Group 10: Xiaoxiang Zhang (xz2631), Minghao Li (ml4025), Ruimin Zhao (rz2390)

a. Title: Route recommendation based on weather and New York Public Transport System

b. Data to be used: NOAA (weather dataset), MTA (transportation dataset)

c. What will you do? Application/Techniques/Systems you want to use

Generally speaking, based on historical weather and traffic information, we want to predict how long the delay will be for some future weather and time(9am,10am,etc) inputs. Then we will recommend optimized route based on the predicted time delay(use GOOGLE map API to get the several routes and our models to predict which route take least time).

1. Firstly, we are going to parse real time weather and traffic data from Internet. The result can be tuples which will be transformed to spark-streaming system for preprocessing and storage.
2. Then we will apply Kernel Regression algorithm to our data to get a model evaluating the relationship between parameters.
3. Finally, we are going to predict the traffic delayed time based on the real-time weather plus time information and offline trained model, as well as the optimized route. All the results will be presented through a web application.

Techniques: Sparking Streaming, Kernel Regression, Flask, Google Map APIs, MTA API

d. How will you show any results?

1. A web application will be presented in the end. Users can see the density and the corresponding predicted delay time of bus stations and subway stations. Besides, user can obtain our recommended public transportation route based on the predicted delay.
2. Data visualization using packages such as Pandas based on the geographic information obtained from MTA data source.

e. What do you think is novel about your approach?

1. Rather than only focusing on one data source, we would utilize two different streaming data sources (MTA public transportation data and weather data) and examine the correlation between them. In this way, hopefully, we are able to obtain more comprehensive and interesting data analysis results such as transportation delay prediction, transportation route recommendation etc.
2. The main purpose of our project is to predict the delay situation of New York public transport system. We noticed that we can only get some vague information from MTA website, such as which line is delayed or which line is in good service. However, people have no idea about the exact delay time. This is the point we will focus on.

f. Any references: [1] http://alert.mta.info/ [2] https://www.dot.ny.gov/wta