Big Mobility Data Analytics for Public Health Data/Code Interview Task

Report your results in this form: https://forms.gle/1GPXFF1b7AKDaBDZ9

Overview: The goal of this exercise is to link two longitudinal spatial datasets, to perform some descriptive analyses and visualizations, and to describe how you would set up a predictive model.

Instructions: The task must be completed in a Jupyter Notebook, uploaded to your GitHub page. You can use the packages and algorithms you prefer. Report all results in the Jupyter Notebook. All steps must be visible. Make it clear and simple. Please include comments, titles, and explanations. The task should take approximately 2 hours to complete. The deadline for this task is **Monday Feb 27th, 11:59pm PST**

1. **Download datasets:**
   a. Download a data set on all recorded NYC taxi trips on 01/15/2015, from this site: https://github.com/uber-web/kepler.gl-data/blob/master/nyctrips/data.csv
      i. This data set, downloaded from NYC Taxi and Limousine Commission (TLC) website, includes yellow and green taxi trip records capturing pick-up and drop-off dates/times, pick-up and drop-off locations (latitude, longitude), trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.
   b. Download a data set on NYC restaurant locations (and inspection results) from this link
      i. The New York City Department of Health and Mental Hygiene (DOHMH) conducts unannounced restaurant inspections on an annual basis in order to check for compliance with policies on food handling. Data on these restaurant inspection results are publicly available at NYC Open Data and are updated daily. The specific data set for download above was accessed on 8/28/17 and contains records from 2014 - 2017. This older data was chosen in order to be more aligned with the taxi data, collected in 2015. This is important, since there is high turnover in restaurants over time (i.e., many go out of business).
      ii. The data set contains information on restaurant name and location, type of food (CUISINE DESCRIPTION), inspection date, and details on violation codes, total scores, and associated grades. The data is longitudinal in nature, with multiple rows per restaurant representing inspections over time. A full data dictionary is available here.
      iii. For the following questions, **we will focus on analyzing the locations of restaurants** (specified by the combination of their building, street, and zipcode features) **and not their inspection results**.

For the following questions, make the assumption that all taxi trip destinations during *lunchtime* (11:30am - 2pm) and *dinnertime* (5pm - 9pm) were to restaurants.

2. **Link the two datasets spatially** by finding restaurants that were (assumed to be) visited at the destination of a taxi ride during lunchtime or dinnertime. You can implement this by finding the closest match to a restaurant address within a 50m buffer (radius of 50m) around a latitude longitude point in the taxi data; that is, count matches only if they are within 50m of the taxi destination.
3. **Create an exploratory map** visualizing the linked data using your choice of packages and/or visual tools (e.g. GeoPandas, or platforms like Kepler.gl, Deck.gl, and Apache Superset). Please include at least one feature from each of the datasets in your map.

4. **Answer the following exploratory data analysis questions, providing a descriptive figure summarizing the results for each**:
    a. How far do people travel based on different types of cuisine ("CUISINE DESCRIPTION")? How does this differ based on the borough where the restaurant is located ("BORO", one of 5 large NYC neighborhoods)? How does this differ by meal time?
    b. What is the average tipping rate for different types of cuisine? How does this differ by borough, by meal time, and by number of passengers in the taxi?
5. **Describe how you would set up a predictive model** of restaurant cuisine type to be visited by a taxi rider based on information present in the two datasets. Feel free to suggest other data sources you would bring in to support your predictive model. Please note you do NOT have to implement this model.
6. **Reporting the results**: Make a Jupyter Notebook explaining all the steps you perform. Upload the results to your GitHub page. Make sure the repository is public and submit the link to the repository here: **LINK TO SUBMIT REPOSITORY**