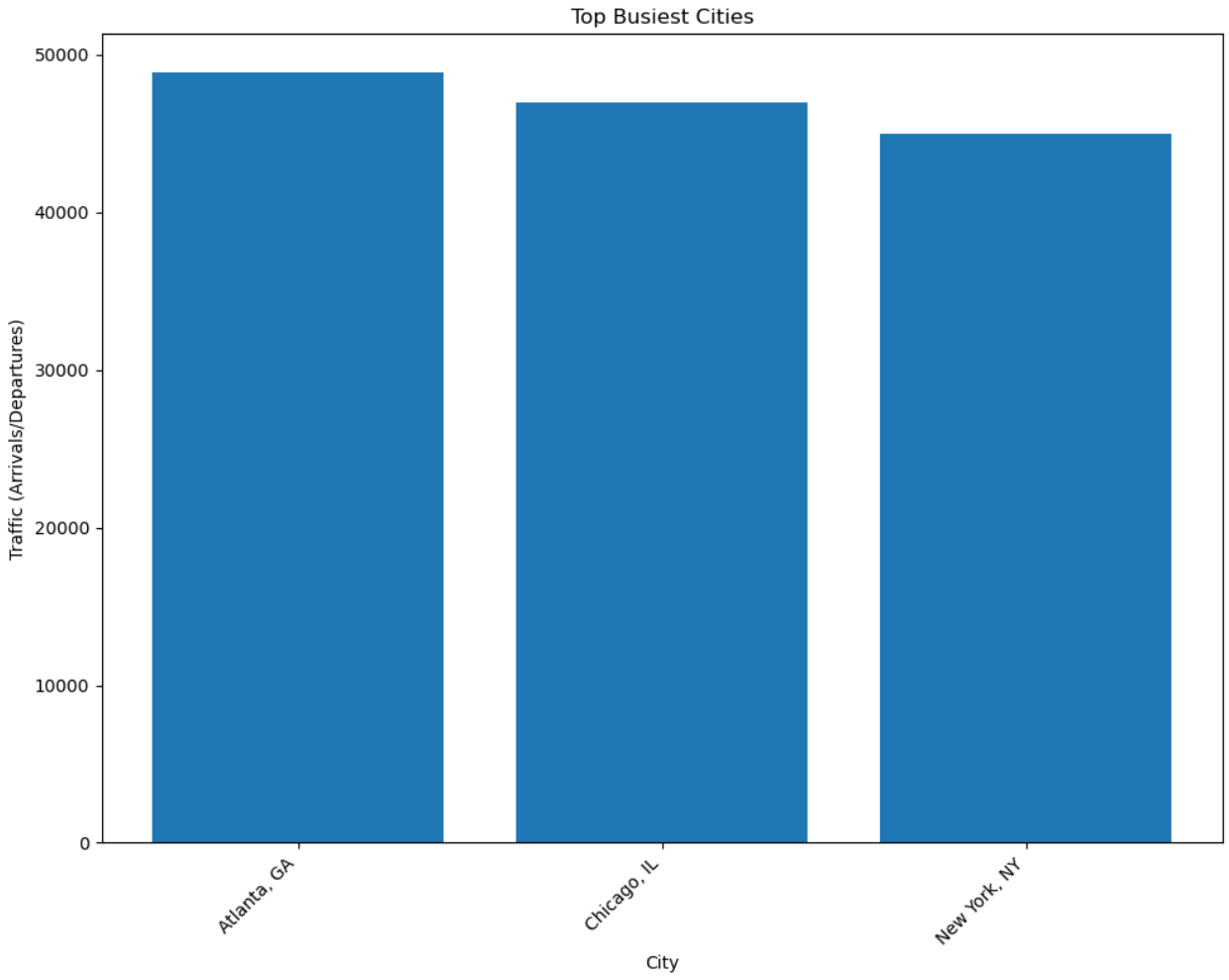
Airport	totalTraffic
ATL	48863
DFW	44056
DEN	38708



From the graph, we observe that Atlanta, Dallas, and Denver all have airports with very heavy air traffic. Considering all three of these cities are very large cities in terms of population, economy, and tourism, this makes sense that these cities are the most popular in terms of air traffic (though it was a shock to see that New York was not included in this top three). The insights from this graph will be used more in depth in the last query concerning the routes with the highest average delay.

The same query was made again, but this time with the city names instead of the airport names. This was done to show the most popular city itself, which now gives more weight to cities that have two popular airports that split the arrivals and departures to that city.

```
+------------+------------+
|        City|totalTraffic|
+------------+------------+
| Atlanta, GA|       48863|
| Chicago, IL|       46994|
|New York, NY|       45020|
+------------+------------+
```



In this graph, the ‘thought-to-be’ most popular airports (Chicago and New York) now show up as the second and third most popular destinations. Atlanta still sits at the most popular destination, which shows that there is only one major airport in Atlanta. The difference in air traffic between the top three cities is much closer than the difference between the top three airports. Again, this insight will be further used in later queries.

Traffic and Time of Day

Next, the delay associated with the time of day will be analyzed. This was done by splitting every flight into four categories:

Morning: 6:00am – 12:00pm

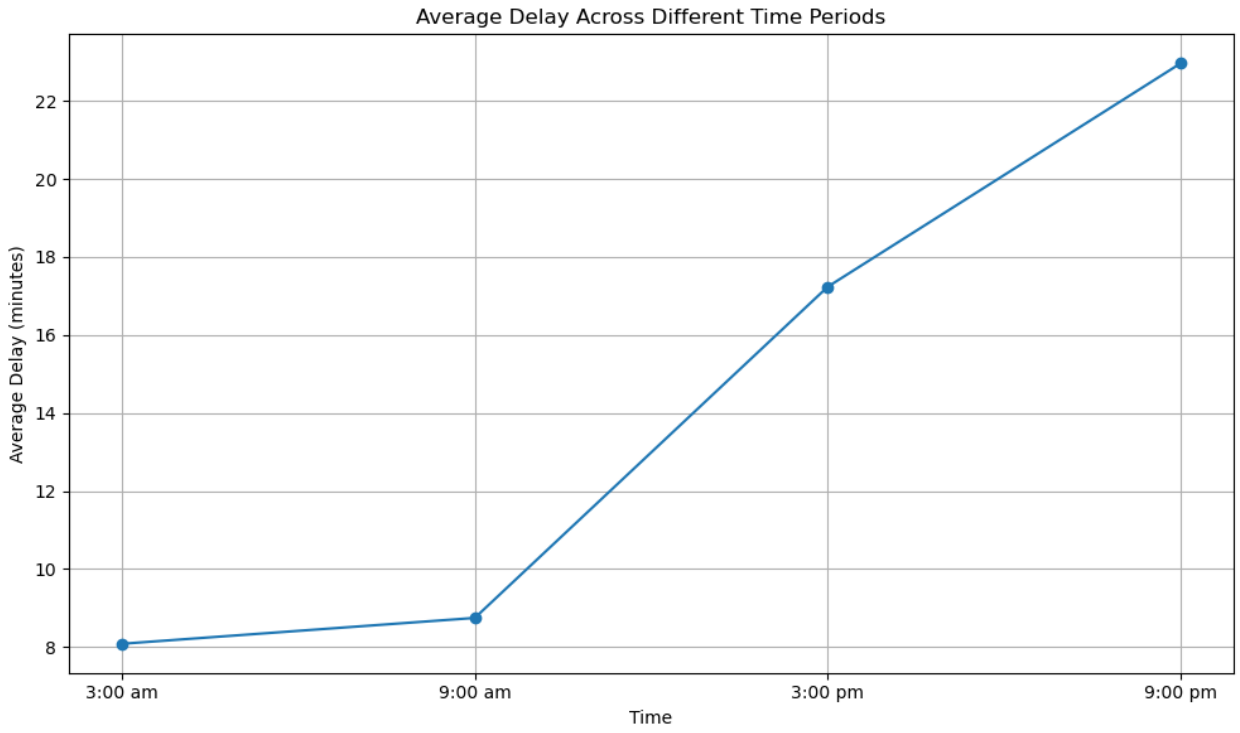
Afternoon: 12:00pm – 6:00 pm

Evening: 6:00pm – 12:00am

Night: 12:00am – 6:00am

A query for the average delay for all the flights falling into each category was made and graphed as seen below.

```
+----------+------------------+
|TimePeriod|          AvgDelay|
+----------+------------------+
|   Evening|22.965202306718552|
| Afternoon|17.237314086922648|
|   Morning| 8.749632061887988|
|     Night| 8.086150490730644|
+----------+------------------+
```

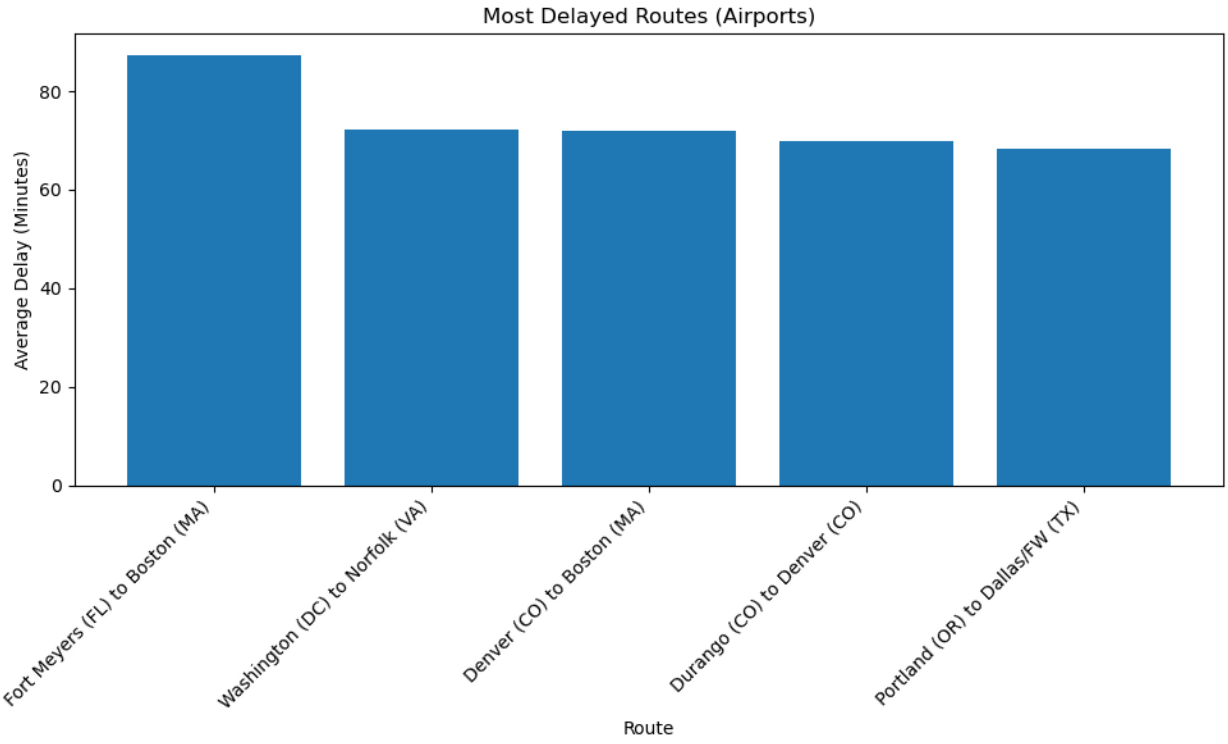


From the graph, it appears that flight delays start the day off with very little delays and sharply increase to high delays by the end of the day. This is intuitive, because a delay for one flight will affect the departure or arrival time of the flight after it, either because another plane needs to use the same terminal that is occupied by a delayed flight, or because that same plane is doing another flight (or just a layover) to another destination as soon as the passengers leave which is very common in air travel. Therefore, the delays throughout the day will naturally snowball and build up until the end of the day when the flight density is much lower. A system to help fix this phenomenon would dramatically lessen delays and make air travel much more efficient.

Average Delay of Routes

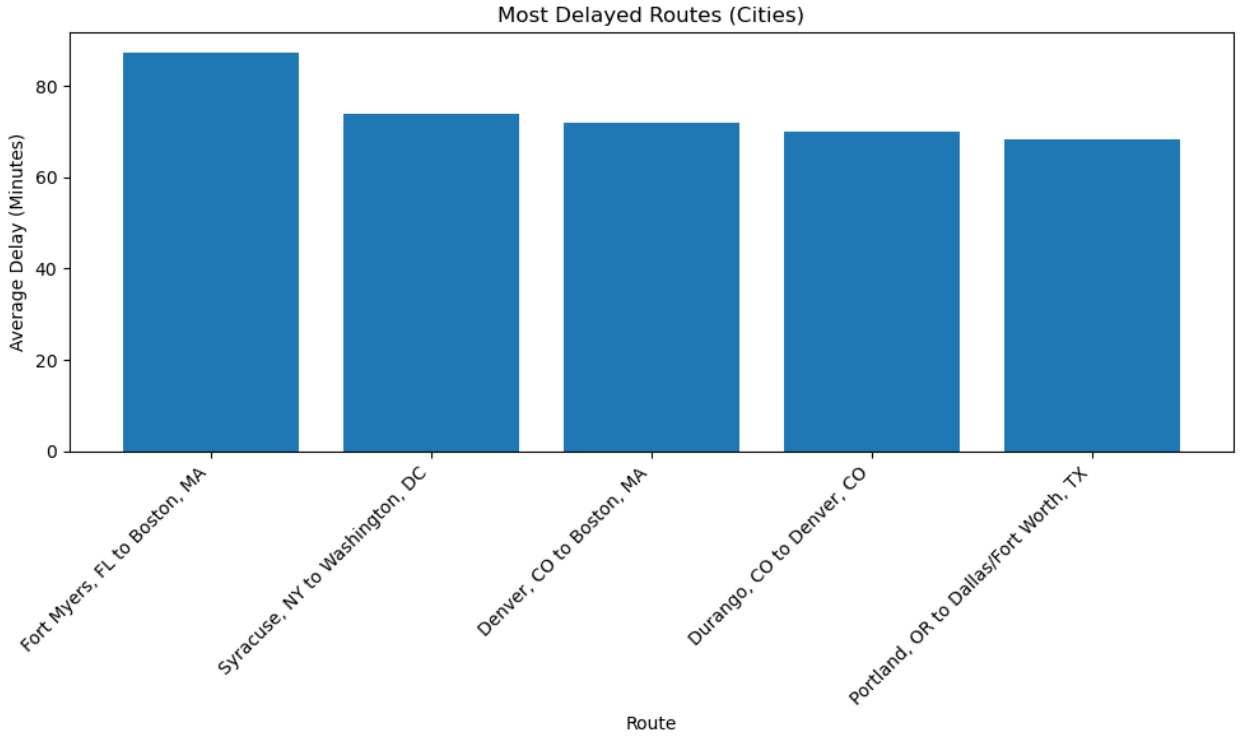
Last, a query was made to analyze the routes (origin and destination airport pairs) that have the longest average delay. This query was limited to routes with over 100 flights to filter out outlier routes that only have had a few flights on them with large delays. The threshold of 100 was determined based off the number of flights to and from airport and city locations analyzed in the 'Busiest Airports and Cities' section of this report and was determined to be an appropriate threshold to minimize outlier data while also not filtering out valid data and phenomena that should be observed. The resulting table and graph can be seen below.

Departing_Airport	Destination_Airport	Average_Delay	Flight_Count
RSW	BOS	87.27922077922078	154
DCA	ORF	72.29807692307692	104
DEN	BOS	71.86705202312139	173
DRO	DEN	69.94117647058823	102
PDX	DFW	68.328125	128



Before this data is analyzed, the same query will be made, but with the city names as departure and destination location instead of airport names. Similar to analyzing high density traffic areas, this is to include any routes that also have heavy delays but might take off from different airports in the same cities. The results can be seen below.

Departing_City	Destination_City	Average_Delay	Flight_Count
Fort Myers, FL	Boston, MA	87.27922077922078	154
Syracuse, NY	Washington, DC	73.91346153846153	104
Denver, CO	Boston, MA	71.86705202312139	173
Durango, CO	Denver, CO	69.94117647058823	102
Portland, OR	Dallas/Fort Worth...	68.328125	128



Analyzing both graphs, it appears that the only difference between the two is the Washington DC to Norfolk route was replaced in the exact same ranking with the Syracuse to Washington DC route. Immediately, it can be observed that there are lots of east coast routes that appear in these graphs. The top two in both graphs are east coast routes, with both the top two containing Washington DC as either an origin or a destination. Weather is one explanation as to why east coast flights tend to see higher delays. Because of proximity to the ocean, storms and weather systems are much more likely to show up and delay flights. This is especially true in places like Fort Myers FL, which is right on the coast of the Gulf of Mexico and notorious for experiencing heavy weather systems daily and is hit hard during hurricane season.

Another observation that can be made is the frequency of Colorado showing up in this query, specifically Denver. The third most delayed route on this list is from Denver to Boston, a combination of a Colorado city and an east coast city. Denver and other cities and Colorado are

also well known for their heavy weather systems, this time in the form of snow and ice storms. These storms can heavily affect flight delays especially in the winter seasons, when so many people travel to Colorado for winter vacations and snow skiing. Considering this dataset was taken in January 2022, it would be correct to assume that the winter delays are heavily affecting Colorado flights in the data.

Denver International Airport was also listed as one of the most popular airports in terms of flight traffic along with DFW international airport, which is also seen in the Portland to Dallas/Fort Worth route. As would be expected, these graphs back the idea that the busier airports tend to have more and longer delays. The absence of Atlanta showing up in any of these routes also shows that Atlanta is one of the more efficient airports out of the large airports. The absence of New York and Chicago on any of the routes also reveals the effect on efficiency when a large city has at least two major airports.

Conclusion

In conclusion, there are many factors that go into airline travel efficiency, with some of them not being easy fixes. JetBlue is a major airline that can improve upon its average delay time to improve overall efficiency among all airlines. Though it seems that one airlines delay would not have a large impact on other flights from other airlines, the idea of snowballing delays that was analyzed with the time range of flights goes to show that it can have a noticeable impact.

It was also discovered that the east coast tends to have a large number of delays. It should also be noted that JetBlue is an east coast based airline (based out of New York), which would further explain the high duration of delays. Airports with a large amount of traffic also tend to have longer delays, as seen in the 'Average Delay of Routes' section. The data shows two possible ways to increase efficiency in these airports in question. The first would be a more in-depth analysis of Hartsfield-Jackson Atlanta International Airport, as this airport is the only one located in a major city without any other major airports that still runs efficiently. The second would be to add a second major airport to the city in question, as has been done in New York and Chicago. Most large cities have a second commercial airport, but most of them are not big enough to take the load of a significant amount of traffic off the main airport. These airports would have to be grown not to be as big as the main airport, but big enough to take a significant amount of traffic to relieve the main airport and make it more efficient. For example, Dallas has its main airport (DWF International Airport) and a smaller airport called 'Dallas Love Field' which only takes flights from Southwest, Delta, and Alaska Airlines[1]. This airport also does not take a large number of flights from these airlines and only consists of 2 terminals and 20 gates compared to DWF International Airport having 5 terminals and 174 gates[2]. If airports like Dallas Love Field in other cities, including Dallas, were grown to take some of the heavy traffic load off the major airport in the city, air travel efficiency would significantly improve.

The last problem that this report will mention is the issue of weather systems. This is almost not worth mentioning, because weather is so difficult to predict and even if it is accurately predicted, there are not many solutions to getting around the system without either delaying, canceling, or rescheduling the flight.

[1]Dallas Love Field: https://dallascityhall.com/departments/aviation/love_field/Pages/love_field_faqs.aspx

[2] DFW Internation Airport: https://www.dfwairport.com/business/about/facts/

Airline Codes: https://www.bts.gov/topics/airlines-and-airports/airline-codes

Airport Codes: https://airportcod.es/#