New York City’s Taxi rides Analysis

Saverio Guzzo

Abstract

The intention of the project is to analyze New York City traffic dynamics through how taxis behave along the hours, days and months. The dataset used it is provided by **Kaggle**, from a past competition aimed at predicting trip durations.

The questions which have been posed regard if it is possible to have a minimum fleet of taxis in NYC and still meet the demand of passengers. In order to understand traffic dynamics, statistical analysis of taxis' demand have been carried out.

The demand increases along the week, to decrease on Sunday. An interesting dynamics of trips towards the areas of entertainment zones (e.g. Williamsburg) is observed during weekends and night hours.

Motivation

Inspired from a project of MIT's Senseable City Lab, what motivated me carrying out this project, was the possibility to discover if there is a minimum (optimal) number of taxis in order to support NYC's demand. Results might be useful for New York City's economists, urban planners and regulations bodies.

Dataset(s)

The dataset has been taken from Kaggle in this repository and contains records of 1,458,644 trips, and features:

- Trip id
- Vendor id
- Pickup datetime
- Dropoff datetime
- Passenger count
- Pickup longitude
- Pickup latitude
- Dropoff longitude
- Dropoff latitude
- Trip duration

Data Preparation and Cleaning

The dataset did not fortunately have any null entry among the feature necessary for carrying out the analysis. Although, a conversion from *string* to *datetime* of the features indicating the pickups and drop-offs times had been done. Following, an application of lambda functions has been carried out, in order to extract more specific information about the datetime.

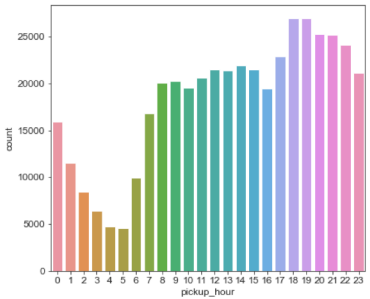
Research Question(s)

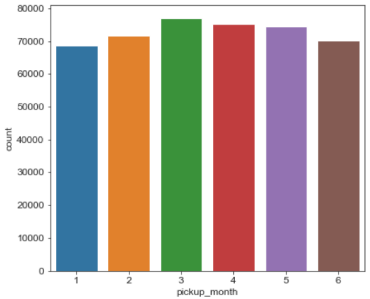
How is New York City traffic shaped? Is it possible to offer an agile supply for a market as volatile as taxi transportation is?

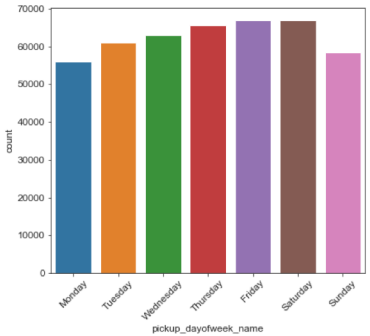
Methods

- Data visualization *(seaborn* library)*,*
- basic statistical analysis (correlation heatmap),
- Support Vector Regression and Random Forest Regression.

Findings

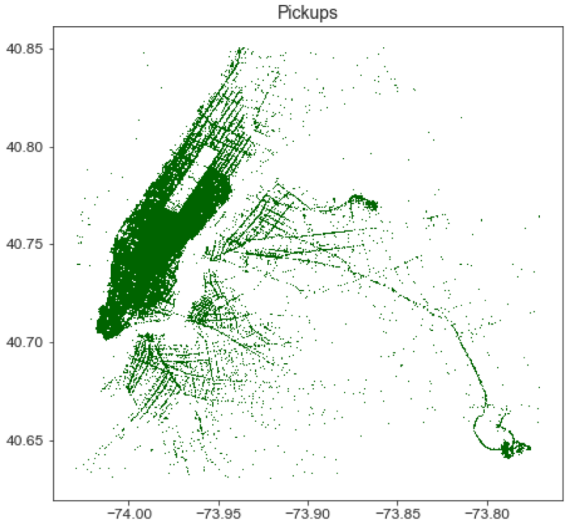


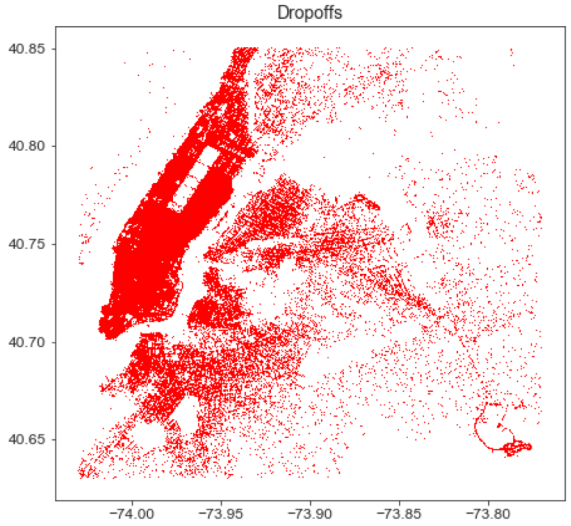




Despite we have just 6 months in our dataset, the demand results pretty stable along these ones. We cannot say the same regarding the demand analised at an hourly and daily level, which appears to have peaks around 6 and 7pm, and on Fridays.

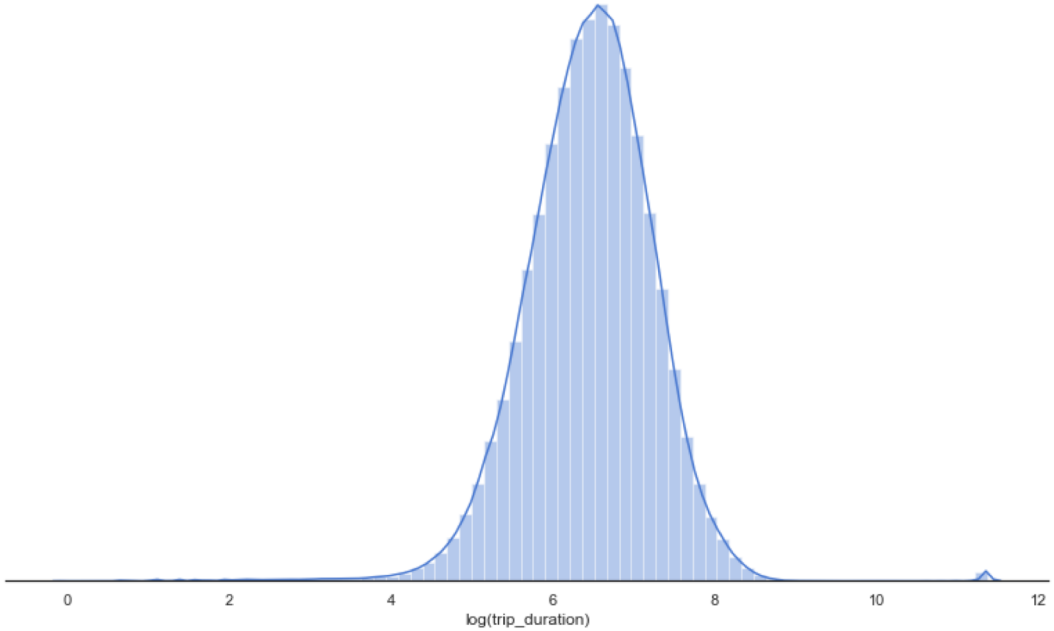
Findings





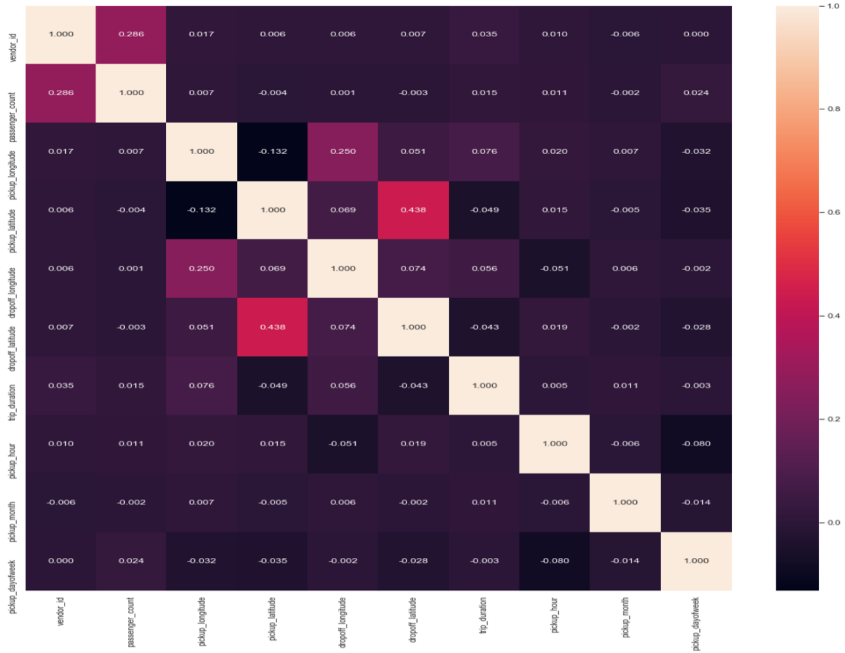
While pickups results mainly concentrated in the borough of Manhattan, dropoffs are spread out all around the area of Brooklyn and Queens, too.

Findings



The image shows the logarithm of trip dirations. Most of trips have a duration between 2.4 (e^5) and 30 (e^7.5) minutes.

Findings



Unfortunately, no strong correlations have been found between features.

Limitations

This analysis may be useful, if done at a much more detailed level, to NYC urban planners and regulators. For common citizens, may be useful for them just to know that it will be more likely to catch a taxi at 5 am rather than at 7pm.

Conclusions

Despite the huge quantity of data we have, making predictions of a trip duration just basing ourselves on pickup, drop-off coordinates and the time in which the trip is done, is not a trivial task. In order to make reliable predictions, more suitable Machine Learning models, as well as, heurists should be used.

Regarding instead the aggregation of travel, clustering analysis may be suitable for further inspections.

Acknowledgements

I didn’t receive any feedback from any friend yet, I reckon I have to perform further analysis before.

References

- https://www.kaggle.com/c/nyc-taxi-trip-duration/data
- https://www.kaggle.com/drgilermo/dynamics-of-new-york-city-animation
- https://github.com/sajal2692/data-science-portfolio/blob/master/911 Calls - Exploratory Analysis.ipynb