# Finding Relaxation in the City

Report submitted for the Applied Data Science Capstone Course of the IBM Data Science Professional Certificate on Coursera

Alberto NUIN GONZALEZ

August 3, 2019

## Table of Contents

# Introduction

## Background

In Ancient times, Romans practiced bathing and enjoyed the rejuvenating waters in magnificent thermal facilities like the Baths of Caracalla in Rome, and many others widely available across the Roman Empire and visited on a daily basis.

Today, finding relaxation in the city can be tricky. No wonder the urban Spa industry is booming, as a growing population of affluent professionals and wealthy individuals seek to relax in the wellness facilities at their doorstep.

## Interest – Target Audience

Our target audience are companies and entrepreneurs targeting the wellness market, who are looking at opening new spas in New York City.

## Description of the Business Problem

While for Ancient Romans, bathing was very popular across a wide variety of social classes, the Spas of today are surrounded by an aura of luxury. But is that really true or just a perception? Are all the spas in wealthy areas or can we find successful spas operating in less affluent areas?

We will first explore the spa market in New York City, the wealthiest city in the world[1], home to the two largest stock exchanges in the world (NYSE and NASDAQ). Its wealth, including all assets (property, cash, equities, business interests) less any liabilities, has been valued at $3,000B, not including areas around New York, like Connecticut and Long Island, that also contain a large amount of wealth.

- Does the market penetration of spas depend on the neighborhood and its wealth?
- How does the presence of spas go hand in hand with other luxury venues like hotels, restaurants and nighlife spots?
- Can Data Science predict areas in New York City where spas can be successful, but are not yet saturated with existing spas?

[1] https://www.visualcapitalist.com/top-15-cities-globally-hold-24-trillion-wealth/

# Data Sources

## Foursquare

We will use Foursquare RESTful API to get the geographical coordinates (latitude and longitude) of spas and other venue types in New York City, London and the other cities we will explore.

Foursquare[2] is an American technology company. Their location platform is the foundation of several business and consumer products, including the Foursquare City Guide app. In both November 2017 and November 2018, Foursquare made it into Deloitte's Fast 500 list of the fastest-growing technology companies. We will use Foursquare Venue Category Hierarchy[3] in our analysis to correlate the presence of spas to other categories of luxury venues.

## Open Data for New York City

NYC Open Data[4] provides free API access to open data for our research. We will use the housing market indicators as a proxy measure for the wealth of different neighborhoods: Summary of Neighborhood Sales for Manhattan[5] by NYC Department of Finance.

We will access NYC Open Data using their Socrata-based Application Programming Interface (API) functionality, which is free for public datasets. This will require installing sodapy[6].

## Nominatim

In order to access information in Foursquare about the venues in the different neighborhoods, we need to pass onto the Foursquare API the geographical coordinates of the locations we are exploring. To do so, we will install geopy in our Jupyter notebook so that we can use Nominatim[7] to get the latitude and the longitude for all the New York City Neighborhoods.

[2] https://foursquare.com

[3] https://developer.foursquare.com/docs/resources/categories

[4] https://opendata.cityofnewyork.us

[5] https://data.cityofnewyork.us/Housing-Development/DOF-Summary-of-Neighborhood-Sales-for-Manhattan-fo/5yay-3jd5n

[6] https://pypi.org/project/sodapy/

[7] https://geopy.readthedocs.io/en/stable/#nominatim

# Methodology

## Access to Data and Research Reproducibility

All the data used in this project is open-source and publicly accessible via APIs. This ensures that this analysis is fully reproducible by others in the Data Science community.

## Data Cleaning and Preparation

### NYC Housing Market Data

We have followed the following steps to access the NYC Housing Market Data:

- Request the housing market data from NYC Open Data using Socrata: as this is an public data set, there is no need for authenticating the client
- Convert the the results from JSON format into pandas dataframe

For each neighbourhood, the dataset provides the following variables, down to the type of home granularity level:

- Average sales price
- Highest sales price
- Lowest sales price
- Median sales price
- Number of sales
- Total number of properties

We have made the following observations, leading to assumptions for our housing market data:

- The type of sales prices columns in the dataframe is object. Therefore we need to convert them into integers to proceed with the analysis (unit of measure: US$)
- The average, highest and lowest sales prices are more likely to be impacted by outliers than the median sales price. Therefore we will use the median sales price as our proxy for neighbourhood wealth.
- There is a relatively low number of sales due to the narrow time window studied in our dataset. This results in some neighbourhoods not having data available for all 3 types of homes (1 family, 2 families, 3 families). We will therefore aggregate the date by averaging across the type of home.

### Geolocation

We use Nominatim geolocator to find the geographical coordinates of each neighbourhood, looping for all entries in the dataframe. A closer look at the coordinates shows that out of 20 neighbourhoods, only 14 have different latitude / longitude pairs.

This is because geolocator cannot interpret the difference between Upper East Side (79-96) and Upper East Side (59-79) or between Harlem-Upper and Harlem-Central, for example.

Due to their geographical proximity and an initial Exploratory Data Analysis (Pareto chart), we conclude there is no significant difference in sale price across those kind of similar neighbourhoods. In order to group them together:

- We clean the neighbourhood names by keeping only the letters up to the first hyphen or parenthesis, whichever comes first. Technically this is obtained by applying the str method to the neighbourhood names in the pandas dataframe, splitting the resulting strings by the pattern ‘-‘and selecting the first of the resulting strings. Same method is applied to split the resulting string by the ‘(‘ pattern.

- We create a pivot table of the dataframe, indexed by neighbourhood name, latitude and longitude. By default the pivot table method aggregates by average.

## Foursquare

In order to ensure that the Foursquare credentials remain private, we hide the cell containing them using # @hidden_cell.

First , we iterate the Foursquare requests to find the number of spas present in a 500 meter radius of the centre of the neighborhoods.
Steps:

- Iterate across all rows of the pandas dataframe,
- Define a specific url request for each neighbourhood,
- Call a request to the Foursquare API
- Convert the information received to a json file
- Read the json file into the pandas dataframe

Second, in order to get the number of different venue categories, we nest 2 for loops over a venue category list and the neighborhoods.
Steps:

- Define a Python list with the following venues: 'Hotel', 'Restaurant', 'Spiritual Center', 'Clothing Store'
- We iterate across all different venues in our list
- For each venue, perform the above request call to the Foursquare API and record the result into the pandas dataframe

## Exploratory Data Analysis

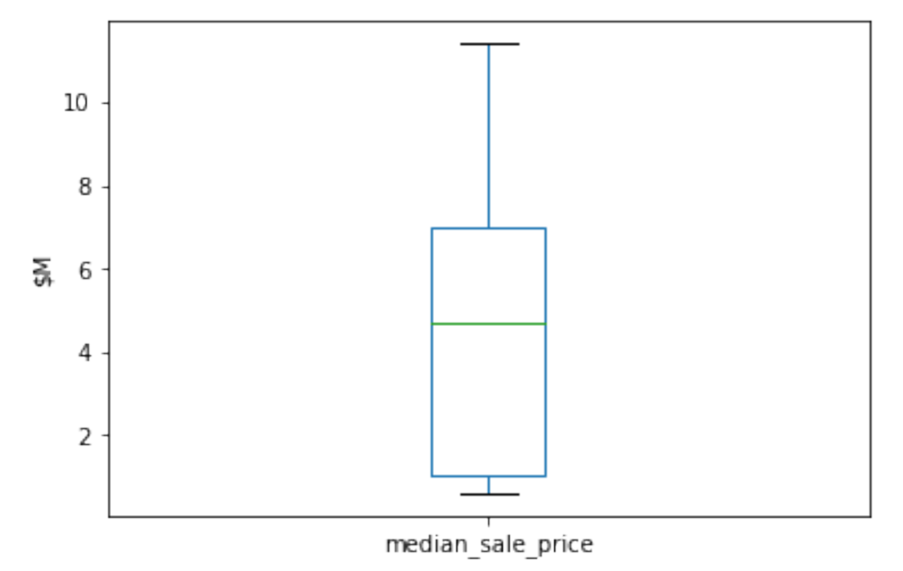
### House selling prices across the 20 neighborhoods

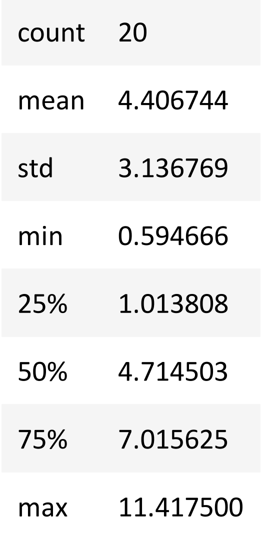
Across the 20 neighbourhoods explored, the central values of the median sales price are:

- $4.7m (median) - $4.4m (mean)

There is a very significant variation as shown by:

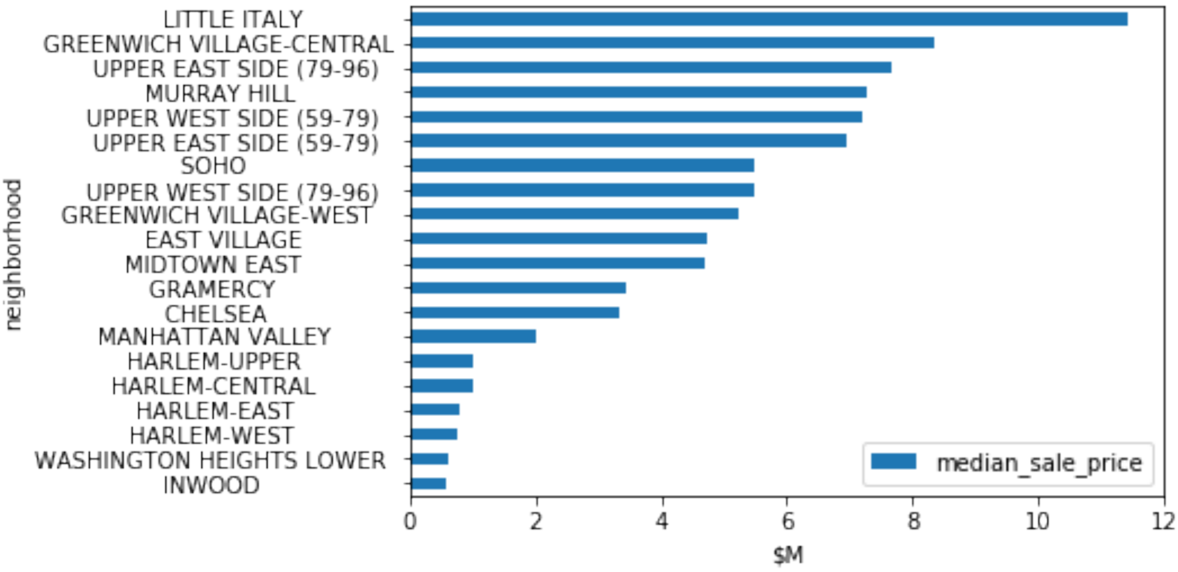
- Interquartile range: $6.0m
- Standard deviation: $3.1m
- Maximum: $11.4m





| | |
|---|---|
| count | 20 |
| mean | 4.406744 |
| std | 3.136769 |
| min | 0.594666 |
| 25% | 1.013808 |
| 50% | 4.714503 |
| 75% | 7.015625 |
| max | 11.417500 |

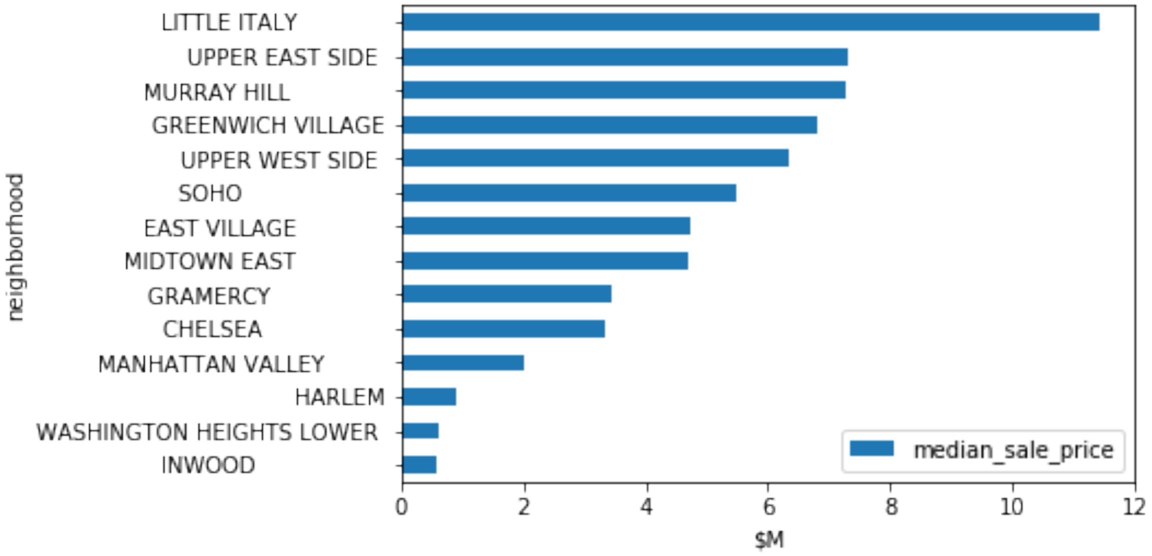
With a Pareto chart, we rank the neighborhoods by their median sale price.



Little Italy looks like an outlier, but before deciding whether to take it out or not, we will explore the presence of spas therein and compare it to the rest of the neighborhoods.

### House selling prices across the grouped neighbourhoods

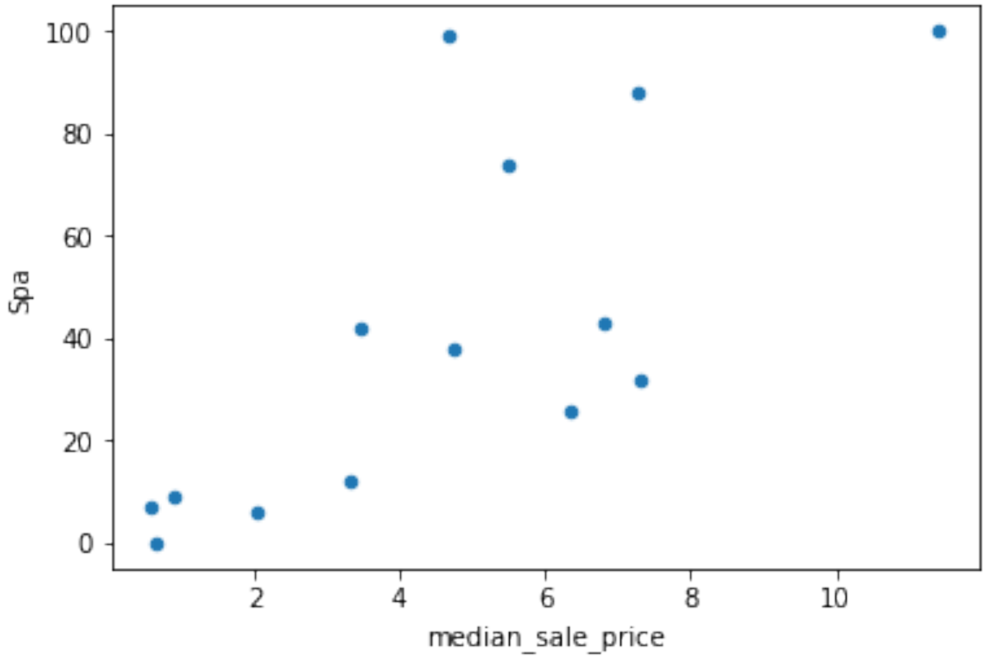
Following the neighbourhood grouping discussed in the Data Cleaning and Preparation section, the median sale price pareto is updated as follows, and we are now ready to explore the relationship between house selling prices and presence of spas in the neighbourhoods.



### Relationship between spas and house selling prices

From visual inspection of a scatter plot displaying the number of spas in the neighbourhood vs the median house sale price, we observe:

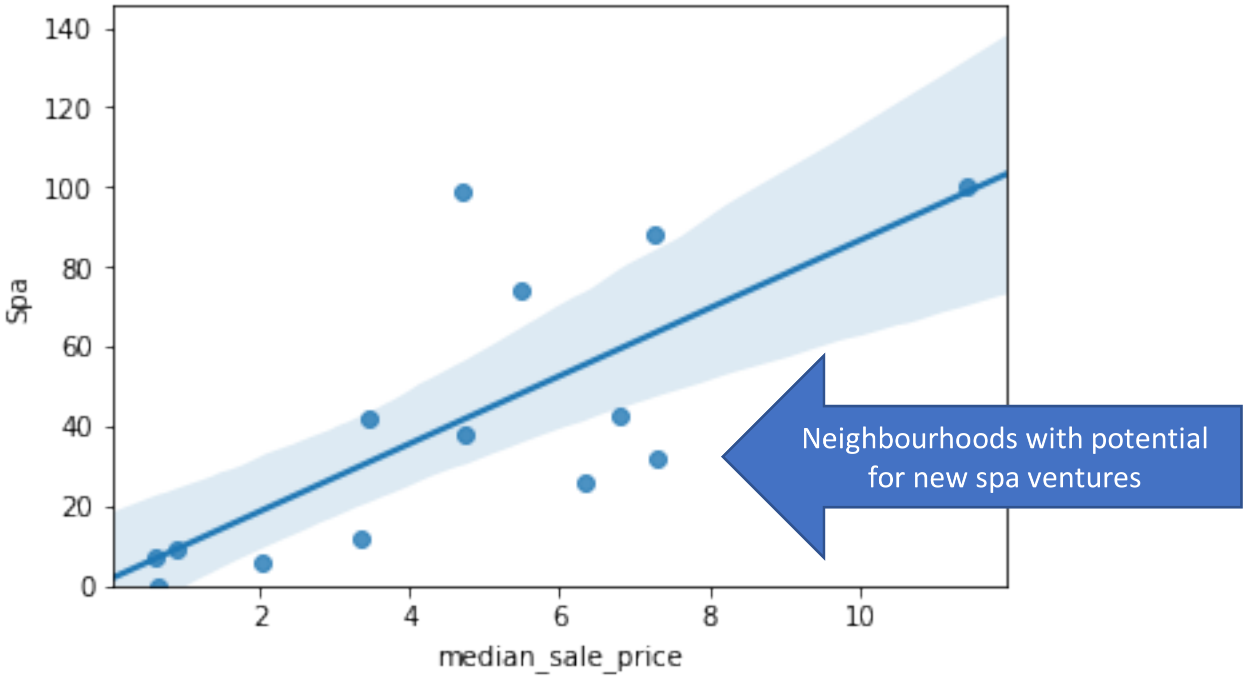
- a positive correlation between both variables
- $3m house selling price looks like a threshold:
    - Below $3m the number of spas is relatively low to the rest of the city
    - Above $3m there is a much more significant presence in the number of spas



Little Italy, which looked like an outlier for the median sale price variable, also has a very high number of spas compared with the rest of the neighbourhoods. Hence, we decide it is not an outlier for the scope of this study and we keep it in our dataset.

A linear regression analysis of the relationship between house selling prices and number of spas, using Seaborn[8], shows a strong 74% correlation between both variables.

From visual inspection of the regression plot below, we can identify a number of neighbourhoods (below the regression line) that are wealthy enough to have more spas in them.



Neighbourhoods with potential for new spa ventures

Even though 74% is a strong correlation we will explore the data further before making a recommendation to the investors.

[8] https://seaborn.pydata.org

## Relationship between spas and other types of venues

In this section, we will analyse the relationship between the number of spas and the number of other types of venues, all related to the theme of luxury and relaxation.

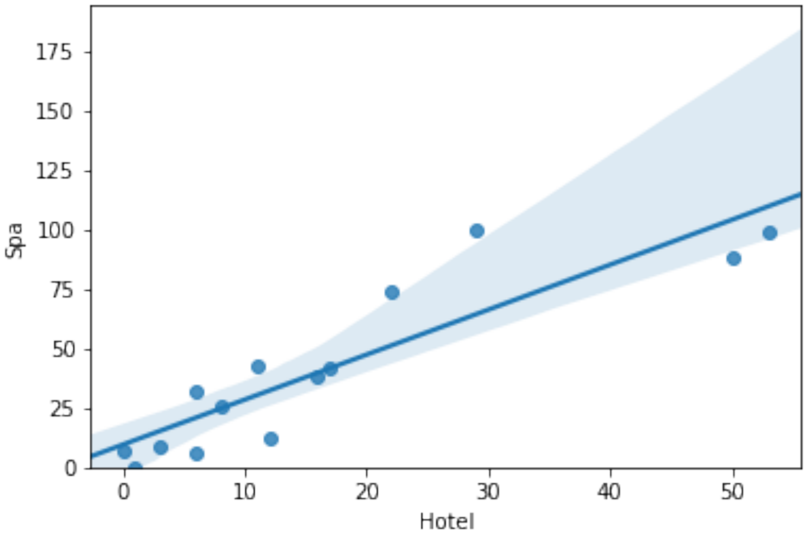
- Hotels
- Restaurants
- Spiritual centres
- Clothing stores

### *Hotels*

We expect the number of hotels to be also a good predictor of the number of spas because:

- In addition to the New York City dwellers, visitors staying in the city may also be enjoying the spas.
- Many luxurious hotels have spas in their list of amenities.

Unsurprisingly, the following linear regression chart shows a strong correlation between the number of hotels and the number of spas: 89%.



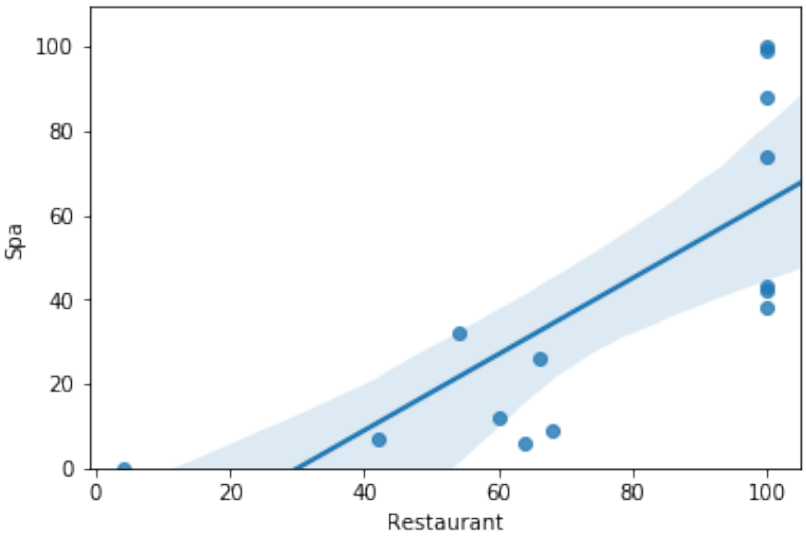
As the correlation between Hotels and Spas (89%) is even higher than the correlation between House selling prices and Spas (74%), we conclude that both variables are of interest when it comes to build a predictive model in the following section.

### *Restaurants*

We expect the number of restaurants to be also a good predictor of the number of spas because:

- Spa customers might want to enhance their spa visit with an enjoyable meal.
- Spas are located closely to hotels, as proven in the previous section, and hotel guests need restaurants nearby.

The following linear regression chart shows a strong correlation between the number of restaurants and the number of spas: 75%.

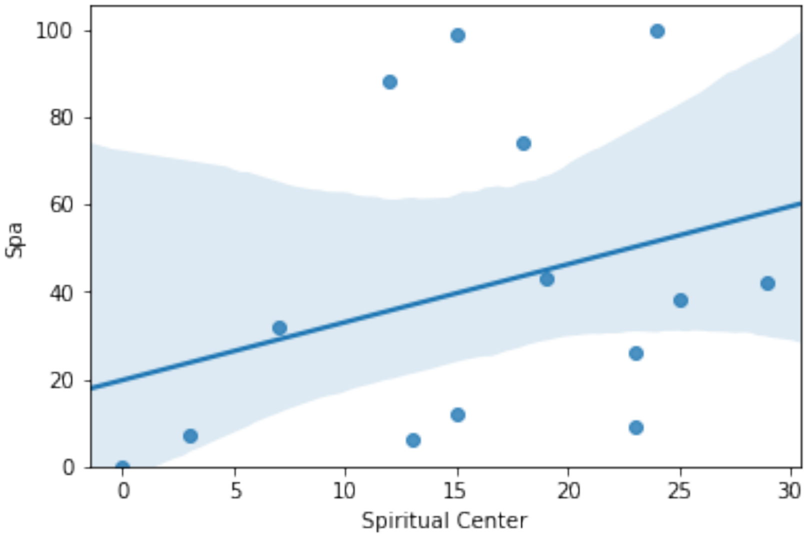


Although not as high as the correlation between Hotels and Spas (89%), the correlation between Restaurants and Spas (75%) is still marginally higher than the correlation between House selling prices and Spas (74%). Hence, we conclude that all three variables are of interest when it comes to build a predictive model in the following section.

### *Spiritual centres*

As the spa experience, in addition to luxury, is also about finding relaxation from the stress of a hectic life, we want to explore whether the number of spas may be related to the number of another venue category that people attend to attend to their spiritual needs. There is a Foursquare category that groups the spiritual centres of all faiths.

The following linear regression chart shows a weak correlation between the number of spiritual centres and the number of spas: 32%.

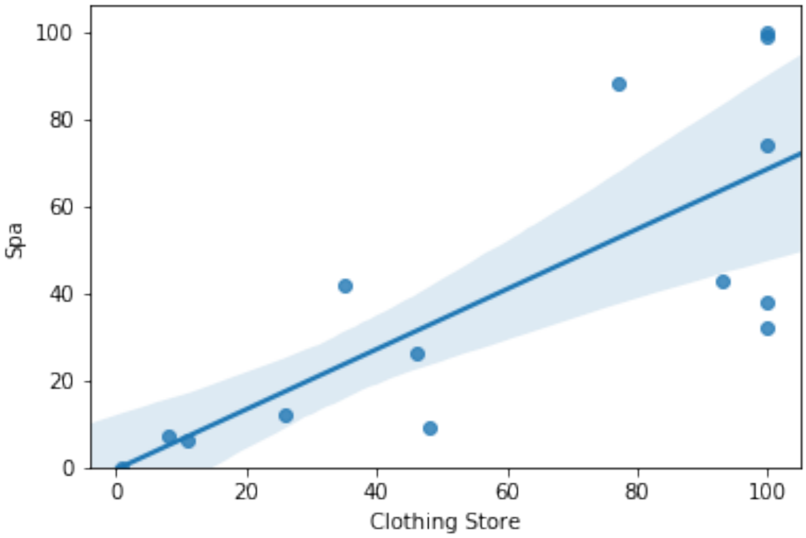


Although quite weak compared the correlation between Spas and the rest of the other variables explored so far, the correlation between Spiritual Centres and Spas (32%) may still provide useful information when it comes to build a predictive model in the following section.

### *Clothing stores*

We expect the number of clothing stores to be also a good predictor of the number of spas because Spa customers, whether New York City dwellers or visitors, might want to enhance their spa visit with a good dose of retail therapy.

The following linear regression chart shows a strong correlation between the number of clothing stores and the number of spas: 76%.



Although not as high as the correlation between Hotels and Spas (89%), the correlation between Clothing stores and Spas (76%) in the same region as the correlation between Restaurants and Spas (75%) and the correlation between House selling prices and Spas (74%).

We will use these variables, and possibly the presence of Spiritual centres, to build a multilinear regression model in the next section.

## Predictive Modeling

Following our Exploratory Data Analysis, we have found strong correlations between the number of spas in a New York City neighbourhood and the following variables:

- House Selling Prices
- Hotels
- Restaurants
- Clothing stores

We will now use Scikit learn[9] to build a multilinear model seeking to predict the number of spas using a linear combination of those variables.

Note that we have found a weak correlation between Spas and Spiritual centres (32%). As such, we do not have yet enough information to exclude them from the model for good. We will explore 2 multilinear models, one taking into account the spiritual centres, the other one not, and explore which one yields the best results.

**Multilinear Model with the spiritual centers:**

```
Value of the intercept: -10.534255688113198
Coefficient for median_sale_price : 3.1430636490315687
Coefficient for Hotel : 1.2908728519943087
Coefficient for Restaurant : 0.13119263346518345
Coefficient for Spiritual Center : -0.1400542807490374
Coefficient for Clothing Store : 0.13012826947726427
The Mean Square Error is: 79.27820029820137
The R-squared is: 0.9322810157134811
```

**Multilinear Model without the spiritual centers:**

```
Value of the intercept: -10.008269817032861
Coefficient for median_sale_price : 3.0940947286452345
Coefficient for Hotel : 1.326347234555379
Coefficient for Restaurant : 0.08186016705774106
Coefficient for Clothing Store : 0.139666474146466
The Mean Square Error is: 79.64540537386391
The R-squared is: 0.9319673512426028
```

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by the independent variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.

Here, the R2 of our model is 0.93, then 93% of the observed variation can be explained by the model's inputs. In addition to strong correlation between the Spas and the independent variables, R2 of 0.93 shows the model is robust.

Both models with and without spiritual centers have very similar R-squared and Mean Squared Errors, albeit the model that takes into account the spiritual centers is marginally better. Therefore we keep that variable in the model.

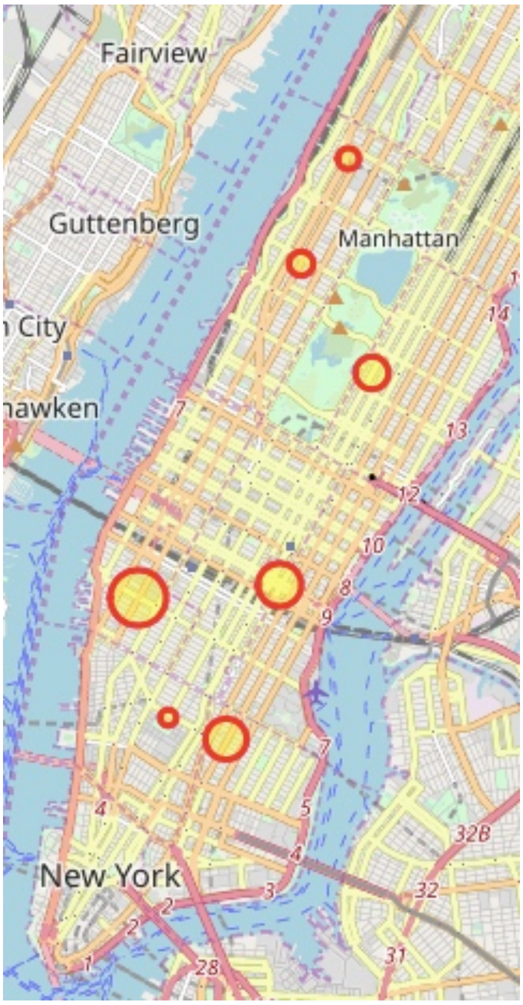
[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

## Results and Discussion

The following table ranks the New York City neighborhoods by their potential for new spas:

- Spas: number of current spas, according to Foursquare
- Prediction: number of spas, according to our multivariate model based on housing selling prices (NYC Open Data) and highly correlated venues (Foursquare)
- Potential for new spas: gap from current number of spas to prediction, whenever positive.

| Neighborhood | Spas | Prediction | Potential for new spas |
|---|---|---|---|
| Chelsea | 12 | 25.0 | 13.0 |
| Murray Hill | 88 | 98.0 | 10.0 |
| East Village | 38 | 47.6 | 9.6 |
| Upper East Side | 32 | 39.7 | 7.7 |
| Upper West Side | 26 | 31.7 | 5.7 |
| Manhattan Valley | 6 | 11.0 | 5.0 |
| Greenwich Village | 43 | 46.8 | 3.8 |



This map of New York City, rendered using Folium[10], shows the locations of the neighborhoods with more potential for new spas, the radius of the marker being proportional to the potential new spas.

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

Our predictive model explains to R-squared of 93% the presence of spas in New York City Neighborhoods based on:

- House selling prices
- Hotels
- Restaurants
- Clothing stores
- Spiritual Centers

We only found a weak correlation between Spas and Spiritual Centers, but we also found that taking into account the Spiritual Centers in the fit of a multilinear model slightly improved the model, albeit very slightly.

Our predictive model enables us to recommend neighborhoods where entrepreneurs can open new spas on the basis of strong underlying independent variables and where the model shows that the number of spas should be higher that the actual.

However, we didn't have enough data to test the predictive model. Future analysis should take into consideration data of different wealthy cities, like London, Berlin and Paris in Europe, to test the model with new data. That data can then be used to refine the predictive model.

## Conclusion

We started this research with the objective of providing data insights to companies or entrepreneurs in the Wellness sector looking at business opportunities to open new spas in New York City.

We have tested a common perception – that spas are closely related to wealth – against the data, and we have proven it correct. Although all social classes in Ancient Roman times enjoyed thermal bathing, today in New York City the presence of spas in different neighborhoods can be closely predicted by neighbourhood wealth – measured by house selling prices – and the presence of other luxury venues like Hotels, Restaurants, and Clothing Stores.

It didn't necessarily need to be like that. Exploring the relaxation and mindfulness dimension of thermal bathing, we have explored whether a correlation could be found between Spiritual Center and Spas, but only found a weak correlation.

We have built a predictive model and obtained a high R-squared of 93%, showing that most of the model variability can be explained by the independent variables rather than due to chance or underlying noise.

We have used the model to predict the number of spas in each neighbourhood, and used it to calculate the potential of new spas that could be opened in non-saturated neighborhoods. Future work should test this model against the data in other wealthy cities, providing an opportunity both to refine the predictive model and to recommend international investment opportunities in the Wellness market.