LAPORAN TUGAS 2
SELEKSI WARGA BASDAT

Eksplorasi Data Science dengan Dataset Perjalanan Taxi

Disusun oleh :

Geraldi Dzakwan 13514065
M. Isham Azmansyah F. 13514014

Teknik Informatika
Institut Teknologi Bandung

2016

Bab I
Penjelasan Dataset

Kami melakukan eksplorasi terhadap dataset perjalanan taxi pada tugas terkait *data science* ini. Data set berisi riwayat data perjalanan taksi di kota Porto, Portugal. Dataset memiliki atribut serta penjelasan sebagai berikut :

1. TRIP_ID: (String) Mengandung ID unik dari setiap perjalanan
2. CALL_TYPE: (char) Mengidentifikasi cara pemesanan taxi dari sebuah trip. Memiliki tiga kemungkinan nilai :
    1. 'A' jika taxi berangkat dari pool utama, dipesan lewat telepon
    2. 'B' jika taxi dipesan langsung ke seorang supir dari stand tertentu
    3. 'C' jika taxi di-stop di sembarang lokasi jalan
3. ORIGIN_CALL: (integer) Jika call_type = 'A', maka origin_call mencatat nomor telepon pemesan. Jika bukan, origin_call bernilai null.
4. ORIGIN_STAND: (integer) Jika call_type = 'B', maka origin_stand mencatat stand taxi. Jika bukan, origin_stand bernilai null.
5. TAXI_ID: (integer): Berisi ID supir untuk setiap perjalanan
6. TIMESTAMP: (integer) Unix Timestamp dalam detik yang menandakan waktu awal perjalanan
7. DAYTYPE: (char) Mengidentifikasi jenis hari dari sebuah trip. Memiliki tiga kemungkinan nilai :
    1. 'B' jika hari libur atau hari spesial lainnya
    2. 'C' jika hari adalah sehari sebelum hari tipe B
    3. 'A' jika hari biasa baik weekday maupun weekend
8. MISSING_DATA: (Boolean) Bernilai salah jika stream data GPS lengkap, bernilai benar jika satu atau lebih lokasi pada stream hilang
9. POLYLINE: (String): Mengandung list dari koordinat GPS dalam format WGS84, dipetakan dalam bentuk string. Awal dan akhir dari string ditandakan dengan kurung siku ([dan]). Koordinat didefinisikan sebagai [LONGITUDE, LATITUDE] pada string. List ini mencatat satu koordinat untuk setiap 15 detik perjalanan. Koordinat pertama adalah koordinat awal perjalanan dan koordinat terakhir adalah koordinat destinasi/lokasi tujuan perjalanan.

Bab II
Langkah-langkah Eksplorasi dan Pembagian Tugas

Karena ukuran file sangat besar dan sulit untuk melakukan pemrosesan secara cepat dengan format file .csv, maka sebagai langkah awal kami terlebih dahulu mengubah file train.csv menjadi file train.hdf5.

HDF merupakan singkatan dari *hierarchical data format*. Format file ini biasa digunakan untuk bekerja dengan data yang besar. Untuk mengubah file train.csv menjadi format ini, kami menggunakan dua script, yakni Convert.py dan ToH5.py. Convert.py dijalankan terlebih dahulu untuk melakukan konversi file menjadi semacam file antara. Kemudian, file antara tersebut dikonversi dengan ToH5.py untuk menjadi file hdf5. Eksekusi convert.py berjalan secara lambat namun eksekusi ToH5.py berjalan cukup cepat. Kedua script ini dikerjakan oleh Isham.

Setelah diubah jadi file HDF5, baris trip pada file otomatis terurut berdasarkan TRIP_ID, namun isinya tetap sama. Kami mengabaikan baris trip yang berisi polyline kosong (tak ada catatan tempat-tempat yang dikunjungi). Untuk trip yang polyline-nya hanya berisi satu koordinat, kami mengasumsikan bahwa koordinat tersebut adalah lokasi awal sekaligus lokasi akhir dari trip.

Kemudian, kami membagi tugas. Saya (Geraldi Dzakwan) bertugas menyelesaikan soal A mengenai tempat yang paling sering dikunjungi di antara seluruh trip taxi yang ada serta soal B mengenai lokasi terakhir setiap perjalanan. Sedangkan, Isham bertugas menyelesaikan soal C mengenai prediksi total waktu tempuh suatu perjalanan taxi berdasarkan lintasan-lintasan awal dari tiap perjalanan.

Tools yang kami gunakan adalah Python. Package yang digunakan adalah sebagai berikut :

1. Untuk mengola data frame, kami menggunakan package Pandas.
2. Untuk visualisasi data, saya menggunakan package matplotlib. Untuk map menggunakan pyplot dan basemap yang ada dalam matplotlib, sedangkan untuk bar chart hanya mengguakan pyplot saja
3. Untuk data analysis, Isham menggunakan package scikit-learn.

Alasan kami menggunakan Python karena kami sudah pernah menggunakan bahasa tersebut sebelumnya untuk pengerjaan beberapa program sehingga tidak harus belajar dari awal. Hanya saja, kami tetap harus mempelajari package-package yang digunakan karena itu merupakan hal baru bagi kami.

Untuk soal A diselesaikan dengan script MostVisited.py. Soal B diselesaikan dengan script DestLocation.py. Cara run file beserta argumen yang diperlukan tertulis dalam komentar script. Kedua script ini masing-masing memerlukan tiga script lain, yakni :

1. CoordinateToPlace.py. Script ini berfungsi memetakan koordinat pada polyline menjadi sebuah tempat berdasarkan file metadata.csv yang diberikan. Walaupun sebenarnya meta data sebenarnya hanya berisi informasi lokasi taxi stand dan tidak mencakup seluruh tempat di kota Porto, namun persebaran taxi stand pada meta data sudah cukup baik untuk merepresentasikan wilayah-wilayah yang ada di kota Porto. Cara pemetaannya yakni mencari tempat yang jaraknya dengan koordinat pada polyline paling kecil daripada tempat yang lain. Hal ini dilakukan karena tidak semua koordinat pada polyline memiliki nilai yang sama persis dengan koordinat tempat pada meta data. Selain itu, script ini digunakan pula untuk menghitung jarak setiap tempat pada meta data terhadap pusat kota Porto untuk analisis lebih jauh.
2. BarChart.py. Script ini berfungsi untuk menggambarkan bar chart / grafik batang dengan sumbu x adalah placeID dan sumbu y adalah frekuensi kemunculannya. Bar disusun terurut bukan berdasarkan placeID di meta data. Namun, bar disusun terurut berdasarkan jaraknya terhadap pusat kota Porto. Bar paling kiri menandakan tempat paling dekat dengan Porto, sedangkan bar paling kanan menandakan tempat paling jauh dari Porto.
3. BasemapPortugal.py. Script ini berfungsi untuk plotting koordinat ke map kota Porto. Untuk soal A, koordinat diberi label nama tempat karena jumlah koordinat sedikit. Sedangkan, untuk soal B, koordinat tidak diberi label nama tempat karena jumlah koordinat sangat banyak.

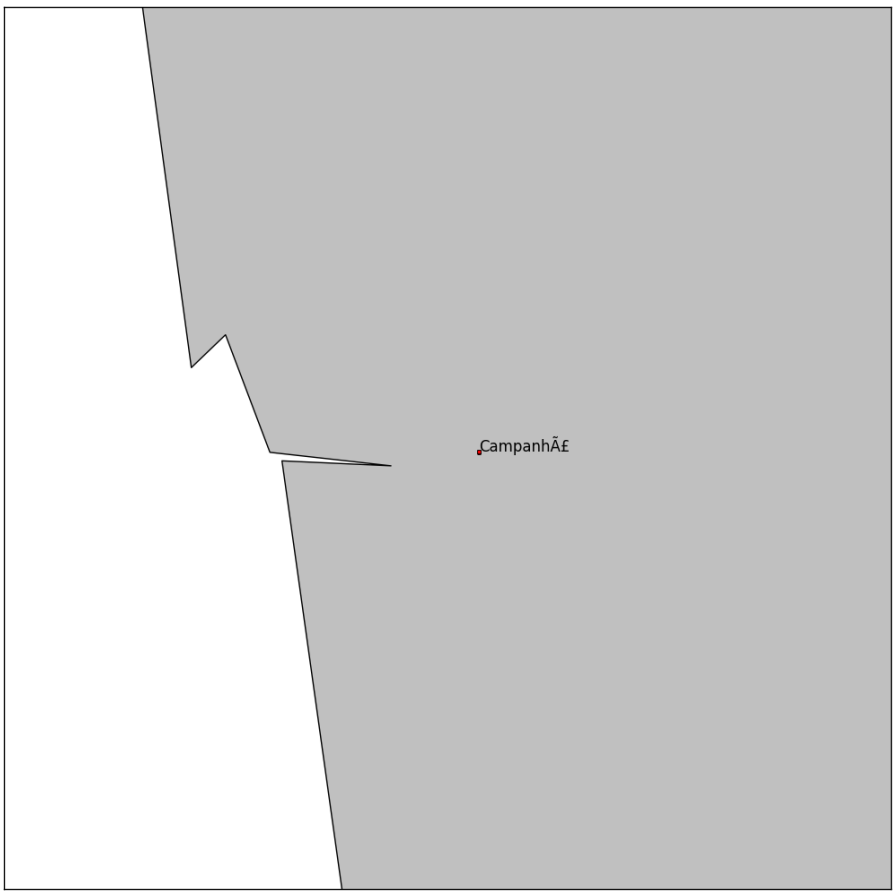
Kelima script tadi merupakan script yang saya (Gerald) kerjakan.

Untuk soal C diselesaikan dengan script Prediction.py dengan bantuan script CreateMLData.py. Script CreateMLData.py melakukan transformasi data untuk mempermudah program, sedangkan script Prediction.py menjalankan algoritma Nearest Neighbor. Kedua script ini beserta script Convert.py serta ToH5.py adalah script yang Isham kerjakan.

Bab III
Hasil Analisis dan Visualisasi

A. Tempat yang paling sering dikunjungi di antara seluruh trip taxi yang ada

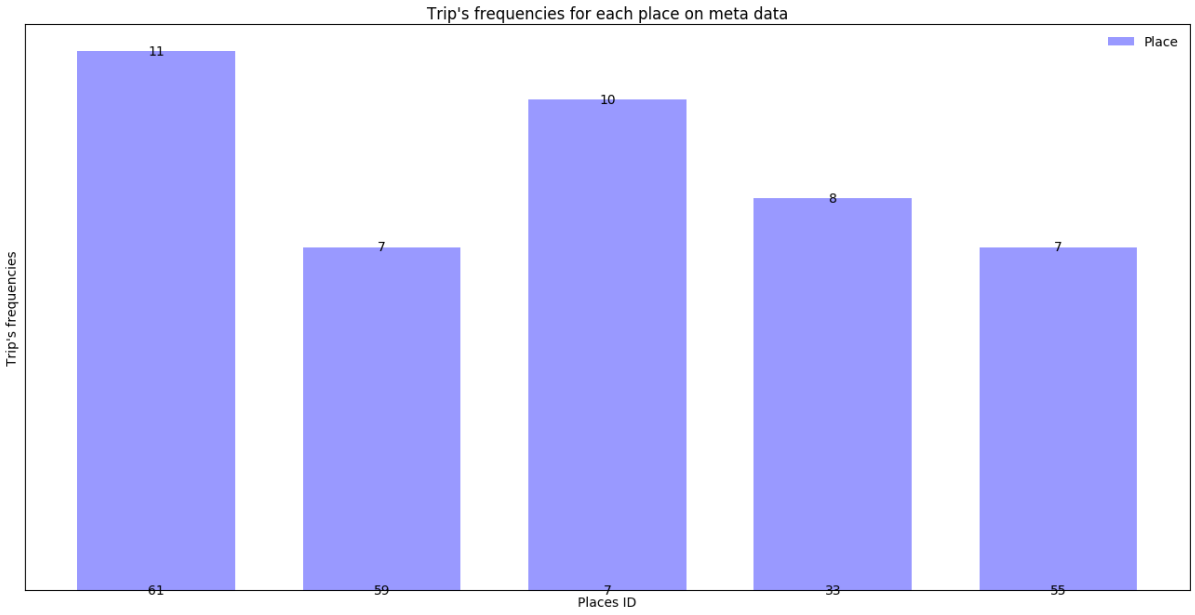
Kami menjalankan script MostVisited.py dengan mengambil seluruh data (tanpa sampling) dan melakukan listing terhadap 20 koordinat yang paling banyak dilewati. Hasilnya, seluruh koordinat terletak di dekat tempat dengan placeID = 15, yakni CampanhÃ£. Tempat ini memiliki koordinat (41.14862751, -8.585876603). Untuk detail 20 koordinat beserta frekuensi kemunculan masing-masing, detailnya bisa dilihat melalui file output csv yang di-upload di github.



Total frekuensi dari 20 koordinat tersebut adalah 17944 (bisa dilihat di csv/bar chart yang di-upload di github).

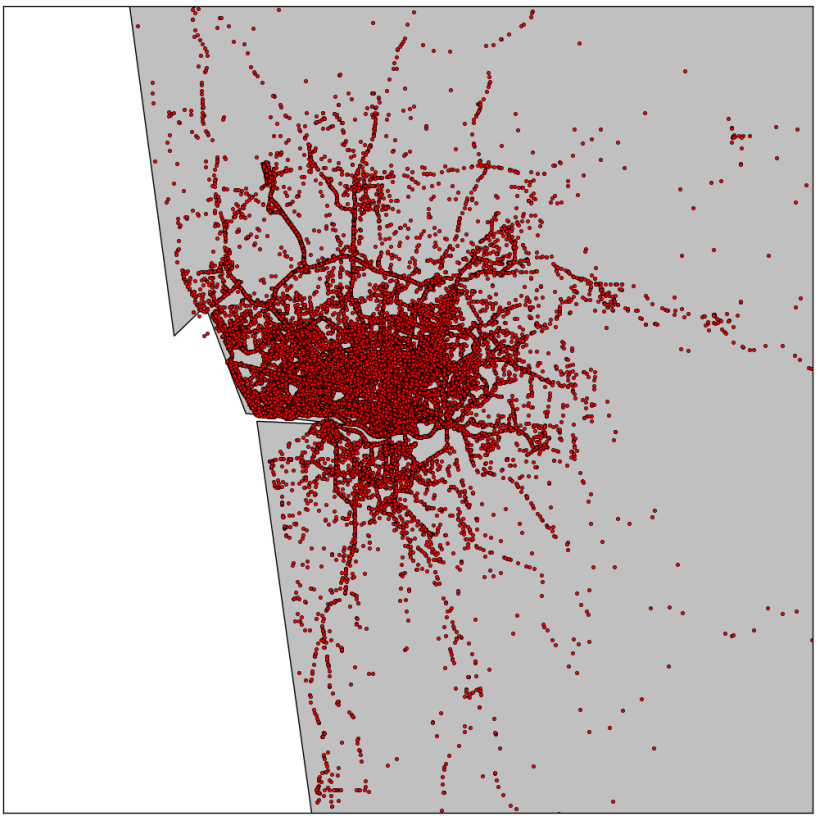
Sebagai perbandingan, kami juga menjalankan script untuk sample 2000 data awal (pertama) dan melakukan listing terhadap 5 koordinat yang paling banyak dilewati.

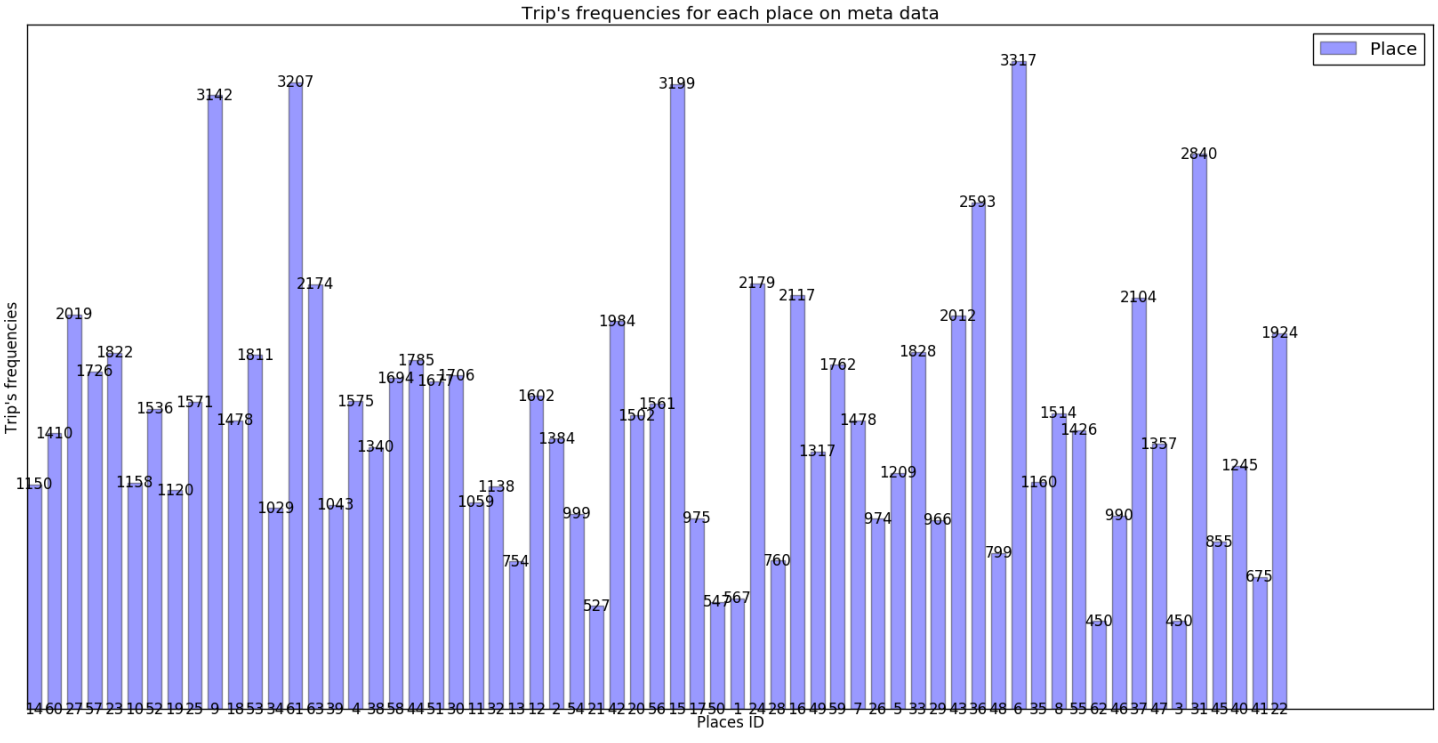
Hasilnya, terdapat 5 tempat yang di-plot di map (masing-masing koordinat terpetakan ke tempat yang unik). Visualisasinya sebagai berikut :



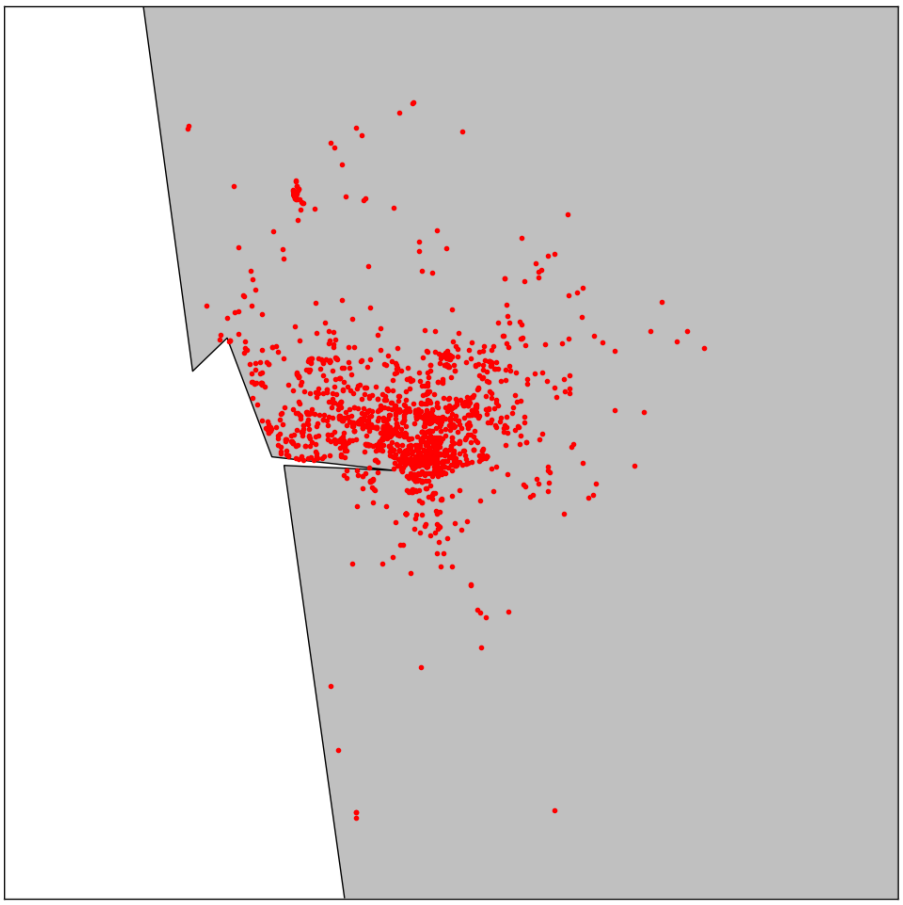
B. Lokasi terakhir pada setiap perjalanan

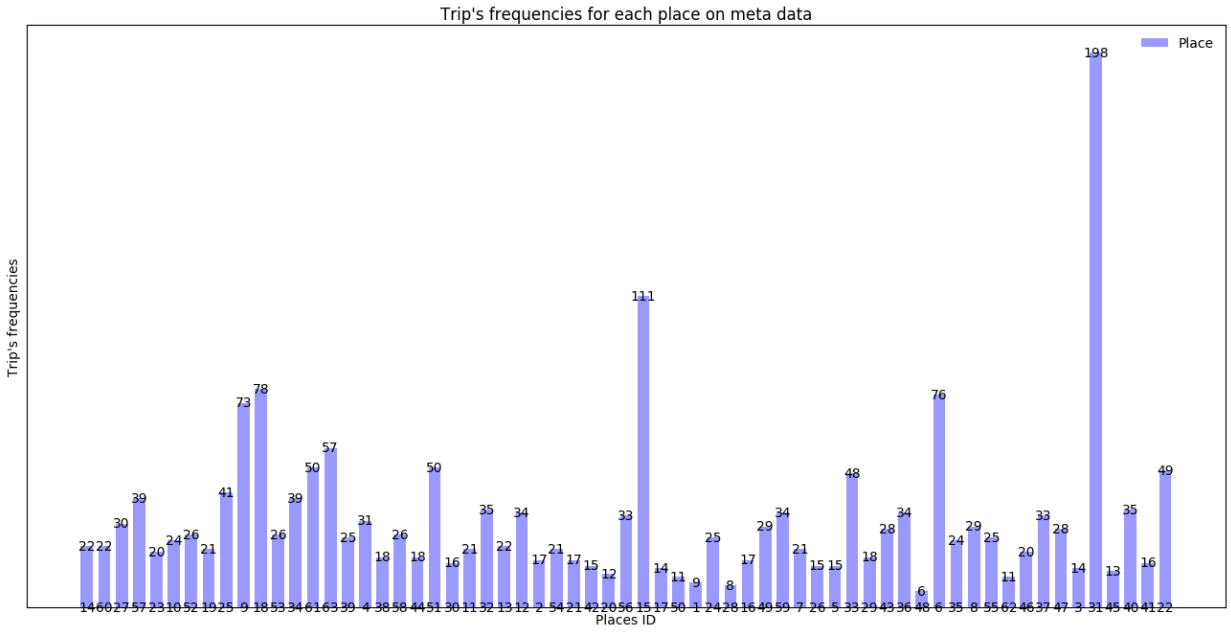
Kami menjalankan script DestLocation.py untuk mengambil semua koordinat akhir dari setiap perjalanan. Dalam script itu, kami juga mengambil 100 ribu sampel titik akhir untuk divisualisasikan. Hasilnya adalah peta yang cukup menggambarkan tujuan akhir taxi di kota Porto. Pada beberapa tempat, dapat terlihat bentuk jalan kota Porto. Berikut visualisasi persebaran di map dan frekuensi masing-masing lokasi akhir pada bar chart :





Sebagai perbandingan, kami juga menjalankan script untuk sample 2000 data awal (pertama). Hasil visualisasinya sebagai berikut :





C. Lokasi terakhir pada setiap perjalanan

Kami menjalankan script CreateMLData.py untuk mengubah seluruh data train dan test menjadi data yang dapat digunakan untuk permasalahan ini. Kemudian, kami menjalankan script Prediction.py untuk memprediksi waktu tempuh dalam detik menggunakan metode Nearest Neighbor. Hasil prediksinya dapat dilihat di file csv yang telah diupload di github. Untuk soal C tidak diberikan visualisasi, karena hasil prediksi sudah cukup jelas pada file CSV.

Sebagai tingkat keakuratan, model prediksi akurat sebesar 36% (115 dari 320) jika perbedaan waktu aktual dan prediksi kurang dari 5 menit. Sedangkan, jika diambil perbedaan waktu aktual dan prediksi kurang dari 10 menit, didapat tingkat keakuratan 73% (235 dari 320).

Tentu, tingkat keakuratannya masih rendah. Oleh karena itu, untuk pengembangan selanjutnya, model yang dibuat harus dikembangkan. Misalnya, tidak hanya melihat lokasi awal, namun juga trayek-trayek awal dan menggunakan algoritma yang lebih baik dari nearest neighbor untuk training.

Bab IV
Kesimpulan

Kami mendukung sebagian pernyataan CEO Uber, Travis Kalanick yang menyatakan bahwa Uber hadir karena mempunyai rencana yang pada akhirnya akan membuat pengurangan jumlah mobil dan peningkatan jumlah orang untuk berpindah tempat. Kami tidak bisa memberikan dukungan atau penolakan terhadap bagian pengurangan jumlah mobil, karena dataset yang diberikan tidak relevan dengan hal tersebut. Tidak diketahui berapa proporsi trip dengan taxi dan mobil pribadi. Namun, kami mendukung bahwa Uber menghasilkan peningkatan jumlah orang yang berpindah tempat. Beberapa alasannya adalah sebagai berikut :

1. Dari hasil analisis tempat yang paling sering dikunjungi, diperoleh CampanhÃ£ sebagai tempat paling sering disinggahi oleh perjalanan taxi. CampanhÃ£ terletak tidak dekat dengan pusat kota. Oleh karena itu, saya menyimpulkan bahwa perjalanan taxi memengaruhi peningkatan jumlah orang untuk berpindah tempat karena sebagian besar perjalanan melewati daerah ini yang notabene tidak dekat dengan pusat kota.

2. Dari hasil analisis lokasi terakhir perjalanan, kita dapat melihat pada visualisasi map bahwa lokasi terakhir perjalanan banyak ditemukan di tempat yang sangat jauh dari pusat kota. Meskipun, mayoritas masih berada di sekitar pusat kota. Detil lebih presisi dapat dilihat dari bar chart untuk 100 ribu sampel lokasi akhir. Bar chart menampilkan tempat urut dari yang dekat ke pusat kota (bagian kiri) hingga jauh dari pusat kota (bagian kanan). Dari bar chart, kita bisa melihat bahwa keduanya cukup berimbang. Bahkan, tempat dengan placeID No.6 yang cukup jauh dari pusat kota memiliki frekuensi terbanyak (3317 kali sebagai destinasi). Artinya, banyak perjalanan taxi memiliki destinasi ke lokasi-lokasi yang jauh dari pusat kota.