Via Data Challenge

Part 1: Aggregation

Data sources:

- NYC taxi data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (also available through BigQuery https://bigquery.cloud.google.com/table/imjasonh-storage:nyctaxi.trip_data, or http://www.andresmh.com/nyctaxitrips/)
- NYC borough polygons: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

1. Propose a metric and/or algorithm to assess the potential efficiency of aggregating rides from many vehicles into one, given the available data. Make realistic assumptions and any necessary simplifications and state them.

2. Implement your proposed method and evaluate Manhattan's overall efficiency using **yellow taxi data from the first full week (Monday-Sunday) in June 2016**. Discuss how your method would scale with more data; in other words, discuss the complexity of your implementation.

3. Based on your implementation in the previous question, use visualizations to show how efficiency varies with time and location. Discuss any potential business implications based on your findings.

Part 2: Statistics

1. Assume Via has a demand prediction algorithm for predicting hourly demand, 24 hours in advance. **Answer the following questions theoretically:**
   a) How could you compare the performance of this algorithm across cities of varying size? Propose a way to predict performance of the demand algorithm in a city we're considering operating in.
   b) A data scientist builds a new demand prediction algorithm, but it is more resource intensive than the current one - we'll only switch if we're sure the new one is better. How would you determine if the new algorithm is worth it?

2. A report claims that between 22.4% and 43.9% of Via rides have excellent music. You can assume that "excellent" is a binary decision at the ride level. What do you think the sample size was?

You will be assessed on:

- The readability of your presentation
- The creativity, validity, and brevity of your solutions
- The quality of your code