Final Project

The final project's objective is to give you the freedom to apply the course's data science tools to improve life in New York City. In this project, your mission is to utilize publically available New York City data sources to solve a problem. Your project should closely relates to real problems in NYC. Here are some examples:

- Finding the optimal location for the new Italian fine-dining restaurant you dreamed to open;
- Finding the best place to live in NYC that is not only safe, but also have the right resources for your family;
- Optimizing Express Bus Routes in Staten Island;
- Where is the best place to plant a tree in NYC;

These are just some examples. You are welcome to investigate any subject related to public interests, such as policy recommendation on how to make the subway, bus and taxi of NYC be more efficient and reliable, how to preserve energy and water, how to promote recycling, how to improve public and private education, etc. The project should explain a given problem, characterize it with data, and offer a possible solution that is backed by data. The solution can be a policy suggestion or a product that helps city users (residents, commuters, tourists, etc).

The outcome of the final project will be a report, your code, and a 5-minute in-class presentation. You can, but are not required, to create software, such as an app or a website, that will make the fruits of the project more accessible.

The grading will be based on the following factors:

- Fulfilling the basic requirements given below
- Impact on life in the city: tackling serious and tough problems
- Creativity in the project and with combining data from multiple sources.
- Meticulous and rigorous data analysis process.
- Sensible and helpful visualization.
- Making the results of the project accessible to policy makers or city users.

Requirements

Data selection: Your project should involve data from at least 2 different datasets. The datasets should be complex and comprehensive enough to support your analysis. Include an overview of each dataset you use. If there are other related datasets that you prefer not to use, include a quick reasoning and comparison if applicable.

Tools: Use Python, Spark (if necessary) and other tools mentioned in the class as the backbone of your project. You are free to use whatever additional package you need.

Soundness of analysis: You should treat the whole data handling process and your report as if it is targeted to experts in the field. Your solution must stand up to scrutiny. Major design decisions, data cleaning, feature selection, approximation and hyperparameter selection needs to be documented and explained. Every analysis and data interpretation needs to be supported by evidence.

Academic integrity: We take the originality of your work very seriously. Include proper citation on the works that you have consulted.

Datasets

Here is a list of useful datasets that might give you an idea of what kind of data you should look for. This is by no means an exhaustive list. You are free to find and use other data that best suit your goal.

- NYC Open Data: https://opendata.cityofnewyork.us/data/
- MTA Developer Data: http://web.mta.info/developers/download.html
- Maps, street view, event registries and any other data sources of your choice.

Submission

Your submission to CMS should be a single .zip file which includes:

1. A report that includes all the contents mentioned above. Include every detail needed to justify the soundness of your analysis.
2. A copy of the presentation presented in class.
3. The code used for your project. Include adequate documentation so that we can reproduce your results.

Presentation

There will be an in-class presentation on May 1st and May 6th. Each group will present for at most 5 minutes. The presentation should cover problem description, dataset selection, analysis process and result. You should assign time to each part in a way that your

presentation highlights the most creative part of your project.

The link for signing up the presentation time slot will be posted on Slack.