Bike shops in the Big Apple

Applied Data Science Capstone project

August 2020

Introduction: The task at hand

Problem

- Problem Definition: What location in New York City presents the best opportunity for opening up a bike shop?
- This question is of interest for new bike shop owners or investors.

Business Understanding

- The success of a bike shop depends on their location, which defines their potential market as well as their costs of operation.

How data will help

- We can leverage demographic and economic data to uncover patterns that favor the number of bike shops in a given area
- By comparing the actual number of bike shops in an area with a data-derived prediction we can recommend areas with high potential

Introduction: Which data is required?

What data is needed?

- We require demographic and economic data for each zip code area along with the number of existing bike shops

Data sources

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

Is the data representative?

- All data is derived from credible sources and collected systematically

Introduction: Manipulation and modeling

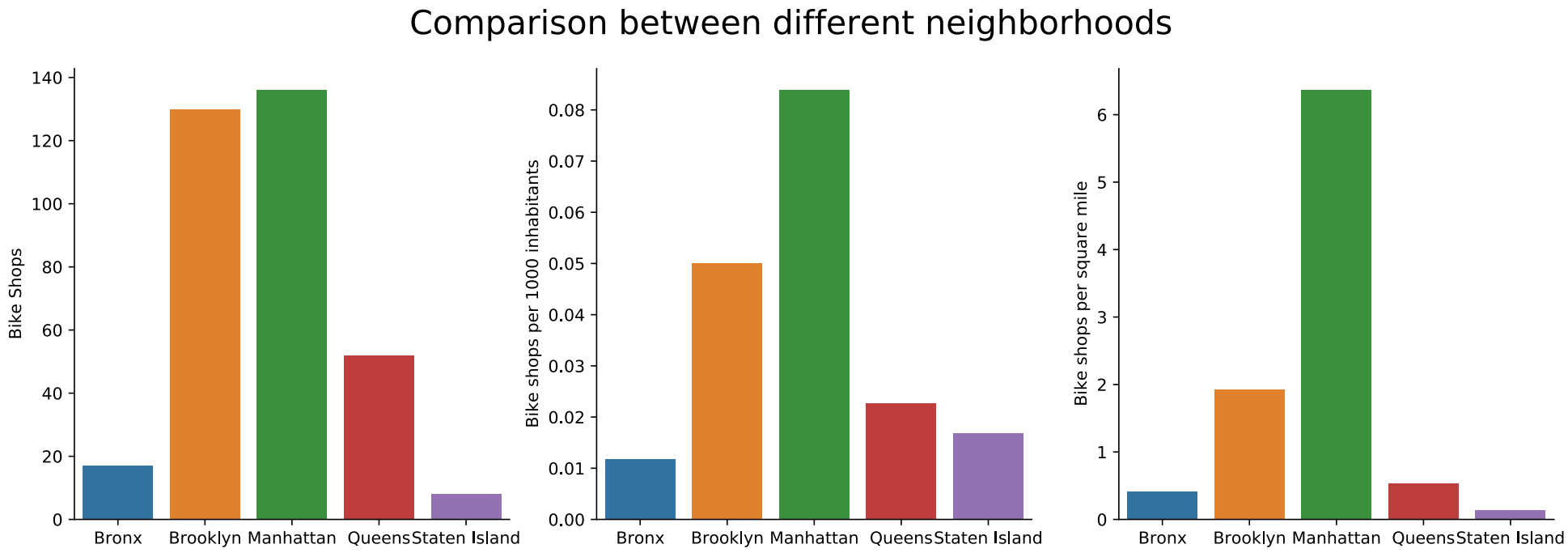
Data preparation

- The data requires manual cleaning and formatting

Deriving the answer

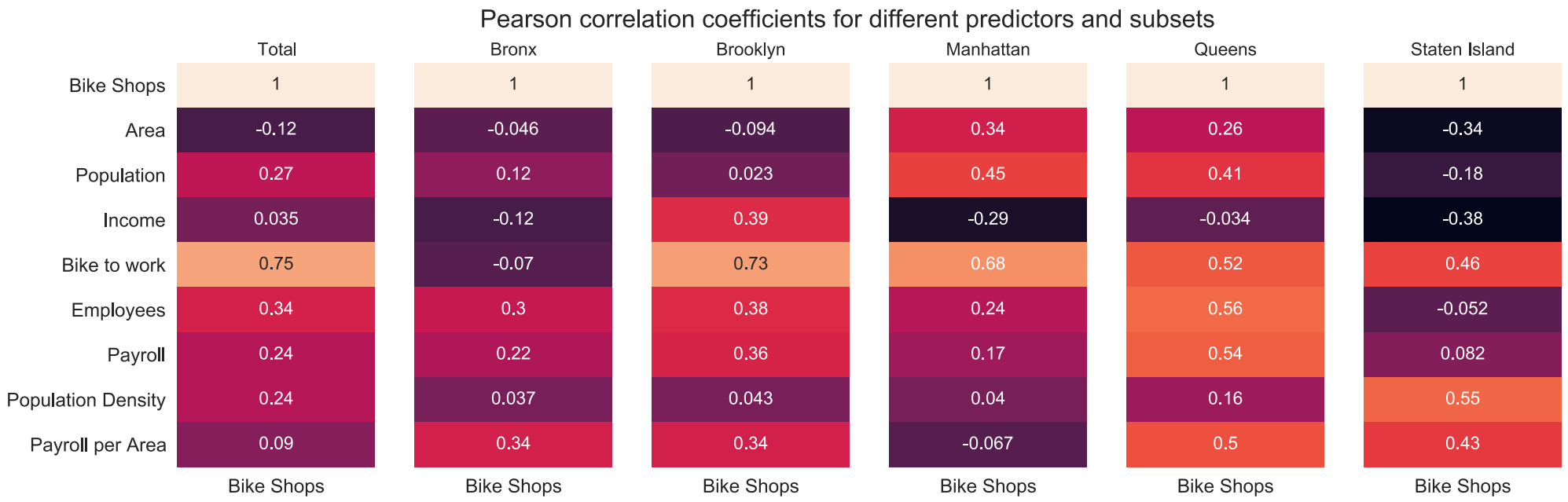
- Exploratory analysis will indicate basic trends and suggest which features are indicative of the number of bike shops
- Different simple and multiple linear regression models are compared
- We use the best model to estimate the number of bike shops in each zip code
- The zip code area with the largest difference between predicted and actual bike shops will be selected as the answer to the original question

Results: Initial analysis



- Brooklyn and Manhattan account for most of the bike shops

Results: Correlation between features and number of bike shops



- The correlation between the different predictors and the number of bike shops varies between different boroughs.
- The number of bike commuters (“Bike to work”-feature) appears to have the highest correlation with the number of bike shops

Results: Linear regression models

	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

- As far as single feature models go, using the number of bike commuters gives the best model
- All other features on its own are unable to predict the number of bike shops

Results: Multiple linear regression models

Using recursive feature elimination we found an improved model that uses the following feature combination:

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

- Using the same 5-fold cross validation this model archives an average R squared value of 0.531, which is slightly better than the best single-feature model (0.497)

Results: The best location

Using the multivariate model we find our ideal location: Zip code area 11221