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 demensional 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 observasions in year of 2016 to identify the hotspot area and time where taxi is mostly in demand.

Question Proposed / Goal

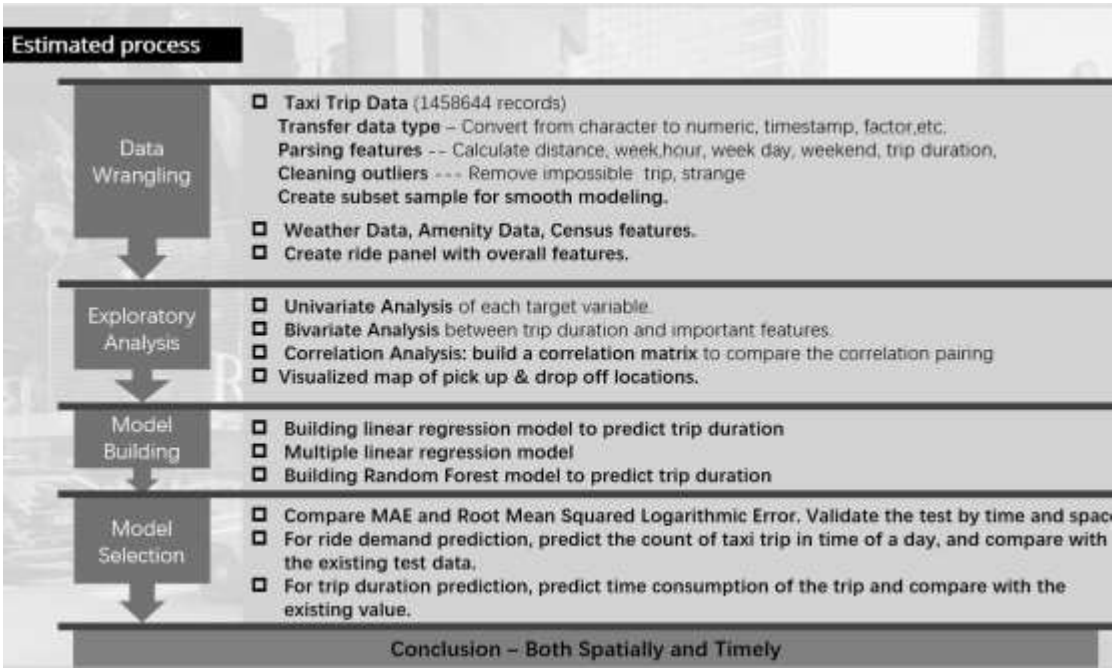
1. With the existing data records with start and end time and location info in 2016 NYC, what's the taxi travel pattern in general? Considering related features such as weather, transition comfort or, demographic characteristics of each census tract of New York City, build a model with the most determinants, 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 consuption 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.

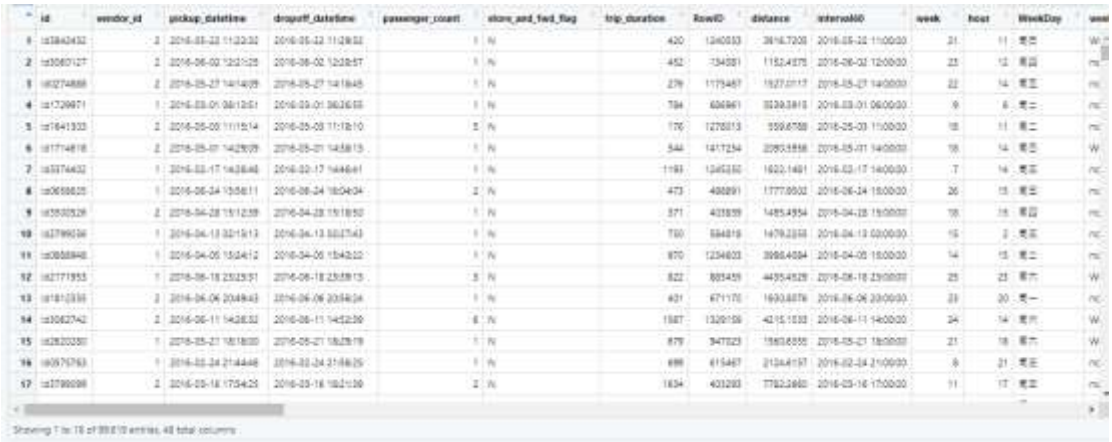
Method

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



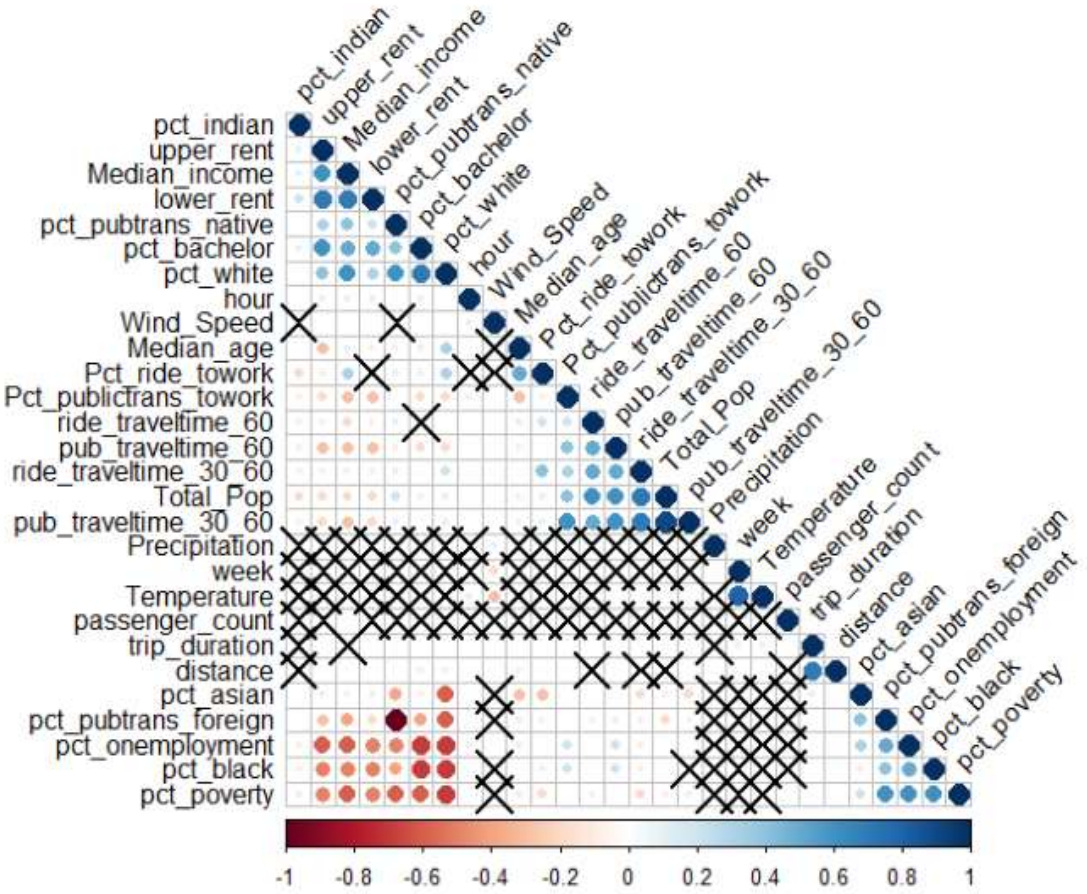
Progress

1. Current feature engineering progress:



Current panel with overall features

2. Correlation matrix between each pair of features:


pct_indian
upper_rent
Median_income
lower_rent
pct_pubtrans_native
pct_bachelor
pct_white
hour
Wind_Speed
Median_age
Pct_ride_towork
Pct_publictrans_towork
ride_traveltime_60
pub_traveltime_60
ride_traveltime_30_60
Total_Pop
pub_traveltime_30_60
Precipitation
week
Temperature
passenger_count
trip_duration
distance
pct_asian
pct_pubtrans_foreign
pct_onemployment
pct_black
pct_poverty
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1