**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 taxi in NYC was deployed in July 1897, when the Electric Carriage and Wagon Company (E.C.W.C.) began running 12 [hansom cabs](https://en.wikipedia.org/wiki/Hansom_cab). Over the years, the taxicabs of NYC have become some of the mostly widely recognized icons in the world, with [over 13,000 medallions issued as of 2014](http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf) (the right to run a yellow taxi). On top of that, numerous green taxis, community cars, and black cars roam the streets under the supervision of the NYC Taxi and Limousine Commission (TLC).

But in 2010, a small company then known as UberCab would make a splash in San Francisco and become one of the largest behemoths in the world, [valued at over $60 billion as of August 2016](http://www.businessinsider.com/ubers-history/#december-2008-kalanick-first-hears-the-idea-for-uber-at-the-leweb-technology-conference-he-thinks-of-it-as-a-way-to-lower-the-cost-of-a-black-car-service-using-your-phone-4). Uber, as it came to be known over the years, has disrupted and is now on the verge of overturning the traditional for-hire vehicle (FHV) industry as we know it.

Uber, and other FHV companies like it, have made life difficult for traditional taxi drivers, as the increasing competition has driven down costs and put many long-time drivers out of business. It has even been tough on Uber’s own drivers, as the lower fares have been mostly passed down to them. With the advent of autonomous vehicles and [California’s recent approval to allow the testing of fully driverless cars on the road](http://learningenglish.voanews.com/a/california-plans-to-approve-testing-of-self-driving-cars-with-no-human-drivers-inside/3764169.html), it seems like the industry will soon look completely different than what we have been used to for over 100 years.

**Your Task**

Your goal is to analyze the NYC Uber trip data (described below), potentially in combination with supplementary datasets, in order to increase understanding of how developments in the NYC for-hire transport industry relate to broader trends in the public and private transportation industries at large.

We have partially pre-cleaned several supplementary datasets for your use. Additional trip data is available, including data from green boro taxis, yellow medallion taxis, and NYC subway turnstiles. We also provide demographic info on age and income, geographic data for the Neighborhood Tabulation Areas (NTAs) of NYC, and daily weather data for NYC.

**You are asked to pose your own question and answer it using the available datasets**. 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 is more important over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict future trip trends. Submissions may also be illuminating, through the 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 is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: NYC has both green boro taxis and yellow medallion taxis. How have trip trends over time differed between these two classes of taxis in relation to Uber ridership?

Sample Question 2: Is there a connection between the demographics of particular NTAs in NYC and the dynamics of taxi or Uber rides involving those NTAs?

Sample Question 3: Investigate the details of pickup times and endpoints. Is there a relationship between these and the NYC weather?

Sample Question 4: Are there interesting similarities/differences in the servicing patterns of Uber, green taxis, and yellow taxis?

Sample Question 5: Explore the interplay between Uber and taxi trip trends and NYC public subway trip trends. What sorts of dynamics exist between these?

**Datasets**

The provided datasets are spread across nine 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.

***uber\_trips\_2014***

Trip data (pickup times, pickup coordinates, etc.) from Uber vehicles in 2014.

*~4.5 million rows & 4 columns.* Size: ~30MB zipped, ~200MB unzipped.

***uber\_trips\_2015***

Trip data (pickup times, pickup location IDs, etc.) from Uber vehicles in 2015.

*~14 million rows & 4 columns.* Size: ~65MB zipped, ~550MB unzipped.

***demographics***

Demographic data (population, age, income, etc.) organized alphabetically by NTA.

*188 rows & 33 columns.* Size: ~0.1MB.

***geographic***

Data about the shape of each NTA (latitude and longitude coordinates, in order).

*9,302 rows & 195 columns.* Size: ~4MB.

***green\_trips***

Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC green boro taxis. *Note: in order to keep the dataset size manageable, the provided data is a 20% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 5 to approximate the actual data.*

*~3.5 million rows & 9 columns.* Size: ~140MB zipped, ~400MB unzipped.

***mta\_trips***

Trip data (time intervals, entries, exits, etc.) from NYC public subway turnstiles.

*~7.5 million rows & 10 columns.* Size: ~50MB zipped, ~700MB unzipped.

***weather***

Temperature and precipitation data for three areas in the NYC metropolitan area.

*2,190 rows & 10 columns.* Size: ~0.1MB.

***yellow\_trips***

Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC yellow medallion taxis. *Note*: *in order to keep the dataset size manageable, the provided data is a 5% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 20 to approximate the actual data.*

*~8 million rows & 9 columns.* Size: ~260MB zipped, ~800MB unzipped.

***zones***

Information about each ride pickup zone in the NYC metropolitan area.

*263 rows & 5 columns.* Size: ~0.1MB.

**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.*

* 1. A clear statement of the question you will be answering for this project and your rationale for choosing this
  2. Your plan or methodology that you will use to solve the problem (indicating software technology that you plan to use)

1. 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.*

* 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 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.
  3. 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.

1. 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.