CMSC 12300 - Spring 2018

Group name: Taxis in town

Members:

- Vu Phan
- Evan Mata
- Nanut Chaichanawanich

Project Description:

We hypothesize that when the stock market significantly changes there will be noticable effects on people's' desires and behaviours. We attempt to uncover some of these changes through the medium of taxi rides and data on both taxi tips and destinations, where we will be taking destinations to be a proxy for people's desires. We will be looking at a variety of different destinations: major banks, comfort/fast food restaurants, relaxation/parks and recreational areas, and international relations hotspots such as the UN building. We will also be looking at a sampling of random data over the same time period to provide a control to compare with.

That is, we will observe the percentage increase or decrease to these locations over a T-length time period (in the range of 1 hour to 1 day, as determined by the specific destination) every time the stock market changes by X% within Y amount of time (values TBD - any stock market changes over 2 standard deviations in rarity on the order of an hour likely work). These percentage increases will then be compared with the long term average to the same types of destinations.

Data Source:
Our primary source of data is http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml .
It records taxi rides between 2009 and 2017 with separate csv files for each month. The csv file for each month is about 2 Gb in size, so depending on the time range we want for our analysis, this is definitely a big data problem. Data on yellow rides is available for all the months. Data on green rides is available for most months and data on FHV rides is available for the most recent years.

The data set has 17 variables. For January 2009, there are over 14 million recorded trips.

Running head() on the pandas dataframe for January 2009:

```
In [3]: df = pd.read_csv('taxi09.csv')

In [4]: df.head()
Out[4]:
  vendor_name Trip_Pickup_DateTime Trip_Dropoff_DateTime  Passenger_Count  \
0         VTS  2009-01-04 02:52:00   2009-01-04 03:02:00                1
1         VTS  2009-01-04 03:31:00   2009-01-04 03:38:00                3
2         VTS  2009-01-03 15:43:00   2009-01-03 15:57:00                5
3         DDS  2009-01-01 20:52:58   2009-01-01 21:14:00                1
4         DDS  2009-01-24 16:18:23   2009-01-24 16:24:56                1

   Trip_Distance  Start_Lon  Start_Lat  Rate_Code  store_and_forward  \
0           2.63 -73.991957  40.721567        NaN                NaN
1           4.55 -73.982102  40.736290        NaN                NaN
2          10.35 -74.002587  40.739748        NaN                NaN
3           5.00 -73.974267  40.790955        NaN                NaN
4           0.40 -74.001580  40.719382        NaN                NaN

     End_Lon    End_Lat Payment_Type  Fare_Amt  surcharge  mta_tax  Tip_Amt  \
0 -73.993803  40.695922         CASH       8.9        0.5      NaN     0.00
1 -73.955850  40.768030       Credit      12.1        0.5      NaN     2.00
2 -73.869983  40.770225       Credit      23.7        0.0      NaN     4.74
3 -73.996558  40.731849       CREDIT      14.9        0.5      NaN     3.05
4 -74.008378  40.720350         CASH       3.7        0.0      NaN     0.00

   Tolls_Amt  Total_Amt
0        0.0       9.40
1        0.0      14.60
2        0.0      28.44
3        0.0      18.45
4        0.0       3.70
```

Proposed timeline

Task	**Estimated Deadline**
Identify reasonable numeric values for everything.	End of Week 4
Find the locations we count as destinations for our categories (ie between lat x, x' and long y, y' would be classified as recreational)	End of Week 6
Compute data - first averages so we can identify the stock market times we wish to target, then the relevant data.	End of Week 9