Data Management Plan

**Logical Collections <NEEDS REVISION>**

All datasets, after being processed, will be named according to the dataset they represent. Each column of both datasets will have their respective data values. Each row will represent a data point for each dataset.

**Physical Data Handling**

There were two datasets that were utilized for the transportation weather study. NYC Green Line taxi dataset was acquired from the NYC Taxi and Limousine Commission after being processed through a C++ script named condense.cpp. Weather data near JFK Airport for 2016 was exported from The Weather Company’s Wunderground application and reformatted to accommodate for the header structure for each month. Both datasets were processed and archived in comma separated value (CSV) file formats with their appropriate dataset names.

**Interoperability Support**

The files containing each individual dataset were stored in close proximity using hierarchical file-folder layout to enable interoperability. All scripts, python notebooks, and python files could access the appropriate dataset files with ease due to standardized naming conventions and single storage location within a GitHub repository.

**Security**

Datasets were managed and maintained by outside vendors, more specifically, NYC TLC and The Weather Company. Data is secured and changed on the NYC government and IBM servers, respectively. Data were acquired through an export tool which was processed and stored in a secured GitHub repository. Access to processed data can only be attained by authorized Group 5 members via their GitHub accounts.

**Data Ownership**

NYC TLC data has strictly noted that their data collected is not used, sold, or exchanged for commercial or marketing purposes. Data is automatically collected and quality assured by the NYC government. The Weather Company is operated under IBM with little to no restrictions on access to their data. Data is automatically collected and quality managed from weather devices and applications used by IBM.

**Metadata <NEEDS REVISION>**

Metadata for each dataset were noted within the file labeled metadata.json. Each file name was provided a description identifying where the data was obtained from and what processes had been taken to refine them and a fields section noting the column names and appropriate definitions.

**Persistence**

Raw datasets for both NYC TLC and Wunderground are found on their respective websites noted in the metadata. The processed versions of each dataset is published on the NYC Transportation Weather Study GitHub public repository hosted by Ian Gross.

**Discovery <NEEDS REVISION>**

The processed datasets can be found via the NYC Transportation Weather Study GitHub public repository to enable outsiders to gain access to each dataset.

**Dissemination**

A project description and appropriate topics will be provided to create connections between interested parties and our datasets. GitHub allows for all changes and additions to be publicized to all outside parties and enables them to track via GitHub issues. GitHub enables the team to disseminate and connect external users to our data project management.