Bike shops in the Big Apple

Applied Data Science capstone project

August 2020

By Christoph Ensinger

Table of contents

Introduction

Cycling, both as a means of transport or a leisure activity, plays a big role in New York City. It is estimated that 24% of adult New Yorkers (about 1.6 million) ride a bike at least once a year. Almost 800 000 people are considered to be regular bike riders. Cycling offers a solution to a city where the ever-increasing demand for mobility can no longer be handled by other modes of transport, particularly if the ecological costs of transport are considered. Looking at current political initiatives to encourage bike riding, it is reasonable to assume that the number of bike rides in New York City is only going to increase. And while some of those additional rides will handled by bike sharing services, this trend also indicates possible business opportunities for bike shops.

Business problem

As stated in the previous section, the future upside potential for bike shops in the New York City area is evident. However, the question on *where* to realize that potential in the form of a bike shop in NYC is decidedly non-trivial. There are vast economic and demographic disparities between different neighborhoods, resulting in very different environments for retail business. In an effort to maximize the economic potential of a bike shop, a stakeholder might be interested in a model that predicts the number of bike shops for an area, based on reported demographic an economic data. Afterwards this model can be used to discover those areas where the predicted demand for bike shops exceeds the current number of bike shops. A potential bike shop owner could use this information to determine promising locations for a new bike shops. Investors on the other hand could use this model to examine investment opportunities. Because of the large disparities even in the same neighborhood, the model should give a prediction not just for a neighborhood, but rather a zip code area.

Data

The following table gives an overview of the data used in this project:

Data	**Source and comment**
Zip codes	GeoJSON with all zip code areas for the state of New York
Neighborhoods, coordinates and land areas	uszipcode database (Python library)
Data about existing bike shops	Foursquare API, *search* endpoint
Demographic data	American Community Census API, *acs5* endpoint
Economic data	American Community Census API, *zbp* endpoint

The final notebook uses a GeoJSON file that contains all zip codes from the state of New York along with area information that is used for visualization purposes. In the next step, the uszipcode python library is used to filter out the relevant zip codes for all five counties of New York City (New York County, Bronx County, Queens County, Kings County and Richmond County). This library also provides the land area and latitude and longitude information for each zip code.

The Foursquare API is used to obtain all bike shops in New York City. After making an API call for each of the zip code areas using the zip code area coordinates. Most of the results returned by Foursquare contain information about the zip code a bike shop is in, and if required, we use the uszipcode library to get missing zip code information.

We obtain demographic and economic information from the publicly available data from the American Community Census. For each zip code, we get the population, the average income and the (estimated) number of inhabitants that use a bike to get to work. Amongst the available indicators, these three features are best suited to describe the potential market size for a bike shop. Ideally, we could contrast this market potential with information about the cost of running a bike shop in a certain zip code area. Unfortunately, detailed information about commercial rent per square feet or other running costs is not available on a zip code area level. We therefore make use of the American Community Census again which provides data about the total salaries of all employees as well as the total number of employees in a given zip code area. We use this data to construct an average salary for each zip code area. This average salary is used as a proxy for the cost of running a business in a zip code area since it is reasonable to assume that higher average salaries are more likely in areas where overall business operation is more expensive.