LAPORAN TUGAS 2

SELEKSI ASISTEN BASIS DATA 2016

Oleh: 13514033 dan 13514025 (kelompok 6)

A. Penjelasan dataset

Dalam tugas 2 ini, kami memilih dataset dengan judul "TAXI TRIP". Dataset ini berisi data-data mengenai rekam jejak perjalanan dari taksi di suatu wilayah. Khususnya untuk data ini, diceritakan rekam jejak perjalanan beberapa taksi UBER pada suatu wilayah di Portugal.

Kumpulan data yang kami dapatkan mengandung atribut yang terdiri dari TRIP_ID, CALL_TYPE, ORIGIN_CALL, ORIGIN_STAND, TAXI_ID, TIMESTAMP, DAY_TYPE, MISSING_DATA,dan POLYLINE. Setiap bagian menunjukkan informasinya masing-masing. Penjelasannya ialah sebagai berikut.

- TRIP_ID menunjukkan ID dari setiap perjalanan
- ORIGIN_CALL menunjukkan nomor telepon yang dihubungi dari permintaan customer
- ORIGIN_STAND berisi sebuah ID yang dihubungkan dengan ID kota pada 'metadata' dan menunjukkan posisi awal taksi saat dihubungi *Customer*
- TAXI_ID menunjukkan ID taxi yang telibat pada suatu perjalanan
- TIMESTAMP berisi informasi waktu dimana perjalanan dimulai
- DAYTYPE mengidentifikasikan keadaan hari pada perjalanan, misalnya pada saat liburan atau hari spesial lainnya (tipe B), sehari sebelum liburan (tipe C), atau hari-hari biasa dan weekend (tipe A)
- MISSING_DATA menunjukkan apakah lokasi yang ditunjukkan oleh GPS hilang atau tidak. Jika tidak hilang, bernilai *False* dan sebaliknya.
- POLYLINE berisi koordinat-koordinat (dalam *latitude* dan *longitude*) yang menunjukkan *tracking* perjalanan taksi pada suatu trip. *Tracking* posisi taksi tersebut dilakukan setiap 15 detik sekali. Koordinat pertama menunjukkan posisi awal dari customer, koordinat akhir menunjukkan destinasi sebuah perjalanan.

B. Langkah Analisis

Pertama-tama, kita harus mengetahui informasi-informasi mana saja yang penting untuk dipakai dalam analisis data sesuai dengan kasus yang diberikan. Hal-hal yang diminta pada soal diantaranya.

1. Tempat yang paling sering dikunjungi pada trip yang ada
2. Lokasi akhir pada setiap perjalanan
3. Prediksi total waktu tempuh rata-rata pada setiap lintasan perjalanan

Maka, hal-hal tersebut membutuhkan informasi yang terdapat pada kolom 'POLYLINE'. Kolom yang berisi informasi tentang *tracking* posisi taksi pada rentang waktu terssebut. Secara general, untuk mengakses *value* dari suatu atribut POLYLINE, kami menggunakan sebuah converter yang dimiliki *library* json, yaitu langsung mengubahnya menjadi *array of coordinates* sehingga lebih mudah untuk diakses.

Secara general, script atau kode program kami dituliskan dengan Bahasa Python. Lalu, kami memakai beberapa library untuk membantu pengelolaan data dari csv. Yang pertama, kami memakai "csv" dan "pandas" untuk membaca file ekstensi .csv sebagai data asal. Library "csv" juga kami gunakan untuk menghasilkan file eksternal dalam menulis jawaban pertanyaan ke dalam file berekstensi .csv. Untuk mengakses koordinat-koordinat POLYLINE, kami menggunakan "json" yang memiliki converter khusus dalam mengakses string berbentuk array. Untuk pengoperasian data, kami gunakan "numPy". Lalu, untuk visualisasi kami gunakan "matplotlib" yang keluaran gambarnya berupa file berekstensi .png dan "folium" yang keluarannya peta berekstensi .html.

Untuk penyelesaian masalah per kasus, dijelaskan sebagai berikut.

- Kasus 1
  Pada kasus ini, kami menambahkan suatu asumsi. Yakni, tempat yang dikunjungi pada setiap perjalanan ialah destinasi perjalanan yaitu koordinat terakhir pada POLYLINE. Karena jika dilihat, koordinat-koordinat lain yang terdapat pada sebuah data POLYLINE merupakan *tracking* taksi yang sedang bepergian sehingga bukan merupakan 'tempat yang dikunjungi'.

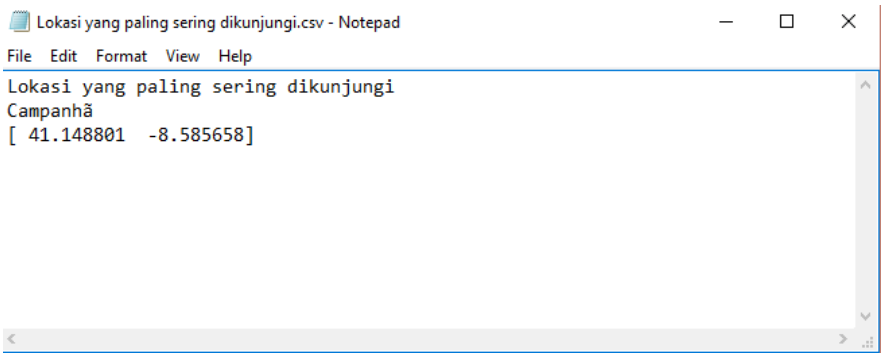
  Maka, tempat yang paling sering dikunjungi berarti titik atau koordinat akhir yang paling sering muncul dari seluruh data. Disini, kami menggunakan library collection yaitu Counter, yang dapat menentukan koordinat dengan kemunculan paling banyak pada keseluruhan data.

- Kasus 2
  Untuk lokasi akhir, kita langsung saja ambil koordinat terakhir pada atribut POLYLINE. Untuk nama lokasinya, kita dapat mengakses file meta-data yang tersedia dan membandingkan koordinat *latitude* dan *longitude* nya. Untuk membandingkannya, kami menentukan sebuah titik apakah berada dalam rentang radius tertentu dari setiap nilai koordinat di meta-data.

- Kasus 3
  Pada suatu perjalanan, 'jarak' antara dua koordinat nilainya 15 detik. Maka, kita cari untuk setiap data POLYLINE berapa panjang (len) array nya. Waktu tempuh total pada suatu perjalanan adalah (len-1)x15. Setelah semua waktu tempuh diakumulasikan dengan dijumlah, akhirnya dibagi jumlah data/ *tuple* yang ada. Hasilnya adalah rata-rata waktu tempuh. Namun masih ada kolom yang harus dipertimbangkan, yaitu pada MISSING_DATA. Jika *value* nya TRUE, maka tidak akan sertakan dalam jumlah data (bagian pembagi). Karena jika *value* nya TRUE, maka lokasi tidak terdeteksi dan kami menganggap hal tersebut tidak termasuk dalam "perjalanan".

C. Hasil Analisis

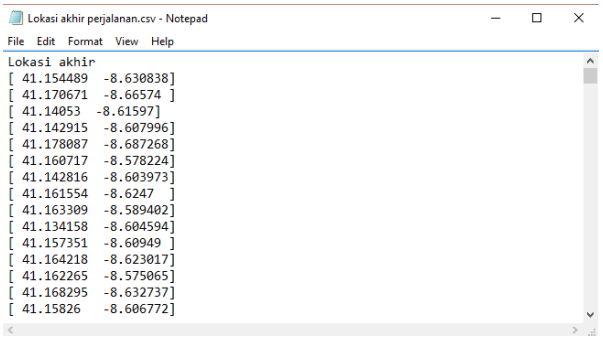
- Kasus 1



Gambar 1. Screenshot file hasil pemrosesan kasus 1

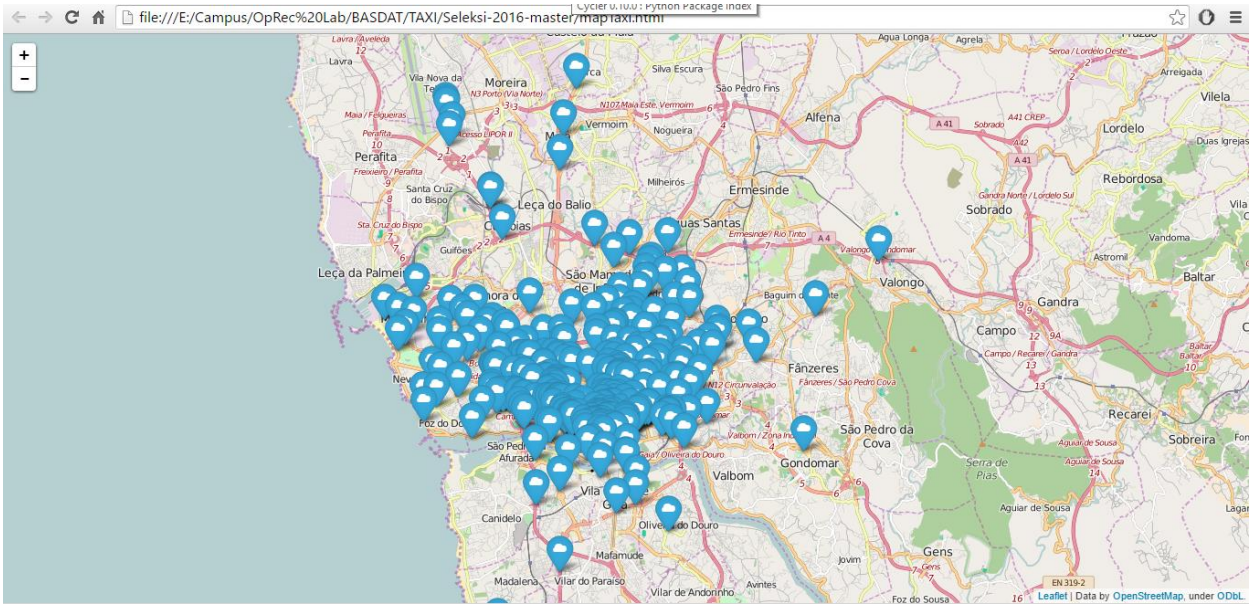
Dengan keluaran file csv seperti ini, kita mendapatkan jawaban bahwa lokasi yang paling sering dikunjungi dalam dataset train.csv adalah kota Campanha.

- Kasus 2



Gambar 2. Screenshot file hasil pemrosesan kasus 2

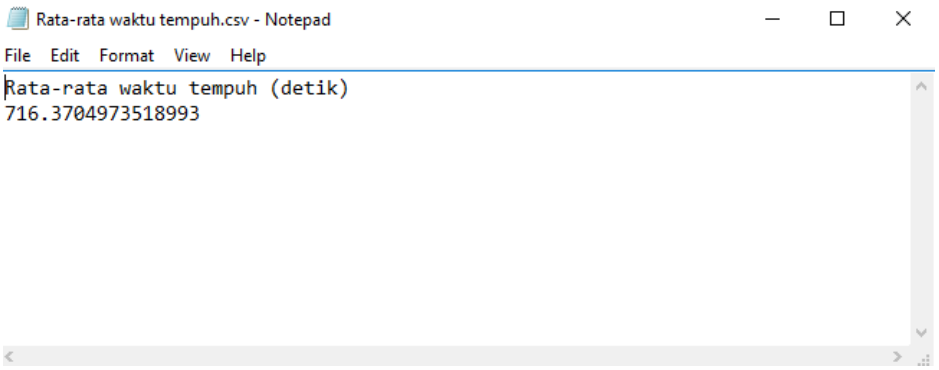
Gambar 3. Visualisasi hasil pemrosesan dalam file .png



Gambar 4. Visualisasi dengan folium

Dari visualisasi-visualisasi di atas, lokasi-lokasi yang menjadi tujuan setiap perjalanan sudah dapat digambarkan dengan jelas.

- Kasus 3



Gambar 4. Screenshot file hasil pemrosesan kasus 3

Seperti yang terlihat dari file keluaran ini, rata-rata waktu tempuh yang dilakukan pada dataset adalah 716,370 detik. Namun kalau dilihat satu per satu dari dataset asal, masih timmpang antara perjalanan yang satu dengan yang lainnya. Ada yang menempuh jarak cukup jauh, ada pula yang hanya berisi satu koordinat (kolom POLYLINE).

D. Penutup

Setelah dilihat dari jawaban-jawaban pertanyaan dataset tersebut, pernyataan CEO Uber untuk mengurangi jumlah armada mobil masih belum bisa disimpulkan lebih dalam. Namun dari pertanyaan kedua, terlihat pola lokasi yang menjadi destinasi akhir tiap perjalanan memperlihatkan tujuannya tidak jauh berbeda. Maka mengurangi armada Uber dapat dilakukan bisa saja dilakukan. Tetapi tetap harus ada solusi jika pemesanan customer tetap banyak. Salah satunya yaitu fitur baru Uber yaitu UberPool. Fitur ini dapat memperjalankan sekaligus beberapa customer (hingga 4 orang dalam satu mobil) yang arah tujuannya berdekatan dan tempat penjemputannya juga searah.

Untuk memperkuat argument tersebut, masih bisa dilakukan eksplorasi lebih lanjut supaya keputusan tersebut memang efektif. Banyak aspek yang harus dipertimbangkan. Sehingga dataset tersebut dapat dimanfaatkan maksimal dan memunculkan banyak terobosan baru guna memperbaiki pelayanan dan meningkatkan profit perusahaan tersebut.