Classifying NYC venues based on traffic speed

[steffen.janssen@gmail.com](mailto:steffen.janssen@gmail.com)

# Introduction

## Background

New York City has a significant problem with traffic congestion. It is a significant factor that is hampering growth for businesses throughout the city. Businesses in particular can be affected very strongly when their venues are along roads or avenues that cannot be accessed anymore: their customers may seek to visit other businesses instead that they can get to quickly by car.

New York City is monitoring the flow of traffic and have published a section online where one can use and download this data in real-time. By now it has collected many years of data and it´s a very large dataset that can be used to pinpoint problem areas regarding traffic flow. We can by now create a pretty good baseline of the traffic situation in the city by using this overwhelming amount of datapoints.

At the same time we can utilise the Foursquare API and Foursquare service to identify which businesses are located where. By combining this data we can at least get a good understanding of the problem and we can start to identify which businesses and what kinds of businesses are located in busy or less busy areas.

## Problem and Stakeholders

The problem is the traffic congestion itself. Now it is a tall order to try to resolve that by ´only´ doing Data Science, but we can start to identify correlations and determine what types of businesses are located where and try to better understand if they could have a factor to play in resolving the problem.

Interested parties in this investigation / Data Science research project are the NYC government and Business owners so that they can better understand baseline traffic situation in NYC. The business owners can use this to determine where to expand their business if they feel that traffic situations are limiting their potential.

# Data

There are two main sources of data, both of which are publicly available, that I have used in this analysis project.

1. Foursquare: Foursquare API can be used to gather information about venues near a certain location. It also contains latitudes and longitudes.
2. Traffic data contains speed, latitude, longitude and a number of other variables that can be used to determine the traffic throughput in that road/area.

## Data links

1. Foursquare API: [https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}](https://api.foursquare.com/v2/venues/explore?&client_id=%7B%7D&client_secret=%7B%7D&v=%7B%7D&ll=%7B%7D,%7B%7D&radius=%7B%7D&limit=%7B%7D)
2. Traffic Data: <https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq>
3. Road segments: <https://www.kaggle.com/crailtap/nyc-real-time-traffic-speed-data-feed#linkinfo.csv>

Now for a comprehensive description of these various data sources:

**Foursquare**:

* Location: An object containing none, some, or all of address (street address), crossStreet, city, state, postalCode, country, lat, lng, and distance. All fields are strings, except for lat, lng, and distance. Distance is measured in meters.
* Some venues have their locations intentionally hidden for privacy reasons (such as private residences). If this is the case, the parameter isFuzzed will be set to true, and the lat/lng parameters will have reduced precision.
* Category: An array, possibly empty, of categories that have been applied to this venue. One of the categories will have a primary field indicating that it is the primary category for the venue.

Acknowledgements: <https://developer.foursquare.com/docs/api/venues/search>

**Traffic Data:**

* This data contains 'real-time' traffic information from locations where NYCDOT picks up sensor feeds within the five boroughs of NYC, mostly on major arterials and highways.
* NYCDOT uses this information for emergency response and management, see Acknowledgements.
* NYC Real Time Traffic Speed Data Feed for the year 2016, separated in monthly files of 5 minutes intervals of 'real-time' traffic information within the five boroughs of NYC. Each row represents a given street section (link), for which the average speed, travel time, timestamp and an id of the street section (link) is given. For each link id, information about this link is given in the linkinfo.csv file, e.g., geo coordinates.

Acknowledgements: <http://data.beta.nyc/dataset/nyc-real-time-traffic-speed-data-feed-archived> <https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/xsat-x5sa>

## Examples of what the data consists of

The traffic file (1) contains the following columns (example follows behind the column name)

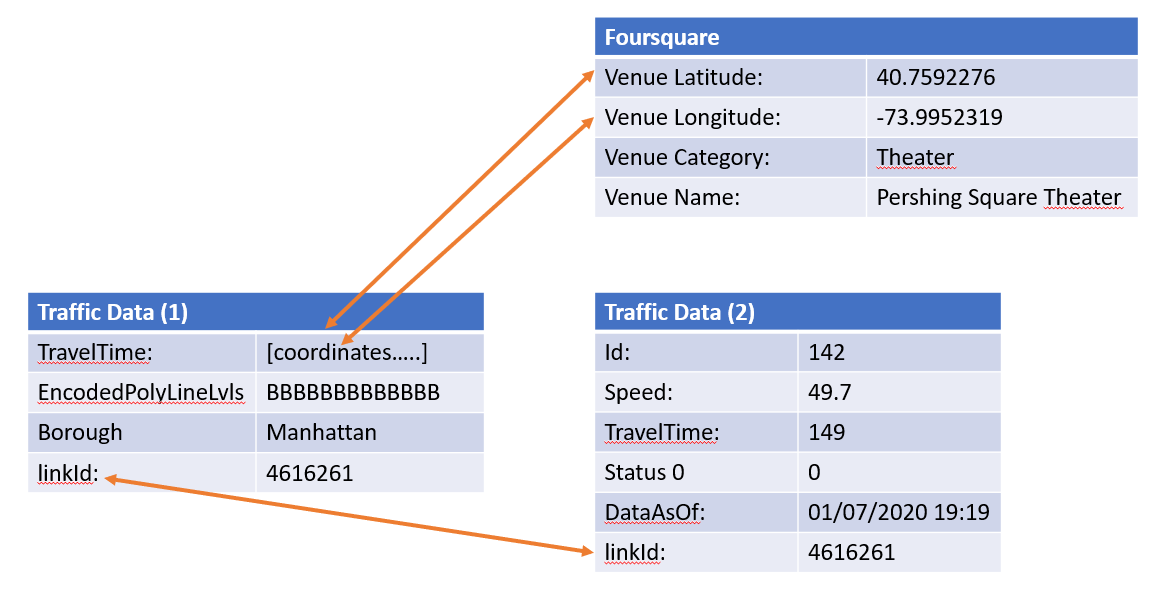
* Id: 142
* Speed: 49.7
* TravelTime: 149
* Status 0
* DataAsOf: 01/07/2020 19:19
* linkId: 4616261

Traffic Data file (2) contains these columns (with examples):

* linkId: 4616337
* linkPoints: 40.74047,-74.009251 40.74137,-74.00893 40.7431706,-74.008591 40.7462304,-74.00797 40.74812,-74.007651 40.748701,-74.007691 40.74971,-74.00819 40.75048,-74.008321 40.751611,-74.00789 40.7537504,-74.00704 40.75721,-74.00463 40.76003,-74.002631 40.7607405,-7
* EncodedPolyLineLvls: BBBBBBBBBBBBB
* Borough: Manhattan

The Foursquare API contains these columns (with examples)

* Venue Latitude: 40.7592276
* Venue Longitude: -73.9952319
* Venue Category: Theater
* Venue Name: Pershing Square Signature Theater



## Data acquisition and cleaning:

**Foursquare**: I decided that it would not be feasible to query the data in real time always due to the limits in the free Foursquare API and the Real Time trafficData API was too slow.

Therefor I have prepared functions in Python to facilitate the retrieval of data from Foursquare, which is then saved in a local csv file which I have since reused during development of the model.

**Traffic Data:** When it comes to the Traffic Data there is a direct link where the full dataset can be downloaded from (<https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/xsat-x5sa>). This dataset is 15GB in total in csv format.

For performance reasons it was impossible for me to build the analysis on the full 15GB of traffic data which is why I cleansed the data to limit it to only the information from 2020. This data cleansing was done in Excel. Otherwise the raw data was ingested into the python model and further data cleansing was applied there.

These are the steps of Data Preprocessing that were required for the traffic data:

1. ["TRANSCOM\_ID", "LINK\_NAME"] columns were dropped because they do not contain data we can use. Link\_Name is identical to LINK\_ID, TRANSCOM\_ID is a coded version of LINK\_ID.
2. “ENCODED\_POLY\_LINE\_LVLS “is an encoded value for the length of the road segment, this is converted to an integer for easy processing.
3. “LINK\_POINTS” is split up into individual featues for each coordinate of the road segment. The original is a chain of coordinates (longitude and latitude), this is split up into individual tuples for each coordinate for easy access.
   * The last two tuples are dropped due to the fact that they do not reliably contain valid data (valid coordinates). Often times the coordinates are cut off halfway when it comes to the last two segments. I expect that there is a length problem with the original data collection where it is cut off at a certain length.
4. “DATA\_AS\_OF” value is harmonized across the different measurements. Different link segments submit date information in different formats. This needs converting to a universal format that is consistent.

## Feature Selection

Although the problem that is being investigated is complex, the features that are in play are rather limited. The primary feature for the analysis is based around the coordinates (longitude and latitude) of the Venues and the road segments.

A secondary feature from the Foursquare data is the Venue Category. We are not interested in the Borough, Neighbourhood, Venue Name. The neighbourhood and Borough depend entirely on the Coordinates, whereas the name does not add any value to our analysis.

The secondary features from the Traffic Data set are the speed, status and the encoded length of the road segment. We will be using these features in building our model and predicting the properties of coordinates or venues.

# Methodology

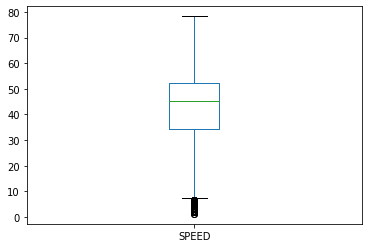
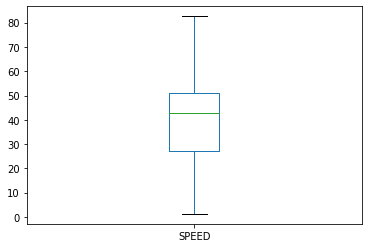
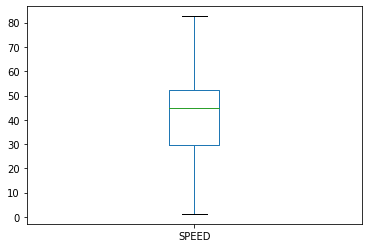
The high level approach for the actual modelling is as follows:

1. Extract data
2. Cleanse and pre-process Data
3. Combine Traffic Data (1) and Traffic Data (2) by matching on the LINK\_ID
4. Create a baseline by writing a function to score “traffic speed performance”
5. Use the baseline function against an average of the dataset for each of the time periods and road segments to score that date/time/segment using road speed
6. Build a classifier against the date/time/segments (X features) in order to predict the performance (y value)
7. Classify Venue locations using classifier
8. Analyse results

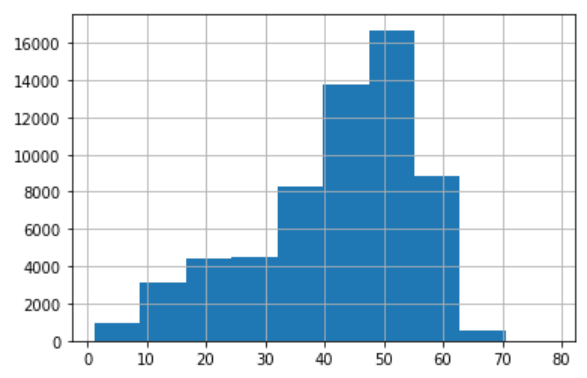
### Exploratory Data Analysis

Speed on average is very varied, but morning/evening/night differences aren’t that pronounced

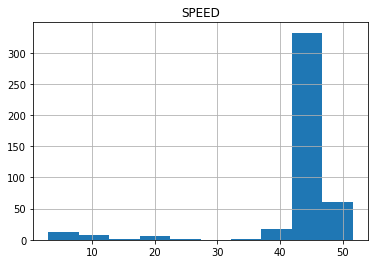
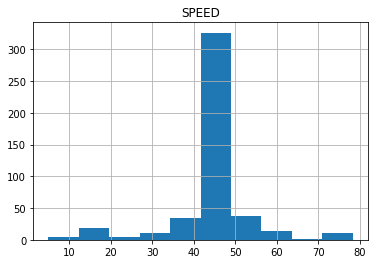
**Average** **Morning** **Night**

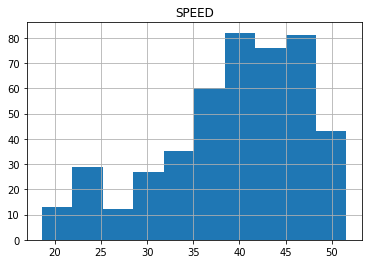
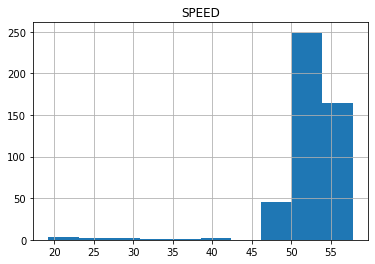


The histogram shows this distribution over all measurements for all road segments:



But as expected there is a large difference between the individual segments. This is a snapshot of some of these histograms for different road segments (not all of them):





|  |  |
| --- | --- |
|  | When we plot the road segments on the map of New York City we can visualise exactly what data has been captured. This will automatically lead us to notice the massive data sparsity when you consider the entirety of NYC. |
| Besides the traffic data we also have a Foursquare DataSet. When we plot these points on the map they track the road segment outlines, due to the fact that I only captured the data in relative proximity to the roads for which I have data: | C:\Users\STEFFENJanssen\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\597B25FE.tmp |

## Analysis

The principle behind the creation of the baseline (see point 4 in the methodology) is to define a set of rules that can derive a classification on the training set to indicate the overall average traffic flow in that point.

We need to take a couple of things into account:

* The speed in a segment by itself is the average traffic flow and is a pure indicator of the absolute performance of the traffic flow in that point.
* The perceived traffic flow however can differ: 60 mph on a highway is considered to be normal traffic speed, at the same time 30 mph in a downtown alley is also considered to be good although the absolute speeds differ.
  + Conversely 30 mph on a highway is not particularly good whereas 60mph in a downtown alley is an outlier.

For this reason I have ensured to define a baseline against which we can compare the current situation.

The baseline is therefor the 50th quantile over all speed measurements in the chosen time period for each of the road link segments. To account for the perceived traffic flow I also defined other business rules to derive the performance feature. These are evaluated in order:

1. Actual speed > 0.75 \* Max speed : traffic flow is **good**
2. Actual speed > 75th quantile : traffic flow is **good**
3. Actual speed > 35 mph : traffic flow is **ok**
4. Actual speed > 50th quantile: traffic flow is **ok**
5. Otherwise: traffic flow is **bad**

So to perform the analysis on the current traffic flow we can evaluate the current speed against these baselines and business rules.

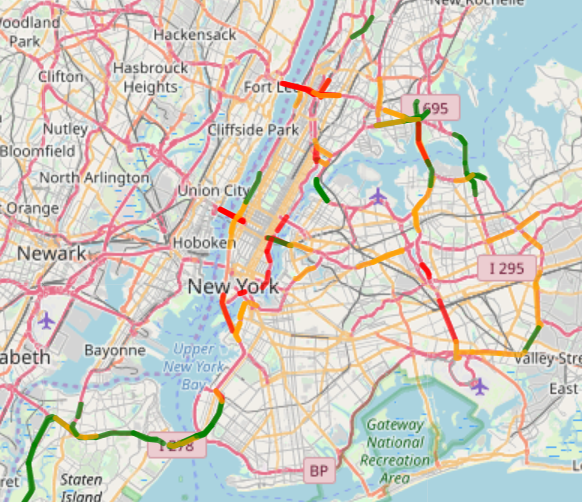
However when it comes to evaluating venues, which do not in itself have a velocity, there is a need for a different evaluation metric.

1. Average speed is above 50 mph: traffic flow is **high**
2. Average speed is above 25 mph: traffic flow is **medium**
3. Otherwise: traffic flow is **low**

# Results

Unsurprisingly the worst traffic is found more central to the New York City Downtown. The below plot superimposed on the NYC map shows that the road highlighted in red are particularly poor in terms of traffic flow as a baseline, with measurements averaged over the full dataset.

What does jump out is the fact that we actually have a full picture because there are just 135 road segments that we are dealing with in this dataset whereas there is so much more that we would be interested in analysing.

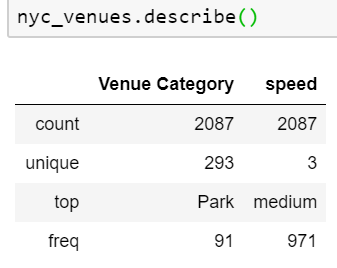


## Predictive modelling

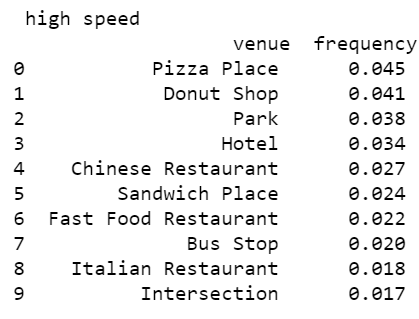
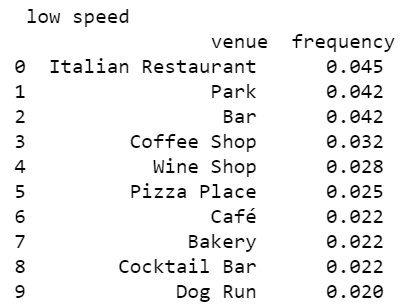
Using 80% of the training data to build a KNN classifier and testing it on the 20% of the test data provides a reasonably good classifier.

|  |  |
| --- | --- |
| C:\Users\STEFFENJanssen\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E9C37644.tmp | The idea is to have a heat map that indicates the traffic flow in an area and to use that heat map in predicting the traffic flow in other areas using separate coordinates.  With 3 or 5 Nearest Neighbours the classifier successfully classifies over 90% of the training points. |

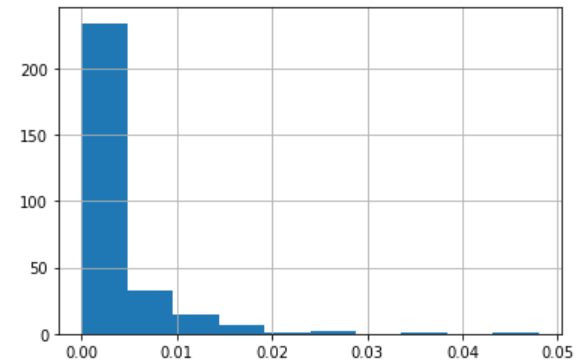
The actual business question for the project was to determine how we can score venues in terms of accessibility. Thanks to the predictor we built there is a way to classify venues. The result is shown below:



The top 10 venue categories for each class are:

The top 10 usually contain a variety of different restaurant types as well as parks. With a distribution of frequencies for the unique venue categories very much skewed to the low end on the frequency spectrum:



# Discussion

There are two parts to the results: One the one hand we are presented with information about what roads and what areas have poor traffic flow compared with the average and in absolute terms, on the other hand we used that data to classify venues.

To start with the results for the traffic data are interesting in terms of visualizations and also allow us to do even further analysis based on the different timeframes for which the data is captured. However due to the lack of data of smaller streets we are unsure if we can really be certain to assume that traffic in that general area is the same as the traffic flow on the main roads. Still in general it does seem to suggest that the more central and the more dense the area is the worse the traffic flow is. However there are simply too many factors that we have not taken into account such as the maximum speed on the road, the maximum potential throughput of the road and alternative connections such as subway/rail.

Still we have shown that traffic is generally orse in the morning and in the evening as compared with the rest of the day. This is, once again, most pronounced in Manhattan.

The results for the classification of venues shows:

* There is hardly any overlap between venues in slow speed and high speed areas, other than Parks and Italian restaurants.
* Fast food type restaurants are located near roads that have fast flowing traffic.
* Tradtitional restaurants are located in areas that have slower traffic.
* Parks are the most common venues in New York City.

# Conclusions

This projects study presents an outcome that does seem to suggest that there is a difference in venue types that are typically situated close to faster or slower moving traffic. Empirical observations also teach us that fast food joints are more commonly placed near larger roads, and that chique or boutique restaurants will be more likely to do business from slightly quieter downtown neighbourhoods.

Overall I think that it is safe to say that we need more and more precise data to be able to do more or better correlations between the two data sets.

* The first step would be to make sure that the NYC Traffic Data is complete and actually contains all the coordinates of the road segment, and not just the 12 first sets of longitudes and latitudes.
* Secondly the NYC Traffic Data should be expanded to more roads and should also include smaller roads besides the main arteries that are currently included.
* Thirdly the categorisation of the Venue types needs improving to be able to collate different venue types into a parent group