CLUSTERING AND UNDERSTANDING TOY STORES IN LA COUNTY

IBM Data Capstone Project for Professional Certificate

Abstract

Research, findings, and discussion on using FourSquare Venues API with Python Machine Learning Libraries to Cluster and Analyze Data

By Joshua Pascascio

Joshua.pascascio@gmail.com

TABLE OF CONTENTS

INTRODUCTION

This project is designed to provide new categories and ways to group toy stores in Los Angeles as to help provide a more detailed, personal insights for both toy vendors and toy buyers.

BUSINESS PROBLEM

Playing with toys is nearly every person's most cherished memory from his or her childhood.
While some stop playing with toys as they get older and approach adolescence or adulthood. Many more continue to collect and appreciate them as adults. This is because toys extend much further then the cheap action figures or dolls we might remember from our childhood; many can be intricate, handcrafted pieces of art that can be worth a fortune. This makes the vendors, especially those with brick-and-mortar retail shops, vary greatly in their merchandise and what kind of audience and clientele they cater to.

This can make trying to understand toy stores and their customers very difficult because the products and price points can vary so widely. Many toy store customers whether children or parents looking to buy toys for their kids to the seasoned collector with shelves of figures worth more thousands need a better way to separate and categorize all these stores so they can find the store that has the toy for them. Vendors may want to understand how to categorize themselves in relation to other stores so they can find their best customers and optimize their business and location strategies.

AUDIENCE

The audience for this report includes both toy store vendors, meaning companies who have physical stores that sell, sometimes exclusively, some sort of toy products as well as the customers. The customers that are the primary focus are toy collectors especially those that collect specific products like FunKo Pops or anime figurines from a franchise. For vendors, these insights, especially when it comes to what categories

and related category of store they belong to, can give guidance on to their marketing and what products they should focus on for promotion. For customer it can provide easier ways to search and navigate stores based on their interest.

DATA

DATA SOURCES

The primary data source will be data retrieved from the FourSquare API on venues and categories.

FOURSQUARE API

There will be a total of 3 different endpoints that will be queried to fetch the dataset. One will be the *search* or *explore* endpoint to get a list of venues that are toy store related and in Los Angeles. The following input will be to search each endpoint by the VENUE_ID in each result to get more venue-specific data including related categories, store hours, visitors and other data that can be used as features for clustering. The final one would be one to get all the categories and related venue ids that will be used later for clustering.

Data Retrieval

To create a dataset for cluster analysis our dataset these characteristics are needed:

- Venue ID or Name
- All the listings the individual venue belongs to
- The category that its parent venue, if present, belongs to [i.e Shopping plaza, Mall, Theme Park, etc]
- Metrics for customer or social media engagement such as photos, likes, ratings, tips on Foursquare
- All the venue categories the venue belongs to

There is also a need for each venue's geographic information to visualize how these clusters and how they might relate to the neighborhoods they might belong to in LA county.

The list of venues can be fetched through an API request to the */search* endpoint of the FourSquare venues API and then joining that data with data retrieved from individual API calls to each venue's detail endpoint.

Once the data is joined it can then be scaled, wrangled, and preprocessed further so it can be used to develop a clustering model with various Python libraries.

METHODOLOGY

API Requests

To begin, a list of all toy stores in the LA county area must be found. The most accurate way to achieve this is to issue a request to the ***search*** API endpoint in Foursquare. The search radius will have the city of **Los Angeles** as the center with a radius of 50,000 miles, sorted by popularity. The search will be limited to venues that belong categories that are in the **Toy / Game Store** category.

In the search request we cannot search by category name, but by *category id* instead. The category id can be retrieved by making an API call to the **venues/categories** endpoint and finding the category result that has the **Toy / Game Store** name.

After the initial API call and searching through the results, the category id was found to be `'4bf58dd8d48988d1f3941735'`.

Once the category id is retrieved, the search can be made to **/search** endpoint and the list of venues can be retrieved and attributes like the venue's name, geographic coordinates, address, and venue-id can be fetched.

Those attributes are not enough however to make meaningful clusters, more details such as social media metrics, categories, and listings need to gathered as well. This means for *every venue*, an additional API cal l to **venue/VENUE_ID** endpoint must be made to gather additional details on each venue.

These additional details include:

- Number of Tips
- Number of User Ratings
- Average Rating

- Number of Likes
- Associated Listings
- Verification Status
- Chain Status
- Associated Categories
- Parent Category, if applicable
- Number of Photos posted to Foursquare

Below is a snapshot of the assembled dataset, that was retrieved using and stored into **Pandas DataFrame**

	name	id	address	latitude	longitude	categories	photos	parent-categories	listings	verified	chain	tips	likes	rating	ratingSignals
0	Anime Jungle 1st	4b85b253f964a5207f6d31e3	319 E 2nd St Ste 103	34.049408	-118.240749	[Toy / Game Store, Hobby Shop]	91	Shopping Mall	[L.A. to do, Vinyl Figures and Toys, Favorite ...	0	0	16	54	7.7	77
1	Game Chest	5faf537db2e12c113624a98c		34.135820	-118.052040	[Toy / Game Store]	0	Shopping Mall	[]	0	0	0	0	0.0	0
2	Mind Games	5d1804bc4dcbca0023d181ec		34.133780	-118.049045	[Toy / Game Store]	1	Shopping Mall	[]	0	0	0	0	0.0	0
3	The Dinosaur Farm	4af33343f964a520b1eb21e3	1510 Mission St	34.115978	-118.151199	[Toy / Game Store, Bookstore]	9	None	[Los Angeles VI, The 10 Most Fun Toy Stores in...	0	0	6	10	7.3	15
4	Chalice Collectives	5d16e75b3e8ac40024229a1e		34.134762	-118.050716	[Toy / Game Store, Hobby Shop, General Enterta...	3	Shopping Mall	[]	0	0	0	0	0.0	0

Request Processing

The process of creating the data frame from the API response was simple since it mostly involved Just iterating through the venue lists and flattening and joining records by ***venue-id*.**

- If a venue is verified, it is encoded in a binary category [0,1]
- If a venue belongs to a chain of stores is also encoded into a binary category as well: [0,1]
- If a venue belonged to parent venue, such as a plaza or shopping mall, that venue's category would be stored into the **parent-categories** column, otherwise None.
- All categories a venue can belong to are collected and stored into a **list** label in the **categories** column
- All listings a venue can belong to are collected and stored into **list** label in the **listings** column.

Preprocessing

In many datasets, especially those with many features, it is common for certain numerical features to have widely different scales, meaning their raw variances, minimums, maximums, and medians can have Different magnitudes. This can create an issue with calculating the dissimilarity measure to form clusters. This can be resolved by using a Standard Scaler on these numerical features.

Another thing to consider is that some numerical features can be equivalent, meaning their values are very highly correlated, usually above 0.95 between two features. This can be checked with a correlation heatmap

During this process it was discovered that the **likes** column had high correlation with the other columns, meaning it could be removed.

The **parent-categories** column was originally listed as string names but were then ordinally encoded into n umbers.

The scaled numerical features can be viewed below:

	name	id	address	latitude	longitude	categories	photos	parent-categories	listings	verified	chain	tips	rating	ratingSignals
0	Anime Jungle 1st	4b85b253f964a5207f6d31e3	319 E 2nd St Ste 103	34.049408	-118.240749	[Toy / Game Store, Hobby Shop]	91.0	2.0	[L.A. to do, Vinyl Figures and Toys, Favorite ...	0	0	16.0	7.7	77.0
1	Game Chest	5faf537db2e12c113624a98c		34.135820	-118.052040	[Toy / Game Store]	0.0	2.0	[]	0	0	0.0	0.0	0.0
2	Mind Games	5d1804bc4dcbca0023d181ec		34.133780	-118.049045	[Toy / Game Store]	1.0	2.0	[]	0	0	0.0	0.0	0.0
3	The Dinosaur Farm	4af33343f964a520b1eb21e3	1510 Mission St	34.115978	-118.151199	[Toy / Game Store, Bookstore]	9.0	0.0	[Los Angeles Vl, The 10 Most Fun Toy Stores in...	0	0	6.0	7.3	15.0

Encoding Categories and Listings

In this dataset, a venue can have many different combinations of listings or categories. A venue can belong no categories, one category, multiple categories, or all the categories or listings available. The issue is that these combinations need to be encoded in a way that can represent all these combinations for all these diffe rent categories. This can be achieved by making each category / listing its own *binary* feature and assigning each data point a 1 if it contains that feature and 0 if it does not.

This approach involves using a **MultiLabelBinarizer** on our dataset, leaving a new data frame or data table looking like the one below:

verified	...	Welcome to the Tragic Kingdom	West Coast 2019	What should I do today? Oh I can go here!	dca	favorite places	my places	toys and games	لوس انجلس	マンガやアニメの画像 Best Manga & Anime Images	好きなお店
0	...	0	0	0	0	0	0	0	0	1	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0

DATA ANALYSIS

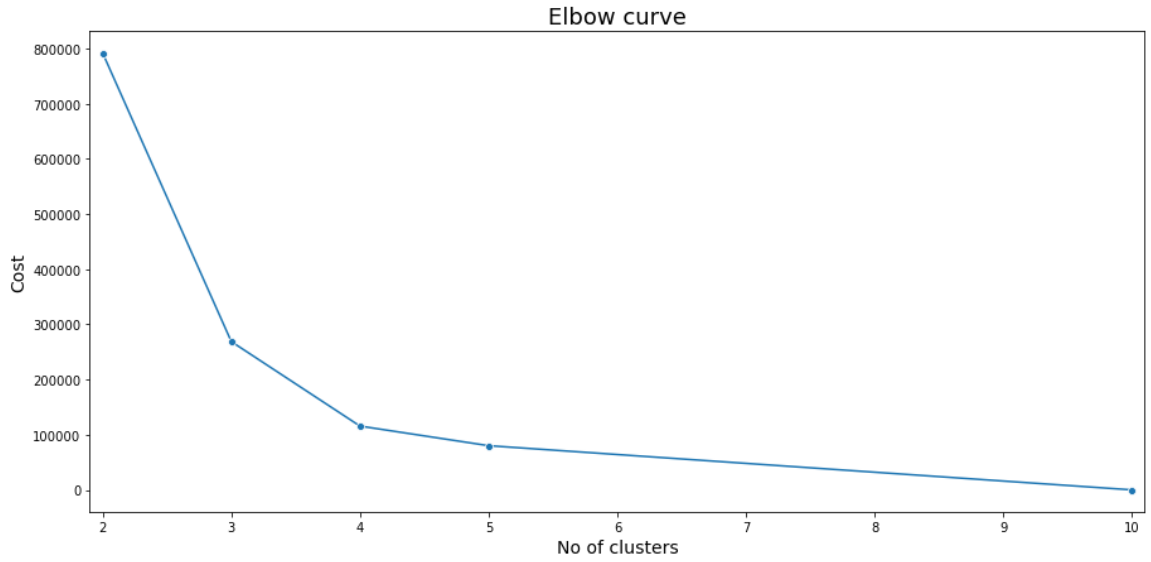
Choice of Clustering Model

Since the categories and listings are likely a huge influence as to how similar venues are to each, it can be very fitting to use a hierarchical clustering model where the grouping of stores can be segmented and visualized to see and understand how some clusters may be more similar towards each other and can be used to provide insights for businesses and customers to see how their specialties tastes fit into the greater whole.

The hierarchical clustering model that will be used is an agglomerative clustering model. It is a bottom up Approach that focuses on grouping observations that are closest to each and recursively joining them all until they are merged into the top of the hierarchy.

Calculating Optimal K

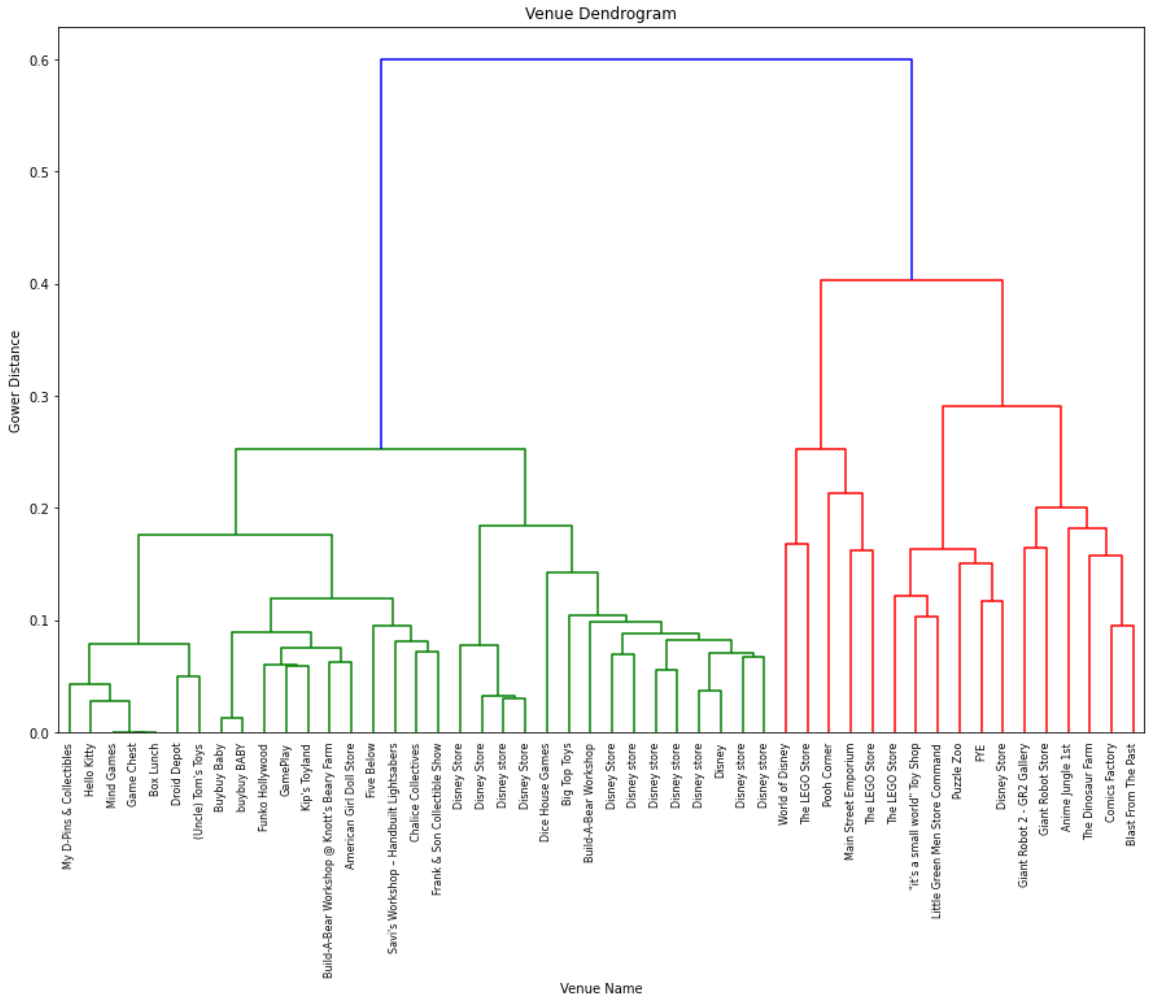
The optimal K was calculated using the *elbow* method with K values ranging from **2 to 10**. From inspection , of the *elbow graph,* it was determined that the optimal K was **K=5**.



Calculating Similarity between Venues

In many standard clustering, the features used are either *only numerical* or *only categorical*. This dataset on the other hand is a *combination of both numerical and categorical variables.* As a result, a custom distance algorithm known as the **Gower distance** was used to calculate dissimilarity. Using a Python Gower's distance package, a *distance* matrix was calculated which can be used to link the dendrogram for visualizing the hierarchy as well as predicting the cluster labels in an agglomerative clustering model.

The hierarchy can be visualized with the below dendrogram:



RESULTS

The results of the clustering resulted in a total 5 clusters between 50 venues in the LA county area.

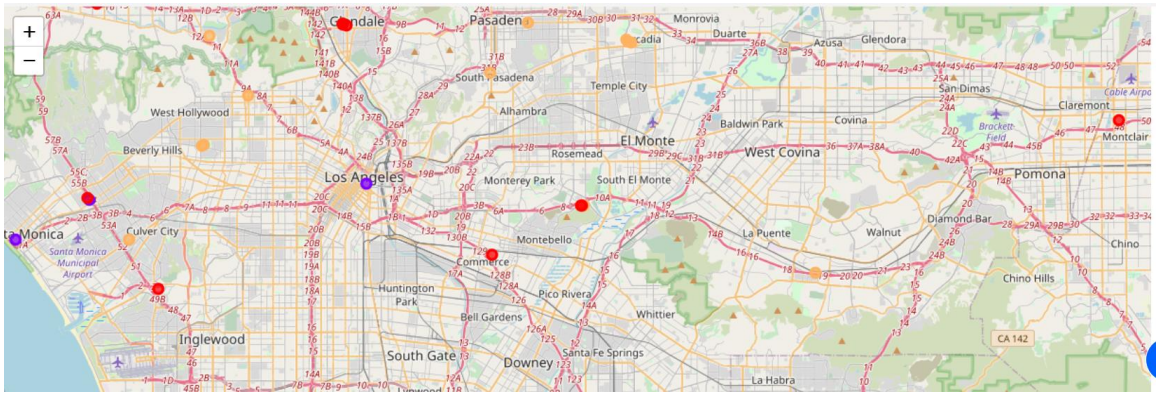
]:

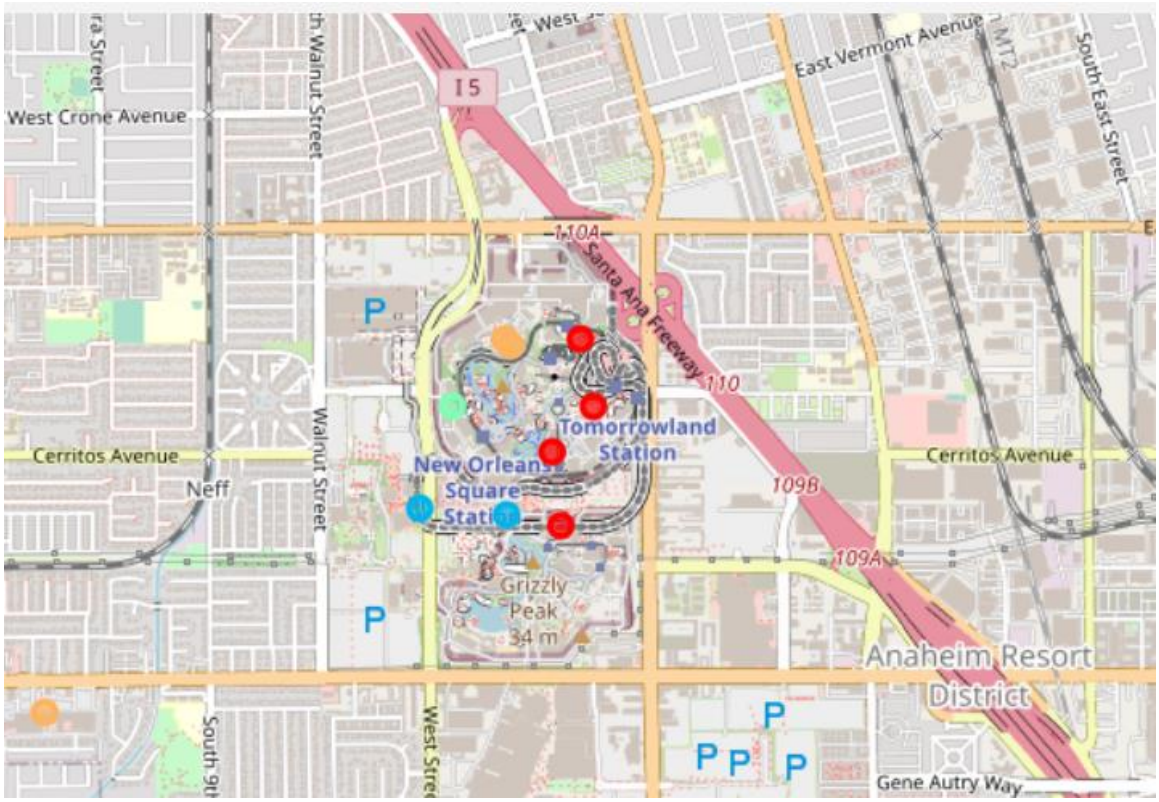
	id	name	address	latitude	longitude	Cluster
0	4b85b253f964a5207f6d31e3	Anime Jungle 1st	319 E 2nd St Ste 103	34.049408	-118.240749	1
1	5dec504dbed40a00072a3bbd	Box Lunch		34.035905	-118.083326	4
2	4af33343f964a520b1eb21e3	The Dinosaur Farm	1510 Mission St	34.115978	-118.151199	4
3	5e3082776b6ff00007fd9429	Build-A-Bear Workshop @ Knott's Beary Farm	8039 Beach Blvd	33.843169	-117.998639	4
4	5d16e75b3e8ac40024229a1e	Chalice Collectives		34.134762	-118.050716	4
5	57cf5325498e3025508b44c5	My D-Pins & Collectibles	1648 W. Katella Ave. Suite B	33.802089	-117.939232	4
6	5d1804bc4dcbca0023d181ec	Mind Games		34.133780	-118.049045	4
7	5faf537db2e12c113624a98c	Game Chest		34.135820	-118.052040	4
8	5c9ad42eff0306002c2c7ebd	Hello Kitty		34.137822	-118.355035	4
9	60373e244814104b3f3b2c7a	Five Below	18309 Brookhurst St Ste 5	33.696298	-117.955702	4
10	54a70415498e9c8f4ca1d078	buybuy BABY	22999 Savi Ranch Pkwy	33.874955	-117.734874	4
11	55b55c5b498e550d3f47424e	Buybuy Baby	6621 Fallbrook Ave Unit B	34.190315	-118.624932	4
12	4a88ca4ef964a5209d0720e3	Puzzle Zoo	1411 3rd Street Promenade	34.015402	-118.495747	1

- Cluster 0 has 23 venues
- Cluster 1 has 3 venues
- Cluster 2 has 2 venues
- Cluster 3 has 1 venue
- Cluster 4 has 21 venues

Visualizing Clusters

Using the middle latitude and longitude of all venues respectively, (33.94838954508498, -118.08 76358), all the clusters are rendered visually onto a map of LA County.





Cluster 0 is rendered as a **red dot**, cluster 1 as a **purple dot**, cluster 2 as a **bluish dot**, cluster 3 as a **green dot**, cluster 4 as an **orange dot.**

From zooming out at inspecting, cluster 0 has the furthest reach and is associated chain stores that are targeted to a wide audience of young kids, namely the Disney store, where the merchandise is non-specific. Cluster 1 upon further research belong to more unique stores with a wider, more niche merchandise, which can cater to adults, namely Anime Jungle in Little Tokyo. Cluster 2 an d 3 are both centered in the Disney with 2 relating to chain stores and 3 relating to the specific Po oh Store in Disney. Cluster 4 also has wide geographic reach with a seeming focus on games and collectibles such as Funko Hollywood.

DISCUSSION

Use Cases for Cluster Hierarchy

A cluster hierarchy can be used by businesses to see where they relate in the family of toy stores, they belo ng to in their area. This can make product assortments and business plans easier to implement because with this understanding, businesses can use the insight to still focus on their main clientele, but not isolate custo mers that may shop at related Toy or game stores. Also knowing the hierarchy can help businesses with allo cating the proper amount of budget for products and advertising to certain customers that may correspond t o clusters closer in their hierarchy. If a prospective business notices that there is not a business that is a specific type, or cluster, in their area, it can make that area a prime area for business so they can fill that market void.

With store clusters, eventually a customer hierarchy can be formed which can make future promotions easier to prepare as well as help businesses isolate which customers can be most loyal to their store or most likely to buy their most premium merchandise.

Limits

The biggest limit was the limited data available for venues in the Toy / Store category. If there we re more venues available or more venues that could be retrieved using the free version of Foursquare API, more clusters could be found, thus creating a more in depth and detailed hierarchy. In the future a larger dataset will be needed along with data regarding each stores indi vidual sales merchandise which can give even more insight to determining and quantifying similarities between venues. This type of information when joined to the original dataset can give even more useful information to separate stores from the customer's, especially a hobbyist's, perspective. Another thing that could be helpful would be to track the different types of social me dia engagement as many unique stores that sell collectibles are more likely to promote more heavily on mediums like Twitter and Instagram.

CONCLUSION

In this project, research was gathered, and data was retrieved, organized, and aggregated to separate 50 stores in the Los Angeles area into 5 clusters or 5 unique categories. This process involved joining responses from many different API requests and cleaning and preprocessing those responses. The clusters made use of the many different combinations of categories and listings the stores have in common. Using those shared features as well as using metrics to track customer engagement, an initial hierarchy was formed to make these clusters. The clusters revealed strong separation between small businesses and trade stores as well as stores in bustling locations like Downtown Disney and Little Tokyo versus the suburbs in shopping plazas. While the data may seem initially limited, this model and approach can be used, once given larger datasets, to make detailed custom categories of toy stores that can allow for more insightful marketing for vendors and more personalized shopping for customers.

REFERENCES

1. Foursquare Developers Documentation. *Foursquare.* Retrieved from https://developer.foursquare.com/docs/

2. What is Gower's Distance. *Statistical Odds and Ends.* Retrieved from https://statisticaloddsandends.wordpress.com/2021/02/23/what-is-gowers-distance/