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 [1]. 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.

Methodology

For the scope of this exercise, we use a regression model to predict the number of bike shops in each zip code. We refer to this number as $\hat{y}$ and compare this number to the actual number of bike shops y and define $\hat{y} - y$ as the potential of a zip code area. The best location for a new bike shop consequently becomes the zip code area that maximizes $\hat{y} - y$.

We utilize six features to build the regression model:

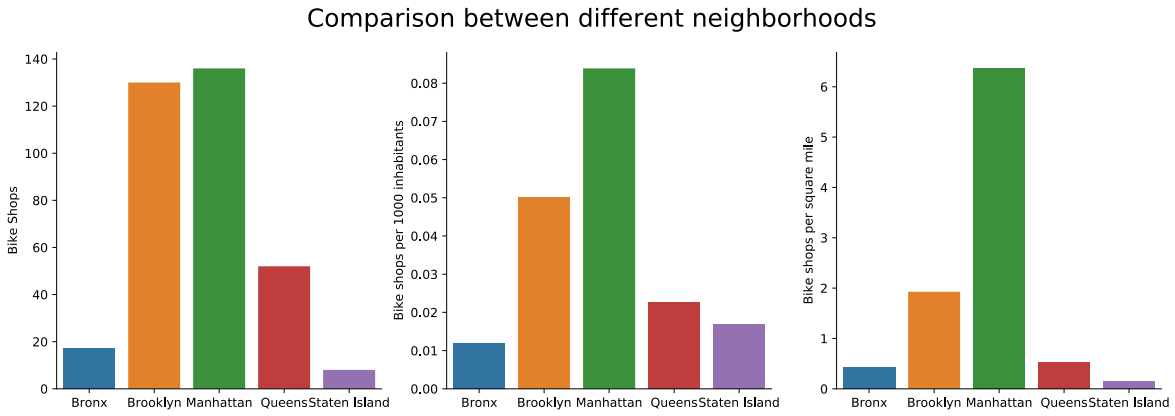
Feature	Unit	Description
Area	square mile	Land area of the zip code area
Population		Number of inhabitants in a zip code area
Income	$	Average household income across all households in a zip code area
Bike to work		Number of adults who use a bike to get to work
Payroll	$	Total payroll of all companies in a zip code area
Employees		Number of employees

We also introduce the population density as a synthetic feature to further differentiate the different zip code areas. Correlation analysis between the features indicate that the number of employees and the total payroll are highly correlated (we show the correlation matrix of all features in the code notebook). We therefore resort to the ratio of total payroll and area to estimate running costs of a business in a zip code area.

After collecting and cleaning of the data we use a five-fold cross validation to compare the average R squared value as well as the mean squared error on the test sets of different models. The next section presents the results of both different single feature linear regression models as well as a multiple linear regression model that we obtained via recursive feature elimination. We then fit the best model on the entire dataset and compute the predicted number of bike shops $\hat{y}$ and subsequently derive a recommendation on where to open a new bike shop.

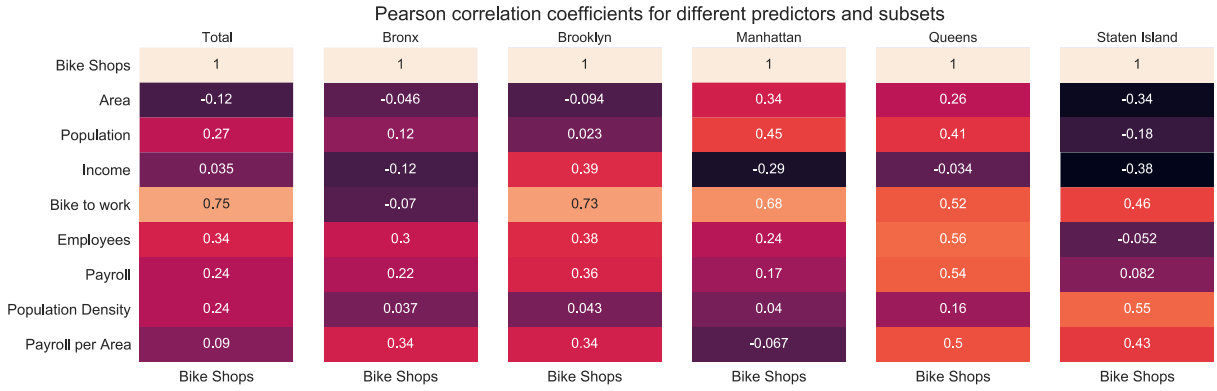
Results

The final dataset contains 179 zip code areas within New York City and 343 bike shops. The following figure gives an overview about the distribution of the bike shops with regards to the different boroughs.

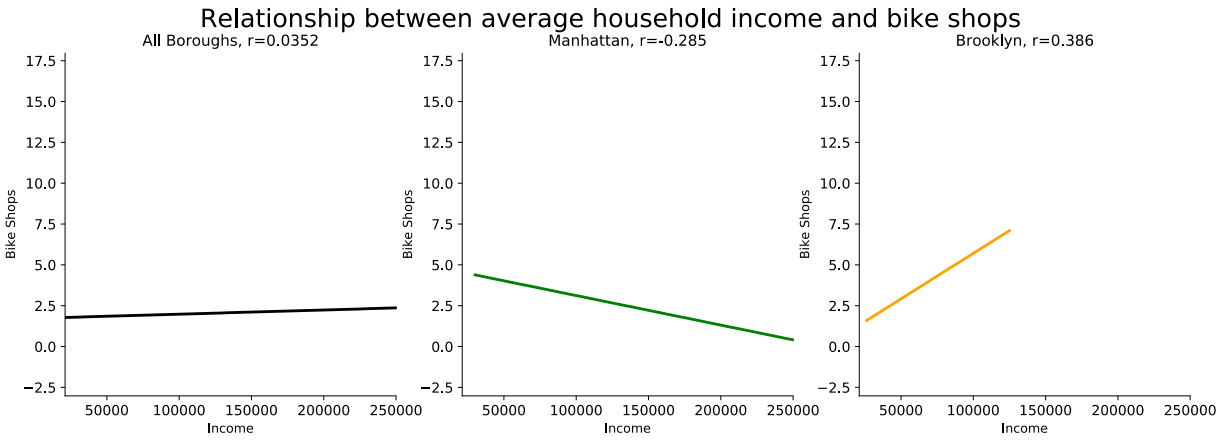


Brooklyn and Manhattan account for the majority of bike shops. Manhattan also features the highest number of bike shops per inhabitant as well as per square mile. The latter figure just mirrors the fact that Manhattan is by far the densest borough in NYC in terms of companies per land area. In any case this distribution already indicates how the potential of a bike shop is influenced by the economic and demographic

A first assessment of the potential explanatory power of our selected features is depicted in the figure below. It shows the Pearson correlation coefficient between each feature and the number of bike shops for each borough as well as for the entire New York City.



There are three main takeaways from this figure. First, we note that the number of bike commuters (“Bike to work” feature) has by far the highest correlation with the number of bike shops when looking at the entire data set (leftmost column). Secondly, we find that some features have varying degrees of correlation with the number of bike shops in different boroughs. The third observation is the fact that some features exhibit a positive correlation in one borough and a negative correlation in another borough. This behavior is demonstrated in the next figure. We clearly see that the income feature has a negative correlation with the number of bike shops in Manhattan and a positive correlation in Brooklyn. It is also evident that the income feature on its own has almost no correlation when used to estimate the number of bike shops across the entire dataset.



The results of the regression models that use only one feature reinforces the observation from the correlation analysis: The number of bike commuters is by far the best predictor of the number of bike shops. All other single-feature models are unable to properly predict the number of bike shops:

	Feature	MSE on test data	R squared on test data
3	Bike to work	-3.393838	0.496652
4	Employees	-6.611817	0.062539
5	Payroll	-7.058879	-0.003102
6	Population Density	-7.145004	-0.037086
1	Population	-7.168853	-0.048781
0	Area	-7.408103	-0.076339
2	Income	-7.555376	-0.092047
7	Payroll per Area	-7.558334	-0.088121

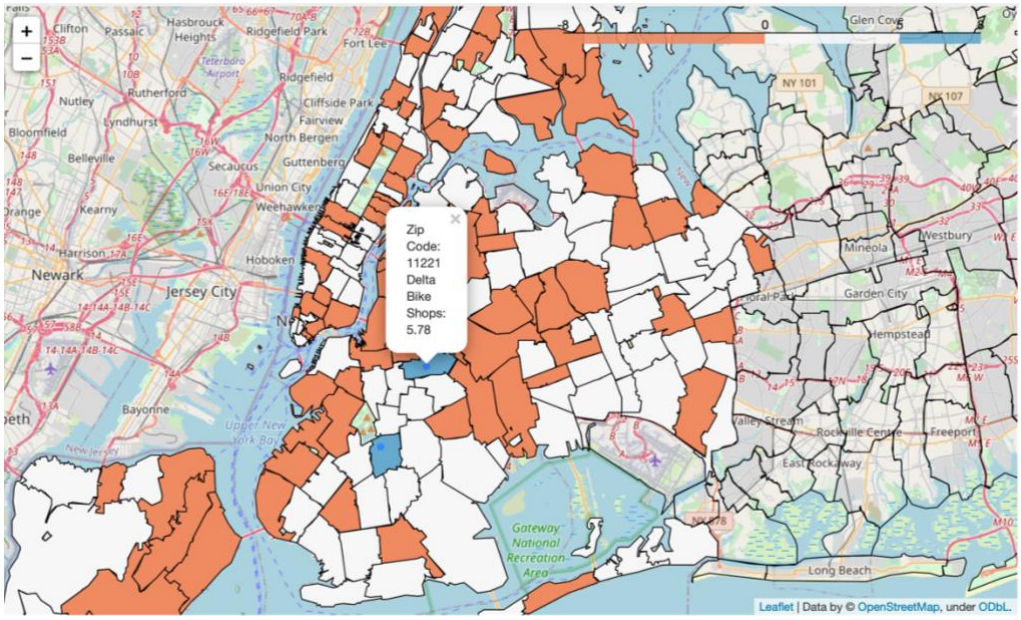
Using a multiple regression model in combination with a recursive feature elimination search we found a model that improves on the single-feature linear model:

Variables	Features
x3	Employees
x0 * x2	Population * Bike to work
x0 * x3	Population * Employees
x0 * x4	Population * Payroll
x0 * x6	Population * Payroll per Area
x3 * x9	Employees * Manhattan

The average R squared value on the test set of this model is 0.531, slightly higher than the 0.497 of the best single feature model. We therefore selected this multiple regression model as our final, retrained it on the entire dataset and obtained the prediction for each zip code. Here are the five zip code areas with the highest difference between predicted and actual bike shops:

	Zip Code	Borough	Bike Shops	Predicted number of bike shops	Delta Bike Shops
103	11221	Brooklyn	4	9.784708	5.784708
33	11226	Brooklyn	1	6.281121	5.281121
131	10467	Bronx	0	3.135995	3.135995
197	10010	Manhattan	0	3.069263	3.069263
167	10004	Manhattan	0	2.988777	2.988777

We see that the zip code area 11221 in Brooklyn has the highest potential for additional bike shops. The second-best choice is also located in Brooklyn. We visualize the other predictions of this model using the following choropleth map:



Areas where our model predicts fewer bike shops than exist are marked red, whereas those areas that have fewer bike shops than predicted are marked white. The two best locations mentioned above are marked in blue.

Discussion

After presenting our findings and giving a recommendation for the optimal location for opening a bike shop we close this report with a brief discussion of our methods, results and other observations.

We limited our investigation to New York City. This decision was primarily driven by the available data and other practical reasons. The fact that cycling plays such a huge role in NYC suggested that we could use rather basic (and therefore readily available) economic predictors to estimate the number of bike shops in a zip code area. We expect our model to perform worse in cities where cycling is less prevalent (i.e. Los Angeles) and the number of bike shops in a given area is therefore more dependent on other factors or even random effects. The result however remains relevant for those stakeholders that are limited to the NYC market.

The second potential issue lies in the modeling itself. We defined the potential of a zip code solely by the current number of bike shops in said area. It appears reasonable to expect

some correlation between the number of bike shops in a given area and the underlying market potential for a bike shop. This assumption foregoes any other important metric such as the revenue of a bike shop or how often new bike shops fail. Including these metrics would have required more advanced economic modeling as well as obtaining very specific data. Both of those tasks are unfortunately beyond the scope of this investigation. Even our simplified model could have benefited from better (economic) data such as the average retail rent. It is also worth noting that the number of bike commuters, a very important predictor in our models, is only an estimate provided by the ACS. This uncertainty in our model is also reflected in the distribution of the residuals, which we present in the online notebooks.

Conclusion

In this exercise we set out to find the ideal location for a new bike shop in New York City. Given the constraints in terms of domain knowledge and available data we derived a model that predicts the number of bike shops in a zip code area given a set of demographic and economic predictors. Based on this model we recommend a potential new bike shop owner to choose a shop in the zip code area 11221 or 11226.

Appendix

Sources:

[1]: http://www.nyc.gov/html/dot/downloads/pdf/cycling-in-the-city.pdf