# Problem Statement

Welcome to Correlation One's 2019 Data Science for All: Women's Summit! This document explains the topic of the project and important details about the datasets you'll be using. You will find all the necessary files in the project folder that was downloaded with this problem statement.

## 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 the insightfulness and depth of your conclusions and analysis. **You need not be comprehensive; quality data analysis is more important 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 the 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 is encouraged; **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 spread across seven tables. 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.

***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.
*~600,000 rows & 24 columns.* Size: ~130MB zipped, ~550MB unzipped. 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.
*~353,864 rows & 12 columns.* Size: ~48MB. Source: National Centers for Environmental Information.

**Other Materials**

We also provide you the schema for each of the data tables. You'll find them in the same project folder.

**Submissions:**

Submissions should have the following components and be uploaded on the portal:

1. Proposal – this should include the following:
   *A draft of this should be emailed to your mentor and then a final version submitted to your portal.*
   a. A clear statement of the question you will be answering for this project and your rationale for choosing this
   b. Your plan or methodology that you will use to solve the problem (indicating software technology that you plan to use)

2. Presentation – this should have three main sections:
   *A draft of this should be emailed to your mentor and then a final version submitted to your portal.*
   a. 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?
   b. Non-Technical Executive Summary – 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.
   c. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process. Again, use of visualizations is highly encouraged.

3. Code – please include all relevant code that was used to generate your results.
   *Code should be submitted to Correlation-One. Specific instructions will be provided by Correlation-One through email.*

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

Your submission 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.

**Presentation Details:**

A good presentation should have the following elements:

- **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?

**Submission Format:**

**All submissions MUST be in a universally accessible and readable format (DOC, DOCX, PDF, PPT):**

- Proposals should be submitted in DOC, DOCX format
- Presentations should be in PPT format
- Code should be submitted in a single zipped collection of files separate from your proposal and presentation

**Ask for Help**

The DS4A technical 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.