High Income US Neighbourhoods Category Analysis

Applied Data Science Capstone – Final Report

-Edoardo Romani-

1.Introduction

1.1 Topic of Interest

In this final project, I am interested in exploring the distribution of venue categories for the 65 most affluent Neighbourhoods in the USA.

Venue categories are defined according to the customer purpose that they serve (eating out- restaurants, entertainment, sports amenities, bars and cafés, etc).

In this project, I group venue categories into the following macro-buckets:

- Restaurants
- Shops
- Sports
- Cafes
- Entertainment
- Art

1.2 Target Audience

The aim is to map each high-income neighbourhood according to the relative presence of each venue category bucket. The following analysis can then help identify over/undeserved neighbourhoods with respect to each category.

Ultimately, this can help small-business owners' value relative segment competition, and can also be supportive in the identification of target segments, for particular areas, in which a business launch or store opening would make sense, given the venue category distribution of the segment and the ability to serve an affluent pool of customers.

2. Data

Data sources used are the following:

2.1 List of highest-income urban neighborhoods in the United States

a. (source: Wikipedia, updated December 2018)

Sample image below

Urban Neighborhood data contains name of neighborhoods, race distribution, income data (mean), as well as location references (Metropolitan Area, State).

The neat thing about this dataset is that is displays good quality data in terms on location, which allows for a good use of the Foursquare API to retrieve venue data for all of the Urban Neighborhood listed.

The data was imported using the requests and BeautifulSoup packages.

65 Highest-income urban neighborhoods

Urban [1] Neighborhood	MSA	St.	Income (mean)[2]	No. of Homes	% Black	% Asian	% Hisp.	% Non-Hisp. White
Grove Isle-Bayshore	Miami	FL	$206,683	758	0.00%	0.00%	20.20%	79.80%
Beekman Place	New York	NY	$201,623	803	0.00%	1.20%	5.20%	93.50%
Sutton Place	New York	NY	$176,980	6,822	0.20%	3.50%	2.80%	92.10%
Old Town (Alexandria)	Washington	VA	$169,658	2,057	0.70%	1.40%	1.70%	95.10%
Tribeca	New York	NY	$163,425	4,013	5.70%	6.20%	4.80%	80.90%
Pacific Heights	San Francisco	CA	$158,937	8,794	0.40%	8.40%	2.90%	86.70%
The Gold Coast	Chicago	IL	$153,358	6,781	1.90%	2.50%	1.70%	92.70%
Georgetown	Washington	DC	$152,209	2,724	3.20%	2.00%	4.00%	89.40%
Marina	Los Angeles	CA	$151,934	2,449	2.40%	2.00%	4.20%	89.50%
Battery Park	New York	NY	$150,075	4,440	2.80%	15.90%	4.30%	74.40%

The goal now is to construct another dataset using the present Wikipedia table. The aim is in fact to fetch, for each neighbourhood, up to 50 recommended venues using the Foursquare API.

To do this, we need latitude and location data for each of the Neighbourhoods. To get these, the Geopy package is used, and the following dataset is rendered.

State	Mean_Income	Homes	%_black	%_Asian	%_Hispanic	%_White	Neighborhood_Lat	Neighborhood_lon
NY	124960	96146	6.5	5.5	9.2	76.8	40.787045	-73.975416
NY	201623	803	0.0	1.2	5.2	93.5	40.753314	-73.964811
CA	151934	2449	2.4	2.0	4.2	89.5	37.799793	-122.435205
CA	124750	5413	0.3	8.1	2.8	86.8	37.799793	-122.435205
PA	101471	4267	15.3	2.7	1.8	78.2	40.077055	-75.207400

(Same view as before, with the addition of Latitude and Longitude coordinates for each neighbourhood)

With the following information, I can now make a call to the Foursquare API.

2.2 Venue and Location data for high income neighbourhoods

(source: Foursquare API)

This data will be fetched from the API based on data source number 1. The resulting dataset is venue data (50 venues for each neighbourhood) of the recommended locations in the most affluent USA neighbourhoods:

Sample image below:

	Neighborhood	City	Latitude	Longitude	Venue	Category	Venue_Latitude	Venue_Longitude
0	Beekman Place	New York	40.753314	-73.964811	Ideal Cheese Shop	Cheese Shop	40.755040	-73.965347
1	Beekman Place	New York	40.753314	-73.964811	Ethos	Greek Restaurant	40.754526	-73.966048
2	Beekman Place	New York	40.753314	-73.964811	Ophelia	Cocktail Bar	40.753318	-73.966438
3	Beekman Place	New York	40.753314	-73.964811	Deux Amis	French Restaurant	40.754648	-73.966044
4	Beekman Place	New York	40.753314	-73.964811	Peter Detmold Park	Park	40.753599	-73.963648
5	Beekman Place	New York	40.753314	-73.964811	Mario Badescu	Spa	40.755540	-73.966961
6	Beekman Place	New York	40.753314	-73.964811	Jubilee	French Restaurant	40.755194	-73.964817
7	Beekman Place	New York	40.753314	-73.964811	Japan Society	Museum	40.752287	-73.968431
8	Beekman Place	New York	40.753314	-73.964811	Peter Detmold Park Dog Run	Dog Run	40.753607	-73.963630
9	Beekman Place	New York	40.753314	-73.964811	Pathos Cafe	Greek Restaurant	40.754590	-73.965355
10	Beekman Place	New York	40.753314	-73.964811	Sip Sak	Turkish Restaurant	40.754517	-73.968842
11	Beekman Place	New York	40.753314	-73.964811	The Smith	American Restaurant	40.755376	-73.968243
12	Beekman Place	New York	40.753314	-73.964811	Crave Fishbar	Seafood Restaurant	40.755026	-73.968576
13	Beekman Place	New York	40.753314	-73.964811	Nishida Shoten Ramen	Noodle House	40.754123	-73.968698

Last 4 columns represent Venue Data fetched from Foursquare; key columns are Category and location (lan,long)

3. Methodology

Once the dataset is in place, the aim is to analyse it by Neighbourhood, and to extract the distribution of the different venue categories to which each neighbourhood venue belongs.

To do this, the first step consisted in an assessment of the various categories that are automatically assigned/rendered to the data points by Foursquare, which are the following:

```
French Restaurant                17
Italian Restaurant               14
Hotel                            12
Park                             11
Wine Bar                         10
Pizza Place                      10
American Restaurant               9
Coffee Shop                       9
Spa                               9
Boutique                          9
Bakery                            8
Cosmetics Shop                    7
Jewelry Store                     7
Gym / Fitness Center              6
Sushi Restaurant                  6
Sandwich Place                    5
Mediterranean Restaurant          5
Steakhouse                        5
Seafood Restaurant                5
Gym                               5
Café                              5
Salon / Barbershop                5
Trail                             4
Clothing Store                    4
Salad Place                       4
Mexican Restaurant                4
Women's Store                     4
Men's Store                       4
Art Gallery                       4
New American Restaurant           3
                                 ..
```

As one can tell, these are too granular and not meaningful. Therefore, I re-coded each data point's Category Value to appear in exclusively one of the following categories, introduced at the beginning of this report;

- Restaurants
- Shops

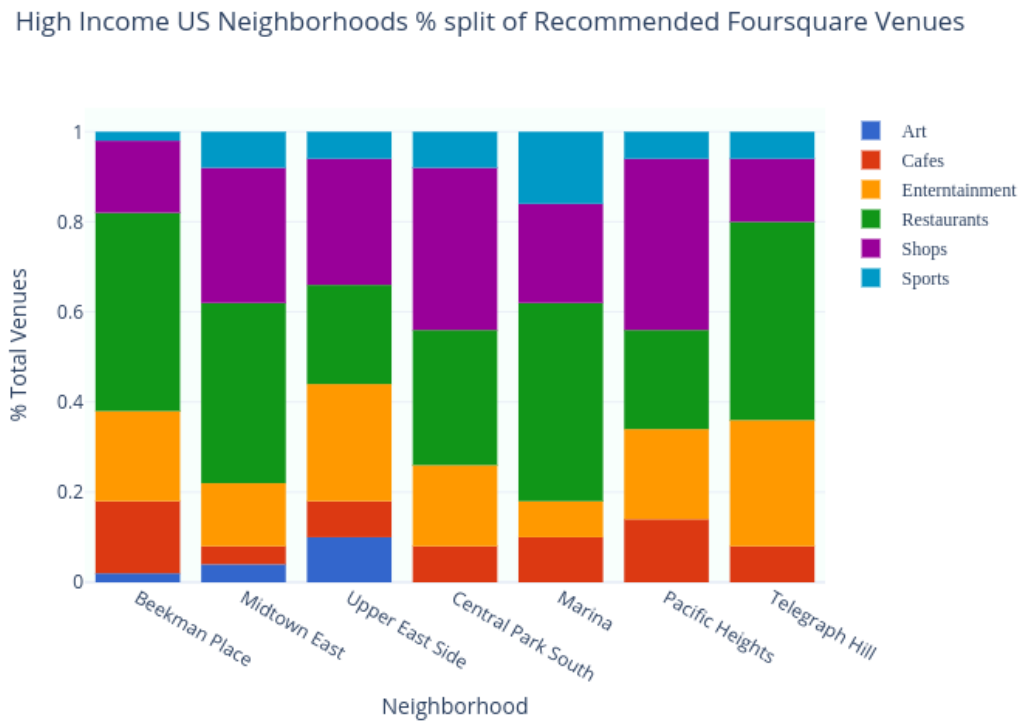
- Sports
- Cafes
- Entertainment
- Art

After doing this, the total dataset distribution by category looked like this:

Restaurants	124
Shops	92
Enterntainment	68
Cafes	33
Sports	25
Art	8

The majority of venues are either restaurant locations or Shops, followed by Entertainment, Cafes, Sports venues, and Art. Knowing that, for each neighbourhood, 50 recommended venues were fetched from the API (the maximum amount allowed under the basic/free account, one can calculate the proportion of each category within a certain neighbourhood.

Doing this for all neighbourhoods, I obtained the following result:



4. Results and Recommendations

The venue category penetration rate shows an interesting result that can lead to further discussion about potential business implications for owners looking to set-up shot in one of these top income areas in the US.

Results highlight:

- Relatively stable representation across the Restaurant, Cafes, Shops and Entertainment categories
- Skewed results across the Art, Sports and Cafes (especially underrepresented in Midtown east)

Potential Business implications (for simplicity, I am assuming this analysis as the only reason for taking an implications, without further investigating into the structural nature of those neighbourhoods, where I realize other confounding factors are here not accounted for)

- Offering Artistic venue solutions to neighbourhoods underrepresented in the Art Venue category as a viable source of revenue with initially low market competition

To further investigate penetration rate differences from a statistical perspective, one could opt for the following as additional model recommendations:

- ANOVA (Analysis of Variance)
- Clustering or Classification Algorithms

5. Conclusion

In this final assignment I have tried to show that it is possible to utilize Foursquare Venue data to identify potentially attractive business opportunities in high-income areas.

The API is very powerful and easy to use, and it is definitely a go-to tool for any geo-based analytical project.

FULL ANALYSIS LINK (NOTEBOOK):

https://github.com/Eromani1/Location_Data/blob/master/Venue_Analysis_%20Top%2010%20US%20Neighborhoods.ipynb