Explore the New York yellow taxi data

Task 1

1.1 Get the data

(1) taxi data

Included in the New York yellow taxi data dataset (2012/01/01-2016/06/30) are each taxi trip's pick-up and drop-off times, pick-up and drop-off GPS coordinates, distance, passenger count, fare, tip and manner of payment, inter alia. You can choose to download taxi data from the NYC website link or the Baidu network disk link below.

NYC link: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
(Note: only yellow taxi trip records are needed).
Baidu network disk link:
Link address: https://pan.baidu.com/s/1iTjE52t0DBHA1fLJ-cL5TQ
Extraction code:sg4q

(2) Job postings data and investment bank locations.

You can find *Job postings data.csv* and *Investment bank locations.csv* in the zip file.

***Job postings data.csv* :** We collected recruitment information for these investment banks in New York and counted the number of job postings that the investment bank posted on the recruitment website monthly.
Month column: for example, "12-Jan" means January 2012.
The total number of job postings column: The total number of job postings released by a particular investment bank monthly.
The company column: Name of the investment bank.

Investment bank locations.csv: We identified 16 investment banks in New York (including branches). GPS data comes from Google map, which you can use to locate these investment banks.

(3) Other data

Feel free to use the datasets you find to discover interesting stories. For example, you can download stock price data from Yahoo Finance to reflect the company's performance.

1.2 Clean the taxi data

Clean guide

Our goal is a dataset consists of investment bank originated night ride observations. All credit card transactions contain tips information (the ratio of credit card transaction increases gradually from 20% in 2009 to 50% in 2013). Cash transaction does not include tips. Each ID corresponds to a unique credit card transaction. We differentiate between asset management, global markets, and investment banks similar to Finer 2018. JPMorgan has two offices very close to each other, so one ID may correspond to more than one bank office.

Outlier removal (based on Finer 2018 and Saia 2019):

(1) Keep the period of 9:00 pm – 4:00 am (the next day).
(2) When drop-off time precedes the pick-up time, swap pick-up, and drop-off times.
(3) Remove rides with duration equal to zero or longer than 3 hours.
(4) Remove rides associated with a total payment lower than the minimum fare ($2.5 entry fee) or over $1000
(5) Keep rides with at least one passenger and no more than five passengers.
(6) Keep rides whose trip distance is at least 0.1 miles and less than 25 miles.
(7) Drop duplicates of identical pick-up time and pick-up latitude and pick-up longitude.
(8) Drop rides whose latitude and longitude are not between -90 and 90.

New variable construction based on distances:
The new variables added are the "Pick-up bank" and "Drop-off bank," which is based on pick-up and drop off location <0.06mile distance to bank location. You can add coordinates of investment banks based on <0.06mile distance to any bank office.

Task 2
Let's analyze something interesting!
Proposed independent variable:
Pick-up count represents the level of overtime working for a particular investment bank. We assume that the more people return home from investment banks late at night, the more common the overtime working of this company.
Tip percent represents the average ratio of the tip to the total fare of who returns home from the investment bank. It measures the mood of a person after working overtime. The lower the tip percent, the worse the mood.
Pick up count*tip percent represents the enthusiasm of employees, the greater the value, the better the working overtime mood, the small value means even though no working overtime, the mood is still not good.

Proposed dependent variable:
Total number of job postings: We collected recruitment information for these investment banks in New York and counted the number of job postings that the investment bank posted on the recruitment website monthly. When an employee leaves the company, the vacant position may be posted online.

Tips
*You can use panel regression to control fixed effects (year, company, etc.)
*You can explore other variables to find interesting associations. For example, the variable that reflects the company's performance: stock return.

Task 3:
Explain your findings using an economics story. You are encouraged to use visualization tools!

Please write your results in a PDF report and package it with your code (Python, R, etc.). **The submitting deadline is two weeks from your starting date on this test.**