CS210 - Introduction to Data Science
Individual Project

Due Date: 15.03.2019 23:55

In this project, you will be exploring and analysing a real world dataset which includes taxi trips in a span of two weeks in New York City.

NYC Taxi Trip Dataset

Each row in the dataset corresponds to a taxi trip. The attributes and their explanations can be found below.

Attribute	Explanation
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
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_store_and_fwd_flag	indicates whether the trip record was held in vehicle memory
trip_duration	duration of the trip in seconds

Project Description

The project consists of two parts; **data exploration** and **hypothesis testing**. In data exploration, you will extract and present insights about the data. And in the second part, you will evaluate two hypothesis regarding trip distances.

Data Exploration

- Give basic information regarding the dataset such as shape, data types and descriptive statistics that summarize columns.
- Create two new columns named **"pickup_district"** and **"dropoff_district"** by applying reverse geocoding [1] [2] to associated coordinates.
- Extract the top 5 districts where passengers prefer to leave and arrive.
- Create a new column named **"distance"** by utilizing pick up and drop off coordinates [3].
- Create a new column named **"time_of_day"** by aggregating timestamps in **"pickup_datetime"** into 5 different categories.
  - 7-9 AM: "rush_hour_morning"
  - 9 AM - 4 PM : "afternoon"
  - 4-6 PM : "rush_hour_evening"
  - 6-11 PM : "evening"
  - 11 PM - 7 AM : "late_night"
- Show how the average distance varies as time of the day changes.
- Show how the trip duration varies as time of the day changes.

Hypothesis Testing

1. Does passenger group size affect the distance?
   - Null hypothesis: passenger group size has no effect on the distance.
   - Apply a suitable statistical test and show the results.
2. Do trip distances increase in weekends?
   - Null hypothesis: The day of the week has no effect on the distance.
   - Again, apply a suitable statistical test and show the results.

[1] `https://developers.google.com/maps/documentation/geocoding/intro`
[2] `https://github.com/thampiman/reverse-geocoder`
[3] `https://pypi.org/project/geopy/`

Submission

As discussed earlier, we expect you to implement your projects on a notebook environment. Name your .ipynb file as ***name_surname*_indv_proj.ipynb** and upload to Sucourse. In addition, publish your notebooks online with nbviewer [4] as described in the first recitation. Put your url in a text file named ***name_surname*_url.txt** and upload it to Sucourse as well. As the final remark, please do not forget to write your name and ID in the first cell.

Policies

- This is an individual project. Please, work on your own!
- TAs do not have official office hours. If you need help, please send them an email and arrange a time slot.

[4] `https://nbviewer.jupyter.org/`