MICHIGAN ROSS
CENTER FOR VALUE CHAIN INNOVATION
UNIVERSITY OF MICHIGAN

# Problem Statement

Welcome to the 2018 Michigan Ross Datathon, in partnership with Correlation One! 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, *Deutsche Luftschiffahrts-Aktiengesellschaft*, with government assistance, began operating airships manufactured by The Zeppelin corporation. 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. 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, 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 both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight will be rewarded over breadth of the question posed.**

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

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question will factor into judges' assessment of your submission:

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 receive at registration, and are spread across six 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.

***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.

***events_US***
Public events from around the US throughout 2017.
*~1,151 rows & 4 columns.* Size: ~0.1MB. Source: Shore Fire Media.

***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.

***flight_traffic***
Information about delays for US domestic flights in 2017. *Note: in order to keep the dataset size manageable, the provided data is a 10% unbiased sample of the raw data. If using flight count metrics, remember to multiply quantities by 10 to approximate the actual data.*
*~600,000 rows & 24 columns.* Size: ~40MB. Source: Bureau of Transportation Statistics.

***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.

***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.

**Additional Datasets**

Below is a table mapping each airline ID in the above datasets to an airline name:

| | |
|---|---|
| UA | United Air Lines Inc. |
| AA | American Airlines Inc. |
| US | US Airways Inc. |
| F9 | Frontier Airlines Inc. |
| B6 | JetBlue Airways |
| OO | Skywest Airlines Inc. |
| AS | Alaska Airlines Inc. |
| NK | Spirit Air Lines |
| WN | Southwest Airlines Co. |
| DL | Delta Air Lines Inc. |
| EV | Atlantic Southeast Airlines |
| HA | Hawaiian Airlines Inc. |
| MQ | American Eagle Airlines Inc. |
| VX | Virgin America |

Additionally, 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).

**Submissions: Content**

Submissions should consist of a single report, with three main sections:

1. Topic Question – What is the question that your team set out to answer? Why is it an important question? What datasets did you use to answer your question?
2. Non-Technical Exposition – What were your key findings, and why are they important? It is crucial that you communicate your insights clearly and substantiate them with sound logical analysis. Summary statistics and visualizations are also encouraged.
3. Technical Exposition – What was your data-driven methodology/approach towards answering the questions? Describe your data manipulation and exploration process. Again, use of visualizations is highly encouraged.

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

Until the finalist Q&A session begins, 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: 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.

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 3:45 PM and begin uploading your submission at that time. **Submissions MUST be received by 4:00 PM. Any submission received after 4:00 PM will NOT be evaluated by the judges**.

## Tips & Recommendations

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.