Analyzing relationships between median household income and prevalence of commercial venues in San Francisco neighborhoods

Author: Alex Hsiao
Date: May 6, 2021

Introduction

Before a new business opens in a neighborhood, the business owner often performs an analysis of the neighborhood demographics to better assess his/her business' potential for success. This can include an analysis of the age distribution of residents or the evaluation of the zoning in the neighborhood (proportion residential vs. commercial). All of this is imperative so that the business owner has confidence that the product or service provided will have sufficient demand in the neighborhood market.

Another important parameter is median household income. Median household income is a significant economic indicator that is not only indicative of the economic growth in an area, but is directly related to individuals' spending power. The greater the spending power (also referred to as purchasing power or disposable income), the more likely individuals' in a neighborhood will patronize local venues. Therefore, it is also imperative that business owners evaluate the median household income within their target neighborhood to see if it aligns with their product or service offerings. For example, it may be unreasonable for a business owner to open a five-star restaurant in a neighborhood with median household income well below the national average.

San Francisco is a vibrant and diverse city in the state of California. It is home to over 800,000 residents spread over more than 25 neighborhoods. Each of the neighborhoods varies in style and demographics; ranging from the notoriously rough Tenderloin to the trendy SOMA (South of Market). While some neighborhoods like Chinatown pay homage to the past, others are home to some of the largest high-tech companies in the world.

In this study, the various neighborhoods of San Francisco will be evaluated and an analysis performed to find relationships between the median household income of each neighborhood and the prevalence of venues, along with venue types, per capita. While this does not indicate whether or not each venue is financially successful, this will help indicate the level of demand for each venue type within a neighborhood. Similarly, the lack of a venue type within a neighborhood would indicate poor demand and inform a business owner to target a different neighborhood. Analysis will be conducted to see if the venue type correlates with median household income and if similar neighborhoods by venue type are also similar by median income. Overall, results of this study will help inform new business owners of the potential demand and competition of their product or service within each San Francisco neighborhood as correlated to median household income. This will give business owners an additional tool to use when they assess locations to open their new businesses.

Description of Data

For this study, San Francisco neighborhoods are defined by zip codes. Zip code information is easily accessible online from websites such as zipatlas.com (http://zipatlas.com/us/ca/san-francisco/zip-code-comparison/median-household-income.htm). An excerpt of the zip code data from the website is shown in **Figure 1**. As seen in the table, the website also provides location data with latitude and longitude coordinates, as well as population and average household income statistics. It should be noted that the population and household income data from this particular website is not current. Thus, the key data to be scraped from this website will be the zip codes and latitude/longitude coordinates. Detailed descriptions of how this data is scraped and processed will be discussed in the Methods section below.

#	Zip Code	Location	City	Population	Avg. Income/H/hold	National Rank
1.	94127	37.736535, -122.457320	San Francisco, California	20,624	**$95,313.00**	#350
2.	94105	37.789168, -122.395009	San Francisco, California	2,058	**$88,976.00**	#488
3.	94123	37.800254, -122.436975	San Francisco, California	22,903	**$84,710.00**	#633
4.	94130	37.820894, -122.369725	San Francisco, California	1,453	**$80,959.00**	#785
5.	94131	37.746699, -122.442833	San Francisco, California	27,897	**$76,044.00**	#1,042
6.	94114	37.758085, -122.434801	San Francisco, California	30,574	**$75,727.00**	#1,062
7.	94129	37.797526, -122.464531	San Francisco, California	2,228	**$73,571.00**	#1,212
8.	94116	37.744410, -122.486764	San Francisco, California	42,958	**$66,627.00**	#1,928
9.	94117	37.770533, -122.445121	San Francisco, California	38,738	**$63,983.00**	#2,270
10.	94121	37.776718, -122.495781	San Francisco, California	42,473	**$61,776.00**	#2,604

Figure 1. San Francisco zip codes and location data. Excerpt of San Francisco zip code data from the zipatlast.com website. The full table has 27 zip codes with data for latitude and longitude, as well as population and median household income from 2000 census.

Updated 2019 population and median household income data for each zip code (neighborhood) is obtained directly from the U.S. Census website (https://data.census.gov/). Specifically, population data is obtained by accessing the Table ID DP05: ACS DEMOGRAPHIC AND HOUSING ESTIMATES and further refining the data by geography. The population data is filtered down to only the zip codes in San Francisco. The data table can then easily be downloaded as a csv file and analyzed in Python simply by importing/reading the data. An excerpt of this csv data is shown in **Figure 2.**

GEO_ID	NAME	DP05_0001E
id	Geographic Area Name	Estimate!!SEX AND AGE!!Total population
8600000US94102	ZCTA5 94102	31392
8600000US94103	ZCTA5 94103	30703
8600000US94104	ZCTA5 94104	429
8600000US94105	ZCTA5 94105	10916
8600000US94107	ZCTA5 94107	31461

Figure 2. Sample data of Population by zip code. Excerpt of San Francisco zip code data for median household income from the U.S. Census website. Data obtained from Table ID B19013 and filtered to San Francisco zip codes.

Similarly, the median household income data is accessible in Table ID B19013: MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS. The same geographic filtering is done directly on the website and the data downloaded in csv format. Alternatively, one could use the U.S. Census API to access the data; however, for this study it was simpler to simply download the data in csv format. An excerpt of the median household income by zip code data is shown in **Figure 3**.

The corresponding neighborhood name (or names) is obtained for each zip code from the website for San Francisco Burden of Disease and Injury Study (http://www.healthysf.org/bdi/outcomes/zipmap.htm). The website page has an HTML table with San Francisco zip codes and neighborhood names. This data is scraped into Python and the neighborhood names easily associated to the aforementioned data. An example of the table on the webpage is shown below in **Figure** 4. In the case that a particular zip code is not listed on this website, the neighborhood information is obtained from the website Zip Data Maps (https://www.zipdatamaps.com/).

GEO_ID	NAME	B19013_001E
id	Geographic Area Name	Estimate!!Median household income in the past 12 months (in 2019 inflation-adjusted dollars)
8600000US94102	ZCTA5 94102	46372
8600000US94103	ZCTA5 94103	75764
8600000US94104	ZCTA5 94104	51500
8600000US94105	ZCTA5 94105	213987
8600000US94107	ZCTA5 94107	166985

Figure 3. Sample data of Median Household Income by zip code. Excerpt of San Francisco zip code data for median household income from the U.S. Census website. Data obtained from Table ID DP05 and filtered to San Francisco zip codes.

Zip Code	Neighborhood
94102	Hayes Valley/Tenderloin/North of Market
94103	South of Market
94107	Potrero Hill
94108	Chinatown
94109	Polk/Russian Hill (Nob Hill)

Figure 4. Neighborhood name(s) by zip code. Excerpt of the data from the San Francisco Burden of Disease & Injury Study website (http://www.healthysf.org/bdi/outcomes/zipmap.htm). Table shows each zip code with its corresponding Neighborhood name (or names).

Finally, venue data is accessed through the Foursquare API using the 'explore' query. Venues for each neighborhood are requested using the associated latitude and longitude coordinates for each zip code/neighborhood. The venues will be categorized and grouped by the 'venue category' field returned from the Foursquare venue database. This will then provide a simple table for each neighborhood with counts for each venue category. Moreover, the data can be normalized by the population data described above to get a per capita data table. Analysis on a per capita basis will allow for more effective comparison of the neighborhoods. More detailed analysis and utilization of the Foursquare data will be described in the Methods section.