**Dilshat Uteshev**

Python Data Engineer Assignment

*General information:*

Size of analyzed NYC taxi dataset: **240 GB (197 csv files)**

Hardware: **Intel(R) Core(TM) i5-7300HQ, 8GB RAM CPU**

Language used: **Python 3.6**

*Instructions on installing:*

Before you start, please make sure that your environment fits the requirements, which you may find in the enclosed **requirements.txt** file.

Also, please check and adapt configurations of the program in the enclosed **config.yml** file:

* Main section **“paths”**
  + PATH\_CD - path to the shapefile of NYC districts
  + PATH\_BB - path to the shapefile of NYC boroughs
  + PATH\_AP - path to the shapefile of NYC airports
  + PATH\_DATA - path to the NYC taxi dataset (all csv files)
  + PATH\_RESULT - path to folder, where you want to store filtered files
  + PATH\_TAXI\_ZONES - path to the file mapping of location IDs and their names, file “taxi\_zone\_lookup.csv”
  + PATH\_TMP - path to the folder, where you want to store files generated while analysis

Now you are ready to start running the program.

**Task #1** - *“Select the rides which started from Manhattan and ended in John Fitzgerald Kennedy International Airport. The filtering does not need to be perfect but 80% correct”*

1. First, we need to clean and preprocess our dataset:
   1. There are some error lines in the yellow\_tripdata\_2010-02.csv and yellow\_tripdata\_2010-03.csv files. Let’s remove them, for this please use the following enclosed script:

./remove\_error\_lines.sh

* 1. In addition, file **filtering\_utils.py** has a number of useful scripts for preprocessing and filtering input data (function descriptions are available).

1. Now, to select the rides that started from Manhattan and ended in John Fitzgerald Kennedy International Airport, please use the following command:

python main.py

This script will filter all the available taxi rides to retrieve only required rides.

1. If you want to filter only one file, you may use **filter\_one.py** script:

python filter\_one.py <filename>

**Task #2** - *“Identify whether there is a correlation between the number of rides within a day and the weather, and more specifically the precipitation. The answer can be provided as free text, graphs, spreadsheets, csv files or any other way you find the most appropriate”*

**Bonus tasks:**

* *Visualize all of the filtered rides.*
* *Provide a way in your visualization where the filtered rides can be distinguished between precipitation levels.*

Solutions to the all above tasks you may find in the enclosed “**Analysis and Visualization.html**” document.

The code is available in the enclosed “**analysis and visualization.ipynb**” jupyter notebook, and some helpful utils you may find in the **viz\_utils.py** file (function descriptions are available).

Prerequisites:

1. Before running the code, please make sure that you have file **“noaa.csv”**, weather-related dataset with meteorological measurements from a National Climatic Data Center weather station located in Central Park, in the same folder as “**analysis and visualization.ipynb”**
2. As well as the folder you specified in **PATH\_RESULT** in configuration file contains processed files (i.e. only filtered rides).