Applied Data Science Capstone

An analysis of arcades in New York City

A report by:

Scott Wiley

Introduction

Arcades have been on the decline in the United States for some time now. As home video consoles become more powerful, and with the advent and continued improvement of online multiplayer, one can now play high quality games with friends while never leaving their own chair. While this has for the most part been seen as a positive thing for the industry as a whole, it does beg the question: What position do arcades have in the modern technological world?

For people who enjoy retro gaming and still enjoy making an occasion of going somewhere and spending some quarters in order to play games, arcades still hold a valuable, but gradually disappearing position. This project aims to reveal where in New York City the arcades are, and perhaps most importantly, what sorts of things arcades are doing in order to survive.

I chose New York City since it's a fairly populous area, so I thought it might be a good barometer of sorts for seeing a sample of the arcade scene in a large city. It also, admittedly, is a place where I already possessed the geospatial and neighborhood data, so I could spend more time on the analysis approach and figuring out what I wanted to say and do rather than spending a lot more time trying to consolidate and gather the data while relying on somewhat unreliable Python libraries.

Business Problem/Target Audience

As stated before, the objective of this project is to see where specifically in New York City arcades are surviving. This may be useful for people who are looking to open an arcade in the area and want to know where to go to take part in an existing arcade culture, but maybe want to avoid so much of the competition. This project may also be useful for those who are visiting New York City and want to know where to go in order to see the greatest amount and variety of arcades during their stay.

In a more broad sense, this project is meant for people who love arcades and perhaps are just as worried as I am that they are disappearing. Hopefully, we can gain some insights into the distriction of arcades and what they've become and evolved into, so to speak.

Data

The data I will be using to answer the aforementioned questions is as follows:

- New York City location data, including boroughs, neighborhoods, as well as the latitudes and longitudes of each. This data was provided by the Cognitive Class skills network for a previous lab, though I believed it to be accurate enough to be able to be used here as well.
- Venue data drom the Foursquare API, specifically the names and Ids of venues, their latitudes and longitudes, and even their categories. While it may seem silly to pull the categories since we're only going to be looking at arcade locations, keep in mind that one method people have used to help arcades survive is to combine them with another type of business, and I want to see what types of businesses people are combining.

Location data and venue data will be mapped using Folium, and there will be some data that is graphed using matplotlib. Sklearn will be used when we cluster neighborhoods according to their venues.

Methodology

I started this process, naturally, by downloading the new york data from the IBM Developer Skills Network. I then used the JSON library in order to turn that JSON into something we can actually use for our purposes. I then created a Dataframe using Pandas in order to load the data into something claner and more legible. This Dataframe included columns for Borough, Neighborhood, Latitude, and Longitude.

```
[5]: neighborhoods.head()
```

[5]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
[6]: neighborhoods.shape[0]
```

```
[6]: 306
```

Illustration 1: As you can see, we have our location data, and it contains 306 rows corresponding to the 306 neighborhoods

I then leveraged the FourSquare API in order to get venue data for every neighborhood, however, I did not pull every venue. FourSquare organizes venues into different categories, and so what I did was use the category ID corresponding to arcades in order to pull only arcades from the area.

```
[11]: arcade_venues.head()
```

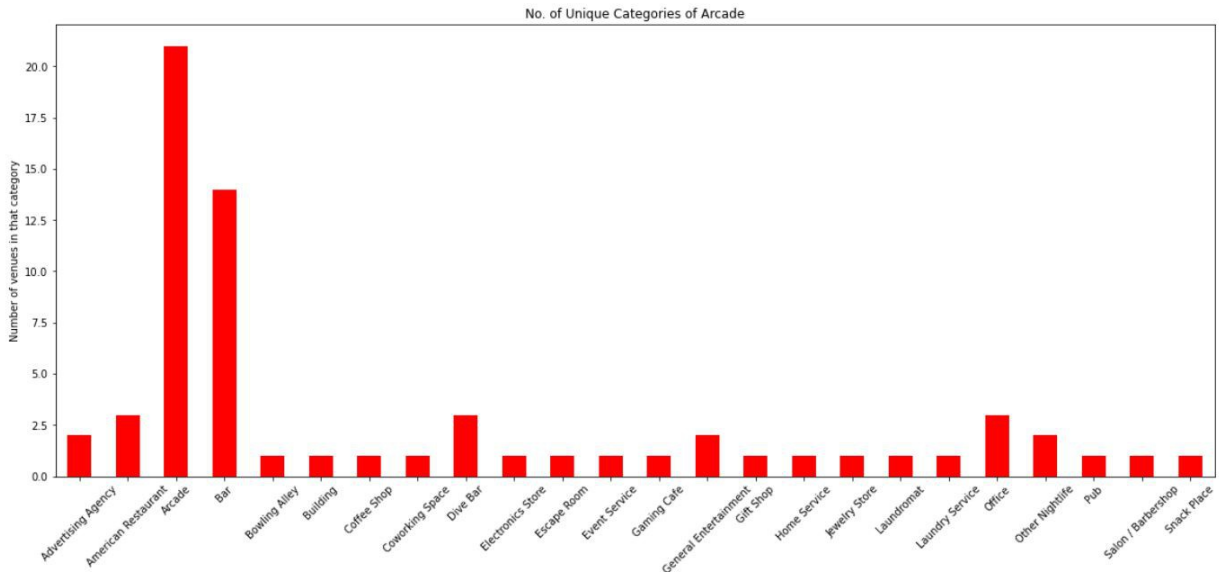
[11]:

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Baychester	40.866858	-73.835798	Chuck E. Cheese	4b6f201bf964a520b8de2ce3	40.863216	-73.833268	Arcade
1	Bronx	West Farms	40.839475	-73.877745	River Park	4b006a75f964a520f43d22e3	40.843062	-73.877520	Arcade
2	Brooklyn	Greenpoint	40.730201	-73.954241	Sunshine Laundry & Pinball Emporium	4c018f2e716bc9b6c319bc55	40.729318	-73.953564	Laundry Service
3	Brooklyn	Greenpoint	40.730201	-73.954241	Black Rabbit	46869096f964a52048481fe3	40.730057	-73.956588	Bar
4	Brooklyn	Greenpoint	40.730201	-73.954241	Farnum & Son Interactive	4d61a796ef378cfac5a68da6	40.733700	-73.953003	Arcade

Illustration 2: Our venues and what neighborhoods they are in.

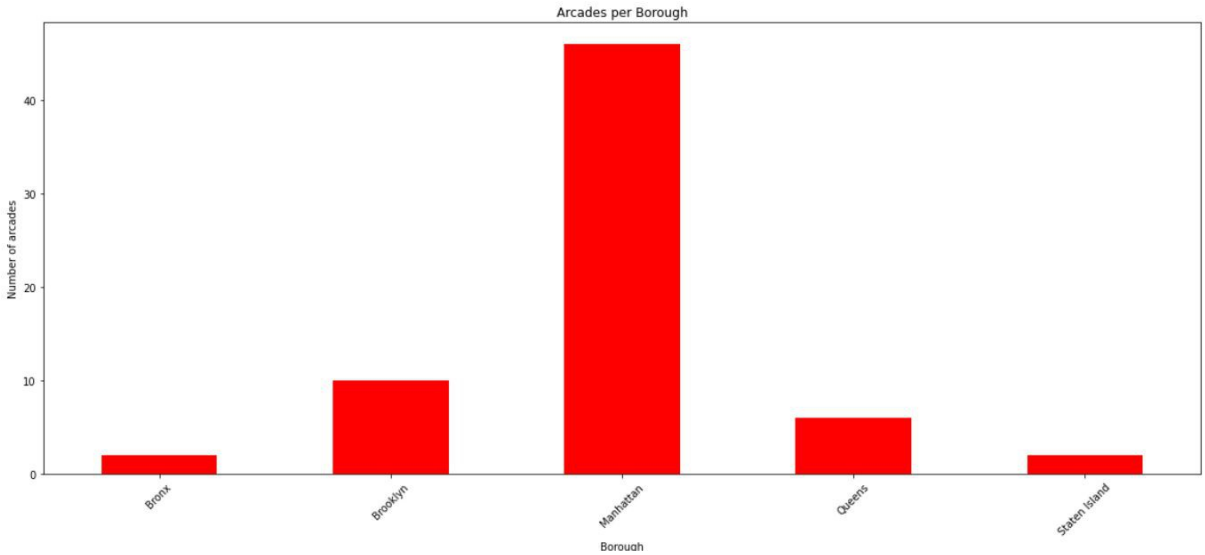
I'd like to draw your attention towards the venue category column. Notice how while most of these are labelled as arcades, some of them are labelled as bars and laundry services? These constitude a location that is a combination of both an arcade and some other type of venue.

Once all this data had been gotten, I decided to create a bar graph in order to demonstrate how many of these arcades had combined with another type of venue, as well as see just how many different types of venues arcades had been combined with.

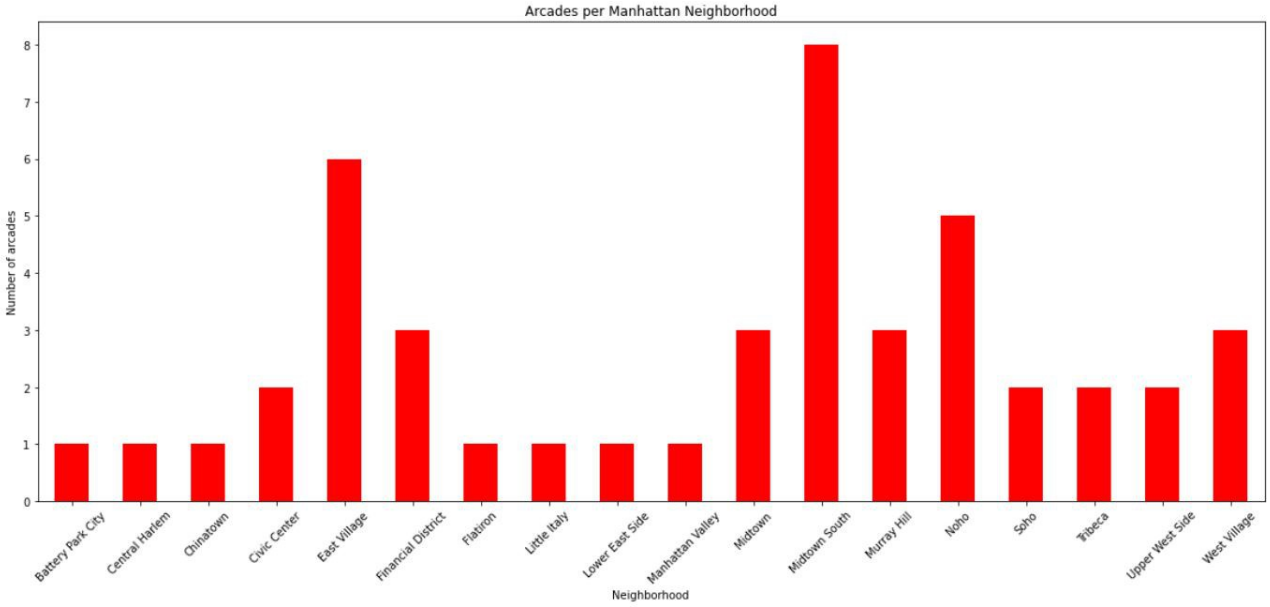


It's actually quite a variety! While pure arcades are still in the highest amount, bars follow fairly closely behind. Besides those two, arcades have been combined with a variety of other venues as well, including laundry services and American Restaurants.

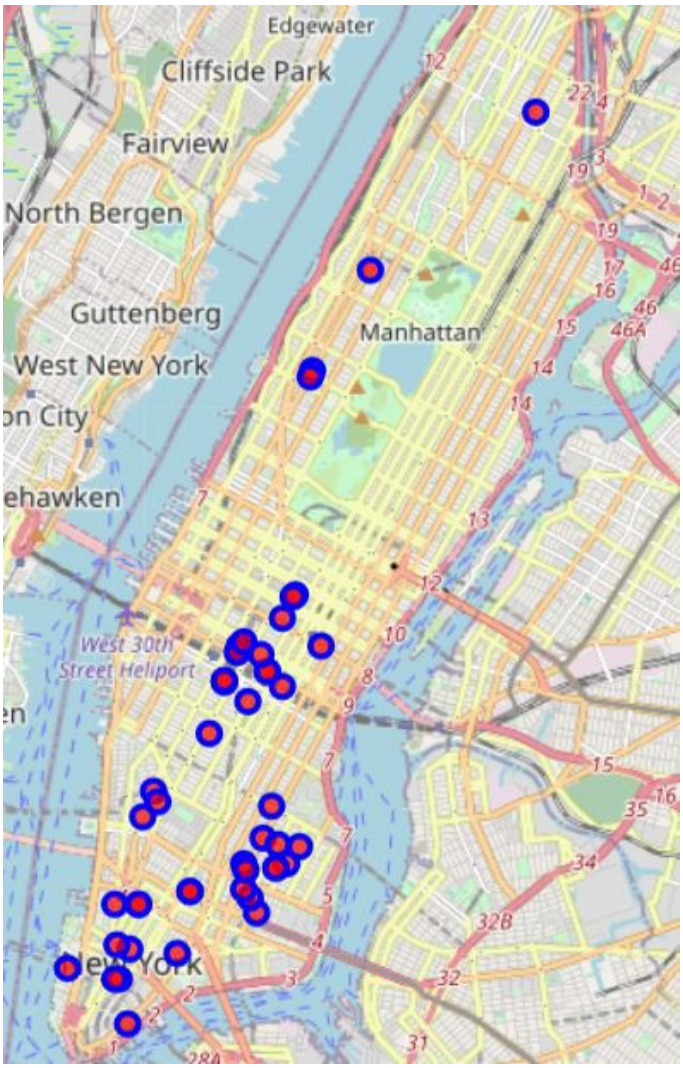
Once this was done, I wanted to see which borough had the most arcades in it, so I created a bar graph for that as well.



As you can see, Manhattan is clearly the winner here, having over 40 arcade venues while the others stay around the 10 mark. From here, I was curious as to where in Manhattan these venues were, if they stayed concentrated in one neighborhood or if they were spread out evenly. To this effect I created another, you guessed it, bar graph.

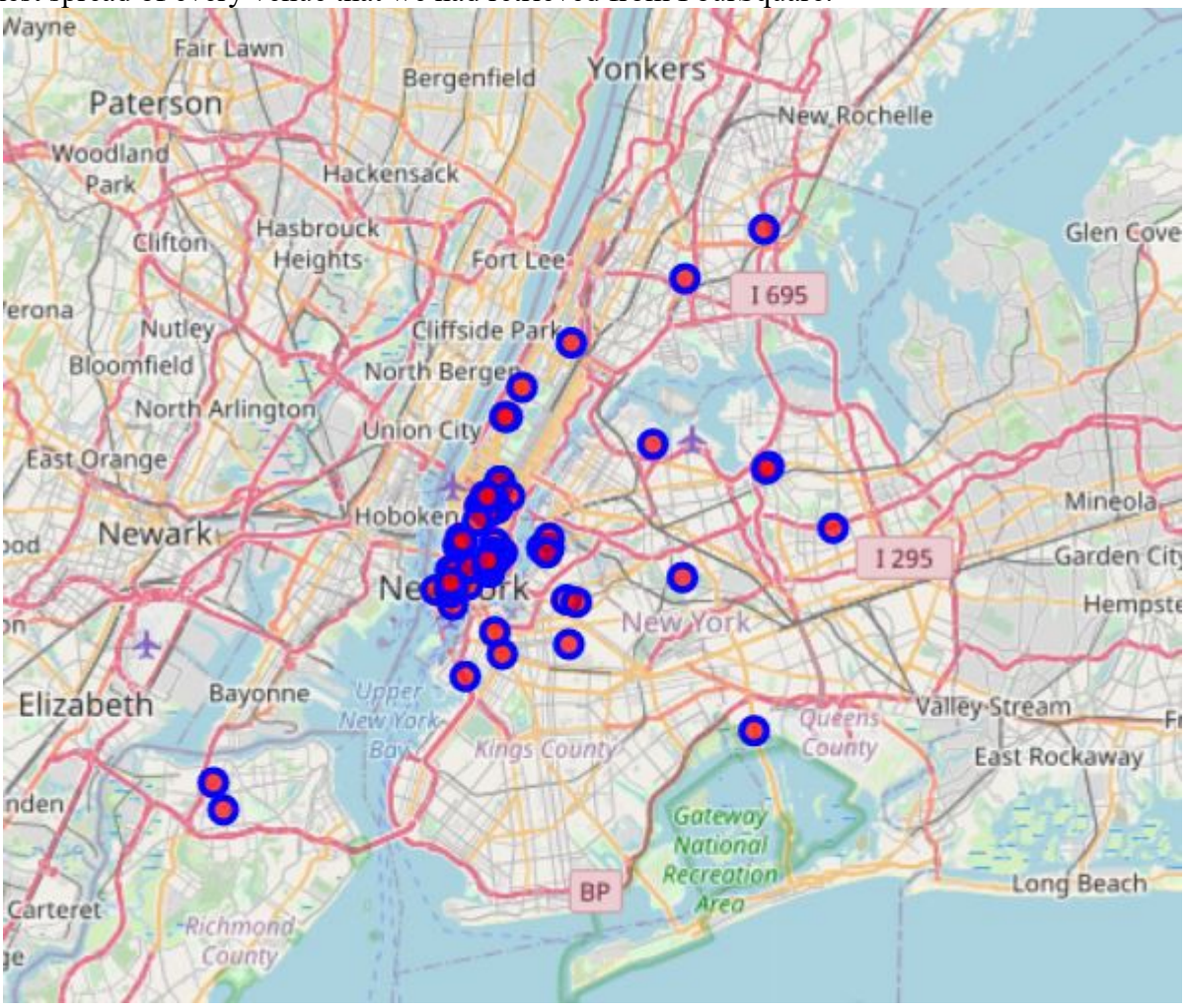


It seemed Midtown South had the most arcade venues in it, having 8 in its area. After that was East Village at 6, and Noho at 5. The rest only had 1-3 arcade venues in them. To demonstrate this point further, I created a map of Manhattan using Folium in order to see where in physical space these arcades were.



Most of our Manhattan venues are clustered in the southern area, with a few strays up by Central Park.

After this, I decided to simply map the entire list of venues I had from all over New York City. I started with Manhattan simply because it had the most venues, but now I wanted to see a more honest spread of every venue that we had retrieved from FourSquare.



Even with all the arcades recorded by FourSquare, we can see that there's more of a cluster in the Manhattan area, while ther others are far more spread out across the city.

Once this was done, I dediced to begin the process of forming clusters of neighborhoods using Onehot encoding and k-means clustering. Through the onehot encoding, I was able to see the frequency of the top 5 of certain types of venues in each neighborhood. Most of our neighborhoods only had a single arcade in it, while a select few had many different types. I will share a few here just to demonstrate what I'm talking about.

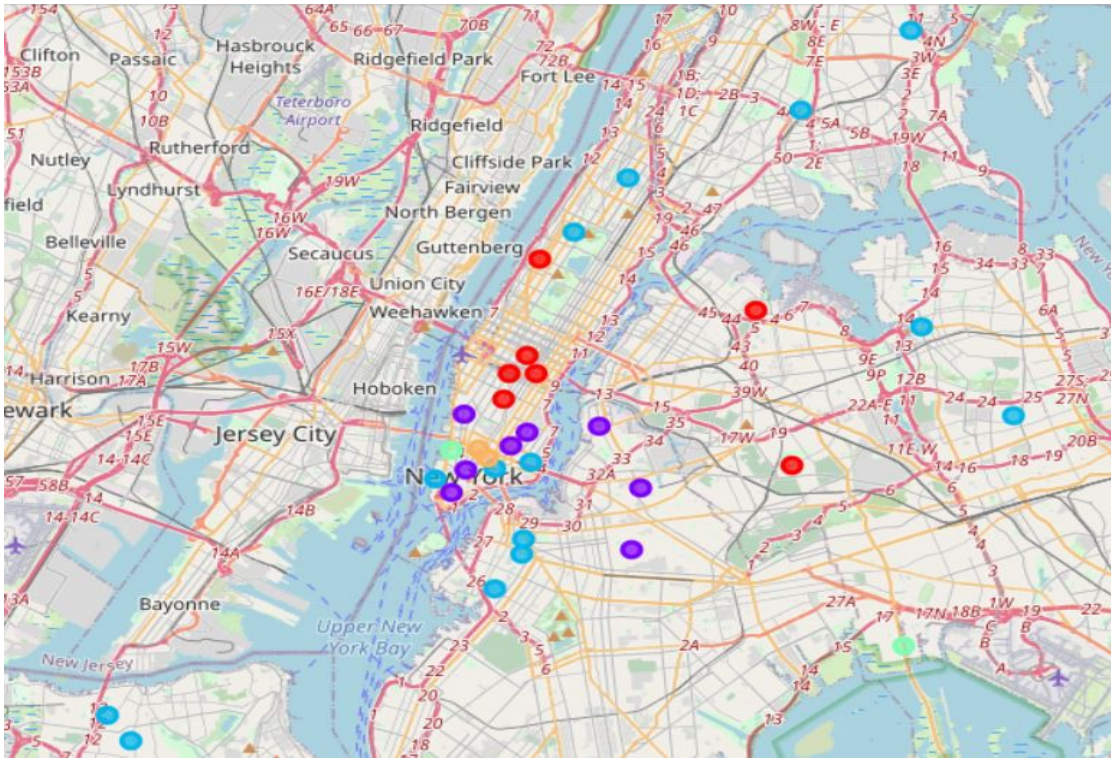
```
----Astoria Heights----
                   venue  freq
0          Bowling Alley   1.0
1     Advertising Agency   0.0
2  General Entertainment   0.0
3     Salon / Barbershop   0.0
4                    Pub   0.0

 ----East Village----
                  venue  freq
 0                  Bar  0.50
 1            Dive Bar  0.33
 2               Arcade  0.17
 3  Advertising Agency  0.00
 4           Gift Shop  0.00
```

Astoria Heights, as you can see, contains a single bowling alley as its sole source of an arcade venue. East Village, meanwhile, contains a few bars, some dive bars, and pure arcades.

Once this was done, I turned these results into a pandas dataframe, which admittedly, was not as much help as I might have hoped, but I'll talk about that in the Discussions section later in this report.

It was here that I assigned the cluster labels, organizing the data into five different clusters arbitrarily. Once again, a decision I will talk about in the discussion section. From these formed clusters, I created a folium map to show where they were and perhaps give insight into how they were clustered.



Finally, I examined the contents of each cluster to see how the algorithm decided to cluster our venues.

Results

Overall, I found the results of the project to be quite satisfying. We now know that southern Manhattan has the most arcades in it, and that while pure arcades are still the most prevalant type of arcade, barcades follow closely behind and beyond that, there's a variety of different combination arcades that exist.

I would now like to go over the results of our clustering, which provide some interesting insights into the data, as well as some issues which I will go over in the Discussion section.

```
arcades_merged.loc[arcades_merged['Cluster Labels'] == 0, arcades_merged.columns[[1] + list(range(5, arcades_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Upper West Side	Snack Place	Dive Bar	Salon / Barbershop	American Restaurant	Arcade
11	Midtown	Jewelry Store	Coffee Shop	Salon / Barbershop	Snack Place	Escape Room
12	Murray Hill	Arcade	Building	General Entertainment	Snack Place	Escape Room
27	Midtown South	Office	Event Service	Electronics Store	American Restaurant	Bar
30	Astoria Heights	Bowling Alley	Snack Place	Salon / Barbershop	American Restaurant	Arcade
31	Flatiron	Escape Room	Snack Place	Salon / Barbershop	American Restaurant	Arcade
32	Middle Village	Home Service	Snack Place	Escape Room	American Restaurant	Arcade

The first cluster was a bit all over the place. I quickly determined that Snack Place to Barbershop to American Restaurant to Arcade was the default order the algorithm ut things in when there was only a single venue in a neighborhood. Cluster one seemed to simply be a variety cluster.

```
arcades_merged.loc[arcades_merged['Cluster Labels'] == 1, arcades_merged.columns[[1] + list(range(5, arcades_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Greenpoint	Bar	Arcade	Laundry Service	Snack Place	Escape Room
3	Bedford Stuyvesant	Bar	Snack Place	Salon / Barbershop	American Restaurant	Arcade
7	East Williamsburg	Office	Bar	Snack Place	Escape Room	American Restaurant
13	East Village	Bar	Dive Bar	Arcade	Snack Place	Salon / Barbershop
18	West Village	Bar	Pub	Snack Place	Escape Room	American Restaurant
21	Financial District	Bar	Laundromat	Coworking Space	Snack Place	Escape Room
25	Noho	Arcade	Bar	Gaming Cafe	Snack Place	Escape Room
26	Civic Center	Arcade	Bar	Snack Place	Salon / Barbershop	American Restaurant

Cluster two, on the other hand, was much clearer. Cluster two seems to be where most of the barcades ended up. Once again, Snack Place to Barbershop or Snack Place to Escape Room seemed to be used as default values when there was a zero in that section for the neighborhood. This is better seen in the notebook itself.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Baychester	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
1	West Farms	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
4	Gowanus	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
5	Downtown	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
6	Boerum Hill	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
8	Chinatown	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
9	Central Harlem	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
14	Lower East Side	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
19	Manhattan Valley	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
20	Battery Park City	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
23	Flushing	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
24	Westerleigh	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
28	Elm Park	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar
29	Utopia	Arcade	Snack Place	Salon / Barbershop	American Restaurant	Bar

Once again, it's clear what's happening here. All of these places each have one arcade venue in them. This is easily demonstrated by taking a look back at our frequency data.

```
----Utopia----
                   venue  freq
0                 Arcade   1.0
1    Advertising Agency   0.0
2  General Entertainment   0.0
3    Salon / Barbershop   0.0
4                   Pub   0.0


----West Farms----
                   venue  freq
0                 Arcade   1.0
1    Advertising Agency   0.0
2  General Entertainment   0.0
3    Salon / Barbershop   0.0
4                   Pub   0.0
```

Both Utopia and West Farms, listed in the table above, each only have one arcade in them. The rest are default values for zero.

The last two clusters are very small, and both come to a similar conclusion, and so I will be inserting them both and then explaining what's happening afterward.

```
arcades_merged.loc[arcades_merged['Cluster Labels'] == 3, arcades_merged.columns[[1] + list(range(5, arcades_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	Tribeca	Advertising Agency	American Restaurant	Salon / Barbershop	Arcade	Bar
22	Howard Beach	American Restaurant	Snack Place	Salon / Barbershop	Arcade	Bar

```
arcades_merged.loc[arcades_merged['Cluster Labels'] == 4, arcades_merged.columns[[1] + list(range(5, arcades_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
16	Little Italy	Other Nightlife	Snack Place	Escape Room	American Restaurant	Arcade
17	Soho	Advertising Agency	Other Nightlife	Escape Room	American Restaurant	Arcade

These two clusters seem to be formed based off one key venue contained within each. The top one has American Restaurants, while the bottom both have Other Nightlife venues contained within them.

Discussion

Let's start off with our basic questions.

- Where should I open an arcade in New York City?
- What type of arcade should it be?
- If I were in New York on a trip, where should I go to hit the most variety of arcades?

The answer to the first is a bit tricky. Manhattan is the clear winner when it comes to number of arcade venues, meaning they must be popular in some capacity in order to continue thriving there. This also means more competition. If you wanted to be in a popular area with less competition, I might recommend opening an arcade in East Village. If you wanted very little competition, but perhaps a riskier venture, I would recommend opening one in a different borough, or in an area like Utopia which only has one arcade in it.

Based off the earlier recommendation of opening an arcade in East Village, the type of arcade is also a bit open. The one thing you'll notice is that many of the arcades in East Village are barcades, so if you wanted your arcade to stand out, consider opening something like a pure arcade or some different combination of arcade with another venue.

The last question has a simple answer. For variety and number of venues, you'll want to go to Midtown South in Manhattan.

Overall, I think the project the was successful in determining what we wanted to know. The main problems come from data quality and cleaning quality, especially in the clustering. I tried for days to find a way to eliminate or lessen the impact of nonexistent venues in certain neighborhoods, but couldn't come up with an effective way to do it, and the more I ttried, the more overcomplicated everything became, so I decided to simply stick with what worked and what was at least understandable with explanation. The clustering could have also been done a bit better. My default value for k-means clustering was five, though I feel that with the way the last two clusters ended up, the best value might have been three. It doesn't change our ultimate result, it's more something to note in case someone else wants to improve this experiment or if I decided to do something like this again.

Conclusion

While Manhattan seems to be doing well enough in the arcade scene, I cannot say the same for the boroughs around it. New York City, if the FourSquare data is to be believed, only has 66 venues in it that are considered to be arcades, and considering how populous New York City is, I feel like that number has the potential to be much higher. As a gamer and game designer myself, I understand why arcades are on the decline, but that doesn't mean I necessarily like seeing it happen. Hopefully this project gave you some insight if you were looking to open an arcade of your own in the area, or if you were simply looking to visit and wanted to give your business to some of these locations in order to help them survive, because I fear that arcades will only continue to become rarer and rarer.