****

**Problem Statement**

Welcome to Correlation One’s 2020 London Regional 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 earliest bike share initiatives were meant as charitable projects or to promote bicycles as a non-polluting form of transport. In 1965, fifty unlocked white bicycles were left on the streets of Amsterdam for everyone to use. However, most of them were stolen or vandalized and the whole project ended in less than a month. Many years later, bike docking stations were conceptualized to overcome these problems and this system was first implemented by the University of Portsmouth in 1996.

As of 2014, over 1000 bicycle sharing systems have been launched under the same model, spanning 50 countries, 712 cities, 37,500 stations, and 806,200 bicycles. The bike share systems have been particularly popular in urban areas because they provide affordable access to bicycles for short-distance trips that are too far to walk but too short for taxi/subway. Further, they reduce congestion, noise, and air pollution as well as providing an efficient way to connect people to the public transit hubs (the “last mile” problem).

The Citibike bike share system was introduced to New York City in the summer of 2013, starting with just the borough of Manhattan. Since then, there was a [major expansion](https://nyc.streetsblog.org/2015/07/24/citi-bike-will-start-rolling-out-139-new-stations-august-10/) in the summer of 2015 to Brooklyn and small parts of Queens. Most recently, there are plans for [another expansion](https://www.citibikenyc.com/blog/major-citi-bike-expansion-map-revealed) that will be completed in 2023, with the goal of expanding coverage to the Bronx and Queens (as well as the more distant neighborhoods of Brooklyn).

**Your Task**

Your goal is to analyze the NYC bike share trip data, potentially in combination with supplementary datasets, in order to increase understanding of how developments in the bike share infrastructure of metropolitan cities (e.g. NYC, Boston, SF) relate to broader trends in the public and private transportation industries.

We have partially pre-cleaned several supplementary datasets for your use. Additional trip data is available, including data from the Boston and SF bike share systems, NYC rideshare trips (i.e. Uber, Lyft, etc.), 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 bike share 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: Given that bike share systems are a fairly recent development, new bikes and stations are periodically added as part of expansion efforts. How do these additions impact trip volume and other factors?

Sample Question 2: Are there interesting similarities/differences in the servicing patterns of the bike share systems of NYC, Boston, and San Francisco?

Sample Question 3: How does the growth of the NYC bike share system impact the trip trends of other forms of transportation, spanning public (subway) and private (taxi, rideshare) options?

Sample Question 4: How are aggregate measures (e.g. ride volume) affected by NYC weather?

Sample Question 5: Is there a connection between the demographics of particular NTAs in NYC and the dynamics of bike share rides involving those NTAs?

**Datasets**

*Please refer to the Citibike Data Schema document for more details about the datasets.* The provided datasets are spread across eleven tables. Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted but we encourage you to use our data since they have been organized and cleaned to “play nice” with each other.

**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.

**Submissions: Content**

Submissions should have three components:

1. Report – this should have two main sections:

a. 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.

b. 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, the 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 technical report without your team there to explain it; therefore, **your submission must “speak for itself”**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

**Submissions: Evaluation**

The competition will have multiple rounds of evaluation. The most important component of this evaluation will be your Report, which will be judged as follows:

● **Non-Technical Executive Summary**

o *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**

o *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? Please describe your process in detail within your Report.

o *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?

o *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 sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 3:30PM. Any submissions received after that time will NOT be evaluated by the judges**.

**Tips & Recommendations**

You will have ~12 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, ~7:00PM – 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.

We’ve compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not pertain to every team, so we recommend that you and your team apply any tips accordingly:

|  |  |
| --- | --- |
| **Tips for Success** | **Try to Avoid** |
| **1.** Focus on hypothesis testing when brainstorming your research question | **1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy |
| **2.** Spend at least 3 hours on your report to ensure strong communication through visualizations and writing | **2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient |
| **3.** Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality | **3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile |

**Ask for Help**

Correlation One’s technical product 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.