An Exploratory Investigation of Bookstores in San Francisco

Report written by We Toh

for

Coursera-IBM Data Science Specialization Capstone Project

January 12, 2019

Table of Contents

Introduction

I often find it fascinating when visiting a bookstore, I'm usually able to locate a coffee shop nearby. It makes me wonder if there could be a connection between bookstores and coffee shops. Is this a universally applicable, or, scalable observation? Putting it in the perspective of business intelligence, could this be a useful observation to, say, assist a prospective bookstore owner to decide whether opening a bookstore near a coffee shop is a good thing to do? Just like Starbucks partnering with Barnes and Noble. What about if there's another bookstore nearby already? Further, could similar observations be applied to, not just for bookstores and coffee shops, but for other kinds of business also? These are questions I think worth exploring.

As a first step to this journey of discovery through the use of data, this project shall focus just on bookstores.

A Jupyter python notebook accompanies this report. I have deployed the Foursquare API to build what I call a "venue profile" for the bookstores in the city of San Francisco. I have chosen San Francisco because it is where I reside in. With the venue information gathered and with the help of the Folium python library, the data will be visually presented on a map that shows where the bookstores and coffee shops are located. We will go further to quantitatively compare San Francisco's profile for the bookstores against New York City's[1]. Finally, we will run a few statistical tests to compare the significance of differences in the venue profiles.

Methodology to the Analysis

The Business Problem

As mentioned in the Introduction, we will be focusing on the bookstores in San Francisco. The goals of this project are to find out the following:

1. how common are the bookstores near one another.
2. how common are the bookstores and coffee shops close to one another.
3. how do the bookstores in San Francisco compare with New York City.

[1] New York City has sufficient number of Foursquare venues to justify the data comparison.

In terms of quantitative goals, we want to:

1. measure the proportion of bookstores near one another.
2. measure the proportion of bookstores near coffee shops.
3. compare the proportions applicable to San Francisco and New York City.

The investigative nature of the business problem means that the analysis carried out in this project will follow a descriptive approach.

Data Requirement

The following data will be needed:

1. Venue location data
2. Proximity data

The venue location data applies to the bookstores and the coffee shops. They will be collected by making Foursquare API calls[2]. The proximity data is an engineered feature using the venue location data. Each venue will be checked if another bookstore is nearby. Each venue will then be checked if a coffee shop is nearby.

For consistency in the data collection, the following parameters are also specified:

City center latitude and longitude	Use Foursquare's defined values for the city itself
Venue coverage radius	4000 meters
Proximity radius	250 meters (equivalent to about 1 to 2 city blocks)

Data Workflow

The workflow of involving the data can thus be summarized as such:

1. **Gathering/storing** – Gather venue data from Foursquare and store in a dataframe.
2. **Preparation** – Carry out logical tests ('True'/'False') to check if there are other bookstores or coffee shops nearby for each bookstore, and append the results to the dataframe.
3. **Analysis** –
    a. *Statistical computation*: Calculate the proportion of bookstores near one another/coffee shops.
    b. *Data visualization*: Place the bookstores and coffee shops on a map to provide a visual sense of their closeness to one another.

[2] https://foursquare.com/

We will obtain one dataset for San Francisco, and another dataset for New York City, and we will carry out data visualization on San Francisco venues as part of the project focus. Data visualization for the New York City dataset is not included in this project.

Confidence Intervals for the Population Proportions

The formula for the confidence interval of the population proportions is given as

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion and n is the sample size. Z is the Z-value at the chosen level of confidence. At 95% confidence level, $Z = 1.96$.

The formula is valid when we hold the assumption that the sample includes at least 5 ‘success’ events and 5 ‘failure’ events. Here, a ‘success’ event is defined as

- When the response that a randomly chosen bookstore from the sample is near another bookstore is *‘True’.*
- When the response that a randomly chosen bookstore from the sample is near a coffee shop is *‘True’.*

Statistical Tests

To compare the results between San Francisco and New York City, we will conduct statistical tests. Since we are working with proportions, and we'll be sampling two sets of data, it will be appropriate to carry out two-sample proportions Z-test on the datasets. In order for the statistical test results to be considered meaningful, the following assumptions are taken:

1. The samples are independent.
2. Each sample includes at least 5 ‘success’ events and 5 ‘failure’ events, with the same definition of ‘success’ as before.

Results – The Analysis of the Collected Data

We have made the assumption that the results returned from Foursquare are good, accurate and consistent, and the data collected do not need additional cleaning/scrubbing. It should be noted that Foursquare app is a crowd-sourced platform. This app heavily relies on the contribution of data from its users, which means that the data collected can only be as good as what Foursquare users have contributed.

Note also that when Foursquare frequently updates its database. The collection of the dataset was carried out using Foursquare database version *'20181201'*.

Datasets

The venue datasets collected for San Francisco and New York at the time when this project was carried out are available as reference, respectively stored in csv file format with the filenames *'df_SanFrancisco.csv'* and *'df_NewYork.csv'*.

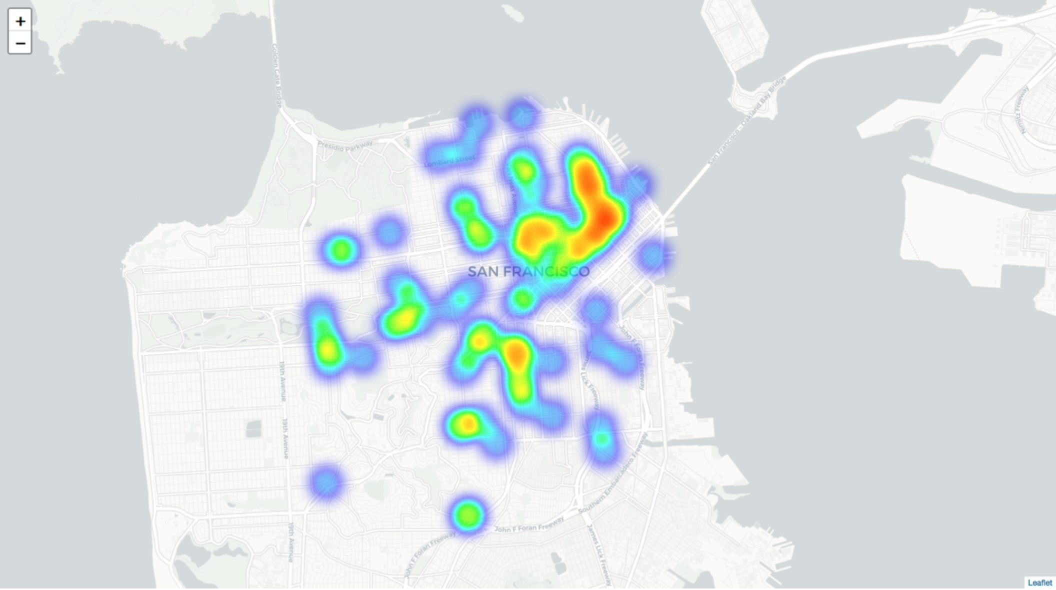
Measurement of proportions

Below is a tabulated summary of the statistical computations.

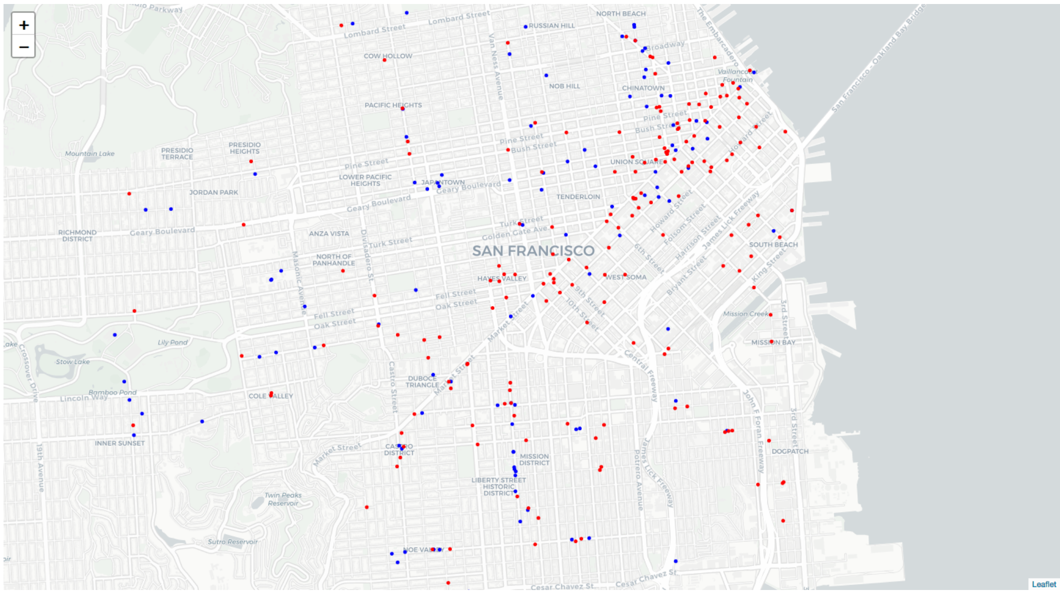
	San Francisco	New York City
Sample size	131	203
Number *(and proportion)* of bookstores near another bookstore	90 *(.687)*	152 *(.749)*
Number *(and proportion)* of bookstores near a coffee shop	93 *(.710)*	134 *(.655)*

Visualizing Bookstore Venues: San Francisco

The following “heatmap” is created with the use of Folium python library[3]. The red color shows where the bookstores are concentrated in the city of San Francisco, while the blue-purple indicates that a few bookstores are located away from the city center.



Proximity of Bookstores and Coffee Shops



The bookstores are shown as blue dots (●) and the coffee shops are shown as red dots (●) on this map. The street grid may give a rough indication on the number of city blocks one venue is separated

[3] https://python-visualization.github.io/folium/

from another. This map does indicate to me that many bookstores do indeed have coffee shops within one or two blocks.

Discussion

Confidence intervals for the Population Proportion

At 95% confidence level, the table summarizes the C.I. of the population proportion of bookstores in San Francisco (sample size of 131):

Proportion of bookstores	Confidence interval at 95% level
Near another bookstore	(.608, .766)
Near a coffee shop	(.632, .788)

From what I can tell, there is a high level of confidence to state that more than half the bookstores in San Francisco are near another bookstore. Likewise, more than half of them are near a coffee shop.

Comparing San Francisco and New York

Proportions of bookstores near another bookstore

The level of statistical significance of .05 is chosen. Based on the two-sample proportion Z-test, using the Foursquare data collected with a coverage radius of 4000m, the proportion of bookstores that are within 250m from other bookstores in San Francisco *(.687)* is not significantly different from that of the bookstores in New York City *(.749), z = -1.233, p = .217*.

Proportions of bookstores near a coffee shop

Based on the two-sample proportion Z-test, using the Foursquare data collected with a coverage radius of 4000m, at .05 level of significance, the proportion of bookstores that are within 250m from coffee shops in San Francisco *(.710)* is also not statistically significantly different from that of the bookstores in New York City *(.655), z = 1.044, p = .296*.

Personal Thoughts

This project has taken a brief look at the venue profile of bookstores in San Francisco and New York. While the data collected are sampled proportions, I would argue that we may use the proportion values as estimates in predicting the probability of finding another bookstore or a coffee shop nearby. For business owners, I think the proportion values may be useful indicators on two things.

1. The level of competition from neighboring stores offering the same service.
2. The level of complementary services around a target location.

If we have profit/loss data and foot traffic data to corroborate the data collected, we could inspect further at what level of competition is beneficial to the service provided, or, what kind of complementary services are good to have around.

It should be noted that this analysis only holds under the conditions of a 4000m coverage radius and a proximity distance of 250m. Note also that it is subjected to the limitations held within the Foursquare data.

Conclusion

This project has taken a brief look at the venue profile of bookstores in San Francisco and New York City. In San Francisco, the proportion of bookstores near another bookstore is *.687*, while the proportion of bookstores near a coffee shop is *.710*. In New York City, the proportion of bookstores near another bookstore is *.749*, while the proportion of bookstores near a coffee shop is *.655*.

We have found through the statistical tests that, the venue profiles of bookstores in these two cities are not significantly different – both cities have similar proportions of bookstores that are near another bookstore, and both have similar proportions of bookstores that are near a coffee shop.

Supporting Documents

1. Jupyter python notebook: Capstone-Final-Project-FINAL.ipynb
2. San Francisco dataset: df_SanFrancisco.csv
3. New York dataset: df_NewYork.csv