

Applied Data Science Capstone

Capstone Project - The Battle of Neighborhoods (Week 2)

Project: Data Base of demographic and economic characteristics of neighborhoods in Manhattan.

Author: Felipe Fassi Pinto

1. Introduction

New York is one of the most diverse cities in the world. Different cultures, different people and environments, living together in a single geographical and economic space. The economic space, which in its magnitude, made the city nicknamed "the city of business".

All this diversity and predominant economic factor create the scenario for companies to become extremely competitive, a fact that cheers consumers, and discourages some new entrepreneurs or business that wish to install their operations to new parts of this giant city.

Companies and entrepreneurs are increasingly looking to use data to establish business plans and operations. And it is in this scenario that the present project emerges.

1.1 Problem Description

The real estate company 'NEWPLACE', very well known in New York, intends to expand its services to the Manhattan area. The problem is that it is an extremely competitive area, and that the company has no knowledge about the neighborhoods, prices and the characteristics of the properties and in which locations they must act to reach a good target audience.

They then decided to turn to data science in order to understand more about Manhattan, and make their investment in expansion much safer.

The company also made it clear that, if it were possible, it would like to use these analyzes for other areas, thus changing only the database / city they are looking for to generate another area knowledge report.

1.2 Project StakeHolders or Target audience:

This project aims to reach the company "NEWPLACE", with an analysis of property prices, number of property sales and characteristics of Manhattan to assist in its business plan and action.

However, other companies selling real estate and tourism to understand the characteristics of the neighborhoods, thus improve their offerings to their customers, and even expand their markets and investments can use the same project.

1.3 Success Criteria:

We can say that we were successful in the project if the final result is a base with the information clustered by segmentation of prices of properties sold, number of sales and business / environment of the neighborhoods of Manhattan, ready to be analyzed by the business due to understand where they can act in their expansion.

2. Data Sources:

2.1 Data 01 - Manhattan Property Sales:

In order to carry out the project, information on all property sales that took place in Manhattan in recent years is required, containing information such as price, type of housing, amount of sales and also the neighborhood in which the property was sold. For this, we selected the database of property sales information in Manhattan, made available by the New York government, from which the following features were used:

- NEIGHBORHOOD;
- TYPE OF HOME;
- NUMBER OF SALES;
- AVERAGE SALE PRICE;

Link: https://www1.nyc.gov/assets/finance/downloads/pdf/rolling_sales/neighborhood_sales/2019/2019_manhattan_sales_prices.xlsx

2.2 Data 02 - FourSquare Api Information:

Venue information from the FourSquare API will also be used for the neighborhoods of Manhattan. The purpose of using them is to enrich the final

base with information about the characteristics of that neighborhood (businesses, parks, gyms and others), to give an even more comprehensive view of the physical structure of the area.

The base will contain:

- Business;
- Neighborhoods;
- Locations.

Link: <https://foursquare.com/>

2.3 Data Extraction

To extract all the information mentioned, the present project use the package *urllib* from the python library and the FourSquare API. Using these methods, the script can execute in any computer, and could be altered in order to be executed for different scenarios.

2.4 Data Formatting

After the extraction of the Manhattan property sales information from the New York site, it was noted that the data had some differences in the formatting between the year's sheets that was collected. For example, in the column "**TYPE OF HOME**", in some sheets the pattern of that information come as "01 ONE FAMILY DWELLINGS" and in others, the pattern is "01 ONE FAMILY HOMES". For that column, the present project had considered only the first number of the beginning ("01"), that way, it creates a pattern and also facilitates all the posterior calculation, because now we converted a string do a number.

In the present project, the types of the sheet mentioned also were changed, as follows:

Figure 01 – Manhattan Property Sales Original

```

NEIGHBORHOOD      object
TYPE OF HOME      object
NUMBER OF SALES    object
LOWEST SALE PRICE  object
AVERAGE SALE PRICE object
MEDIAN SALE PRICE  object
HIGHEST SALE PRICE object
Type of Home for Alg object
dtype: object

```

Figure 1 - Screenshot taken from the project notebook.

It was changed for:

Figure 02 – Manhattan Property Sales Changed

```

NEIGHBORHOOD      object
TYPE OF HOME      object
NUMBER OF SALES    int32
LOWEST SALE PRICE  int32
AVERAGE SALE PRICE int32
MEDIAN SALE PRICE  int32
HIGHEST SALE PRICE int32
Type of Home for Alg int32
dtype: object

```

Figure 2 - Screenshot taken from the project notebook.

The “**Type of Home for Alg**” is the column that had the number of family dwelling formatting as mentioned before. This is the new dataframe then:

Figure 03 – Manhattan Property Sales Dataframe Changed

	NEIGHBORHOOD	AVERAGE SALE PRICE	HIGHEST SALE PRICE	LOWEST SALE PRICE	MEDIAN SALE PRICE	NUMBER OF SALES	Type of Home for Alg
0	ALPHABET CITY	35448605	36513700	34383509	35448605	9	15
1	CHELSEA	101401186	131163815	78070744	97425250	31	27
2	CLINTON	18767392	18767392	18767392	18767392	4	10
3	EAST VILLAGE	47918900	56126400	42551400	44538900	14	13
4	GRAMERCY	65243125	76697500	49450000	68962500	18	19

Figure 3 - Screenshot taken from the Python Notebook

3. Methodology

The project has three parts: Exploratory analysis, Machine Learning and output results. The following pages will discuss these topics.

3.1 Exploratory Analysis

The first act in the Exploratory Analysis was to create a histogram. The goal is to understand the “Average Sale Price” distribution of the sales made in the last six years in Manhattan.

3.1.1 Histogram Analysis

Figure 04 – Histogram of Average Sale Price in Property Sales of Manhattan

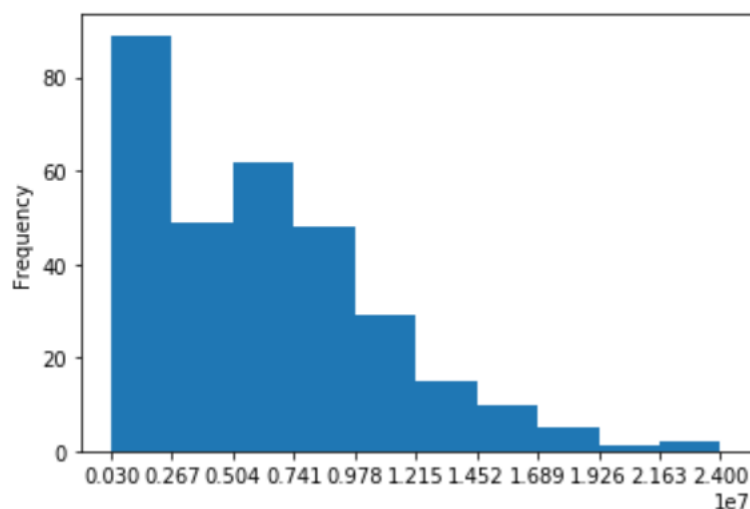


Figure 4 - Screenshot taken from the Python Notebook

In the histogram, the major number of sales come from properties between USD \$30,000 and USD \$267,000. Fact that makes sense, after all, properties with lower prices will inevitably sell more than expensive ones. In that way, we can confirm that our data is following a logical pattern, and the column “Average Sales Price” could be use in other analysis with confidence.

Note that there is other concentration of sales between USD \$504,000 and USD \$1,215,000. That is also intuitive, as Manhattan, being part of the greatest economic center in the world, with important places and well-located ones, have a higher regular property prices than many other places in U.S.A.

After USD \$1,215,000, we have the lowest number of sales. Another logical conclusion is to think that higher prices could be only accessible for people who has higher money, which are a small percentage of the general population.

The distinctions identified in the histogram will be also important when the Machine Learning step of clustering begin. The expectation is that the same type of division occurs in the cluster later.

3.1.2 Bar Chart Analysis

Figure 05 – Bar chart of Number of Sales per Neighborhood

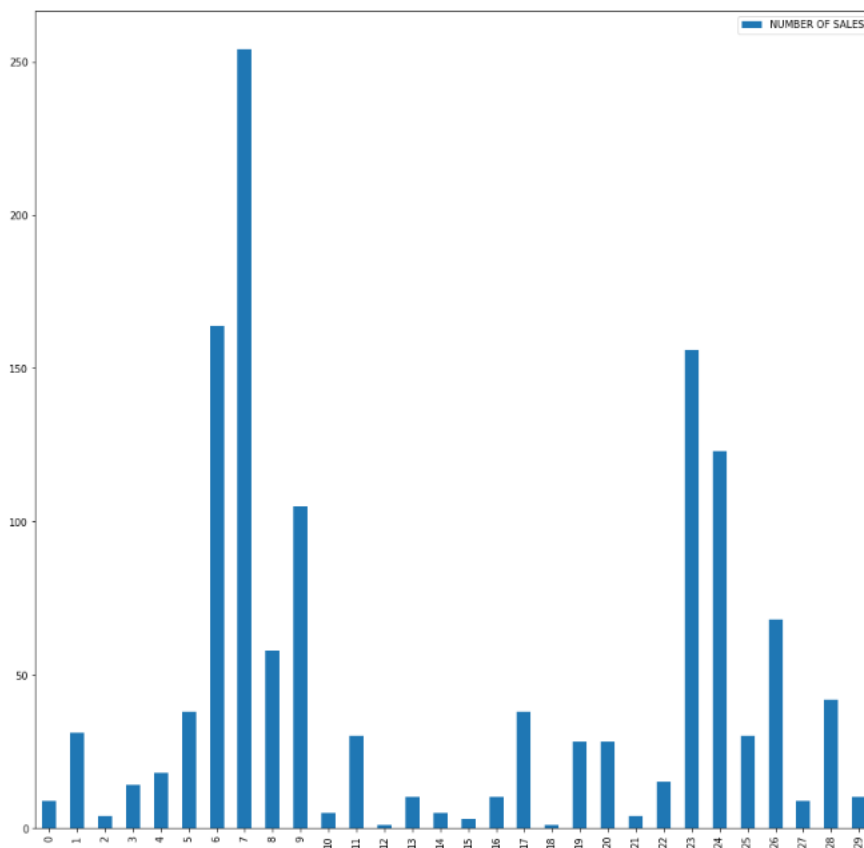
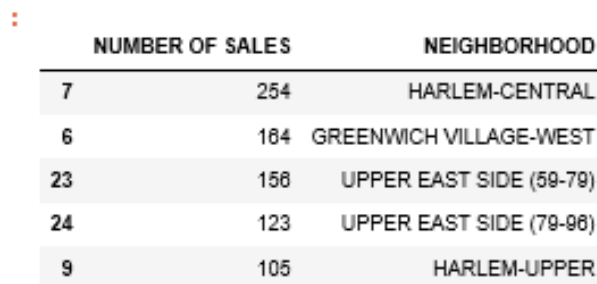


Figure 5 - Screenshot taken from the Python Notebook

As shown in the graphic below, there is a clear discrepancy between the numbers of sales of each neighborhood in Manhattan, that could be see better in the following sheets:

Figure 06 – Five Neighborhood with highest number of sales in the last six years.

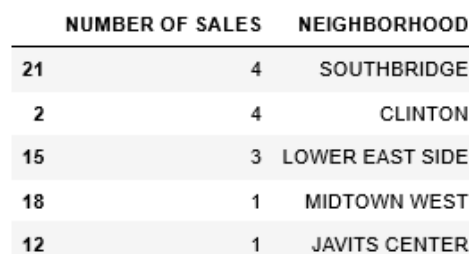


	NUMBER OF SALES	NEIGHBORHOOD
7	254	HARLEM-CENTRAL
6	184	GREENWICH VILLAGE-WEST
23	158	UPPER EAST SIDE (59-79)
24	123	UPPER EAST SIDE (79-96)
9	105	HARLEM-UPPER

Figure 6 - Screenshot taken from Python Notebook

In the list above, we could see the central Neighborhoods (Harlem – Central) and other culturally important ones (Greenwich), which is accurate because they are places that people generally want to live, and have a higher diversification of prices, to attend many people incomes.

Figure 07 – Five Neighborhood with the lowest number of sales in the last six years.



	NUMBER OF SALES	NEIGHBORHOOD
21	4	SOUTHBRIDGE
2	4	CLINTON
15	3	LOWER EAST SIDE
18	1	MIDTOWN WEST
12	1	JAVITS CENTER

Figure 7- Screenshot taken from Python Notebook

In that list otherwise, we have the commercial centers and also the neighborhood with expensive prices and less variety of prices to get more sales. That conclusion is also be observed in the next following topics of the present project.

3.1.3 Boxplot Analysis

Figure 08 – Boxplot of price per dwelling distributions of Neighborhoods

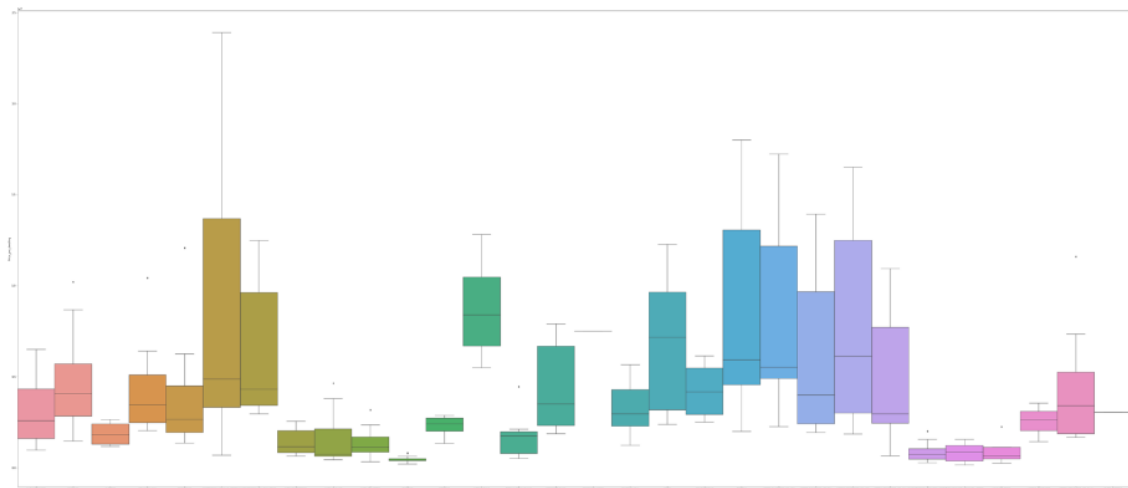


Figure 8 - Screenshot taken from Python Notebook

In the boxplot the higher sized bars represent the neighborhood that has great variety of prices per Dwelling, and the position that the bars are positioning in the “y” axis, influence in the prices, from highest to lowest.

Many neighborhoods presents variety of prices per dwelling. That is a good thing, as the NEWPLACE Company could embrace many types of costumer in many types of neighborhoods. However, defining what the best place to act is, it is not the scope of this project. This project goal is to create a geographical report of prices and locations for then the company select what it will be their strategy to operate in Manhattan.

Another fact to note is that we will have some cases where fewer sales were made for that specific neighborhood, which will alter for example, the diversity of prices per dwelling. This will not affect much our analysis, but it is a fact that is worth to mention.

If the reader wants to see more about the boxplot chart, please look at the python notebook uploaded on GitHub, but to summarize the boxplot insights:

Figure 09 – Dataframe with the neighborhood with the highest price per dwelling on Manhattan.

	NEIGHBORHOOD	Price_per_dwelling	NUMBER OF SALES_x
18	MIDTOWN WEST	7.500000e+06	1
12	JAVITS CENTER	3.050000e+06	1
14	LITTLE ITALY	2.193344e+06	5
21	SOUTHBRIDGE	1.067708e+06	4
15	LOWER EAST SIDE	7.800000e+05	3

Figure 9 - Screenshot Taken from Python Notebook

As it shown in the sheet above, most of the expensive neighborhoods per dwelling are places with low number of sales. That confirms the previous analysis of the present project, that the highest number of sales will always be the cheapest property and that low number of sales will increase the price if compared to other neighborhoods with high number of sales on the previous years.

Figure 10 – Dataframe with the neighborhood with the lowest price per dwelling on Manhattan.

	NEIGHBORHOOD	Price_per_dwelling	NUMBER OF SALES_x
7	HARLEM-CENTRAL	63351.797840	254
8	HARLEM-EAST	62416.915225	58
9	HARLEM-UPPER	58452.888889	105
28	WASHINGTON HEIGHTS LOWER	50394.216837	42
11	INWOOD	32763.963211	30

Figure 10 - Screenshot Taken from Python Notebook

The lowest price per dwelling in the city have higher number of sales, again, as expected. The interesting thing about the comparison between the higher and lowest prices for price per dwelling in Manhattan is the range difference between them. From USD \$30.000 until USD 7,000,000 per dwelling. Which shows that Manhattan has a great variety of houses and prices, and for a company like NEWPLACE operates there, is a great choice.

3.2 Machine Learning and Base Enrichment

As the present project has as its goal deliver a base with the information already defined and well described to the company target, the projects author select the K Means Clustering method of Machine Learning to cluster the information in categories to then, represent them in the final base. In this project, two clustering executions were made, and a extraction from the FourSquare API, in order to enrich the final base with the information of places and business of the neighborhoods, which will be discussed in the following topics.

3.2.1 K Means Clustering for price per dwelling

In that execution, the project used the price per dwelling as the data to cluster the information. The goal in that execution is to concatenate the Neighborhoods from the less expensive to the highest ones.

Four Clusters were created. They are:

- **Cluster 0:** Further changed in the final DataFrame to “Highest Prices Per Dwelling”

- **Cluster 1:** Further changed in the final DataFrame to “Lower Prices Per Dwelling”
- **Cluster 2:** Further changed in the final DataFrame to “Highest Prices Per Dwelling”
- **Cluster 3:** Further changed in the final DataFrame to “Regular Prices Per Dwelling”

Two cluster were named as “Highest Prices Per Dwelling”, that’s because they were pretty similar, and it will only difficult visualization if the author separate both.

Now that the clustering for price per dwelling is created, the project will move on to the next one.

3.2.2 K Means Clustering for Number of Sales

In that execution, the project used the Number of Sales of property as the data to cluster the information. The goal in that execution is to concatenate the Neighborhoods from the lowest number of sales of property to the highest ones.

Three clusters were created:

- **Cluster 0:** Further changed in the final DataFrame to “Best Selling Regions”
- **Cluster 1:** Further changed in the final DataFrame to “Worse Selling Regions”
- **Cluster 2:** Further changed in the final DataFrame to “Worse Selling Regions”

As the majority of the base of sales presents similar Number of Sales, the best way to visualize the information in the final base, is to give a binary classification, were is easier to see what is the regions that really sells it, and the ones that really do not sell that well.

3.2.3 Base enrichment: Places and Business in FourSquare API.

The goal in this part of the script is to extract information of business and places in the analyzed neighborhoods, in order to enrich the final base with all the information and characteristics of that neighborhood. As it was shown in the **Course Applied Data Science Capstone**, the present project extract all the information, summarized in the most common business and places by region. The final output is the following sheet:

Figure 11 – Dataframe with the neighborhood most common business and places.

NEIGHBORHOOD	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
LITTLE ITALY	Bubble Tea Shop	Chinese Restaurant	Bakery	Mediterranean Restaurant	Spa	Café	Thai Restaurant	Italian Restaurant	Ice Cream Shop	Hotel
LOWER EAST SIDE	Chinese Restaurant	Ramen Restaurant	Park	Cocktail Bar	Café	Art Gallery	Pet Café	Performing Arts Venue	Ice Cream Shop	Pharmacy
TRIBECA	Italian Restaurant	Park	Bakery	Café	Spa	Wine Bar	Hotel	Playground	Steakhouse	Greek Restaurant
EAST VILLAGE	Cocktail Bar	Pizza Place	Bar	Mexican Restaurant	Coffee Shop	Japanese Restaurant	Ramen Restaurant	Juice Bar	Wine Bar	Seafood Restaurant
SOHO	Italian Restaurant	Mediterranean Restaurant	Coffee Shop	Café	Sandwich Place	Gym	Clothing Store	Ice Cream Shop	French Restaurant	Spa
CLINTON	Theater	Coffee Shop	Gym / Fitness Center	Gym	Hotel	Sandwich Place	Cocktail Bar	Pizza Place	Spa	Italian Restaurant
GRAMERCY	Italian Restaurant	Coffee Shop	Pizza Place	Bar	Playground	Bagel Shop	Mexican Restaurant	Sandwich Place	Taco Place	Park
CHELSEA	Art Gallery	Coffee Shop	Café	Ice Cream Shop	American Restaurant	Market	Seafood Restaurant	Boutique	Cupcake Shop	Cycle Studio
MURRAY HILL	Hotel	Sandwich Place	Coffee Shop	Steakhouse	Pizza Place	Japanese Restaurant	Gym / Fitness Center	Chinese Restaurant	Juice Bar	Indian Restaurant
MANHATTAN VALLEY	Coffee Shop	Spa	Pizza Place	Bar	Mexican Restaurant	Grocery Store	Playground	Park	Noodle House	Latin American Restaurant
INWOOD	Mexican Restaurant	Restaurant	Café	Lounge	Pizza Place	Wine Bar	American Restaurant	Park	Bakery	Frozen Yogurt Shop

Figure 11 - Screenshot taken from Python Notebook

3.2.4 Final Output: Unifying bases

To form the final base, the script merges the first clustering (Price per Dwelling), the second clustering (Number of Sales) and the data of business and places from the FourSquare. The result, is a complete table with the information

on the prices per Dwelling, it's segmentation, the number of sales segmentation and also the most common business from that neighborhood:

Figure 12 – Final Dataframe

Price per Dwelling segmentation	NEIGHBORHOOD	Price_per_dwelling	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Selling per Region segmentation
Highest Prices Per Dwelling	LITTLE ITALY	2.193344e+06	Bubble Tea Shop	Chinese Restaurant	Bakery	Mediterranean Restaurant	Spa	Café	Thai Restaurant	Italian Restaurant	Ice Cream Shop	Hotel	Worse Selling Regions
Regular Prices Per Dwelling	LOWER EAST SIDE	7.800000e+05	Chinese Restaurant	Ramen Restaurant	Park	Cocktail Bar	Café	Art Gallery	Pet Café	Performing Arts Venue	Ice Cream Shop	Pharmacy	Best Selling Regions
Regular Prices Per Dwelling	TRIBECA	6.543844e+05	Italian Restaurant	Park	Bakery	Café	Spa	Wine Bar	Hotel	Playground	Steakhouse	Greek Restaurant	Best Selling Regions
Regular Prices Per Dwelling	EAST VILLAGE	5.265813e+05	Cocktail Bar	Pizza Place	Bar	Mexican Restaurant	Coffee Shop	Japanese Restaurant	Ramen Restaurant	Juice Bar	Wine Bar	Seafood Restaurant	Best Selling Regions
Regular Prices Per Dwelling	SOHO	4.733399e+05	Italian Restaurant	Mediterranean Restaurant	Coffee Shop	Café	Sandwich Place	Gym	Clothing Store	Ice Cream Shop	French Restaurant	Spa	Best Selling Regions
Regular Prices Per Dwelling	CLINTON	4.691848e+05	Theater	Coffee Shop	Gym / Fitness Center	Gym	Hotel	Sandwich Place	Cocktail Bar	Pizza Place	Spa	Italian Restaurant	Best Selling Regions
Regular Prices Per Dwelling	GRAMERCY	3.815387e+05	Italian Restaurant	Coffee Shop	Pizza Place	Bar	Playground	Bagel Shop	Mexican Restaurant	Sandwich Place	Taco Place	Park	Best Selling Regions
Lower Prices Per Dwelling	CHELSEA	2.682571e+05	Art Gallery	Coffee Shop	Café	Ice Cream Shop	American Restaurant	Market	Seafood Restaurant	Boutique	Cupcake Shop	Cycle Studio	Best Selling Regions
Lower Prices Per Dwelling	MURRAY HILL	2.332665e+05	Hotel	Sandwich Place	Coffee Shop	Steakhouse	Pizza Place	Japanese Restaurant	Gym / Fitness Center	Chinese Restaurant	Juice Bar	Indian Restaurant	Best Selling Regions
Lower Prices Per Dwelling	MANHATTAN VALLEY	1.976954e+05	Coffee Shop	Spa	Pizza Place	Bar	Mexican Restaurant	Grocery Store	Playground	Park	Noodle House	Latin American Restaurant	Best Selling Regions
Lower Prices Per Dwelling	INWOOD	3.276396e+04	Mexican Restaurant	Restaurant	Café	Lounge	Pizza Place	Wine Bar	American Restaurant	Park	Bakery	Frozen Yogurt Shop	Best Selling Regions

Figure 12 - Screenshot Taken from Python Notebook

4. Results

The final output of the project was a success. Now, the company “NEWPLACE” has a great base of information that contemplates most of the information that a Real Estate company needs to start its strategy in a new area. In addition, it could also use the business and places most commons in the areas to improve their offers to possible costumers. They can show for example, the types of business next to the customer new house, and the opportunities of a good and comfortable life that they have when buying it.

This script can be used to different places, not only Manhattan. Getting all this information and analysis in some seconds, for every place that a company

wishes to start operating, is really a great time saver, and many risks are mitigated in the process.

5. Conclusion

The project does not need to be used only for companies in Real Estate market. In reality, tourism companies that want to have better offers for the customers, companies that need to study a city before creating a new branch, and even any individual that wants to establish a new business, or even a new house in a certain area.

That is why Data Science is so valuable in today's world. Using data to help business and individuals is something that should (and it its) get more popular every day, and the author of the present project certifies it's satisfaction to be part on this process and help the world.