New York City Taxis

Group 5
Joe Froelicher, Tommy Guess, Michael Huffman

Questions to Answer

Our team is very interested in data visualization. All questions we venture to answer will include graphical components. Some of these questions include:

- Which riders (by origin) tip best? What other factors influence tipping?
- How did COVID impact the industry?
- Given the start point, time of day, and destination can we predict the cost? (Not sure if this is an interesting question or not, gets at the traffic patterns I think)
- Are there districts with high/low taxi traffic? (Maps)
- Are there districts with high/low ride costs? (Possibly exterior districts, people having to ride farther to get into the city)
- Are there district-to-district trips that occur more frequently than proximity alone would predict?
- Do the distributions of Green Cab Taxi rides reflect their mandate?

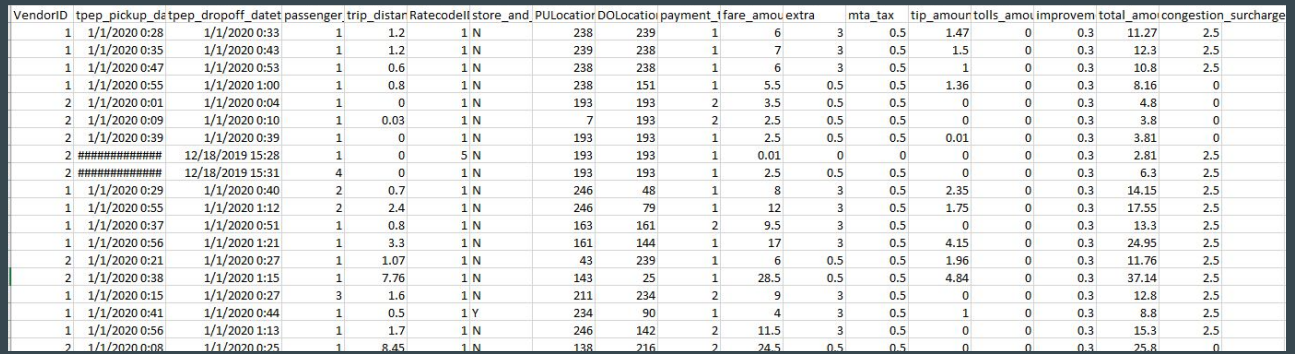
Prior Work

The TLC Trip Data Record is a robust dataset of an important metro area spanning from 2009 until present, and so has been the subject of much previous work. The following are a few high-profile examples:

- A team mined the data to determine the viability of electric cars as taxis in NYC.
  - [C.-M. Tseng and C.-K. Chau, “Viability analysis of electric taxis using new york city dataset,” in Proceedings of the Eighth International Conference on Future Energy Systems. ACM, 2017, pp. 328–333.]
- Liu and Guo used this dataset in conjunction with Markov decision process in order to optimize ride pick-up locations.
  - Caihong Liu, Chonghui Guo, “Mining top-N high-utility operation patterns for taxi drivers”, Expert Systems with Applications, Volume 170, 2021, 114546, ISSN 0957-4174]
- Buchholz analyzed the spatial equilibrium of taxicabs to show how common taxi regulations lead to substantial inefficiencies as a result of search frictions.
  - [Buchholz, Nicholas. “Spatial Equilibrium, Search Frictions and Dynamic Efficiency in the Taxi Industry.”]

Dataset

- NYC Taxi and Limousine Commission Trip Data, potentially supplemented with NOAA weather records.
- Trip records from Yellow Cab, Green Cab and Ride Share Taxis for each month
- Attributes include pickup/dropoff location (by zone), time, trip distance, number of passengers, fare amount, tipping, etc.
- The dataset is split into month-specific CSV files. We have downloaded and opened multiple files on our machines to ensure we can access it.



VendorID	tpep_pickup_da	tpep_dropoff_datet	passenger	trip_distar	Ratecodell	store_and_	PULocatior	DOLocatio	payment_t	fare_amou	extra	mta_tax	tip_amour	tolls_amou	improvem	total_amo	congestion_surcharge
1	1/1/2020 0:28	1/1/2020 0:33	1	1.2	1	N	238	239	1	6	3	0.5	1.47	0	0.3	11.27	2.5
1	1/1/2020 0:35	1/1/2020 0:43	1	1.2	1	N	239	238	1	7	3	0.5	1.5	0	0.3	12.3	2.5
1	1/1/2020 0:47	1/1/2020 0:53	1	0.6	1	N	238	238	1	6	3	0.5	1	0	0.3	10.8	2.5
1	1/1/2020 0:55	1/1/2020 1:00	1	0.8	1	N	238	151	1	5.5	0.5	0.5	1.36	0	0.3	8.16	0
2	1/1/2020 0:01	1/1/2020 0:04	1	0	1	N	193	193	2	3.5	0.5	0.5	0	0	0.3	4.8	0
2	1/1/2020 0:09	1/1/2020 0:10	1	0.03	1	N	7	193	2	2.5	0.5	0.5	0	0	0.3	3.8	0
2	1/1/2020 0:39	1/1/2020 0:39	1	0	1	N	193	193	1	2.5	0.5	0.5	0.01	0	0.3	3.81	0
2	##############	12/18/2019 15:28	1	0	5	N	193	193	1	0.01	0	0	0	0	0.3	2.81	2.5
2	##############	12/18/2019 15:31	4	0	1	N	193	193	1	2.5	0.5	0.5	0	0	0.3	6.3	2.5
1	1/1/2020 0:29	1/1/2020 0:40	2	0.7	1	N	246	48	1	8	3	0.5	2.35	0	0.3	14.15	2.5
1	1/1/2020 0:55	1/1/2020 1:12	2	2.4	1	N	246	79	1	12	3	0.5	1.75	0	0.3	17.55	2.5
1	1/1/2020 0:37	1/1/2020 0:51	1	0.8	1	N	163	161	2	9.5	3	0.5	0	0	0.3	13.3	2.5
1	1/1/2020 0:56	1/1/2020 1:21	1	3.3	1	N	161	144	1	17	3	0.5	4.15	0	0.3	24.95	2.5
2	1/1/2020 0:21	1/1/2020 0:27	1	1.07	1	N	43	239	1	6	0.5	0.5	1.96	0	0.3	11.76	2.5
2	1/1/2020 0:38	1/1/2020 1:15	1	7.76	1	N	143	25	1	28.5	0.5	0.5	4.84	0	0.3	37.14	2.5
1	1/1/2020 0:15	1/1/2020 0:27	3	1.6	1	N	211	234	2	9	3	0.5	0	0	0.3	12.8	2.5
1	1/1/2020 0:41	1/1/2020 0:44	1	0.5	1	Y	234	90	1	4	3	0.5	1	0	0.3	8.8	2.5
1	1/1/2020 0:56	1/1/2020 1:13	1	1.7	1	N	246	142	2	11.5	3	0.5	0	0	0.3	15.3	2.5
2	1/1/2020 0:08	1/1/2020 0:25	1	8.45	1	N	138	216	2	24.5	0.5	0.5	0	0	0.3	25.8	0

Proposed Work

- Data cleaning
  - Dataset appears to be well maintained
  - Check data for invalid entries, etc.
- Data preprocessing
  - Decide which attribute are actually interesting/necessary
- Utilize Python code to process dataset and implement unsupervised learning techniques (multivariate analysis, hierarchical clustering, k-means, local outlier factor)
- Visualize spatiotemporal distribution of taxi rides:
  - Establish baseline behavior for weekdays and weekends
  - Apply methodology to outlier days, and observe differences (New Years, Valentine's Day, major storms, sporting events)

Tools

- Scikit-Learn - ML and Mining Algorithms
- Numpy - Parallel operations on large data
- Pandas - Simple data reading and integration
- Scipy - ML and Mining Algorithms, Optimization routines
- Matplotlib - Visualization
- R (ggmap) - Visualization
- Git - source control

Evaluation

- Based on prior work one of the primary outcomes for taxi data is visualization
  - Visualize between district correlations for price, ride destination, ride time etc.
  - Visualize volume of taxi traffic
  - Visualize monthly comparison (pre shut down, during shut down, after shut down)
- Draw conclusions about the data based on results of exploratory and primary analyses.
- Validate results with previous explorations
- Note any interesting findings for further exploration in the future