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

Author: Alex Hsiao

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 (SF) 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.

Methods

Initial SF neighborhoods data

The key data gathered for each of the SF neighborhoods included: zip code, latitude and longitude coordinates, population, median household income, and neighborhood names. The bulk of this data was found to be readily available on the Zipatlas webpage (http://zipatlas.com/us/ca/san-francisco/zip-code-comparison/median-household-income.htm). An excerpt of the data on the webpage is shown above in **Figure 1**. The data from the Zipatlas webpage was scrapped using the BeautifulSoup Python library and organized into a Pandas dataframe.

Updated census data for neighborhood population and median income

A check of the SF neighborhood data revealed that the neighborhood population and median income data, though accurate, was outdated and sourced from old census data. To update this information, the most recent U.S. Census data from 2019 was obtained from the U.S. census data website (https://data.census.gov). While there does exist an API that could be used to pull data from, for the purpose of this study, it was simpler to access the relevant data tables directly through the web application and download the data in CSV format. The CSV data was then easily imported using the Pandas 'read_csv()' function. Specifically, the 2019 population data was accessed in Table ID DP05 and the median income data was accessed in Table ID B19013. To obtain only the SF neighborhoods data, a geographical filter was applied in the web application to filter by zip code. Then, only the SF neighborhoods were manually selected to be included in the data table. Examples of the downloaded data in CSV format are shown in **Figures 2 and 3** above. The CSV data was imported into Pandas dataframe objects and subsequently merged with the SF neighborhoods dataframe by zip code.

Finally, to ensure meaningful downstream data analysis, all neighborhoods with a population less than 10,000 people was eliminated from the dataframe. Low population neighborhoods often have low populations because the majority of the land may be zoned for industrial or other purposes. Moreover, the low population may be due to the fact that the neighborhood area is small. Additionally, because this study will focus on an analysis of venues per capita, I did not want any data to be skewed by significantly smaller population neighborhoods. In total, 5 neighborhoods had populations less than 10,000 and were eliminated. These include the zip codes of: 94104, 94111, 94129, 94130, and 94158.

Neighborhood names

Neighborhood names for each zip code were scraped from the SF Burden of Disease and Injury Study website (http://www.healthysf.org/bdi/outcomes/zipmap.htm). An excerpt of the data on the webpage is shown above in **Figure 4**. BeautifulSoup was once again used to scrape the data from the table, which was then converted to a Pandas dataframe and merged with the SF neighborhoods dataframe by zip code. The zip code 94105 did not have an associated neighborhood name available from the scraped website data. However, this information was easily obtained via an online search and then manually modified in the dataframe.

With the neighborhood names completed, the final SF neighborhoods dataframe with the most recent census data for population and median income was completed. An excerpt of the completed dataframe is shown in **Figure 5**.

	Zipcode	Latitude	Longitude	Population 2019	Median Income 2019	Neighborhood
0	94102	37.779500	-122.419233	31392	46372	Hayes Valley, Tenderloin, North of Market
1	94103	37.773147	-122.411287	30703	75764	South of Market
2	94105	37.789168	-122.395009	10916	213987	Financial District, South of Market
3	94107	37.768881	-122.395521	31461	166985	Potrero Hill
4	94108	37.791998	-122.408653	14143	63263	Chinatown

Figure 5. Completed SF neighborhood dataframe. Excerpt showing the first 5 rows of the completed SF neighborhood dataframe with data for zip code, latitude and longitude coordinates, population, median income, and neighborhood names.

Neighborhood venues using the Foursquare API

Venue data for each neighborhood was collected using the Foursquare explore call in the API. Each call to the Foursquare API used the neighborhood's latitude and longitude coordinates with a radius of 500 meters. One hot encoding was utilized to convert the categorical variables into indicator variables and filtered by venue category. This generated a count of venues for each category type. Subsequently, the data was grouped by zip code to get a count of all the venue types within each neighborhood.

Custom grouping of venue categories

Foursquare's default venue categories are quite expansive. For example, for restaurants alone, there are separate categories for every type of ethnic food (African, Chinese, Japanese, Italian, German, French, etc.). Analysis on all of these separate categories, while very specific, is too narrow for the scope of this study. Thus, custom higher-level categories were created to group similar types of venues together.

The final set of custom venue categories includes:

- Restaurants
- Cafés/Desserts
- Food joints
- Businesses
- Gym/Sports
- Grocery/Markets
- Health/Wellness
- Pts of Attraction
- Transportation
- Retail
- Home/Garden
- Bars/Nightlife
- Auto/Gas
- Arts
- Lodging
- Finance

Correlation between venues per capita and median income

Once the data for all the venues was organized into the custom venue categories, the venues per capita were calculated for each neighborhood. Each of the SF neighborhoods varies in population size and normalization of the venues data was necessary to ensure there was no bias based purely on population size. For example, a larger neighborhood with more people may inherently have more grocery stores to serve the larger population. Thus, analyzing the neighborhoods by total venue counts vs. median income, while not controlling for these differences in population, could result in skewed results. By normalizing by population, we can truly assess if median income is a contributing factor to increases or decreases in venue types within a neighborhood.

The venues per capita dataframe is then analyzed by determining the Pearson correlation coefficient (r value) of each venue category vs. the neighborhood median income. The Pearson correlation coefficient indicates whether the venues per capita and the median income of the neighborhoods are positively or negatively correlated. Additionally, the absolute value of the r-value is indicative of the correlation strength. The strengths were evaluated with the following cutoff values shown in **Table 1**.

Absolute Value of Pearson Correlation Coefficient (r)	Correlation Strength
$0 \leq r < 0.2$	Very Weak
$0.2 \leq r < 0.4$	Weak
$0.4 \leq r < 0.6$	Moderate
$0.6 \leq r < 0.8$	Strong
$0.8 \leq r \leq 1.0$	Very Strong

Table 1. Threshold values for determining the strength of the Pearson Correlation Coefficient.

Results

SF neighborhoods data

To better understand the SF neighborhoods, a histogram plot of the neighborhood median incomes was generated. The 22 SF neighborhoods were binned into 8 groups. The histogram is shown in **Figure 6**. The histogram plot shows that a majority of the neighborhoods have a median household income below $130,000 per year (14 neighborhood) compared to above this value (8 neighborhoods). Overall, the median household income for all of San Francisco is about $112,000 per year. Ten of the neighborhoods fall below this value and 12 above.

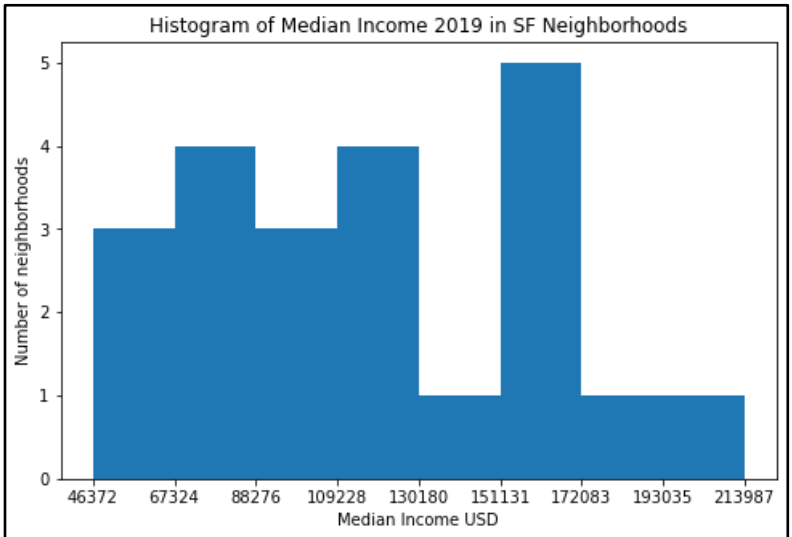

Figure 6. Histogram of neighborhood median incomes. The 22 SF neighborhoods are binned into 8 groups to assess the distribution of neighborhood median incomes.

In further assessing the SF neighborhoods data, the relationship between the neighborhood median annual income and the population was plotted and shown in **Figure 7.** From the scatter plot, it can be seen the population is not distributed evenly amongst the neighborhoods. The most affluent neighborhood has the lowest total population. The neighborhoods with the highest populations tend to be those with median household incomes ranked in the middle. Overall, the distribution appears to follow more closely to a standard normal curve with the most affluent and the more poor neighborhoods tending to have smaller populations than the neighborhoods whose median incomes are closer to the city's average.

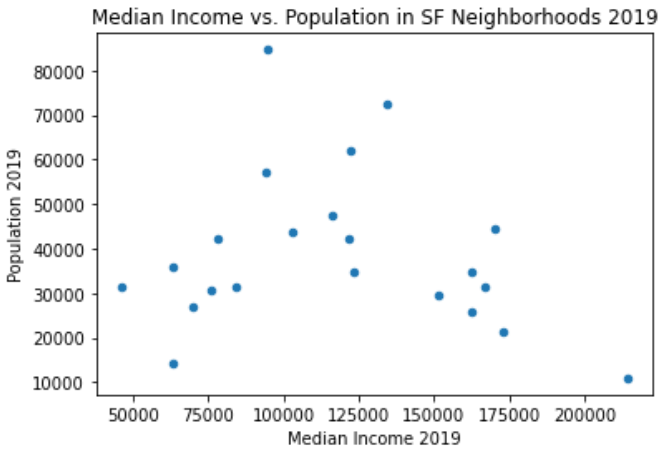

Figure 7. Scatter plot showing relation between neighborhood median income and population. The SF neighborhood data is plotted showing the relationship between the neighborhood median income and the population.

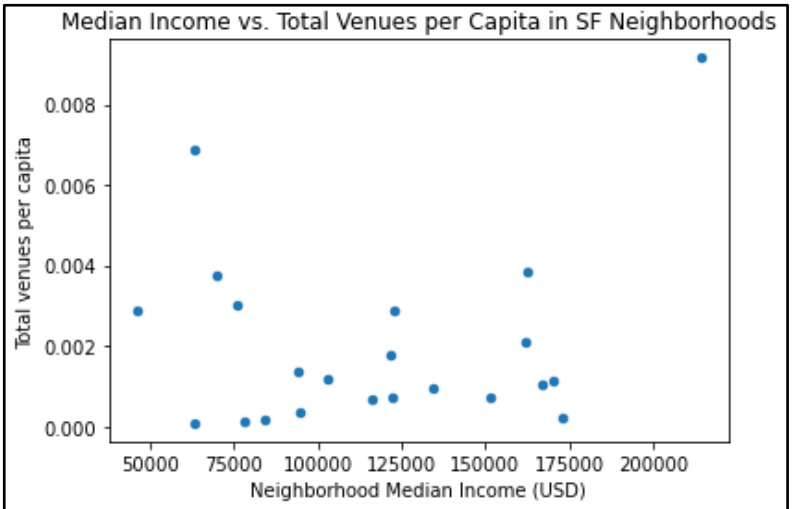

Figure 8. Total venues per capita vs. neighborhood median income. The total venues per capita, regardless of venue type, is plotted vs. the neighborhood median income.

Venues per capita and neighborhood median income

Venues per capita for each venue category were calculated based on the population within each neighborhood. The total venues per capita for each neighborhood was also calculated and the results plotted in **Figure 8**. The resulting scatter plot indicates that the most affluent neighborhood in SF also has the most number of venues per capita, regardless of the venue type. At about 0.0092 venues per capita, this is significantly higher than the neighborhood with the second highest venues per capita (0.0068 venues per capita). The distribution of venues per capita vs. median income does not appear to be linear, but in fact more of a U-shape. The neighborhoods with higher venues per capita are those either at the lower or higher ends of the median annual income range.

Correlation between venues per capita and neighborhood median income

To assess whether a correlation exists between the venues per capita of each venue category vs. the neighborhood median household income, the Pearson correlation coefficient was calculated. A positive value indicates positive correlation, increase in household income is associated with an increase in venues per capita, and negative value indicates the two variables move oppositely. The resulting Pearson correlation coefficients are depicted in **Figure 9**, with negative correlation coefficients depicted in red bars and positive correlation coefficients in green bars. The venue categories are ordered by the Pearson correlation coefficient from most negative to most positive.

From the plot, it is seen that the Finance related venues have the most negative Pearson correlation coefficient values whereas Food joints have the most positive. This indicates that as median income in neighborhood increases, the prevalence of Finance venues per capita decreases while the number of Food joints per capita increases. The venue categories whose correlation coefficients are closest to zero, indicating no correlation relative to median income, are Retail and Restaurant venues. Regardless of income in the neighborhood, these venue types have more consistent venues per capita. Other negatively correlated venue categories include Lodging, Arts, Auto/Gas, Home/Garden, and Bars/Nightlife. Other positively correlated venue categories include Transportation, Points of Attraction, Cafes/Desserts, Health/Wellness, Grocery/Markets, Businesses, and Gym/Sports. Overall, it is nearly even in the number of venue categories that have a negative correlation (7) and those with a positive correlation (9) to median household income.

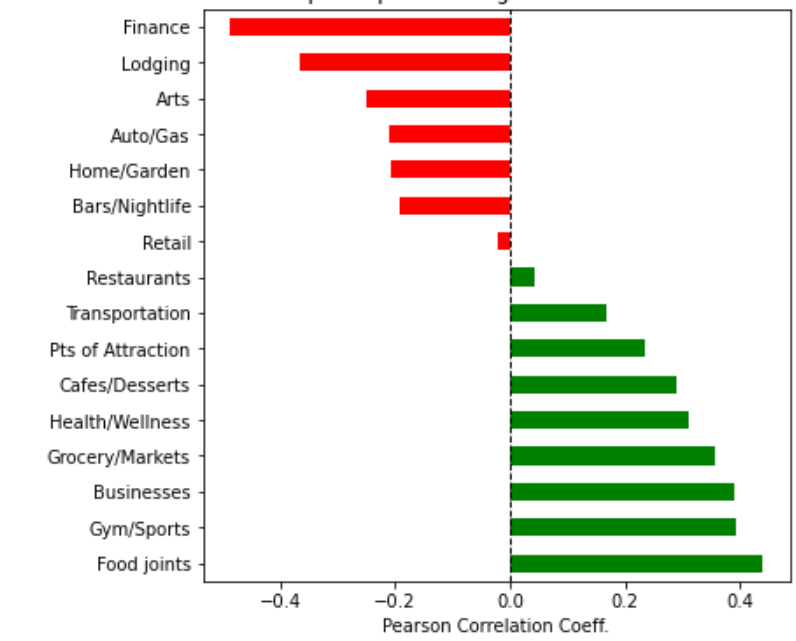

Figure 9. Pearson Correlation Coefficients of venues per capita vs. Median neighborhood income for each venue category. Pearson correlation coefficients are plotted in the bar chart with red bars indicating negative values and green bars indicating positive values.

The Pearson correlation coefficients can further be assessed for their strength of correlation by evaluating the absolute value of the coefficients. The resulting strengths of correlation are shown in **Figure 10**. The results indicate that most venue categories have very weak or weak correlations between the venues per capita and the neighborhood median income. Only two categories show moderate correlation: Food joints and Finance venues, with Food joints having a positive correlation and Finance venues having a negative correlation. It should be noted that Gym/Sports (r=0.393) and Businesses (r=0.388) have correlation coefficients very near the cutoff between moderate and weak correlation strength. Overall, none of the venue categories has a strong correlation, indicating that in the SF neighborhoods, the venues per capita and the neighborhood median income may not be well associated.

	Correlation Coefficient	Correlation Strength
Food joints	0.439616	Moderate
Gym/Sports	0.393195	Weak
Businesses	0.388447	Weak
Grocery/Markets	0.357510	Weak
Health/Wellness	0.310968	Weak
Cafes/Desserts	0.288786	Weak
Pts of Attraction	0.234615	Weak
Transportation	0.166585	Very Weak
Restaurants	0.042191	Very Weak
Retail	-0.021882	Very Weak
Bars/Nightlife	-0.193068	Very Weak
Home/Garden	-0.208350	Weak
Auto/Gas	-0.211488	Weak
Arts	-0.250921	Weak
Lodging	-0.365057	Weak
Finance	-0.487516	Moderate

Figure 10. Strength of correlation for each venue category vs. median neighborhood income. Pearson correlation coefficients are evaluated for correlation strength by the absolute value of the correlation coefficient.

Discussion

In this study, we are evaluating 22 different SF neighborhoods and looking to see if there is a correlation between the prevalence of a venue type (using venues per capita) and the neighborhood median household income.

Initial analysis of the various SF neighborhoods shows there is a rather large range in median household incomes, as shown in **Figure 6**. The distribution of median income in the neighborhoods is non-normal and is skewed to the left. In general, SF

neighborhoods are wealthy relative to the general U.S. population, where median household income is about $68,000 per year. Nineteen of the 22 neighborhoods surpass this national median. Furthermore, 18 of 22 SF neighborhoods surpass the California median household income of $75,000 per year. Thus, it is safe to say that San Francisco, in general, is a relatively wealthy city.

Using the Foursquare API, venue data was gathered for each of the neighborhoods and grouped into custom venue categories. The venues per capita was calculated to normalize the data and remove any effects of the population, thus allowing a more objective analysis of venue types per capita versus the neighborhood median income. A first look at the distribution of venues per capita across the various neighborhoods is shown in **Figure 7.** It is clear from the plot that the most affluent neighborhood has the highest number of total venues per capita, regardless of venue type. Interestingly, the neighborhood with the second highest venues per capita has one of the bottom three median household incomes. In fact, the distribution is clearly non-normal and almost appears more U-shaped, with the most venues per capita in neighborhoods with either low or high median income. There may be multiple reasons for this observation in the data. Higher median household income neighborhoods indicate a population with more disposable income. This could attract business venues looking for a market with money to spend, thereby resulting in higher venues per capita in the affluent neighborhoods. On the other hand, the lower income neighborhoods may present businesses with real estate that is more affordable and more cost effective. The middle tier income neighborhoods could also be zoned more for residential purposes and thus has fewer venues per capita. Future work could investigate each neighborhood's zoning and average business rental rates to corroborate or refute these hypotheses.

Digging into specific venue categories, an analysis of the correlation of venues per capita (by venue category) versus median household income was performed; results are shown in **Figures 9 and 10**. It is evident that for all venue category types, the correlation with household income is mostly weak or very weak. Only two venue categories, Food joints and Finance venues, have a moderate strength correlation. Food joints show a positive correlation with median household income, while Finance venues show a negative correlation. It should be noted that Gym/Sports and Businesses have correlation coefficients very near the moderate strength cutoff. Further analysis of these 4 venue categories is shown in **Figure 11**. The other venues with weak or very weak correlation strength are not further assessed. Because the correlation strength is weak or very weak, it indicates that the venues per capita are relatively uncorrelated to the median household income. Thus, business owners may not find using median household income useful in deciding whether they should open a business in a particular neighborhood. Instead, other factors should be considered.

Looking at **Figure 11** we can try to further assess the correlation of Food joints, Finance, Businesses, and Gym/Sports venues with neighborhood median income. Immediately it is obvious that the most affluent neighborhood has the highest venues per capita for Businesses, Food joints, and Gym/Sports venues. Each of these three categories shows a positive correlation with neighborhood median income. This neighborhood is in fact the Financial District/South of Market (SOMA) neighborhood with a zip code of 94105. Specifically, the high venues per capita of food joints and gym/sports could indicate that these venues target those with high disposable income.

Looking more closely at the scatter plots, it is seen that for Food joints, the venues per capita is relatively consistent across all neighborhoods except for the most affluent one. This indicates that there is clearly a push for business owners to target the affluent community. On the other hand, the moderate strength correlation may be skewed by this outlier data point. In fact, eliminating the most affluent neighborhood reduces the Pearson correlation coefficient for food joints down to $r = -0.09$, which is very weak. It would interesting to see if this phenomenon exists in other metropolitan cities across the world.

For Gym/sports venues, there is a similarly significant difference in venues per capita for the most affluent neighborhood compared to the other neighborhoods. This could point to the fact that gym/sports venues also target those with high disposable income. But interestingly, some of the lower income neighborhoods also have higher gym/sports venues per capita. Further dissecting the data into more specific gym/sports venue types could shed light on differences between the neighborhoods. For example it is possible that the more affluent neighborhoods have more private gyms whereas the lower income neighborhoods have more public sport courts or fields.

Surprisingly, finance related venues have higher venues per capita in the lower income neighborhoods. The Finance related venues includes places like banks, ATMs, and credit unions. One might think these would be equally prevalent or more prevalent in wealthier neighborhoods. Perhaps, the negative correlation with median household income is related in part to rental rates for these types of businesses. Further analysis in future studies may dig into this deeper to understand why.

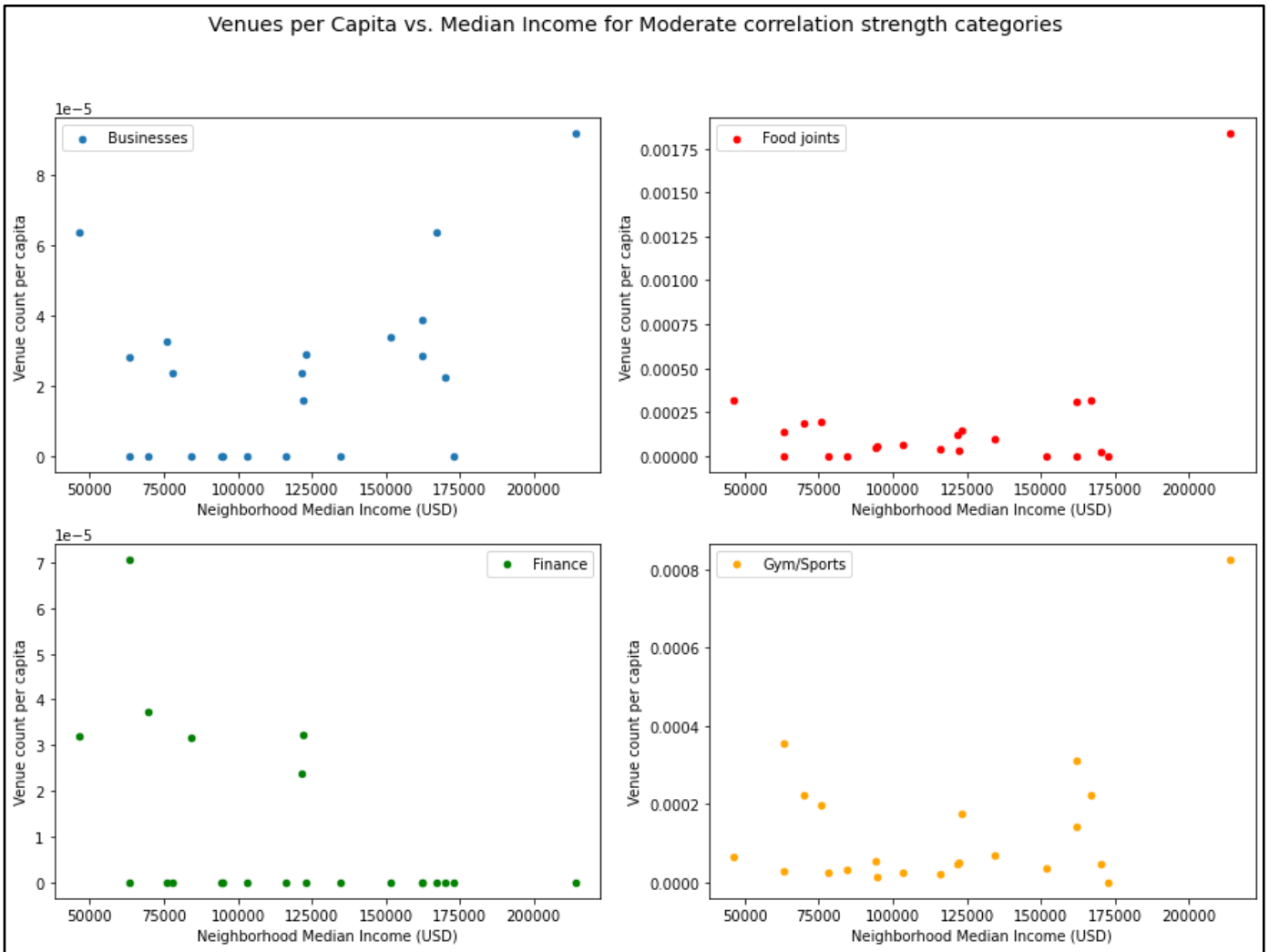

Figure 11. Venues per capita vs. Median household income for Moderate correlation strength venue categories. Scatter plots showing the venues per capita vs. median household income for SF neighborhoods, specifically the Businesses, Food joints, Finance, and Gym/Sports venue categories.

Although the data analysis tends to show there is generally weak correlation between venues per capita and median household income, this could be in part due to the grouping of the venues into broader categories. For example, if Restaurants, Cafes/Desserts, and Food joints were all combined into a broad category called “Food”, then this category

would have a Pearson correlation coefficient of 0.260, which is weak. Only upon separating these 3 categories do we see that Food joints have moderate correlation strength with household median income. Thus, future analysis could further break down each venue category into more specific groups to see if particular venue types have stronger correlations than others. This may be more useful in the instance that a business owner has a niche market that a broad category would not accurately reflect.

It should also be taken into account that the Pearson correlation looks at a linear relationship between the venues per capita and the neighborhood median income. It is very likely, as shown in **Figure 8** or the scatter plots in **Figure 11**, that the relationship between venues per capita and median income is non-linear. Future work may focus on non-linear relationships and modeling to see if there is a better measure of correlation. Additionally, other confounding factors may be contributing to the data, such as zoning within each neighborhood which could restrict the number of commercial, industrial, or residential units.

Conclusion

In this study, we are evaluating 22 different SF neighborhoods and looking to see if there is a correlation between the prevalence of a venue type (using venues per capita) and the neighborhood median household income. Overall, only 2 venue categories, Food joints and Finance venues, showed moderate correlation between venues per capita and neighborhood median household income. Food joints showed a positive correlation whereas Finance venues showed a negative correlation. Thus the data suggests that food joint business owners do tend to target more affluent neighborhoods with higher disposable income, whereas finance venues may be seeking out lower income neighborhoods. However, in general the Pearson correlation values are weak and business owners may not rely solely on evaluating the neighborhood median income to guide whether or not they should open a venue within the neighborhood. Future work could focus on refining the analysis to evaluate non-linear relationships as well as more specific venue categories. Additionally, more parameters could be assessed, such as zoning within a neighborhood or average business rental rates, to better help guide business owners on which neighborhoods to target.