Up in the Air

Uday Datta, Steve Alessandrini, Jimmy Chen, Jack Harris

IST 718
22 December 2018

Project Description

In this hypothetical project, four team members need to meet annually in Denver. They fly from Austin, Dallas, Albuquerque, and Boston. The objective of this project is to identify and predict when would be the best time of the year for all four team members to fly to Denver while minimizing **cost** and **delays**.

BOS
DEN
ABQ
DFW
AUS

Source: http://www.gcmap.com/mapui?P=AUS-DEN,BOS-DEN,+dfw-den,+abq-den&MS=wls

Obtain

We attempted to obtain pricing and delay data from various sources; however, the data was difficult to locate. Ultimately we leveraged data from the Bureau of Transportation Statistics (BTS)[1]. BTS provided the most complete data set of any site we found and it was free to use. We also furthered our understanding data from Openflights[2] as another source to corroborate conceptual definitions such as airport codes, locations, etc.

Other data sources we did explore were Flightaware and Faredetective but there were high costs associated with these websites and aggregating the data was challenging.

BTS Delay Data - 60 months of delay data for the four cities (from 2013 to 2017) where the destination was Denver (reported quarterly/monthly)

BTS Pricing Data - 60 months of pricing data for all cities (reported quarterly)

[1] The BTS Website collects approximately 10% of all ***purchased*** non charter flights for US carriers and is based on the round trip prices (unless the consumer purchased a one way fare).

[2] Primarily used to confirm Airport names, locations, etc.

Scrubbing and Transformation

Our initial analysis and approach was to model the scenario based on our actual locations (Phoenix, Albany, Dallas, and San Francisco); however, as we dug deeper in the data, we had to modify our locations. The BTS's collection of data did not have data available based on these locations. This was identified when we attempt to total the number of quarterly records as we aggregated the pricing data (with the expectation of having 4 cities X 5 years X 4 quarters = 80 records). Instead, we found Albany only had 9 records over the period. Similar issues existed for some of our other original cities, prompting us to modify our points of departure for the purposes of this exercise.

Pricing Data

Initially, the pricing data was "filtered" for origin cities and Denver as the destination, and years in scope. Then the data was grouped by "Year", "Quarter", "Origin City", and "Destination" with fares being averaged. As there were no missing data values, no imputation was necessary.

Delay Data[3]

Similarly, the delay data followed a similar approach to the pricing dataset; however, the main fundamental change in pre/post processing was that several non-key fields were dropped as they were not required. At the individual record level, if a flight was denoted with a 0 or a 1 if it was on time or delayed, respectively. Reasons for delays included carrier problems, weather, FAA, security or late incoming aircraft, and were reported in positive numbers. All delays were considered and if a flight was earlier it was actually reported as a negative number.

Results of Scrubbing and Transformation

Dataset	Scrubbing Step	Number of Records
Pricing	Original Dataset	90021
	Limiting to only Denver	1112
	Limiting to Key Origin Cities	389
	Limiting to 2013-2017 and Average	89 (9 removed for Albany) Final Dataset for Pricing
Delays	Original Dataset	77240
	Limiting to cities, Denver as destination, and years (The	1026

[3] Limited to arrival delays only

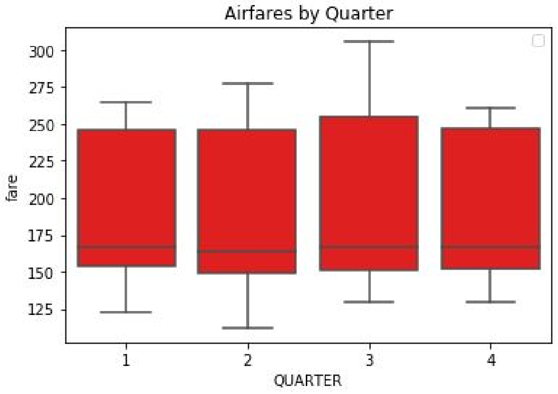
	reason why this did not result in 80 was because it was reported by carrier)	
	Averaging	80 - Final Dataset for Delays

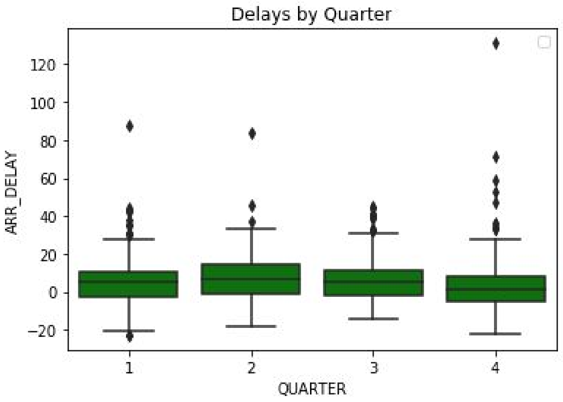
All Data / IPYNB is available here

Explore

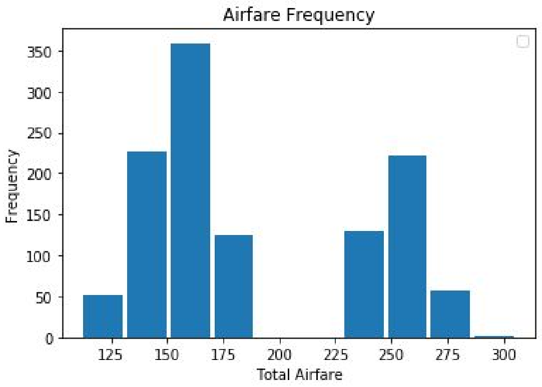
We completed some basic data analysis by calculating summary statistics for fares and delays. By examining histograms, box and whisker plots, line plots, and swarmplots, we were able to make some visual assumptions about the general trends for the fares and delays. Our focus was on the time periods where we would see the least delays, lowest cost, and overall maximization of our time at our meeting point.

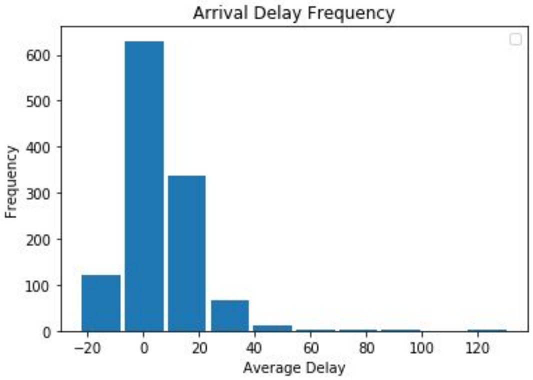
At first glance, we can see some specific airfare spikes in the 3rd quarter, which would be the summer months, and may have the most travel, therefore, we may need to focus our travel on other months. We will let our modeling exercise predict the optimal month for travel.



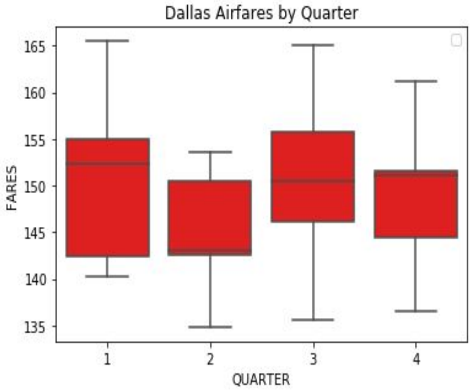


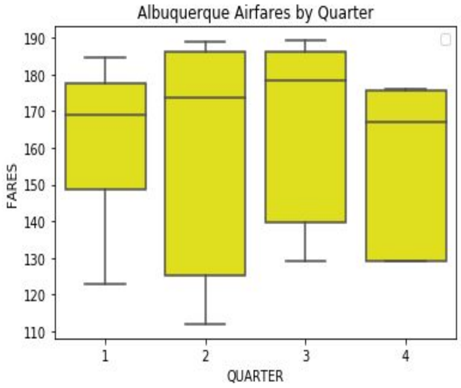
Our frequency charts reveal some interesting trends. We see that with the addition of the long distance flight from Boston, we have a bimodal distribution on the airfares. This makes sense as we have a nearly 1000 mile difference in trip length. Boston to Denver is 1,754 miles and the average distance of the other three cities is around 590 miles. Delays are right skewed with a few outliers representing some significant delay events.

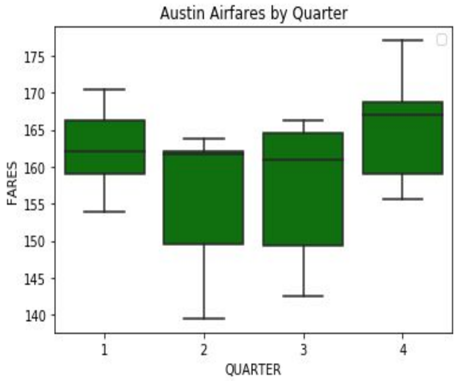


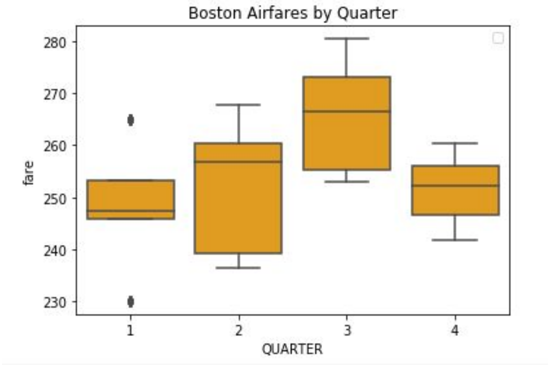


Next we look at the fares by departure site. It is interesting the much wider spread in price in Albuquerque versus other at around $40 overall in the middle quartiles. The other cities see a fairly consistent spread, however much more seasonal volatility in prices.

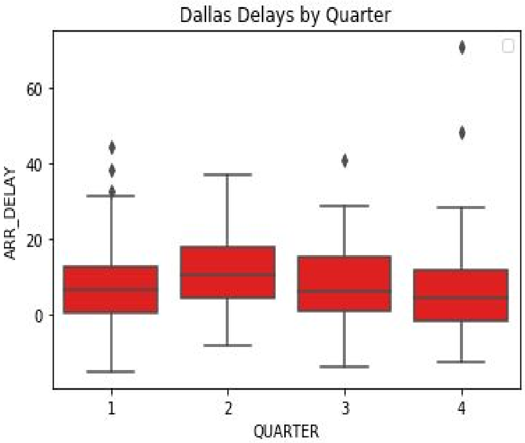


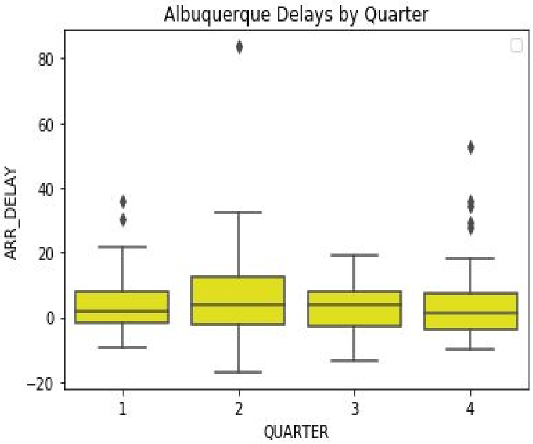


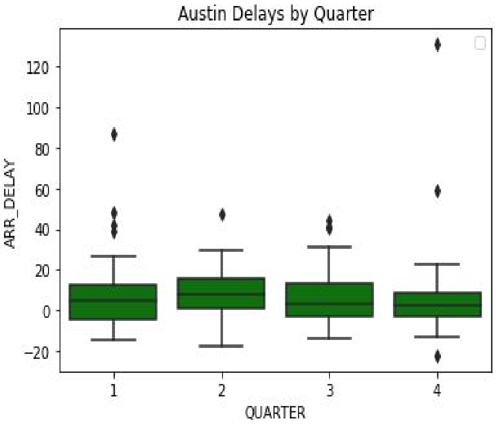


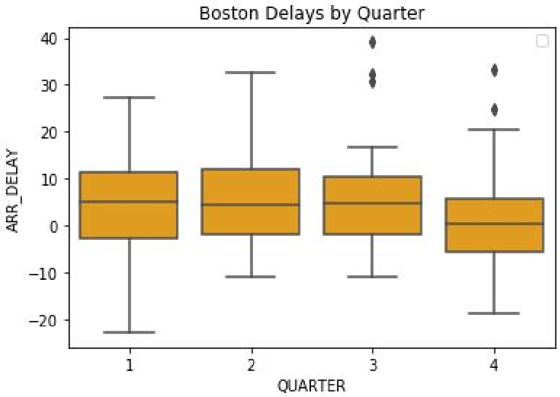


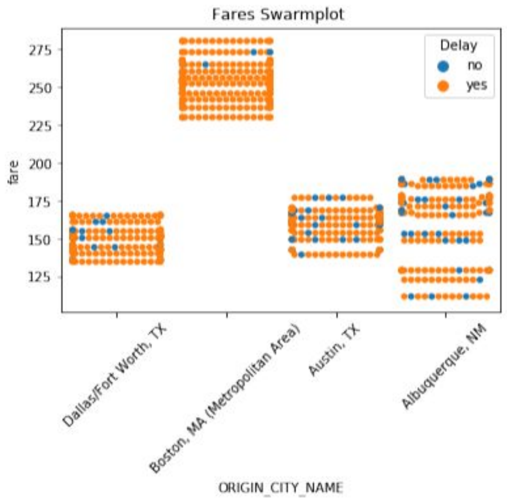
There seems to be a consistent spike in delays across all cities in the 2nd quarter with the exception of Boston, which seems to see an inverse trend of delay times as the year progresses.

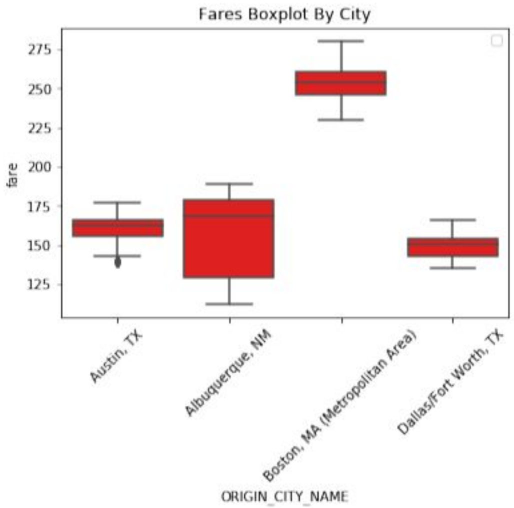

Dallas Delays by Quarter


Albuquerque Delays by Quarter


Austin Delays by Quarter


Boston Delays by Quarter





And our final exploratory views are fare views by city on an aggregated chart showing the individual points in a swarm plot and summary statistics view in a box plot. These clearly exhibit the price disparity between Boston and the other cities and also show that Boston seems to have many more delays that the the mid/western locations.

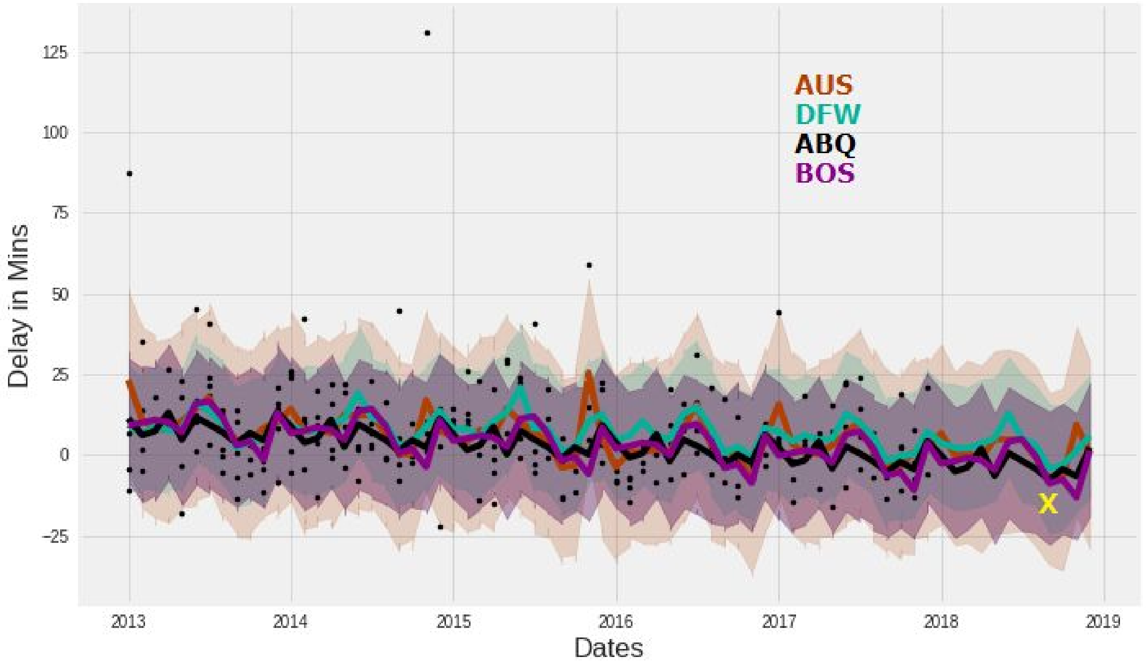
Model

Air travel is often highly dependent on the time of the year and can therefore vary significantly based on the season. Thus, time series analysis can serve a great deal in providing a reference to air fare and delay.

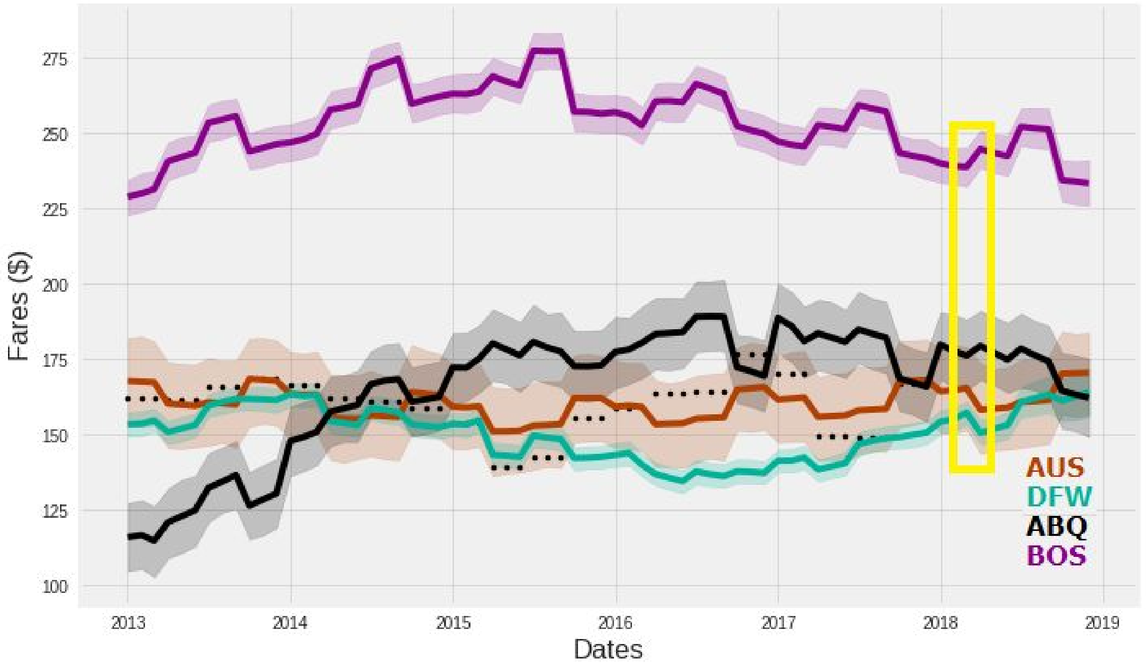
The model of choice is Facebook’s Prophet model. Initially developed for Facebook’s own internal use, the Prophet model is significantly easier to implement compared to other time series models. It has the ability to work with multiple seasonalities as well as seasonalities that occur at irregular intervals. As result, the model is able to achieve accurate forecasting even with default parameters when compared to models developed by high skilled professionals.

Interpret

Using data from 2014 to the end of 2017, the Prophet model was able to forecast the fare and delay information for the 2018 year. The fare and delay data as well as the Prophet model forecast is represented by the graphs below.



The graph above show the arrival delay at Denver from the four cities. The data is from 2013 to the end of 2017. After 2017, the trend line was generated by the prophet model which represent the forecasted values. Upon analysis, there is a consistent pattern of less delay around August of every year. Thus, August is the best time to travel to Denver from the four cities to minimize the delay.



The fares graph is generated in a similar fashion with the prophet model. The best time to travel with the lowest cost will be around February. Unlike the delay pattern however, February is not the cheapest time to fly for all cities. For example, Albuquerque actually tops top around that time. In this case, the best time to fly is determined by the least cost overall for the four cities. Also, delay is perhaps a bigger concern than the fares, because the variance in the fares is only around ten to twenty dollar, which is fairly insignificant compared to having to deal with delays.

Minimizing delays are more important to the team than minimizing cost. The team gave twice as much weight to the delay data as the cost data. Based on the following analysis, the optimal month for the team meeting is **October**.

It is also worth mentioning that the data we had was only up to the end of 2017 and we were forecasting for dates in 2018 which had already passed. However, based on the pattern in the graph, both delay and fares follows a similar trend year after year. Thus, forecast for 2019 will also have similar values.

Comments

This project provided exposure to many aspects of data analysis as well as the difficulties that comes with it. Obtaining and transforming the data seemed to take the most effort, with significant manual work required to download the data. Visualization and modeling were fairly straightforward, which gave us opportunity to further explore and combine graphs from the prophet model, allowing us to conduct a deeper exploration than the default outputs.

The biggest limiting factor for this project was obtaining enough data in the format that would be ideal for answering the original questions: where and when to meet. Since cost and time were limiting factors, we were not able to gather enough data to answer the “where” part of the question. If the data was more readily available, multiple combinations of origin and destination could be analyzed, which would significantly help answering the question.

Overall, the project provided an opportunity to not only analyze a problem with a data approach, but to do so through collaboration with a focused team.

References

https://facebook.github.io/prophet/docs/quick_start.html#python-api

Delays/Fares Chart - adapted from:

https://github.com/facebook/prophet/blob/master/python/fbprophet/plot.py
http://stackoverflow.com - various