Event-driven exploration of airfare prices and routes

American airspace data, 2017 domestic flights

Citadel DataOpen NYC
Team 20

Arjun Lal, Andrew Cui, Yutong Liu, Roger Zhang
University of Pennsylvania

July 21, 2017

Table of contents

Team 20

Executive summary

Why are the same flights more expensive in certain quarters than others? We wanted to examine the relationship between events and flight prices across quarters. Particularly, we chose to study the influence of event variety on fare prices. Here we defined “event variety” as the similarity of quarterly events in one city compared to all events in that city in 2017. The higher the similarity, the more homogenous of the quarterly events are. We used Latent Dirichlet Model (LDA)[1] to model the event topic distributions for each event and compared the distribution of each event with the annual mean to measure similarity.

Our regression analysis indicated that homogenous events in the origination city drives up fare prices (`Spearman = 0.0055, p < 0.01; JSD = -0.016, p < 0.1`), and heterogenous events in the destination city drives up fare prices (`Spearman = -0.0049, p < .01; JSD = -0.0169, p < 0.1`).[2] The psychological underpinning of this result might be that people want to leave the city when events in one quarter are similar to the whole year’s, while eager to travel to another city where different things happened.

Moreover, we analyzed overall and event-dependent flight traffic patterns across the nation. In this, we found a somewhat surprising disparity in cities’ propensities to host events and have flight volume. In addition to flight frequency and routes, we found that in the ~1,200 event data points we had, they matched up with less than 50 cities. Although it makes sense for large-scale events to be held in larger areas (with a larger potential for an audience), this could be a further driver of the increase in flight activity and fares we analyzed when an event is occurring.

Analytical approach summary

We began with exploratory analysis to better understand the airline traffic patterns. Visualizing frequencies, paths, and distributions of traffic helped us confirm an expectation that major metropolitan areas would dominate the airspace. We computed means and variances of each flight in our airfare table, approximated due to the bucketing, to generate our fare response variable for our model. We choose fare / distance as our response variable.

Given the set of public events in the U.S., `events_US`, we used LDA to model 15 different event topics and their corresponding distributions. For each given event, we computed Jensen-Shannon divergence (JSD) scores between the probability distribution of that event and the average across the year for the city in which it occurred. Then we aggregated the measures by quarter and city due to the information provided in `fares.csv`. We also matched origin/destination airport with their cities, and assigned to similarity measures to origin and destination cities.

We computed over cities because in our exploratory analysis we had seen the higher concentrations of flights in primary metropolis areas, which often happen to be hosts for large events. Moreover, given the quantity of flights into and out of those areas,[3] we reduce our risk of having our results biased on one-off events due to sample size. We looked at travel patterns (`flight_traffic`) in addition to textual analysis. To identify changes in flight traffic patterns due to events, we matched flight data to events based on cities and date. Computing normalized flight volumes in the 10 days prior to and after an event occured, we saw changes in flight volumes due to events. With this analysis, event data, and fares data, we created our linear regression model for airfare.

[1] Basically, LDA treat textual documents as mixture of topics. We are generating the top 15 topics in all events and assign distribution to each event. For more information, see Blei (2012), Probabilistic Topic Models.

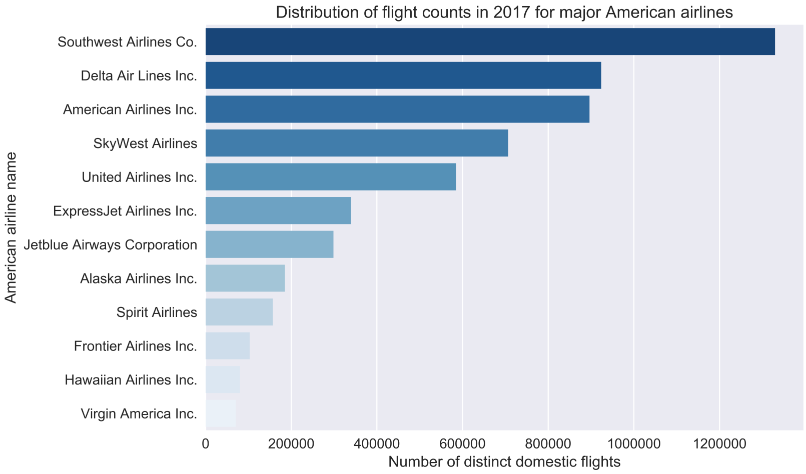
[2] Higher Spearman indicates higher similarity, higher JSD indicates higher differentiation.

[3] See Exploratory analysis, “Airlines and popular airports.”

Exploratory analysis

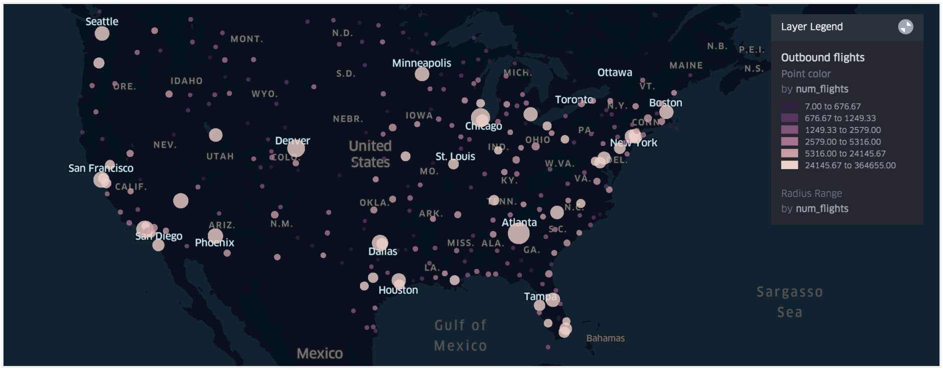
Airlines and popular airports

Our data was a set of air travel routes from 2017 domestic flights, and first we wanted to see the distribution of flights across airlines and cities. Grouping data by airlines and computing total counts, we saw that Southwest Airlines actually carried the greatest volume, instead of prominent global carriers such as American and United. However, noting that `flight_traffic` is actually American-domestic only,[4] the following result becomes more reasonable, given that international carriers have to split their fleets between flights internally and those abroad.



We knew that most airlines, given the industry standard "hub-and-spoke" method of operations, tend to concentrate their flights around major cities where they can maximize volume. We decided to match our `flight_traffic` data to the `airports` dataset, integrating locational (longitude and latitude) data, to process it for visualization. Using Uber's **Kepler.gl**[5] mapping tool, we plotted the major airports and frequencies in 2017 (see next page). We found our greatest frequency of flights located at our metropolitan centers and state capital.

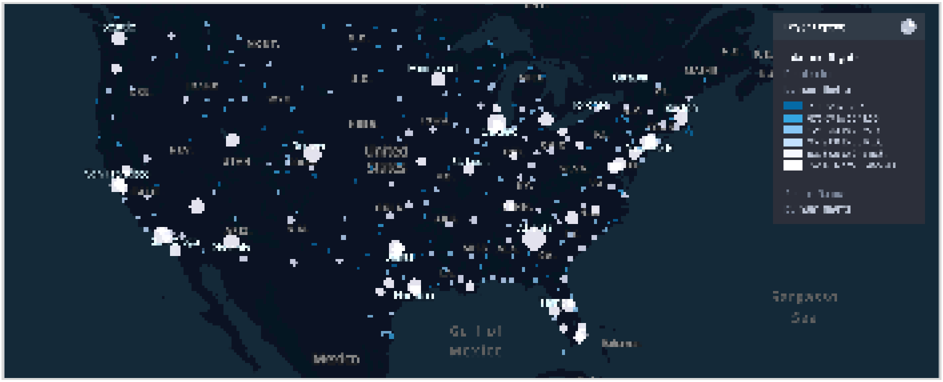
Domestic outbound flights (origin airport)



[4] In fact, even on a global comparison, often-labeled "cheap-flight" carrier Southwest holds its own. As of 2016, according to the World Airline Rankings, Southwest actually carried the third-largest volume out of carriers in the world, behind only American and Delta, despite only starting international service in a limited fashion in 2014.

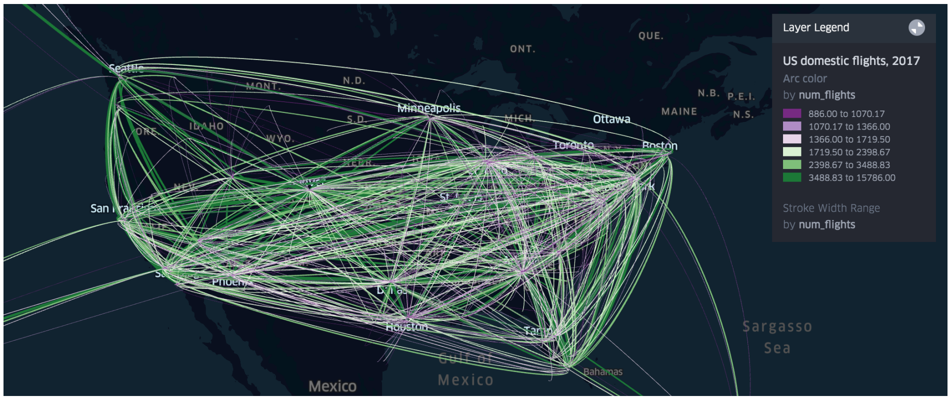
[5] See Appendix, "References."

Domestic inbound flights (destination airport)



American flight paths

Although we saw a disparity in frequencies between the flights, we wanted to confirm that a difference not just in travel, but in routes. Overall, this trend would be akin to tendencies in consumer-centric industries, such as Lorenz curve effects in product sales/heterogeneity.[6] For example, flights out of NYC-JFK could be headed anywhere in the nation, but if we found that the most frequent routes were ones that connected large cities, it would further support our intuition that smaller cities are comparatively neglected in air service, which would make it more difficult for people to get in/out in the case of a major event. Again, using **Kepler.gl** for the top 2000 flight routes by frequency:

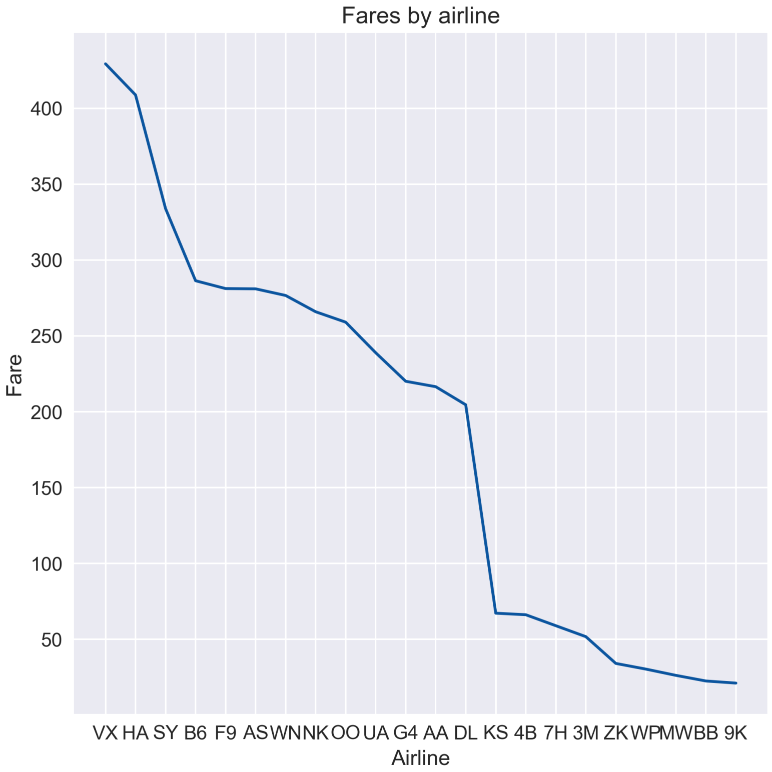


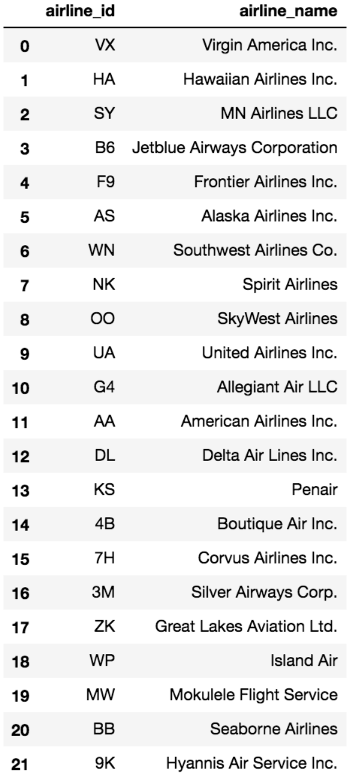
Airfare comparisons

We looked as well at the general exploratory data for airfare, looking to see if we could find any interesting differences across airlines or routes. In our latter analysis, we kept all our airfares/flights regardless of if the `airline_id` was a fully filled in value. This is because we are operating under the assumption that any differences in ticket prices for a certain time period are more impacted by the event than a specific airline's price-gouging potentially could. In this, we assume that if an airline were to do that, which would skew our results later, that they would fail to continuously capture the same sort of business, and would be therefore forced to change. However, since we compared by airline here, we only used the data points that both had an airline and had more than 10 flights, the latter to avoid outliers.

[6] The Lorenz curve looks at the proportion of customers who purchase/use a product as a fraction of its user base. A common example of it is the "80-20 rule" of spending.

We computed the following graph of fares and airlines, with a comparison matrix to see the airlines on the right:





	airline_id	airline_name
0	VX	Virgin America Inc.
1	HA	Hawaiian Airlines Inc.
2	SY	MN Airlines LLC
3	B6	Jetblue Airways Corporation
4	F9	Frontier Airlines Inc.
5	AS	Alaska Airlines Inc.
6	WN	Southwest Airlines Co.
7	NK	Spirit Airlines
8	OO	SkyWest Airlines
9	UA	United Airlines Inc.
10	G4	Allegiant Air LLC
11	AA	American Airlines Inc.
12	DL	Delta Air Lines Inc.
13	KS	Penair
14	4B	Boutique Air Inc.
15	7H	Corvus Airlines Inc.
16	3M	Silver Airways Corp.
17	ZK	Great Lakes Aviation Ltd.
18	WP	Island Air
19	MW	Mokulele Flight Service
20	BB	Seaborne Airlines
21	9K	Hyannis Air Service Inc.

We can see that Virgin America tends to have the most expensive flights, followed by the likes of Hawaiian, Alaska, and JetBlue. We can interpret this with two assumptions and understandings. First, some of the more niche airlines typically fly more specific/longer routes, which can lead to more expensive flights. Hawaiian, for example, will likely fly a more expensive route than the average New York-Philadelphia trip. Second, although these airlines are also less expensive per mile, they fly a more select group of routes, which means the lower-cost shorter-trip flights dominated by large national carries like American, United, and Southwest, will reduce their means and make them appear less expensive. However, at the bottom, we still see our low-cost budget airlines, which due to their limited range and lower costs have the cheapest fares.

Event names and Jensen-Spearman divergence

We pulled out the event names from the highest and lowest three Jensen-Shannon divergence (JSD, from above) scores for three main cities (New York, Los Angeles, Nashville) to validate the legitimacy of our calculation. If high score events are really not typical while low score events are typical, we will have more faith in the topic modeling and score assignment. Some results are listed as follow:

New York High JSD Events:

- National Board of Review of Motion Pictures Awards Gala
- Winter Jazzfest NYC
- Jazz Connect Conference
- APAP conference

- Intelligence Squared U.S. Spring Season Opening Debate
- UJA Federation of New York's REX Gala
- Django a Gogo Music Festival at Carnegie Hall
- Media Summit New York
- CNN Heroes: An All Star Tribute

New York Low JSD Events:

- New York City International Film Festival
- SESAC Pop Music Awards
- Tribeca Film Festival
- OZY Fusion Fest
- Panorama Festival
- New York Comedy Festival
- DOC NYC film festival
- National Book Awards Ceremony
- Macy's Thanksgiving Day Parade
- Gotham Independent Film Awards
- Rockefeller Center Christmas Tree Lighting

Los Angeles High JSD Events:

- Los Angeles Times Festival of Books
- Daytime Emmy Awards
- Electronic Entertainment Expo (E3)
- BET Experience
- Anti Piracy and Content Protection Summit
- Production Music Conference
- Digital Hollywood Fall

Los Angeles Low JSD Events:

- People's Choice Awards
- Producers Guild Awards
- NHL All Star Game
- Screen Actors Guild Awards
- Pollstar Live!
- Directors Guild Awards
- NAACP Image Awards
- The Writers Guild Awards
- GRAMMY Awards
- Guild of Music Supervisors Awards

Nashville High JSD Events:

- Country Radio Seminar
- T.J. Martell Foundation Nashville Honors Gala
- CMT Music Awards
- NASH Country Daily Annual Kick Off Party
- CMA Music Festival
- IMMERSE Conference
- K LOVE Fan Awards

- Summer NAMM
- Barbershop Harmony Society Harmony University

Nashville Low JSD Events:

- Tin Pan South Songwriters Festival
- VenueConnect
- ACM Honors Ceremony
- DIY Musician Conference
- SESAC Nashville Music Awards
- ASCAP Country Music Awards
- BMI Country Awards
- CMA Awards

What we found looking at the events above segmented by their respective cities is that events with high JSD scores are topically atypical and that events with low JSD scores are topically typical.

Qualititatively, this can be seen above. When we look at low JSD events that occurred in New York, we see that the majority of them are ‘Festivals’ or similar, demonstrating that events within this category are topically related and considered typical within New York. In the same vein, we see that low JSD events in Los Angeles and Nashville are primarily centered on ‘Awards’, movie related awards in Los Angeles and music related awards in Nashville.

Conversely, we see quite the opposite when we look at high JSD events, as would be expected. The high JSD events within each of the three cities topically related to only the slightest degree. Though three events within the New York high JSD events are centered on music, the others span from debate to movies. Likewise, the high JSD events within Los Angeles and Nashville cover a wide spectrum of topics ranging from business strategy conferences to book festivals to anti piracy summits.

Quantitative modeling

Textual analysis of event similarity scores

We used nltk and scikit-learn packages in **Python** to process the raw textual data and fit the natural language processing model. We first cleaned out the stop words in each event and then fed all of them into an LDA model with 15 topics. After fitting the model, we assigned topic distributions to each event and computed the mean topic distribution in every city across 2017.

With the annual topic distribution and individual distribution for each event, we computed the Spearman correlation and Jensen-Shannon Divergence for each events by matching and iteration. We aggregated the scores again by city and quarter to match the format of `fares.csv` and then assigned them to `o_spearman` and `o_JSD` if the origin city has any events in given quarter and `d_spearman` and `d_JSD` if the destination city had any events in given quarter. Then we pulled out `o_data` and `d_data`[7] from the original dataset. There are 21, 854 rows in the former, and 21,927 rows in the latter. The subsequent regression are performed respectively on these two datasets.

[7] `O_data` contains valid Spearman and JSD values for origination city, and vice versa for `d_data`.

Regression modeling

We fitted four multiple regression models respectively with origination Spearman correlation, origination JSD, destination Spearman and destination JSD, controlled by quarter. We used `fare/distance` as the response variable to normalize for longer flights being more costly and thus expensive. The estimated model parameters are as follows:

term	**estimate**	**std.error**	**test statistic**	**p.value**
(Intercept)	0.2499	0.0023	109.7864	~ 0
o_spearman	0.0055	0.0019	2.8249	0.0047
quarter	-0.0036	0.0008	-4.4263	9.6335

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	0.2545	0.0026	95.6680	~ 0
o_JSD	-0.0164	0.0100	-1.6532	0.0983
quarter	-0.0037	0.0008	-4.6324	~ 0

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	0.2503	0.0020	122.5642	~ 0
d_spearman	-0.0048	0.0017	-2.7914	0.0052
quarter	-0.0009	0.0007	-1.2536	0.2019

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	0.2458	0.0024	103.03464	~ 0
d_JSD	0.0169	0.0089	1.9036	0.057
quarter	-0.0008	0.0007	-1.0820	0.2793

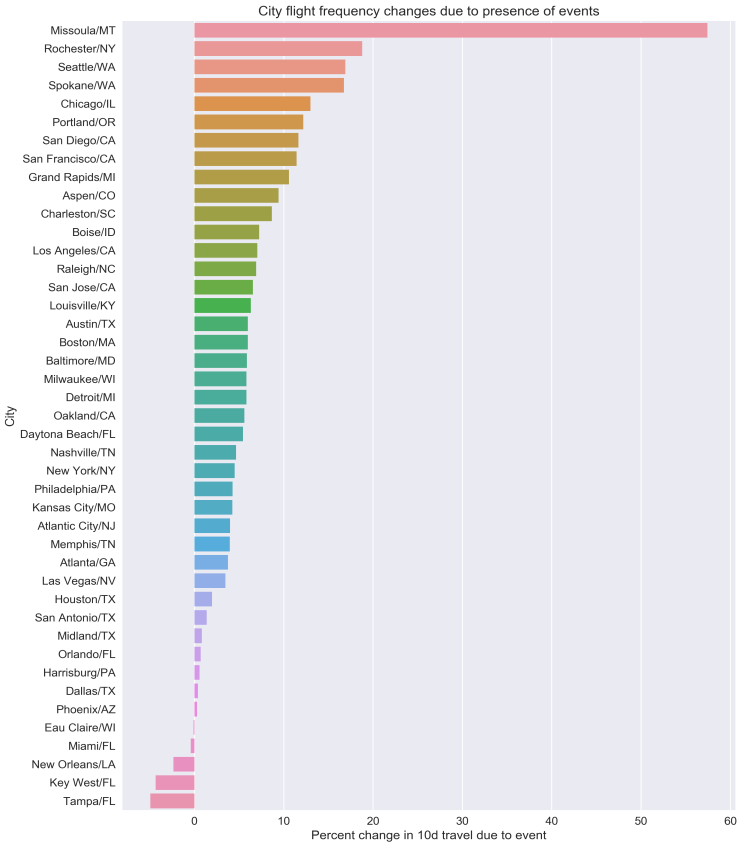
We separately fitted in Spearman and JSD because their functionality are essentially the same: to measure the similarity/difference between two probability distributions. The results show that similar events in origination city in a quarter (compared to yearly event) drive up fare prices and different events in destination city in a quarter drive down fare prices. Maybe people are trying to get out when the events in their city are monotonous, to seek more exciting events elsewhere? We will leave that discussion for further research.

Air traffic evaluation

In addition to our regression model, we wanted to see if there are actual changes in flights' quantities around the time of events, which could indicate the impact of the event on local air transportation as well. For this, we looked at the `events_US, airports,` and `flight_traffic` data in tandem. As we are comparing changes, we set our baseline as the overall average volume of flights throughout the year at a given airport or city, and we compared it to the volume during the event peak seasons for each of the events we observed. In this case, we looked at volumes 10 days before and after the event, allowing adequate time for people to come, attend the event, and then leave if they did so by air. For 1,151 events, we found that the data occurred around ~40-50 cities, similar to the large disparity in quantity of reach that we saw in the flights data.

It is likely that these two are intercorrelated; events are more likely to be held in cities with greater flight volumes, due to the ease of people moving in and out of the city. This would actually support our analysis, we reasoned, given that the converse is that the presence of an event, especially a major one, would to greater service quantity, meaning that the events are a factor that influences the airline industry around these major areas. Our analysis found that the majority of these cities did see, on average, an increase in volume that was correlated with the events.

Given the reasoning that events may be located in larger cities, this would suggest an even more powerful impact, especially for our more-divergent (higher-JSD) events: these more exclusive and desired events create even bigger business for the airlines, and even with an increase in supply to accommodate and capture the customers can still raise prices to generate profits.



Concluding thoughts

Our models and analysis suggest that events have an important impact on the pricing and frequency of air travel. Certainly the air transportation economy is a complex business, yet it is interesting to that there is this sort of association in the way flights behave. Moreover, our analysis of the American airspace indicated a large disparity between travel locations and types of cities - namely, strongly in favor of larger ones. Although there are many other variables to explore, this sheds light on just how difficult it can be to understand this industry.

Appendix

References

For our analysis, we relied on the following data processing languages, tools, and packages:

- **Python** and **Jupyter Notebook** was our overall primary source of data wrangling, analysis, and visualization setup. We used the standard data analysis stack: `pandas, numpy, nltk, sklearn, matplotlib, seaborn.`
- In addition, we used **R** and **RStudio** for some of the analysis. The code for both **Python** and **R** is attached in folder separately (see *Code appendix* below)
- For visualizations, we also used a mapping tool by Uber, known as **Kepler.gl**, designed and open-sourced for latitude-longitude based global map visualizations. We chose this over `matplotlib` primarily due to its efficiency and speed.

We did not use any additional data sources within our analysis. The primary data files that are contained in the above analysis content are:

- Base data: `airlines, airports`
- Event textual analysis: `events_US`
- Flight information and traffic analysis: `fares, flight_traffic`

Caveats, challenges, and future research areas

The analysis of the relation between events (and specifically, when they occur) and flight traffic could be improved by looking into how accurately the given dataset of events represents events in the U.S. Since the number of cities in the dataset which contained events was relatively small compared to the total number of cities (approximately 40 compared to over 300), a more thorough analysis of the impact of events on cities, particularly smaller cities, was not possible. Looking into how rare events are in each city would also be helpful in determining their impact. It would also be useful in the future to directly analyze the impact of the timing of events on airline fare prices, in addition to our current analysis on the timing of events to air traffic.

Another interesting variable that we would want to include would the cities and/or airlines. Namely, are the fares and frequencies we see something that can be explained by individual lines or locations, or does it appear to be something driven largely by the events? Being able to further separate the impact of events would lead to more conclusive determinations about the impacts of the events in our models.

Code appendix

Our analysis was conducted in **Python** and **R**. For the sake of simplicity in this document, we have uploaded our code files in a zipped document folder along with this report. The documents are named aptly towards their purpose in the analysis, many sections of which (outside of visualizations) could be done by different members of the group. Our graphs in Uber's **Kepler** are embedded within the document, but as they were generated in the Uber web app we are unable to have code snippets for it. However, **Kepler.gl** can be found at kepler.gl/#/demo.