COMS4060A/7056A: Assignment #2

Tim Bristow
tim@bristow.za.net

University of the Witwatersrand — October 29, 2021

Introduction

This assignment is based on content covering geospatial and time series data. You will be required to perform some basic data cleaning and exploration techniques on prescribed datasets. The aim is to explore the dataset and make observations. There is no strict requirement on the format of your submission, but any answers should be reasoned, discussed, and relevant data or results should be provided to substantiate this. You might like to submit either a PDF or a Jupyter notebook, for example. You can use any programming language or tool you would like, however.

1 NYC Taxi Data

You will analyse a public dataset from Uber available on Kaggle, available here: https://www.kaggle.com/c/nyc-taxi-trip-duration.

Your primary dataset is one released by the NYC Taxi and Limousine Commission (TLC), which includes pickup time, geo-coordinates, number of passengers, and several other variables for 1.5 million trips between 2016-01-01 and 2016-06-30. Note that for this analysis, just use the *training* sample.

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- trip_duration - duration of the trip in seconds

There is more up-to-date data available from TLC, but the datasets are large (10GB+ per year since 2018). They do include additional fields, however.

1.1 Data Cleaning

There are several outliers in the data. Identify these and give justification for why you can remove them from the analysis. (Hint: look at trip duration, speed, distance, etc). [3 marks]

1.2 Feature generation

This can be done before or after the data cleaning step. Generate additional columns for at least these features (but you're welcome to add more!):

- Distance of trip
- Day of week
- Average speed of trip

[3 marks]

1.3 Time-based questions

Assume pickup time unless otherwise specified.

1. Which day of the week is the most popular? Show plots to motivate your answer. [2 marks]
2. What hour of the day is the most popular on each day? Plot a distributions of the hours and make observations and give possible suggestions for why the data looks like it does. [3 marks]
3. Investigate the differences between weekdays and weekends. What would account for this? [2 marks]
4. How do these patterns change on the major holidays (do they change?): St. Patrick's Day, Easter, Memorial Day, for example? [3 marks]
5. How does the average speed of trips change throughout the day? What time of day are trips fastest? Show plots to motivate your answer. [2 marks]

1.4 Location clusters

Produce a heatmap of all of the trip pickups over (do not do a scatter plot... there are 1.5 million data points and this will almost certainly crash your computer):

1. weekdays and weekends,
2. morning and evening (choose reasonable hours).

Comment on any findings you make. [4 marks]

From the most popular times for Friday night/Saturday morning (around midnight) and Thursday afternoon, find hotspot locations (you will need to perform clustering). If you were to use k-means, you would define the number of clusters. However, here the number of clusters is not at all clear. Using an algorithm like DBSCAN (available in sklearn) determines this for you, and works well on spatial data. DBSCAN has two configurable parameters: ϵ - the maximum distance between any two points, and the minimum number of samples to determine a cluster. Your hotspot location might be defined as at least 15 pickups in that location in an hour, and locations might be required to be within 50 or 100 metres from each other (do motivate your choice of parameters). Using DBSCAN, identify clusters and plot these on a map. How many clusters did you find? [8 marks]

1.5 Airports

Find out how long it takes, on average, to travel to JFK airport from the Empire State Building. Produce a plot showing the travel time by time of day. How does this compare with Newark Airport? [5 marks]

1.6 Vendors

Investigate (ie plot) the difference between the two vendors. Duration of trips, number of passengers, pickup/dropoff locations. One of these represents Yellow Cabs, and the other is the green boro taxis. After reading up on these, can you identify which one is which from your plots? [4 marks]

1.7 Boroughs

You can find the shapefile containing NYC boroughs (basically neighbourhoods) here https://data.cityofnewyork.us/City-Government/2010-Neighborhood-Tabulation-Areas-NTAs-/cpf4-rkhq.

1. Using this shapefile find the neighbourhoods for the trip start and end locations (try geopandas, shapely, or fiona, for example). [3 marks]
2. Plot a chloropleth of all pickups and all dropoffs in NYC. What do you notice about the difference in distribution? [2 marks]
3. Which boroughs have the most incoming trips and the most outgoing trips? [2 marks]
4. Which borough(s) is/are the quietest at night, between midnight and 5AM? (Not everyone wants to party). [2 marks]
5. Which borough(s) is/are the busiest at night, between midnight and 5AM? (Some people party, well, only Rod actually). [2 marks]

1.8 Submission

Work by yourself or in groups of up to four people. Submit your work to Moodle as a PDF or Jupyter notebook.

Deadline: 17 November 2021 (please contact me directly regarding extensions)