The Citi Bike Program(https://www.citibikenyc.com/) is a bike sharing system in New York City. Cyclists can rent a bicycle from one of many stations around the city and return it to any other station. Citi Bike has released ridership information to the public. This contains a list of all rides taken, their start and end locations, their start and end times, and limited demographic information about the rider. Zip files containing the data for each month are available at https://s3.amazonaws.com/tripdata/index.html(https://s3.amazonaws.com/tripdata/index.html). We will be considering the data from 2015 only.

What is the median trip duration, in seconds?

123.4567890

What fraction of rides start and end at the same station?

0.987654321

We say a bike has visited a station if it has a ride that either started or ended at that station. Some bikes have visited many stations; others just a few. What is the standard deviation of the number of stations visited by a bike?

12.34567890

What is the average length, in kilometers, of a trip? Assume trips follow great circle arcs(https://en.wikipedia.org/wiki/Great_circle) from the start station to the end station. Ignore trips that start and end at the same station, as well as those with obviously wrong data.

9.876542310

Calculate the average duration of trips for each month in the year. (Consider a trip to occur in the month in which it starts.) What is the difference, in seconds, between the longest and shortest average durations?

123.4567890

Let us define the *hourly usage fraction* of a station to be the fraction of all rides starting at that station that leave during a specific hour. A station has surprising usage patterns if it has an hourly usage fraction for an hour significantly different from the corresponding hourly usage fraction of the system as a whole. What is the largest ratio of station hourly usage fraction to system hourly usage fraction (hence corresponding to the most "surprising" station-hour pair)?

98.76543210

There are two types of riders: "Customers" and "Subscribers." (https://www.citibikenyc.com/pricing) Customers buy a short-time pass which allows 30-minute rides. Subscribers buy yearly passes that allow 45-minute rides. What fraction of rides exceed their corresponding time limit?

0.123456789

Most of the time, a bike will begin a trip at the same station where its previous trip ended. Sometimes a bike will be moved by the program, either for maintenance or to rebalance the distribution of bikes. What is the average number of times a bike is moved during this period, as detected by seeing if it starts at a different station than where the previous ride ended?

98.76543210

Please provide the script used to generate this result (max 10000 characters).

In what language is the script written?

◯ C/C++	◯ Fortran	◯ IDL	◯ Java
◯ Matlab	◯ Perl	◯ Python	◯ R
◯ Stata	◯ SQL	◯ VBA	◯ Other

Q3: This question is required.

Propose a project to do while at The Data Incubator. We want to know about your ability to think at a high level. Try to think of projects that users or businesses will care about that are also relatively unanalyzed. Here are some useful links about data sources on our blog(http://blog.thedataincubator.com/tag/data-sources/) as well as the archive of data sources on Data is Plural(http://tinyletter.com/data-is-plural/archive). You can see some final projects of previous Fellows on our YouTube Page(https://www.youtube.com/playlist?list=PLOE4k9MRzZanWmZ7MBrJFi7ZekYmVqEIV).

Propose a project that uses a large, publicly accessible dataset. Explain your motivation for tackling this problem, discuss the data source(s) you are using, and explain the analysis you are performing. At a minimum, you will need to do enough exploratory data analysis to convince someone that the project is viable and generate two interesting non-trivial plots supporting this. *The most impressive*