

NYC taxi trip duration and taxi demand time and spatial analysis

Instructor: Jonathan Tannen, Author: Yebei Yao

Background

The on developing Intelligent Transportation system offers more and more opportunities help researchers, companies and planners to identify daily travel habits of local populations. There are a growing number of extensive data sets generated with daily trip from taxi companies, with machine learning model and multi dimensional analysis methods, the spatial and temporal variation of taxi trips can be easily explored.

In this project, I'm going to analyze the data obtained from Kaggle with over 1.5 million taxi trip observations in year of 2016 to identify the hotspot area and time where taxi is mostly in demand.

Question Proposed / Goal

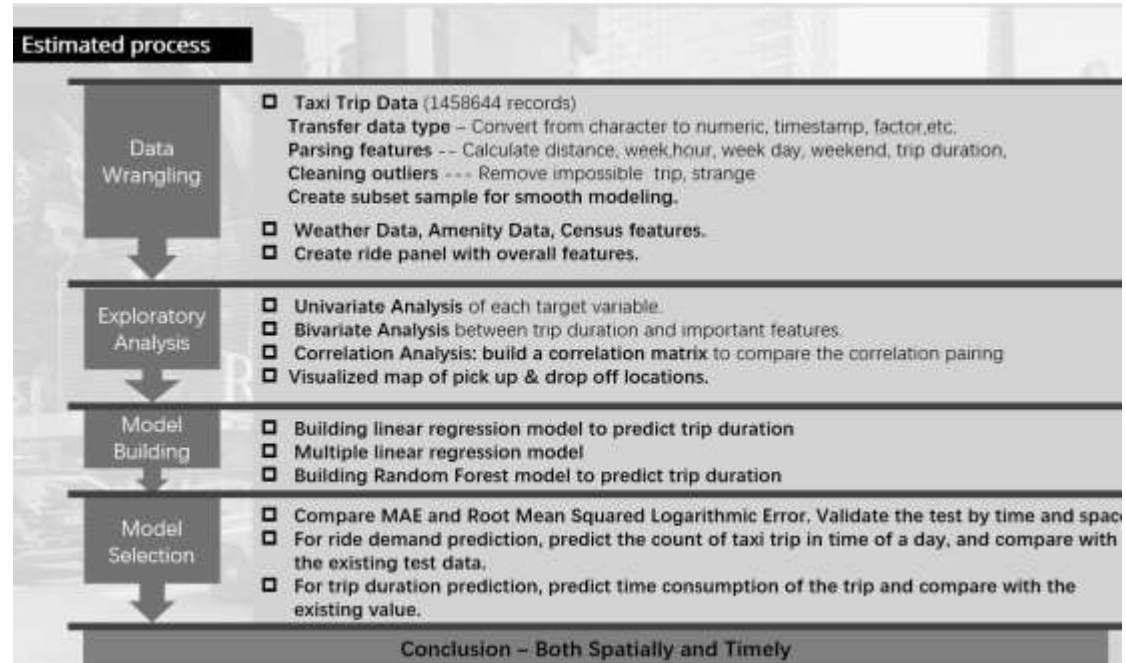
1. With the existing data records with start and end time and location info in 2016 NYC, what's the taxi travel pattern in general? Considering related features such as weather, transition comfort or, demographic characteristics of each census tract of New York City, build a model with the most determinants, to predict the taxi trip in each census tract across time.
2. Considering features like weekday/weekend, time of a day and time spreading between seasons, ect, how the travel pattern and trip time consumption changes with time? Visualize the current temporal pattern and build a model predict the potential time consumption providing the limited condition.

Potential Audience

1. With the fully analysis of the existing taxi trip mobility pattern, we can easily gain the determinants of the taxi trip performance spatiotemporally. This analytical mindset and process can be implemented into other city or districts with identical problem.
2. With the prediction, taxi drivers can go to the most in demand district in advance, so as to decrease both the overall waiting time of passengers and the drivers' driving time without passengers in cars.
3. Besides the traveling pattern and taxi demand prediction, I also would want to predict the time consumption of each trip with the corresponding information like trip start time, destination and weather condition, which acts like the work Google map app currently do. When passengers text down their start and end point, and on a rainy day at noon for example, how much time it will take before the taxi reach them and how long it will take to really gets to the destination? I hope with the model I built with, I can create a dashboard visualize and do the calculation automatically.

Method

1. Feature engineering with provided original dataset, and import other potentially related features.
2. The estimated process is listed below:



Progress

1. Current feature engineering progress:

id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	store_and_fwd_flag	trip_duration	RowID	distance	interrowID	week	hour	WeekDay	week
1	1	2016-05-22 11:22:32	2016-05-22 11:28:32	1	N	420	1240333	2816.7203	2016-05-22 11:00:00	21	11	周二	W
2	1	2016-06-02 12:21:25	2016-06-02 12:28:57	1	N	452	134331	1152.4275	2016-06-02 12:00:00	23	12	周四	W
3	1	2016-05-27 14:14:09	2016-05-27 14:18:45	1	N	279	1179487	1527.0717	2016-05-27 14:00:00	22	14	周五	W
4	1	2016-03-01 06:12:51	2016-03-01 06:13:55	1	N	784	604861	3239.3813	2016-03-01 06:00:00	9	6	周二	W
5	1	2016-05-09 11:15:14	2016-05-09 11:18:10	1	N	176	1276813	559.6708	2016-05-09 11:00:00	18	11	周二	W
6	1	2016-05-07 14:29:09	2016-05-07 14:38:13	1	N	544	1417254	2090.3558	2016-05-07 14:00:00	18	14	周二	W
7	1	2016-02-17 14:38:46	2016-02-17 14:46:41	1	N	1193	1245282	1622.1481	2016-02-17 14:00:00	7	14	周二	W
8	1	2016-06-24 15:56:11	2016-06-24 16:04:04	2	N	473	400891	1777.9902	2016-06-24 15:00:00	26	15	周二	W
9	1	2016-04-28 15:12:59	2016-04-28 15:18:50	1	N	371	423839	1485.4954	2016-04-28 15:00:00	18	15	周四	W
10	1	2016-04-13 02:15:13	2016-04-13 02:17:43	1	N	750	584818	1479.2358	2016-04-13 02:00:00	15	2	周二	W
11	1	2016-04-05 15:24:12	2016-04-05 15:43:22	1	N	970	1234803	3986.4084	2016-04-05 15:00:00	14	15	周二	W
12	1	2016-06-18 23:23:31	2016-06-18 23:28:13	1	N	822	885459	4495.4929	2016-06-18 23:00:00	25	23	周六	W
13	1	2016-04-06 20:48:43	2016-04-06 20:54:04	1	N	421	671170	1693.8876	2016-04-06 20:00:00	23	20	周一	W
14	1	2016-06-11 14:52:33	2016-06-11 14:52:33	1	N	1887	1329158	4215.1533	2016-06-11 14:00:00	24	14	周二	W
15	1	2016-05-21 18:18:00	2016-05-21 18:28:18	1	N	879	347523	1563.6555	2016-05-21 18:00:00	21	18	周六	W
16	1	2016-02-24 21:44:46	2016-02-24 21:46:25	1	N	888	615487	2134.4197	2016-02-24 21:00:00	8	21	周二	W
17	1	2016-03-18 17:54:23	2016-03-18 18:21:30	2	N	1634	403283	7782.2682	2016-03-18 17:00:00	11	17	周二	W

Current panel with overall features

2. Correlation matrix between each pair of features:

