**Data Science Capstone Introduction**

(Please note that much of this text is taken from my notebook. A large part of the point of a notebook, of course, is to mix text and code together, so I originally wrote my report in the notebook.)

My project is going to try to answer the question of where someone should open a comic book store in my area. My audience is largely myself, since I am considering if I should try to start a business or not, but would apply to anyone potentially interested in opening a comic store. I spent most of my life living in the Midwest (United States) and most of the comic stores I knew there were largely dependent on foot traffic: people literally walking past the store and deciding to drop in and see what was there. A few years ago, though, I moved to southern New Jersey where there aren't many places with foot traffic. Most stores are in little shopping centers that are unconnected from each other, and you have to drive most places anyway as the area is not very walkable or public transit friendly. So, there are really two questions to answer: is this a good place to open a comic store at all? And if so, where would you want to place it?

**Capstone Data Section**

To decide if and where to place a comic book store, I decided there would be three steps to the analysis:

* Find other cities like mine
* Analyze how many comic stores they have and what those neighborhoods are like
* Look at my area to see if and which areas are like those in the other cities

**Find Similar Cities**

There are many ways to determine how cities might be like each other. There's population, area, demographics, income levels, and much more. I decided to rely on other people already having done the work. I found two references:

[This](https://www.nytimes.com/interactive/2018/04/03/upshot/what-is-your-citys-twin.html) New York Times article written by someone who works at Indeed. You can look up a city (if it's sufficiently large to be in their list) and it will tell you what other cities are similar to it in terms of job listings. I thought this would be helpful because job listings could be related to income and certain kinds of people who like to buy comics.

[This](https://www.chicagofed.org/region/community-development/data/pcit) tool put together by the Federal Reserve Bank of Chicago, which lets you enter a city and it tells you what other cities are similar to it on a few different dimensions: equity (demographics and income inequality measures), resilience (unemployment, income levels, and labor force participation measures), outlook (population and demographics measures), and housing (home value and rent measures). I thought this would be helpful as more direct measures of what the different cities are like but put together by a different group, and thus resulting in different similar cities.

One limitation of these tool is that my exact city, Egg Harbor Township NJ, is not big enough to be in their databases. So, I will use Atlantic City, NJ as the 'base' city to compare to. Another option would be Philadelphia, which is the nearest large city, but I am not interested in looking at Philadelphia. I want to look at options in southern New Jersey specifically.

Using Atlantic City as the base city, I could use over 20 comparison cities based on these two references. That seems like too many, especially since some of the comparisons won't really be very similar to Atlantic City, so I am going to choose the following:

* Las Vegas, NV from the Times link: Atlantic City is a gambling-based tourist town, so Vegas makes a lot of sense. Vegas is a much bigger city and better tourist draw, though.
* North Port, FL from the Times link: seven of the top ten matches are in Florida. North Port is 4th of the 7, but the similarity scores are very similar and North Port has the smallest population, making it a bit more similar to Atlantic City. Both are coastal towns.
* Cleveland and Dayton, OH: the top matches from the Chicago equity link. They're also on the resilience list.
* Albany, GA; the top match from the Chicago resilience link.
* Port Arthur, TX: the top match from the Chicago outlook link.
* New Orleans, LA: not the top match in the Chicago housing link, but fairly similar. New Orleans was also on the Times list, presumably because both cities are heavily based on tourism.

These seven cities will hopefully give me a good sense of if they can support comic book stores, and thus if southern New Jersey should be able to, while also representing a range of city types. I won't be dependent on comparing my area to one other place and relying on its special characteristics or uniqueness.

**Analyze the Similar Cities**

I will use the Foursquare API to look at these cities in two steps:

1. How many comic stores are there in the area?
2. What kinds of places (Foursquare venues) are near those stores?
3. What kinds of places are **not** near stores?

Specifically, since the cities cover a range of sizes, I will use the lat/long and area listed for each city on Wikipedia to define an appropriate center and range for a Foursquare venue 'search' for comics. I'll have to do some filtering on the results, as I know from doing a Foursquare search of my area that there could be redundant entries (I get my local comic store as well as the mall where it is located, and another business is listed twice under different categories).

Once I have a list of comic stores in each of the comparison cities, I will do a Foursquare venue 'explore' around each store for what other points of interest are nearby. Are comic stores near restaurants and coffee shops? Are they in malls or other retail shopping? Are there any that are near parks or something else? This will be similar to what we did for the week 3 assignment but based around comic stores instead of neighborhoods. Importantly, I will pick a few lat/longs not near comic stores and get their characteristics as well. This is a necessary step to make sure I'm not just picking retail areas or something where any kind of store could be; I want to find what makes a good place for a comic store specifically.

**Compare to My Area**

Finally, I will take the venue types from the comic stores in other cities and see if there's a similar area near me. I will answer this question in a couple ways:

1. How many comic stores are there in the other cities, and how does that number relate to characteristics in the Chicago link? For example, there could be a simple relationship where larger cities have more stores, but perhaps it also varies with income level.
2. How does my area fit into a clustering analysis of those areas? I'll do a similar analysis to our week 3 assignment, adding my area to the comparison cities, and see which cluster my area fits into if any.

At the end of this analysis, I should have an idea of how many comic stores an area might support (since mine already has two) and what kinds of areas they tend to be in as defined by Foursquare venues.

**Methodology**

As noted above, I need to gather data on my area and the seven comparison cities. The Federal Reserve Bank of Chicago site allows for downloads but doesn't let you customize the cities in a list, or combine across their various similarity measures, so I entered the relevant data by hand into Excel and saved it as a CSV.

Next, I need to pull Foursquare data for these cities on what comic stores are in the area for each city. As described before, I used Wikipedia to find a lat/long for each city and its rough area so that I can use that lat/long and search area in the Foursquare API call. I put those into an Excel file as well.

The Foursquare results have a number of repeats or redundancies. On the other hand, some stores could have multiple locations. So I googled any potentially redundant stores and removed them if appropriate.

Once I have the comic store data, I want to do two things. First, count the number of stores per city and put that with the city data. Second, pick another place per each store that can serve as a kind of 'control' - a place relatively nearby but not centered on a comic store. This will be used for distinguishing between locations that would be good or bad for a store in my area.

### Analyzing the Areas[¶](https://render.githubusercontent.com/view/ipynb?commit=ae7ec58bda25962850959356b4856cf7f5d8c467&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f616b6f6e6b656c2f436f7572736572615f43617073746f6e652f616537656335386264613235393632383530393539333536623438353663663766356438633436372f63617073746f6e655f6e6f7465626f6f6b2e6970796e62&nwo=akonkel%2FCoursera_Capstone&path=capstone_notebook.ipynb&repository_id=251710566&repository_type=Repository#Analyzing-the-Areas)

Once I had the venues near the comic stores and their matching locations, it was time to cluster them. I decided to use 15 clusters for the k-means algorithm since many locations ended up in a single cluster and I wanted to get a sense of what the outliers were like.

### Statistical Analysis[¶](https://render.githubusercontent.com/view/ipynb?commit=ae7ec58bda25962850959356b4856cf7f5d8c467&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f616b6f6e6b656c2f436f7572736572615f43617073746f6e652f616537656335386264613235393632383530393539333536623438353663663766356438633436372f63617073746f6e655f6e6f7465626f6f6b2e6970796e62&nwo=akonkel%2FCoursera_Capstone&path=capstone_notebook.ipynb&repository_id=251710566&repository_type=Repository#Statistical-Analysis)

In addition to the clustering analysis, I can use the city data I collected to look at how city-level information informs the decision to open a comic store.

With only seven cities and so many predictors, there isn't much to do in terms of a regression or the like (there will be massive overfitting). But to get a sense of what city-level features are associated with having more comic stores I created a heatmap of the correlations.



I plotted the whole correlation matrix in case there was anything interesting in how the predictors relate to each other, but we are principally interested in the bottom row/right-most column. That row/column has the correlation for number of stores ('store\_count') with the other predictors. For example, the highest correlations are with labor force participation (0.9), labor share manufacturing (-0.6), and population (1.0, perfectly correlated).

I also tried running a LASSO regression on the data since that is one of the few methods that will adjust for having more predictors than observations (it sets the parameters for many predictors to zero, thus selecting only the strongest/most relevant predictors). However, it only selected city population, so that wasn't very insightful.

**Results**

I will break the results into two parts, one for the clustering analysis and one for the correlational/heatmap analysis.

### Cluster Results[¶](https://render.githubusercontent.com/view/ipynb?commit=ae7ec58bda25962850959356b4856cf7f5d8c467&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f616b6f6e6b656c2f436f7572736572615f43617073746f6e652f616537656335386264613235393632383530393539333536623438353663663766356438633436372f63617073746f6e655f6e6f7465626f6f6b2e6970796e62&nwo=akonkel%2FCoursera_Capstone&path=capstone_notebook.ipynb&repository_id=251710566&repository_type=Repository#Cluster-Results)

On different runs of the algorithm, comic book stores were placed largely into a single cluster but occasionally had one or a few stores placed into a second or third cluster. The majority of matching locations were also placed into the same main cluster as comic book stores. However, there is little overlap between the comic cluster(s) and the non-comic clusters. Since the cluster center output from the K-means clustering is very simple for most clusters (putting a weight of 1 or .5 on one or two venue types), I can use that to get a brief description of the critical venue types for each cluster.

In one particular run, for example, clusters were based around ‘pizza place’, ‘lake’, or multiple venues like ‘park, scenic lookout, trail’. The main cluster, in contrast, was not based around one or a few venues; it put a little weight on a large number of venue categories. Looking at the top of the frequent venues table shows why:

| **1st Most Common Venue** | **2nd** | **3rd** | **4th** | **5th** | **6th** | **7th** | **8th** | **9th** | **10th** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pharmacy | Seafood Restaurant | Mexican Restaurant | Southern / Soul Food Restaurant | Frozen Yogurt Shop | Pet Store | Pizza Place | Discount Store | Sandwich Place | Donut Shop |
| Pizza Place | Sandwich Place | Bar | Clothing Store | Tourist Information Center | Sports Bar | Donut Shop | Bank | Dive Bar | Gas Station |
| American Restaurant | Fast Food Restaurant | Pizza Place | Convenience Store | Gym | Farmers Market | Shipping Store | Seafood Restaurant | Fried Chicken Joint | Sandwich Place |
| Music Store | Diner | Auto Workshop | Smoke Shop | Sandwich Place | Liquor Store | Fried Chicken Joint | Bar | Discount Store | Bakery |
| Bar | American Restaurant | Pizza Place | Mexican Restaurant | Sandwich Place | Bank | Shopping Plaza | Donut Shop | Sports Bar | Spa |

The categories vary from row to row (each row reflecting a different comic store or matching location), but they are all basically business areas: restaurants, bars, malls, shopping plazas, etc.

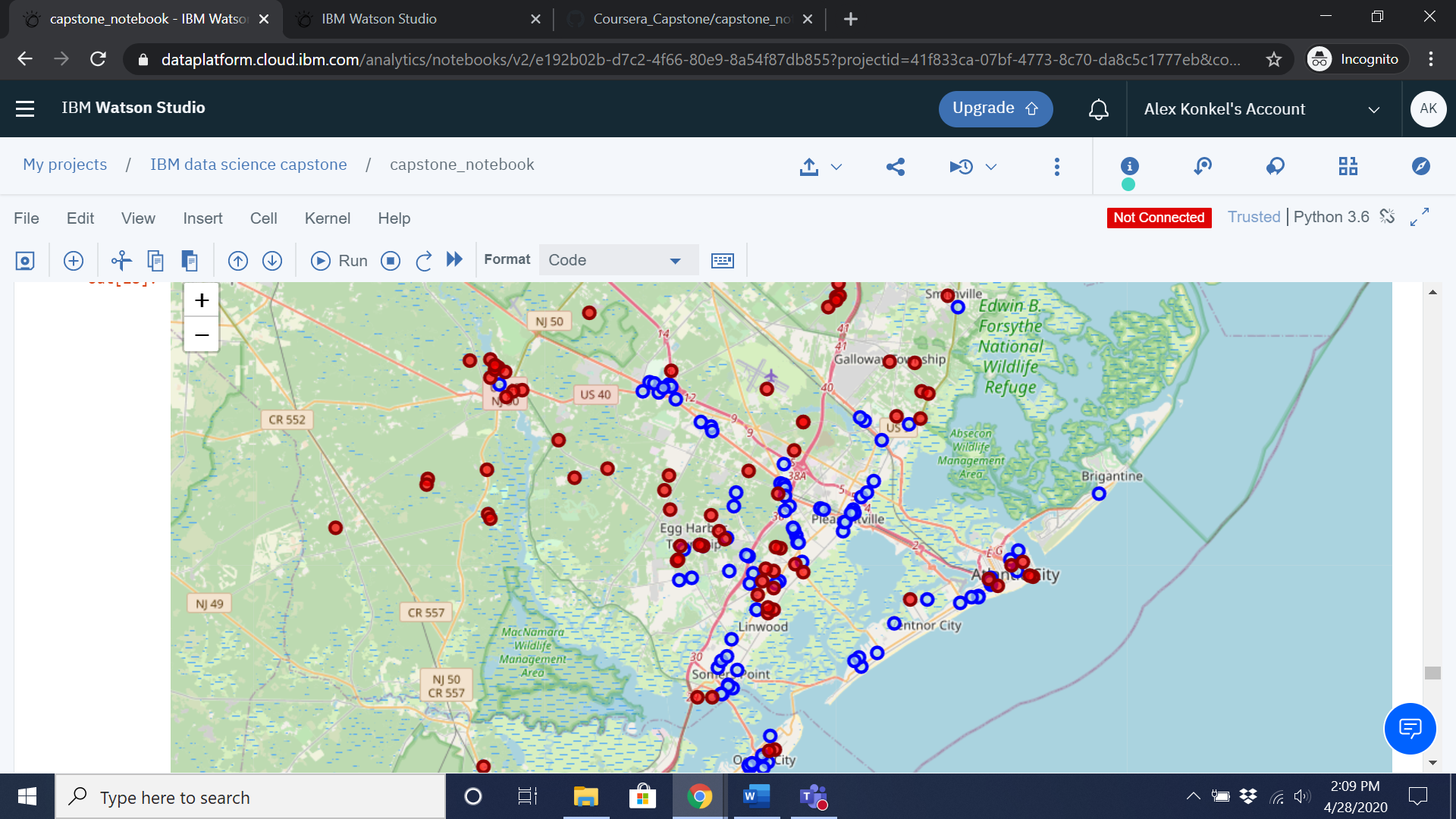
The contrasting clusters, where some matching locations were sorted but not comic stores, are based around bars, parks, lakes, and liquor stores. For example, here is the second largest cluster (excluding the one that comic stores tend to fall in) for non-comic stores.

| **1st Most Common Venue** | **2nd** | **3rd** | **4th** | **5th** | **6th** | **7th** | **8th** | **9th** | **10th** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Trail | Intersection | Vietnamese Restaurant | Fried Chicken Joint | Drugstore | Event Service | Farm | Fast Food Restaurant | Financial or Legal Service | Flower Shop |
| Trail | Scenic Lookout | Vietnamese Restaurant | Food | Donut Shop | Drugstore | Event Service | Farm | Fast Food Restaurant | Financial or Legal Service |
| Trail | Scenic Lookout | Park | Vietnamese Restaurant | Financial or Legal Service | Discount Store | Donut Shop | Drugstore | Event Service | Farm |

These non-comic areas seem to prominently feature open spaces, like parks, trails, and lakes. Comic stores are more reliably in dense business areas.

### Clustering Conclusions

With all that said, what area would then be good near me? I performed a Foursquare search in my area for a few key words like 'restaurant', 'store', and 'bar', along with 'lake' and 'trail', and then mapped them.



Looking at the map, my immediate area (the 'Egg Harbor Township' label) is not a great place for a comic store. It mostly has parks and green space, not a lot of businesses.

To the north, to the right of the US-40 label, would be a possibility. However, that's where the mall is, which already has one of the two comic stores in the area. That's likely a no-go due to competiton.

I see three possibilities near me: Pleasantville, Somers Point, and Absecon Island (the island with Atlantic City, Ventnor, Margate, and Longport). Pleasantville does have a fair density of businesses but is generally a low-SES area. Somers Point is a mix of business and residential with a reasonable SES, and Absecon Island on the whole would be described about the same but with a wide range of SES (lowest in Atlantic City and increasing as you move toward Longport). Ocean City, at the very bottom center of the map, could be another option. It is known as the ‘family friendly’ Atlantic City, with a boardwalk and collection of businesses but no casinos or alcohol allowed on the island.

### Statistical Conclusions

The high (perfect, since there are only seven cities) correlation with population is not surprising as larger cities will generally have more of anything. But there are other city-level predictors associated with having more comic book stores as seen in the heatmap. Atlantic City is low population compared to these other cities but does fairly well on the other measures: for example, labor share manufacturing is negatively correlated with number of stores and AC is low on that, and it has a relatively high 'young adult' population percent and high home value to income ratio.

In summary:

* For my seven comparison cities, I found comic stores in the area using Foursquare's API. The cities ranged from having 1 to 29 comic stores.
* For each comic store and (as much as possible) a nearby comparison location, I found the other Foursquare venues nearby so that I could analyze what kinds of areas comic stores tend to be in.
* Using cluster analysis, comic book stores tend to be in business areas, with restaurants, bars, and so on nearby. Matching areas without a comic store tended to be near parks, trails, or airports; that is, not business-centric areas.
* I mapped restaurants, stores, and bars in my area along with parks, trails, and lakes. This produced a visualization of where a comic store might do well in my area.
* Finally, I found the correlation between number of stores in a city and various city-level characteristics to see what features correspond to having more comic stores. Population was, unsurprisingly, number one but there was also a high correlation with having more working age (20 to 64) adults and a low share of the workforce in manufacturing.

# Discussion

Following on the results, it seems like opening a comic store in Egg Harbor Township, New Jersey (or nearby) would be a dicey proposition. On one hand, there are business areas similar to where comic stores exist in other cities. Atlantic City (the largest city nearby which has city-level data associated with it) also has characteristics similar to other cities that can support multiple comic stores. On the other hand, the number one predictor of having more comic stores is population, and Atlantic City/Egg Harbor Township does not have a large population. Additionally, there are already two comic stores in the area; this could be as many as the area can support. A deeper dive into the area would be needed in order to make a final decision as to if another comic store could thrive here.

# Conclusion

This report aimed to determine if Egg Harbor Township, New Jersey (a suburb of Atlantic City) would be a good place to open a comic store. Using Foursquare data with clustering analysis and city-level descriptors (such as demographics) with correlation analysis, the conclusion is that such a store would be hit or miss. Given that two stores already exist in the area, it might be safer to open a different kind of business or else aim for more of an online presence and less of a brick-and-mortar storefront.