Analyzing Relationships between Median Household Income and Prevalence of Venues in San Francisco neighborhoods

A. Hsiao

Applied Data Science Capstone Project
Coursera/IBM

Introduction

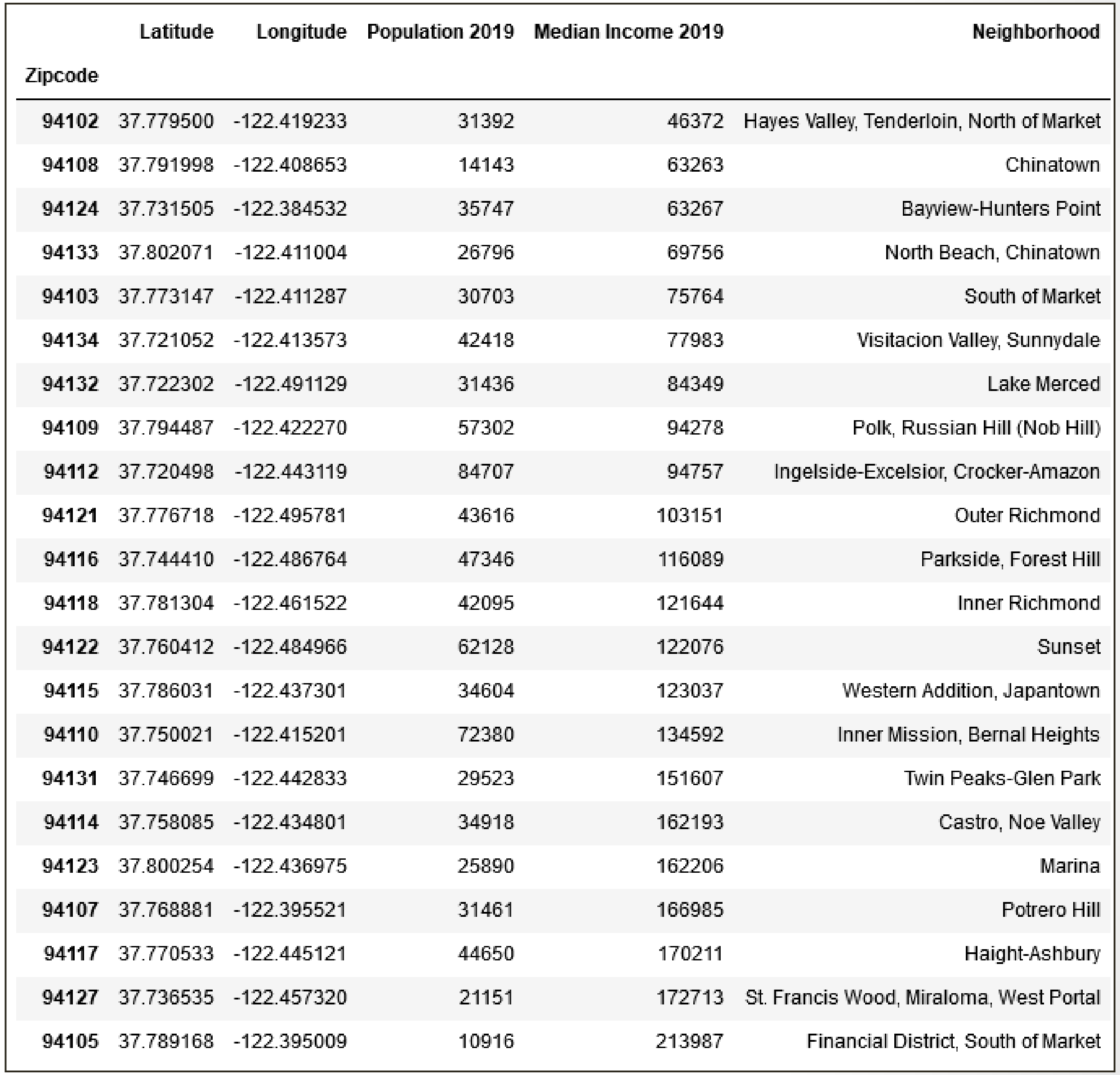
- Analysis of neighborhood demographics/statistics help inform businesses owner about their potential for success
- Median household income is directly proportional to individual spending power
- Prevalence of venue types within a neighborhood can be an indication of high demand within that neighborhood
  - Lack of a venue type may indicate poor demand within the neighborhood

Goal of this study:

- Analyze the correlation between venues per capita vs. median household income in San Francisco neighborhoods
- Determine if any particular venue types could be more successful in higher or lower income areas

SF neighborhood data

- Acquire SF neighborhood data
  - Scrape Zipatlas site to get zip codes and latitude/longitude
    - http://zipatlas.com/us/ca/san-francisco/zip-code-comparison/median-household-income.htm
  - Population and Median income data from 2019 U.S. Census website
    - https://data.census.gov/
  - Neighborhood names scraped from SF Burden of Disease webpage
    - http://www.healthysf.org/bdi/outcomes/zipmap.htm
- Drop all neighborhoods with population < 10,000
- 22 total neighborhoods to analyze



	Latitude	Longitude	Population 2019	Median Income 2019	Neighborhood
Zipcode					
94102	37.779500	-122.419233	31392	46372	Hayes Valley, Tenderloin, North of Market
94108	37.791998	-122.408653	14143	63263	Chinatown
94124	37.731505	-122.384532	35747	63267	Bayview-Hunters Point
94133	37.802071	-122.411004	26796	69756	North Beach, Chinatown
94103	37.773147	-122.411287	30703	75764	South of Market
94134	37.721052	-122.413573	42418	77983	Visitacion Valley, Sunnydale
94132	37.722302	-122.491129	31436	84349	Lake Merced
94109	37.794487	-122.422270	57302	94278	Polk, Russian Hill (Nob Hill)
94112	37.720498	-122.443119	84707	94757	Ingelside-Excelsior, Crocker-Amazon
94121	37.776718	-122.495781	43616	103151	Outer Richmond
94116	37.744410	-122.486764	47346	116089	Parkside, Forest Hill
94118	37.781304	-122.461522	42095	121644	Inner Richmond
94122	37.760412	-122.484966	62128	122076	Sunset
94115	37.786031	-122.437301	34604	123037	Western Addition, Japantown
94110	37.750021	-122.415201	72380	134592	Inner Mission, Bernal Heights
94131	37.746699	-122.442833	29523	151607	Twin Peaks-Glen Park
94114	37.758085	-122.434801	34918	162193	Castro, Noe Valley
94123	37.800254	-122.436975	25890	162206	Marina
94107	37.768881	-122.395521	31461	166985	Potrero Hill
94117	37.770533	-122.445121	44650	170211	Haight-Ashbury
94127	37.736535	-122.457320	21151	172713	St. Francis Wood, Miraloma, West Portal
94105	37.789168	-122.395009	10916	213987	Financial District, South of Market

SF neighborhood venues – Foursquare API

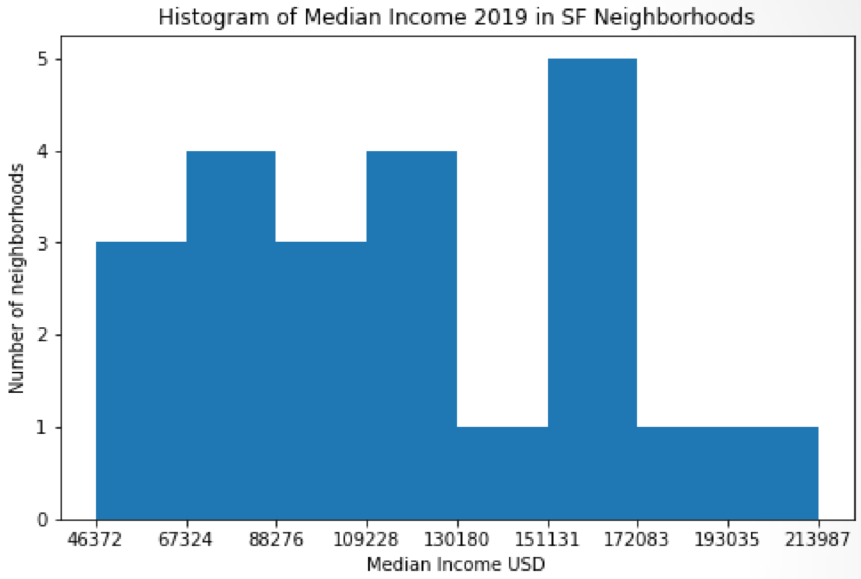
- Get SF neighborhood venues using Foursquare API ‘explore’ call
- Group venues by neighborhood (zip code)
- Organize venues into custom venue categories:

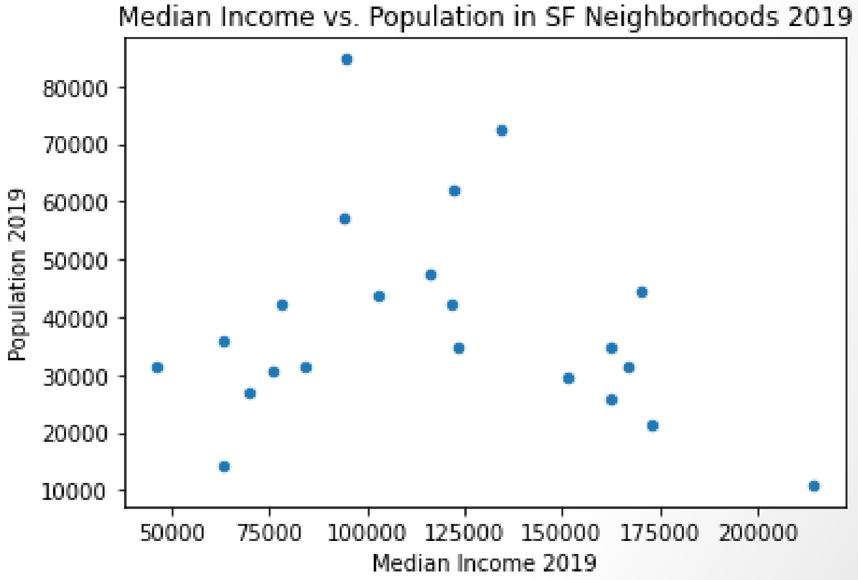
• Restaurants	• Gym/Sports	• Transportation	• Auto/Gas
• Cafés/Desserts	• Grocery/Markets	• Retail	• Arts
• Food joints	• Health/Wellness	• Home/Garden	• Lodging
• Businesses	• Pts of Attraction	• Bars/Nightlife	• Finance

- Calculate venues per capita for each neighborhood
  - Divide venue counts by each neighborhood’s population
  - Normalizes the data
    - Eliminate bias from different population sizes
    - Isolate the effect of median household income

SF neighborhood data visualization

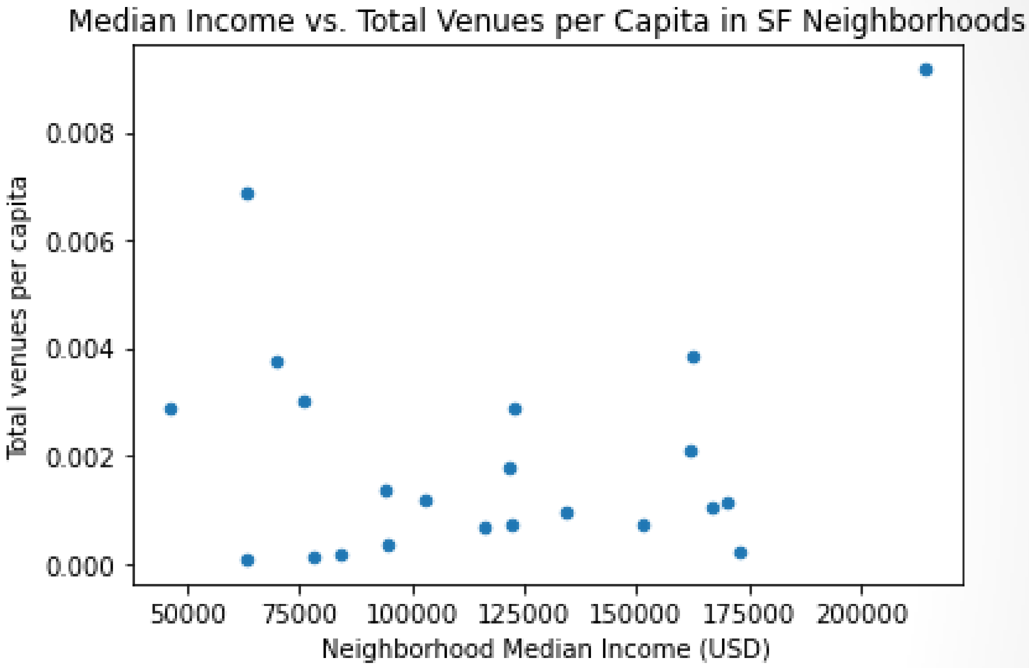
- Histogram of Median Income
  - Left skew
    - 14 neighborhoods below $130,000 annual income
    - 8 neighborhoods above
  - SF neighborhoods are relatively wealthy
    - US household median income: ~$68,000
    - California median household income: ~$75,000
- Median income vs. Population in each neighborhood
  - More normal distribution
  - Most higher and lower income neighborhoods tend to have smaller populations



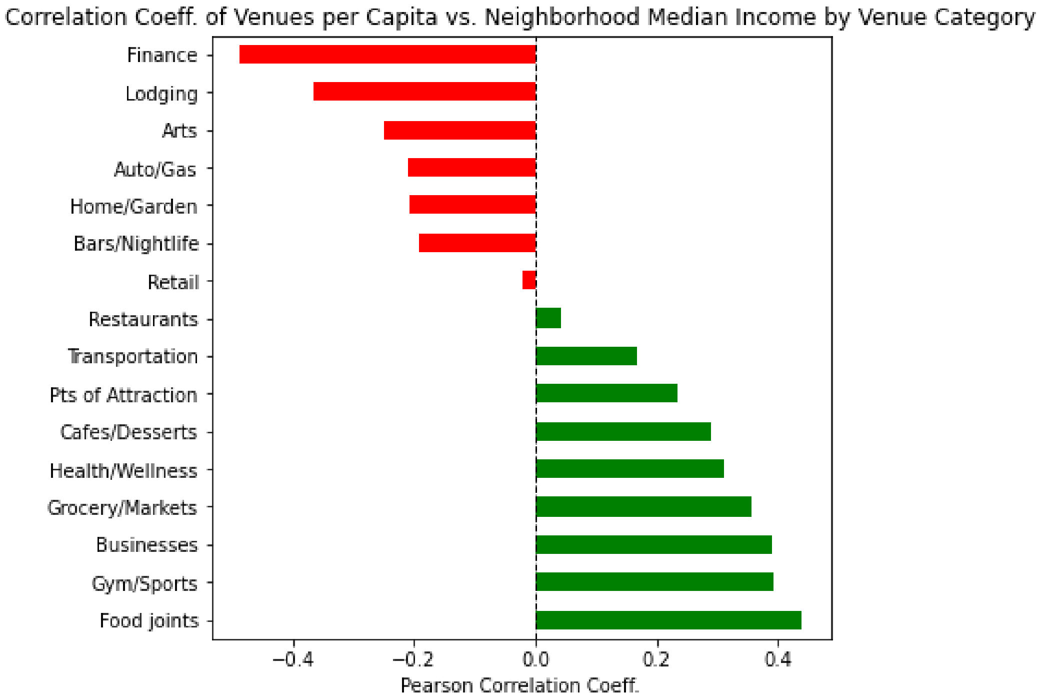


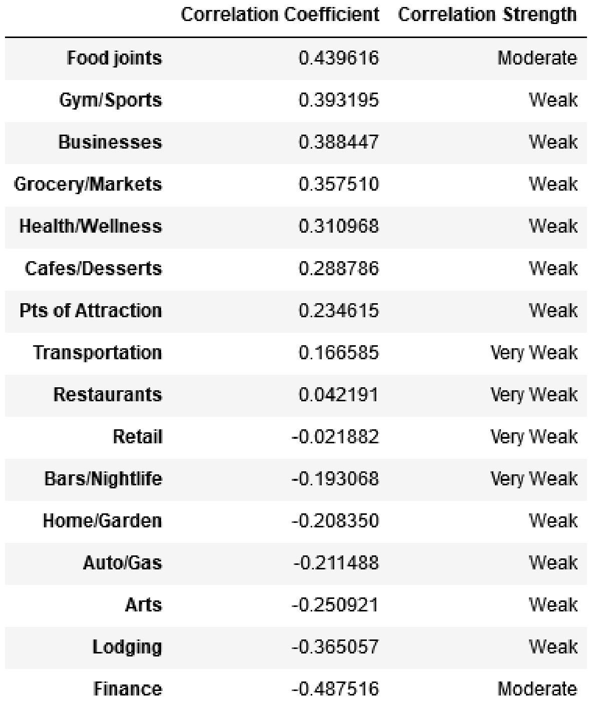
SF neighborhood median income vs. total venues per capita

- Most affluent neighborhood has the highest number of total venues per capita
  - Businesses targeting neighborhood with the most disposable income
- Distribution is non-normal
  - More U-shaped
  - Lower median income neighborhoods also have higher venues per capita



Correlation of Venues per capita vs. Median income





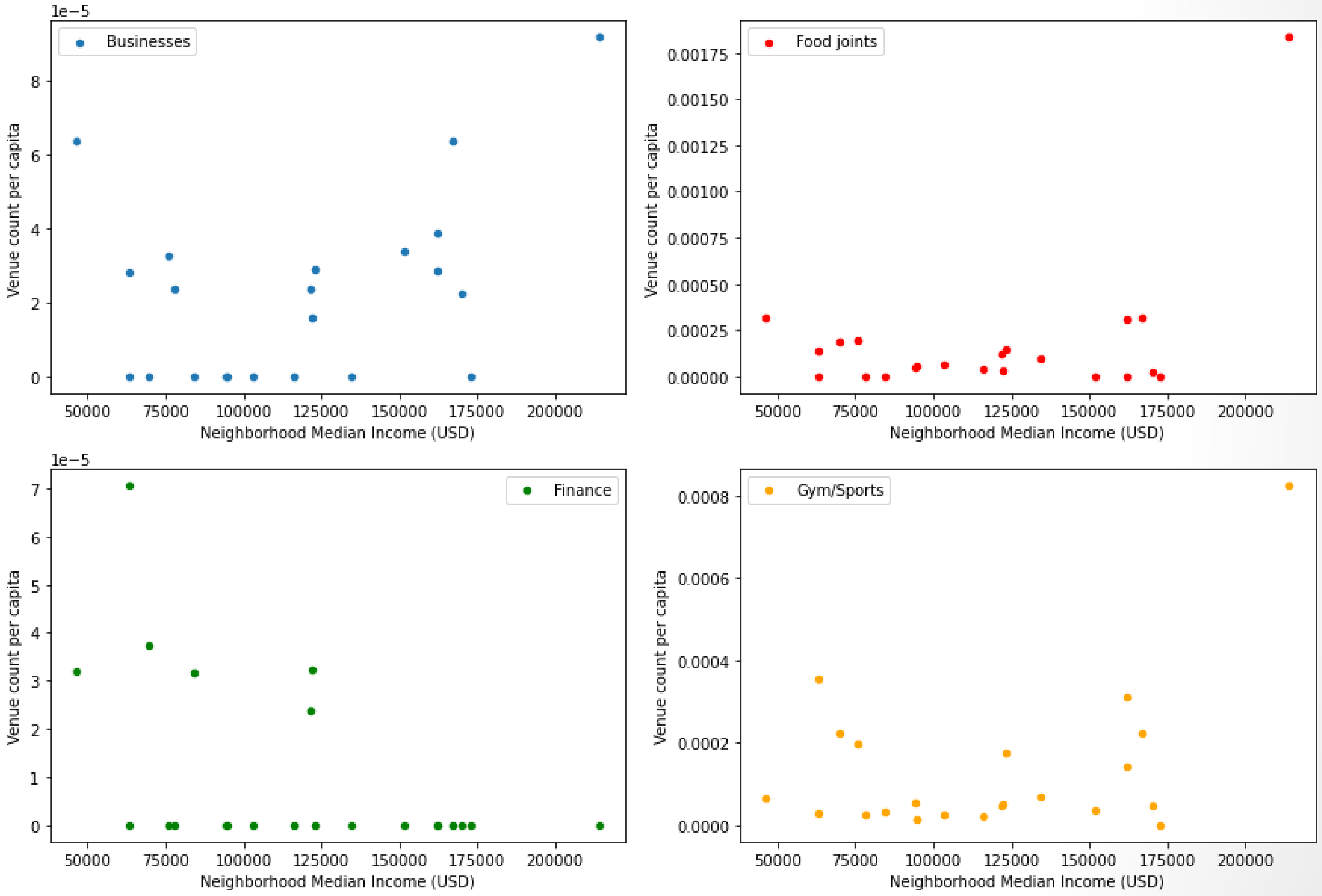
	Correlation Coefficient	Correlation Strength
Food joints	0.439616	Moderate
Gym/Sports	0.393195	Weak
Businesses	0.388447	Weak
Grocery/Markets	0.357510	Weak
Health/Wellness	0.310968	Weak
Cafes/Desserts	0.288786	Weak
Pts of Attraction	0.234615	Weak
Transportation	0.166585	Very Weak
Restaurants	0.042191	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

- Calculate Pearson correlation coefficient for each venue category vs. neighborhood median income
  - Positive correlation – Green
  - Negative correlation – Red
- Correlation strength
  - Most are weak or very weak
  - Only 2 venue types are Moderate strength: Food Joints and Finance venues
  - Gym/Sports and Businesses are very close to Moderate strength

Moderate correlation venue categories

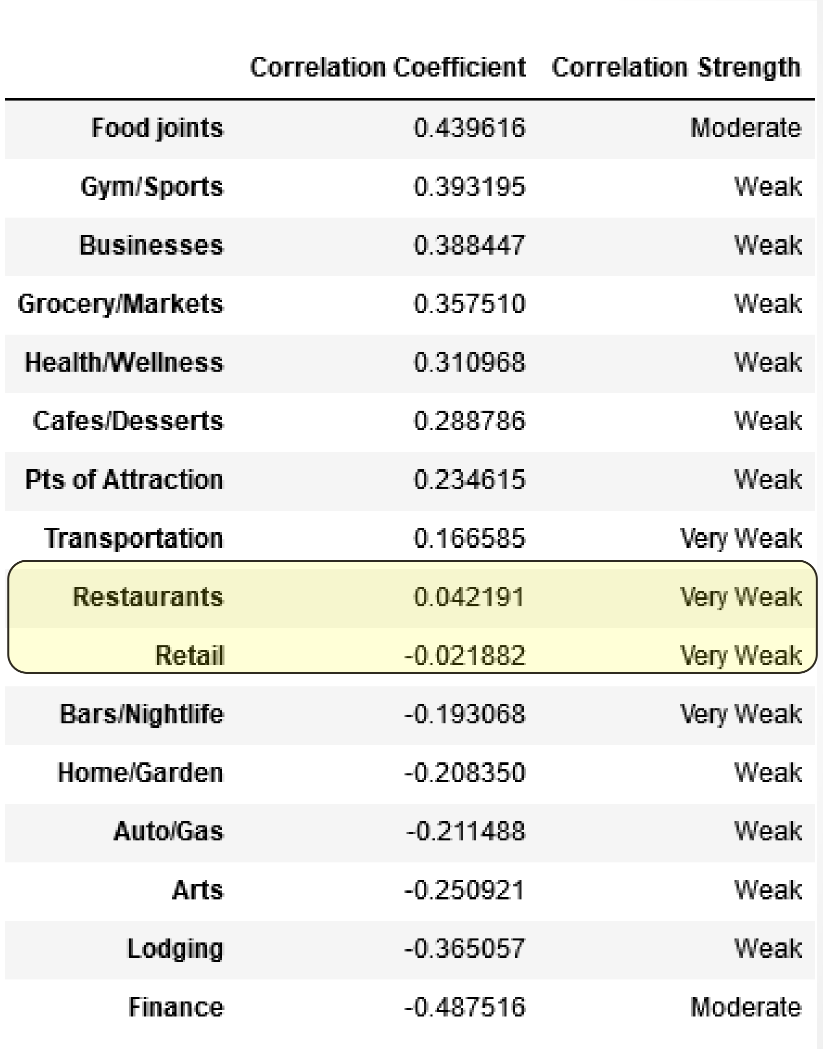
- Most affluent neighborhood has the highest venues per capita
  - Food joints, Gym/Sports, Businesses
  - Target high-income areas
- Finance venues
  - Higher venues per capita in lower income neighborhoods
  - Negative correlation with income

Venues per Capita vs. Median Income for Moderate correlation strength categories



Near-zero correlation venue categories

- Restaurant and Retail venues have near-zero Pearson correlation coefficient values → no correlation
- Restaurants and Retail businesses more equally prevalent in all neighborhoods
  - More resistant to differences in median household income
  - Everyone needs retail (clothing and goods for everyday living)
  - Everyone enjoys restaurants



	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

Conclusions

- Overall, there is a weak correlation between venues per capita vs. median household income for most venue categories
- Moderate strength correlation for Food joints and Finance venues
  - Gym/Sports and Food joints close to Moderate
  - These types of venues could do better in higher income areas
- Most affluent neighborhood has the highest venues per capita
  - Total venues and specifically Food joints, Gym/Sports, Businesses
  - Good for businesses to target neighborhoods with higher disposable income
- Restaurants and Retail have near-zero correlation with median income
  - Can be successful in all neighborhoods
  - All people rely on retail for everyday living (clothes, shoes, goods, etc.)
  - All people enjoy restaurants

Future work

- Look at other factors that may influence venues per capita
  - Neighborhood zoning (residential vs. commercial vs. industrial)
  - Business rental rates in each neighborhood
- Dissect categories into more specific categories
  - Example - Types of restaurants
  - Price level of venues
    - Restaurants in general may not be correlated to median income
    - But different price level restaurants may be prevalent in different neighborhoods
- Analyze businesses metrics and not just venues per capita
  - Net income or profit is a better indicator of a businesses’ success