Data Science Report

Determine The most suitable location in New York City for a Spanish Restaurant

By Steven Bloom

1.0 Introduction & Problem Definition

A client wishes to determine which area in New York City they should open their new Spanish restaurant. The owner is agnostic to exactly where the restaurant should be, so is open to the restaurant being established in all 5 New York boroughs which include the following:

- Manhattan
- Bronx
- Queens
- Brooklyn
- Staten Island

The owner would like the restaurant to be an upmarket establishment; however not on the extreme upper end of the scale (Average entrée will be $25-40 USD). The owner sees other restaurants being the primary competition for this establishing, and not fast-food establishments or coffee shops.

The owner is convinced due to the celebrity chef that they have partnered with that they will be successful if the restaurant is created in a location that has a high level of foot traffic. The cost of the property lease or staff is not an issue, provided the site has a high number of well-to-do patrons living/visiting the area.

The owner does not want to compete with other Spanish restaurants in the areas. The owner is hoping that inserting a Spanish restaurant into a place that already has several dining options; will attract residents away from other establishments due to its novelty.

1.1 Interest Target Audience (Interest)

The report was done specifically for the prospective Spanish Restaurant owner; however the modelling can be applied by any prospective restaurant owner.

The suggested methodology would still hold for other clients, however the last part of the exercise would need to be adapted if a different type of restaurant were to be in scope. This study searches explicitly for other Spanish restaurants in the target area to ensure the selected Zip Code does not have any direct competition. This would need to be adjusted to suit the alternative restaurant in question if this exercise was being done for another client.

2.0 Data acquisition and cleaning

2.1 Key Data Points & Criteria

Important factors for establishing an upmarket restaurant are as follows:

- *Population*: The restaurant should be established in an area that has a high number population living the area. This will provide the restaurant with enough local foot traffic from local residents to ensure its success.

- *Safety*: The area that the restaurant should be established should be safe, so people feel safe traveling to the location. Locations where the average salary of the residents is high generally correlate to areas that have a low rate of violent crime.
- *High Average Income*: The target location should have a local populace with a high average household income. This would entail that the local population would have disposable income to enjoy the services offered by the desired establishment.
- *Foot Traffic*: Certain areas in New York such as West Village, East Village, Soho and Williamsburg are go-to restaurant destinations. This means they enjoy clients who live locally and benefit from having clients who live outside of the area coming to visit. Consequently, areas with many other restaurants are a good place to open a new restaurant as these areas are already on the radar of non-local clients.
- *General Competition*: As the point above illustrates, there are benefits of establishing a restaurant in the same region that has other restaurants. However, regions that have a high number of restaurants, however have a lower restaurant/population ratio would be preferred as it would benefit from high foot traffic and well as lower competition.
- *Direct Competition*: The client does not want to open a restaurant in an area with many other Spanish restaurants. Ideally, no direct competition would ideal; however, the client is prepared to deal with 1 or 2 competitors if other factors are it its favor.
- *Advertisement*: Setting up a restaurant in areas that have high foot traffic provides free advertising to the establishment. Once a location is selected, the restaurant should be on a street with lots of other commerce options to increase the chances of free publicity.

2.2 Data Sources

The following data sources will be used:

- Zip Codes: New York City will be broken down into Zip Codes to determine the best area to establish the restaurant. http://zipatlas.com/ provides a list of zip codes for this purpose
- Population by Zip Code: http://zipatlas.com/ provides population in each zip code. The higher the population, the better the target location.
- Average salary per household: http://zipatlas.com/ provides average salary per household. The higher the average salary, the better the target location.
- Number of Restaurants in Zip Code: FourSquare will be used to determine how many restaurants are in each area Zip Code. A 500m radius will be used as API calls to FourSquare max out at 100. When the results are delivered, only venues that have 'restaurant; in the category name will be counted. Fast Food, burger joints, bakeries, etc. will not be counted.
- Street Addresses: FourSquare will be used to determine which street would be best to establish a restaurant based on other commerce options in the target street.

2.3 Data Cleaning

Once the data is obtained, any zip code that does not have the following other data points will be discarded:

- Population numbers
- Average House Income numbers
- Other venues in zipcode

The results from FourSquare will be filtered to remove non-restaurant food locations (bakeries, coffee shops, etc) and fast food venues are these are not relevant to the exercise.

When the Zip Code has been selected, and a target street is within the area is being located, several string manipulations are required to ensure uniformity for text for the final frequency chart.

3.0 Methodology

3.1. Machine Learning: K-Means Clustering

Once the above data points have been collected and the data has been cleaned, K-means clustering will be used as a form of machine learning to determine which group of Zip Codes will be most suitable for the target restaurant. K-means clustering is being used to group the data points into clusters to identify which Zip Codes have similar factors when it comes to population, income and commercial similarities. We are required to normalize the data set so that a single variable does not influence the clustering more than other variables.

K-means clustering is used to detect patterns in the data that take in a variety of variables and locate the best trade-off between equally important components.

The clusters will be manually analyzed to determine the most suitable group. The data will be split into 12 clusters to provide the necessary granularity required for locating prime zip code locations. The cluster selected will have the highest values in population, average income and restaurant frequency.

3.2 Zip Code Selection in Cluster

Once a desirable cluster is selected, the selected Zip Code will be the one within the cluster that has the lowest [Restaurant] / [Population] ratio. A lower ratio represents less competition as the local population are underserved by dining establishment compared to other areas in the cluster.

Once the target Zip Code is selected, FourSquare will be used to determine that there are no other Spanish restaurants within a 500m radius of the area. If a Spanish restaurant is in the Zip Code, then the next most suitable Zip Code in the cluster will be targeted. This process will be repeated until a Zip Code without a Spanish restaurant is found. If all Zip Codes have Spanish restaurants in a 500m radius, then the Zip Code that has the lowest number of Spanish restaurants will be selected.

3.3 Street Selection

Once the Zip Code has been selected to determine the best street within the Zip Code to establish the restaurant; a frequency chart will be created of each street in the zip code listing the number of FourSqaure ‘Drinks’, ‘Coffee’ and ‘Shops’ venues on each street. The street with the highest number of commerce on it will be selected to be the best place to establish a new Spanish restaurant.

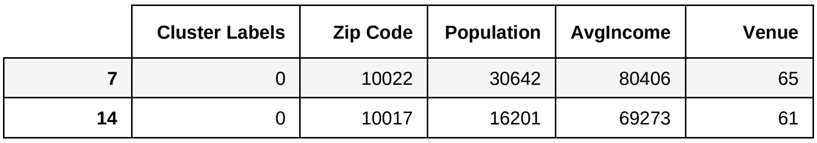
This is to ensure the new restaurant benefits from the foot traffic from the street as well as free advertising.

4.0 Results

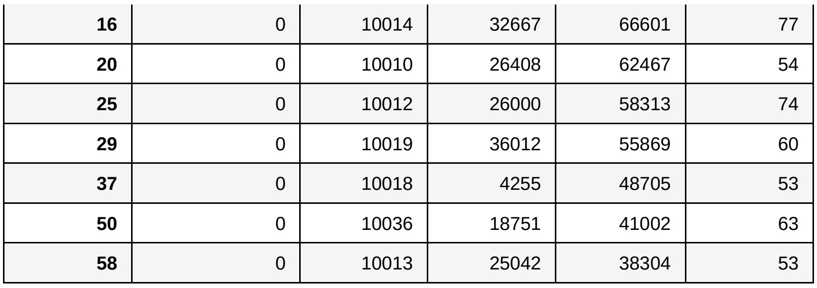
4.1Cluster Analysis - Results

Below are the 12 Clusters with my manual analysis of the grouping:

Cluster 1 - Low Population, Med-High Income, High Venue Count

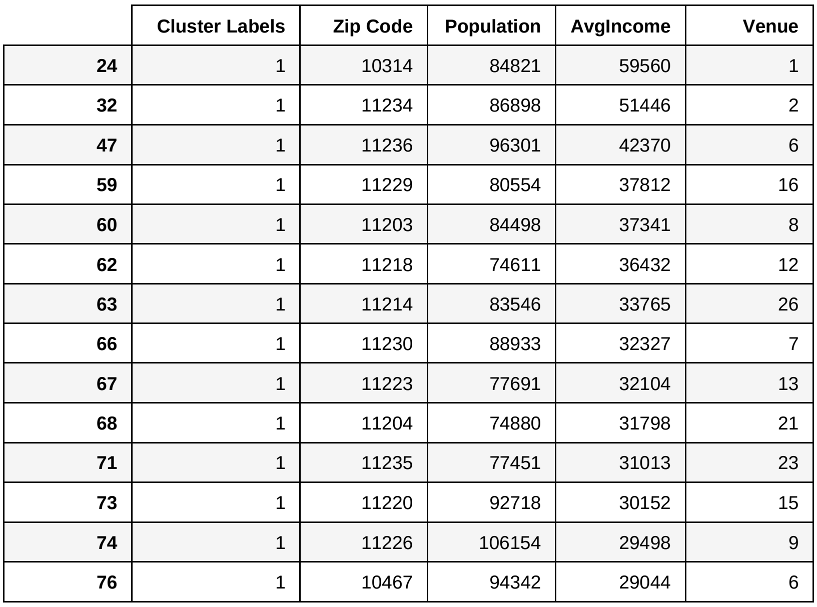


	Cluster Labels	Zip Code	Population	AvgIncome	Venue
7	0	10022	30642	80406	65
14	0	10017	16201	69273	61



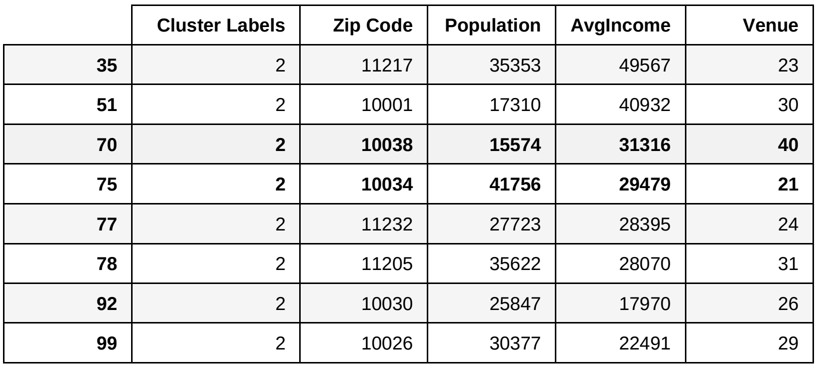
16	0	10014	32667	66601	77
20	0	10010	26408	62467	54
25	0	10012	26000	58313	74
29	0	10019	36012	55869	60
37	0	10018	4255	48705	53
50	0	10036	18751	41002	63
58	0	10013	25042	38304	53

Cluster 2 - High Population, Medium Income, Low Venue Count



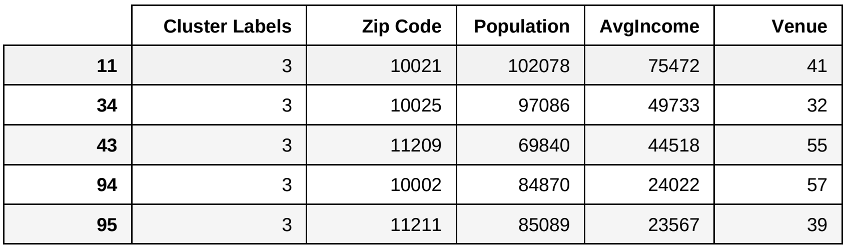
	Cluster Labels	Zip Code	Population	AvgIncome	Venue
24	1	10314	84821	59560	1
32	1	11234	86898	51446	2
47	1	11236	96301	42370	6
59	1	11229	80554	37812	16
60	1	11203	84498	37341	8
62	1	11218	74611	36432	12
63	1	11214	83546	33765	26
66	1	11230	88933	32327	7
67	1	11223	77691	32104	13
68	1	11204	74880	31798	21
71	1	11235	77451	31013	23
73	1	11220	92718	30152	15
74	1	11226	106154	29498	9
76	1	10467	94342	29044	6

Cluster 3 – Low-Medium Population, Low Income, Medium Venue Count



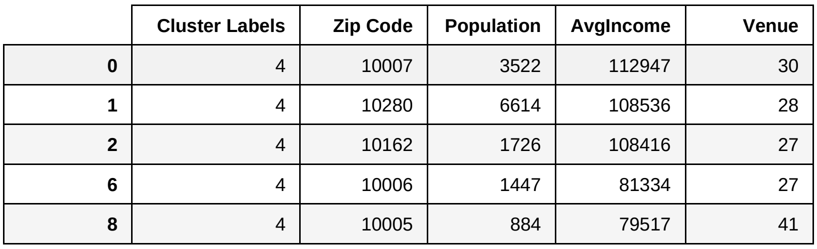
	Cluster Labels	Zip Code	Population	AvgIncome	Venue
35	2	11217	35353	49567	23
51	2	10001	17310	40932	30
70	**2**	**10038**	**15574**	**31316**	**40**
75	**2**	**10034**	**41756**	**29479**	**21**
77	2	11232	27723	28395	24
78	2	11205	35622	28070	31
92	2	10030	25847	17970	26
99	2	10026	30377	22491	29

Cluster 4 – High Population, Low - High Income, Medium Venue Count



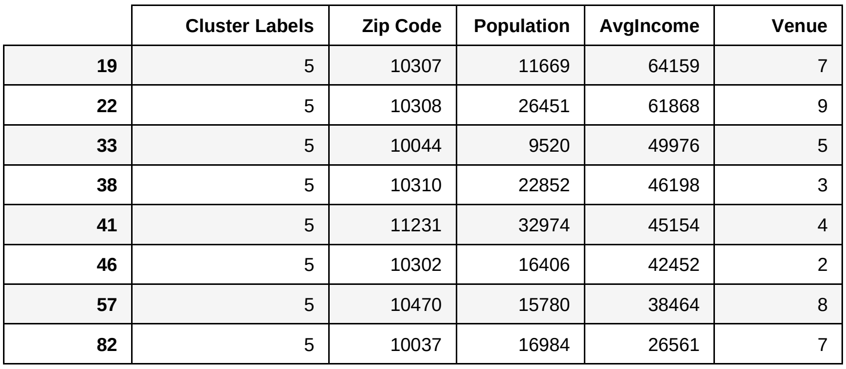
	Cluster Labels	Zip Code	Population	AvgIncome	Venue
11	3	10021	102078	75472	41
34	3	10025	97086	49733	32
43	3	11209	69840	44518	55
94	3	10002	84870	24022	57
95	3	11211	85089	23567	39

Cluster 5 - Low Population, High Income, Medium Venue Count



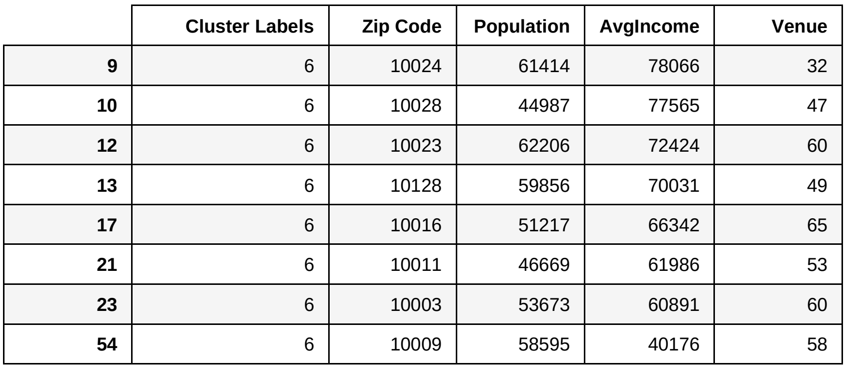
	Cluster Labels	Zip Code	Population	AvgIncome	Venue
0	4	10007	3522	112947	30
1	4	10280	6614	108536	28
2	4	10162	1726	108416	27
6	4	10006	1447	81334	27
8	4	10005	884	79517	41

Cluster 6 – Low Population, Medium Income, Low Venue Count



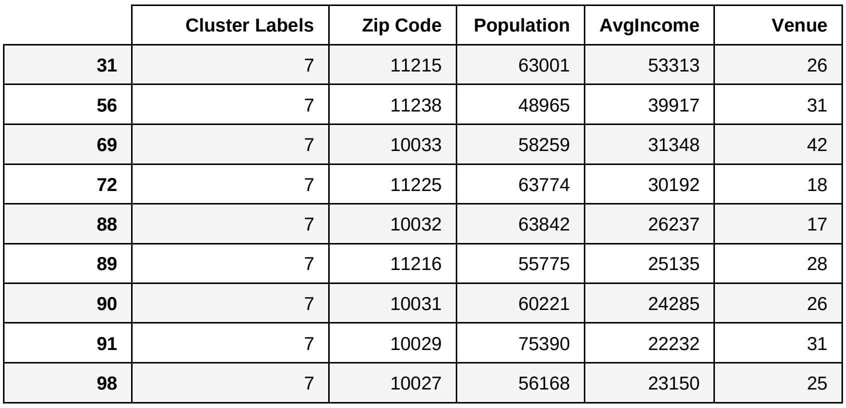
	Cluster Labels	Zip Code	Population	AvgIncome	Venue
19	5	10307	11669	64159	7
22	5	10308	26451	61868	9
33	5	10044	9520	49976	5
38	5	10310	22852	46198	3
41	5	11231	32974	45154	4
46	5	10302	16406	42452	2
57	5	10470	15780	38464	8
82	5	10037	16984	26561	7

Cluster 7 - Medium Population, High Income, High Venue Count



	Cluster Labels	Zip Code	Population	AvgIncome	Venue
9	6	10024	61414	78066	32
10	6	10028	44987	77565	47
12	6	10023	62206	72424	60
13	6	10128	59856	70031	49
17	6	10016	51217	66342	65
21	6	10011	46669	61986	53
23	6	10003	53673	60891	60
54	6	10009	58595	40176	58

Cluster 8 - Medium Population, Low-Medium Income, Medium Venue Count



	Cluster Labels	Zip Code	Population	AvgIncome	Venue
31	7	11215	63001	53313	26
56	7	11238	48965	39917	31
69	7	10033	58259	31348	42
72	7	11225	63774	30192	18
88	7	10032	63842	26237	17
89	7	11216	55775	25135	28
90	7	10031	60221	24285	26
91	7	10029	75390	22232	31
98	7	10027	56168	23150	25

100	7	11237	48910	23104	29

Cluster 9 - Low-Medium Population, Low Income, Low Venue Count

	Cluster Labels	Zip Code	Population	AvgIncome	Venue
65	8	11222	39360	33578	9
79	8	10040	46599	27905	14
83	8	10039	21737	17370	11
87	8	11224	51205	21281	3
101	8	11239	14620	16919	3
105	8	10460	53707	19517	2
106	8	10459	39080	17498	3
109	8	10454	34976	14271	12
110	8	10451	40961	20307	7
111	8	10455	37465	19389	6
115	8	10474	11354	16339	3
116	8	10035	32702	14896	9

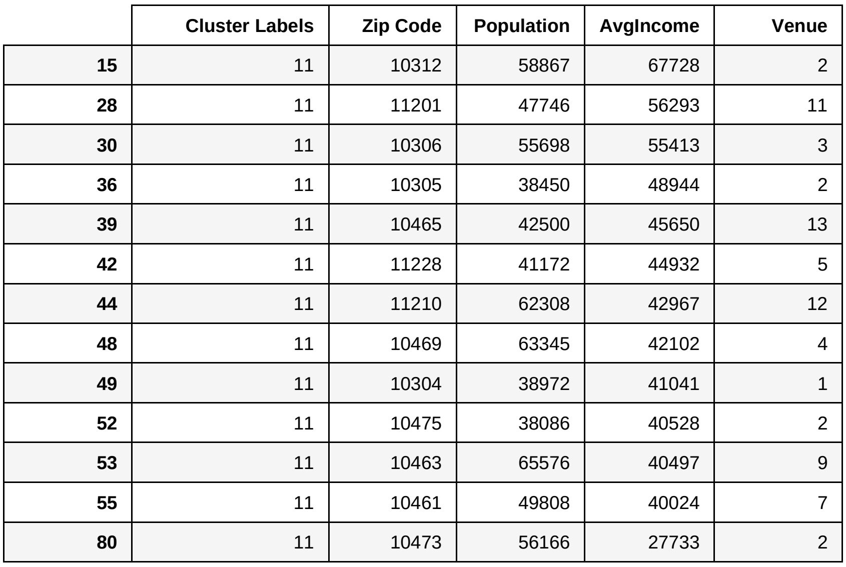
Cluster 10 - Medium-High Population, Low Income, Low Venue Count

	Cluster Labels	Zip Code	Population	AvgIncome	Venue
61	9	10466	68787	37141	3
64	9	10462	72077	33735	8
81	9	11208	87165	27078	5
84	9	10468	78309	26852	11
85	9	11219	85987	26648	3
86	9	11213	65440	26366	12
93	9	11207	86551	24163	5
96	9	11212	85161	20839	7
97	9	10472	64867	23565	7
102	9	11233	61955	22754	4
103	9	11221	76363	22305	7
104	9	10458	77840	22072	5
107	9	10453	76775	21109	16
108	9	10452	72138	20606	9
112	9	10457	69048	19233	5
113	9	10456	76656	16664	5
114	9	11206	69032	18661	15

Cluster 11 - Low Population, High Income, Low Venue Count

	Cluster Labels	Zip Code	Population	AvgIncome	Venue
3	10	10004	1225	101868	4
4	10	10069	1403	88091	3
5	10	10282	1574	86066	4

Cluster 12 - Medium Population, Medium Income, Low Venue Count

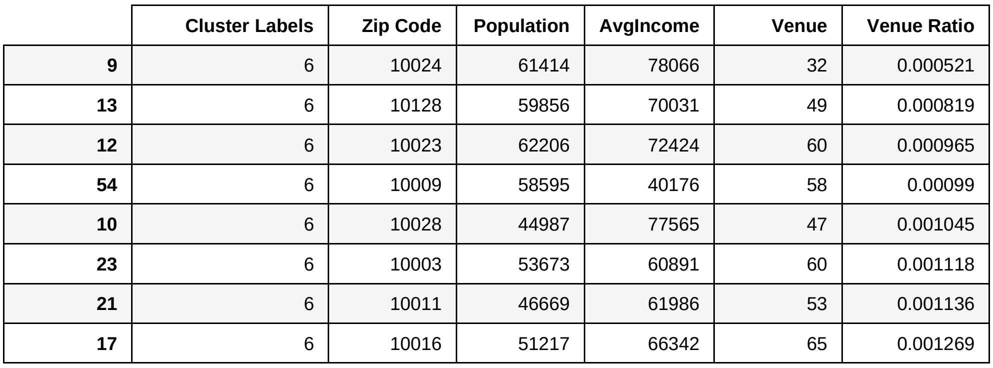


	Cluster Labels	Zip Code	Population	AvgIncome	Venue
15	11	10312	58867	67728	2
28	11	11201	47746	56293	11
30	11	10306	55698	55413	3
36	11	10305	38450	48944	2
39	11	10465	42500	45650	13
42	11	11228	41172	44932	5
44	11	11210	62308	42967	12
48	11	10469	63345	42102	4
49	11	10304	38972	41041	1
52	11	10475	38086	40528	2
53	11	10463	65576	40497	9
55	11	10461	49808	40024	7
80	11	10473	56166	27733	2

From the above data, Cluster 7 is most attractive as it has 'High' values in 2 categories, and 'Medium' in population.

4.2 Population To Restaurant Ration Results – Cluster 7

For Cluster 7 if we are to perform the following to determine the [Restaurant] / [Population] ratio



	Cluster Labels	Zip Code	Population	AvgIncome	Venue	Venue Ratio
9	6	10024	61414	78066	32	0.000521
13	6	10128	59856	70031	49	0.000819
12	6	10023	62206	72424	60	0.000965
54	6	10009	58595	40176	58	0.00099
10	6	10028	44987	77565	47	0.001045
23	6	10003	53673	60891	60	0.001118
21	6	10011	46669	61986	53	0.001136
17	6	10016	51217	66342	65	0.001269

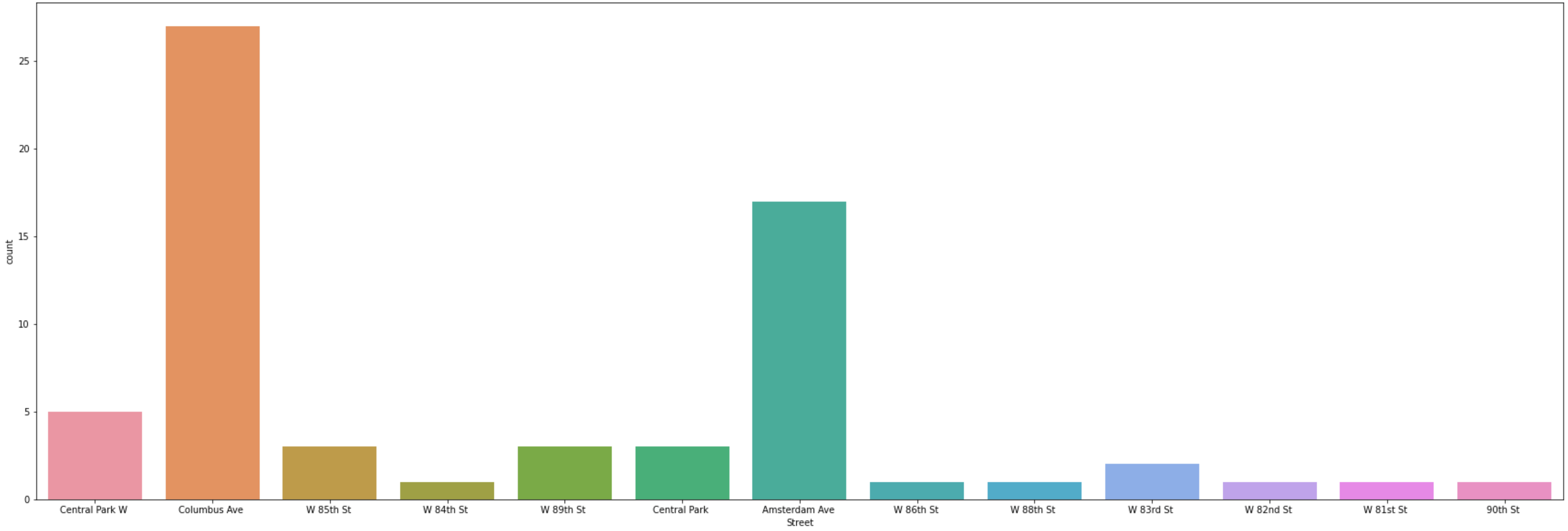
Zip Code 10024 has the lowest ratio, so it would be the best place to establish a restaurant based on how well serviced the local population is with existing dining establishments.

Further analysis of restaurants in Zip Code 10024 reveals no other competing Spanish restaurants in this zip code – which suits the client's requirement regarding direct competition.

Zip Code 10024 is in the Upper West Side of Manhattan, an upmarket area of New York that caters to upper-class families. This information provides confidence that our methodology was successful.

4.3 Most Suitable Street – Zip Code 10024

The following is a frequency street chart of commerce in Zip Code 10024

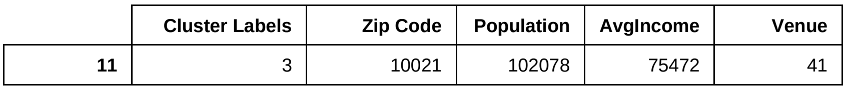


Therefore Columbus Ave is the most suitable street in Zip Code 10024 to benefit from high levels of foot traffic.

5.0 Results Discussion

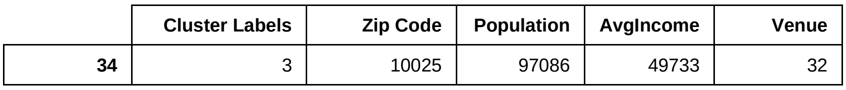
Zip Code 10024 is one of the most exclusive Zip Codes in New York in an excellent school district. There many restaurants in this area, which means it has high foot traffic, however does not have any other Spanish restaurants to directly compete with the client's project.

Through manually examining the clusters, 1 additional Zip Code stood out, which was the following:



	Cluster Labels	Zip Code	Population	AvgIncome	Venue
11	3	10021	102078	75472	41

The very high population and average household income data points would also make this Zip Code attractive. However, although this exercise was run several times, Zip code 10021 was always clustered with a group that had very high local populations but considerably lower incomes compared to the selected cluster. Consequently, whilst this particular Zip Code was suitable, the cluster it was in would not have been appropriate to do the additional analysis suggested by the methodology. The below Zip Code in the same cluster as 10021 would have produced a lower [Restaurant] / [Population] ratio so it would have been selected instead of 10021. The average income of this Zip Code is almost half of Zip Code 10024 so it would not have been a suitable pick.



	Cluster Labels	Zip Code	Population	AvgIncome	Venue
34	3	10025	97086	49733	32

Therefore according to our selected methodology Zip Code 10024 is still the recommended area to establish a Spanish Restaurant. Columbus Avenue is the best street within Zip Code 10024 for this project.

6.0 Results Discussion

New York Zip Code 10024 in the Upper West Side is the most suitable area to establish a new Spanish Restaurant based on the following factors:

- Population
- Average Income
- Foot Traffic
- Competition (No other Spanish Restaurant in area)
- Client Requirements

Columbus Avenue would be the suggested street that the client looks to find a lease for their business.