

**Problem Statement**

Welcome to the Correlation One’s NYC Summer Invitational Datathon! This document explains the topic of the Datathon, important details about the datasets you’ll be using, and guidance on how to submit your results.

**Background**

The first airline was founded in November 1909, when [DELAG](https://en.wikipedia.org/wiki/DELAG), *Deutsche Luftschiffahrts-Aktiengesellschaft*, with government assistance, [began operating airships manufactured by The Zeppelin corporation](https://www.airships.net/delag-passenger-zeppelins/). Not too long after, the US airline found its footing as troves of aviators returning from World War I looked for peacetime work. However, [initial service was extremely limited and often consisted of delivering bags of mail for the U.S. Postal Service](http://www.historynet.com/airmail-service-it-began-with-army-air-service-pilots.htm). It wasn’t until World War II and the aftermath that airlines began investing heavily in civilian air transport, both for passengers and for cargo.

Today, the US airline industry is one of the most critical engines of our economy. Although its constituents have had its ups and downs, from reasonable profitability to [bankruptcy and bailout in the 2000s](http://www.cnn.com/2001/TRAVEL/NEWS/09/29/rec.airlines.financial/index.html), it has survived and remained a mainstay. Additionally, as a key barometer of US commercial travel, it can often serve as a leading indicator of consumer discretionary spending and leisure activity. As the world becomes more interconnected, and we find better and more exciting ways to visualize and explore these connections, the airline industry will continue to be a hotspot of activity and interest.

**Your Task**

Your goal is to analyze 2017 US commercial airline flight traffic data (described below), potentially in combination with supplementary datasets, in order to increase understanding of how developments the commercial airline industry relate to broader consumer trends and global events at large.

We have partially pre-cleaned several supplementary datasets for your use. Additional commercial airline travel data is available, including data about airline passenger fares, airport, and airline stock prices. We also provide info about major US events in 2017, as well as 6-hour weather data from US airports.

**You are asked to pose your own question and answer it using the available datasets in the available time**. What is important is the insightfulness and depth of your conclusions and analysis. **You need not be comprehensive; quality data analysis will be rewarded over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict or model airline travel trends. Submissions may also be illuminating, through use of thoughtfully chosen data visualizations or sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question generally has a positive effect on judges’ assessment of your submission; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: How does an airline’s general flying patterns (e.g. traffic volume, destination choices) relate to that airline’s financial / stock performance? Can any trends be identified to separate top performers vs. bottom performers?

Sample Question 2: How does the severity of weather relate to actual impact on airline flight delays? Is there a breakpoint of weather severity at which flights are more often impacted?

Sample Question 3: Do delay / cancellation patterns impact stock / financial performance at all?  How do airlines financially perform in quarters with worse-than-average weather?

Sample Question 4: How do major US domestic events impact air traffic and passenger fare patterns?

**Datasets**

The provided datasets are stored in your team’s USB drive that you received at registration, and are spread across seven tables. (Alternatively, if you do not have a USB portal, they are also stored in the “Datathon Materials” folder on your team’s Box account (described later).) Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to “play nice” with each other.

***airlines***

Mapping of airline IDs to names.

*61 rows & 2 columns.* Size: ~0.1MB. Source: the Internet, generally.

***airports***

Important details (name, state, identifier, latitude, longitude, etc.) on various US airports.

*322 rows & 6 columns.* Size: ~0.1MB. Source: [US Department of Transportation](https://www.transportation.gov/).

***events\_US***

Public events from around the US throughout 2017.

*~1,151 rows & 4 columns.* Size: ~0.1MB. Source: [Shore Fire Media](https://shorefire.com/).

***fares***

Airline fare distributions for each quarter-route-airline combination in 2017 with a bucket size of $10.

*80,823 rows & 255 columns.* Size: ~44MB. Source: [US Department of Transportation](https://www.transportation.gov/).

***flight\_traffic***

Information about delays for US domestic flights in 2017*.*

*~6 million rows & 24 columns.* Size: ~130MB zipped, ~550MB unzipped. Source: [Bureau of Transportation Statistics](https://www.bts.gov/).

***stock\_prices***

Daily closing stock prices of various US airlines from late-2016 to early-2018.

*380 rows & 10 columns.* Size: ~0.1MB. Source: [Alpha Vantage](https://www.alphavantage.co/).

***weather***

Weather data (temperature, wind, precipitation, cloud cover, etc.) collected at various US airports every 6 hours through 2017.

*~375,000 rows & 12 columns.* Size: ~48MB. Source: [National Centers for Environmental Information](https://www.ncei.noaa.gov/).

**Additional Datasets**

Participants are welcome to scour the Web for their own custom datasets to supplement their analysis. All additional data used should be public and should not exceed 2GB unzipped (consult the technical team if you believe your idea is worthy of an exception).

**Other Materials**

We will provide you the schema for each of the data tables in another packet.

We will also provide you a Datathon manual at registration, which contains a section on using Box. This will show you how to download the datasets (described above) and upload your submissions (described below). It will also provide you helpful tips for the competition, so please read it carefully!

**Submissions: Content**

Submissions should have two components:

1. Report – this should have two main sections:
   1. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
   2. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your work without your team there to explain it; therefore, **your submission must “speak for itself”**. It need not be polished to the level of a final product, but do ensure that your main findings are clear and that any visualizations are functionally labeled.

**Submissions: Evaluation**

You will be evaluated based on your Report, as follows:

* **Non-Technical Executive Summary**
  + *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
* **Technical Exposition**
  + *Wrangling & Cleaning Process.*Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform?
  + *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypotheses tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  + *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

**Submissions: Format**

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

**However, please also include the source file used to generate your report.** For example, if you submit a PDF with math-type, equations, or symbols, please include your raw LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be provided a sheet with your team’s Box account login details when the hacking session begins; you will be using the account to download the datasets as well as to upload your submission content. We recommend that you wrap up your work by 3:15 PM and begin uploading your submission at that time. **Submissions MUST be received by 3:30 PM. Any submission received after 3:30 PM will NOT be evaluated by the judges**.

**Tips & Recommendations**

You will have ~11 hours total to work on the problem statement. However, you will not have access to the actual data until the morning of the competition. As such, we recommend you split your time as follows:

* Friday evening, ~8:00PM – 12:00AM: You will receive a copy of the problem statement, data table schema, and data table heads. This gives you the opportunity to study the available data fields, think about suitable questions to tackle, and plan out your exploration process. Additionally, the data table heads should be sufficient for you to begin putting together some data wrangling & cleaning scripts.
* Saturday, 8:30AM – 3:30PM: You will receive the actual data. If you set up your data munging scripts already, you should be able to quickly apply them and immediately begin working with the data. You should spend most of your day investigating the data, performing qualitative & quantitative analysis, and writing up your process & results.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: <http://jupyter.org/install.html>. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard “terminal + text editor” environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

Finally, **we STRONGLY encourage you to start typing up your final submission AT LEAST three to four hours before the submission deadline**. In the past, many teams have spent a lot of time conducting great analyses, only to realize that they left almost no time for actually writing up and presenting their results. **This cannot be stressed enough – quality data analysis that is incomplete or poorly presented will NOT win one of the top prizes**.

**Ask for Help**

The Datathon team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.