# University of Arizona - Coding Bootcamp - Final Project

## Research Objectives

This project aims to take the [raw data](https://github.com/PaulBernert/DBNA/blob/master/_data/2020-report/data_with_msas.csv) provided from the Doing Business North America report, calculate a rank and score to determine the 'Ease of Doing Business' (see the [Methodology](https://github.com/PaulBernert/DBNA/wiki/Ease-of-Doing-Business-Methodology) for explanation of what Ease of Doing Business measures). The effectiveness of the ranking/scoring process will be tested in an applied setting--using the calculated 'Ease of Doing Business' ranks to see if it correlates with business activity relative to the local population.

Another primary goal of the project is to use Machine Learning algorithms (through Scikit-learn) to test multiple clustering algorithms to see which locations are the most similar in nature. The two clustering algorithms to be tested are KMeans and Affinity Propagation. After comparing these two clustering structures, the next step is to test the relationship between clusters and the 'Ease of Doing Business' ranks to see if there clusters are formed around ranks (whether ranks are a good representative of how locations are clustered).

These tests were chosen not only to test the effectiveness of the data-set in an analytical environment, but to also see whether regulatory burdens truly have an impact on business starts.

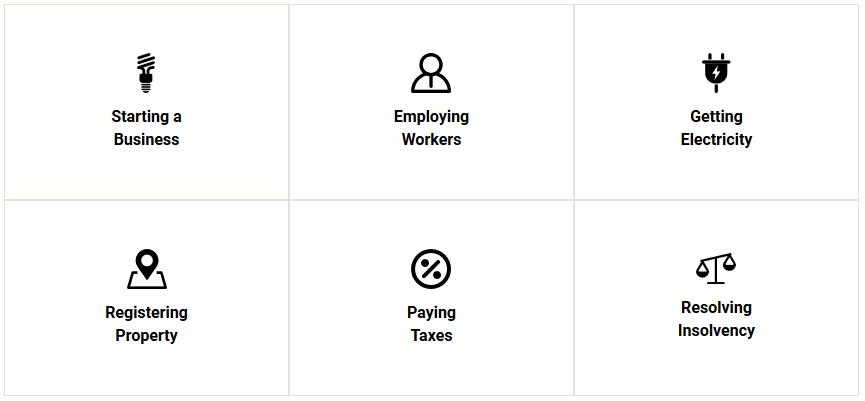
## About the Project

The Doing Business North America (DBNA) project annually provides objective measures of the scale and scope of business regulations in 130 cities across 92 states, provinces, and federal districts of the United States, Canada, and Mexico. It uses these measures to score and rank cities in regard to how easy or difficult it is to set up, operate, and shut down a business.

Over the years, researchers have begun to understand how robust measurement and ranking of regulations that either enhance business activity or constrain it can provide substantial insight into economic outcomes. Objective measurements of those regulations have been vital in this understanding. Unlike many studies that measure regulations at the state level, this annual study measures the impact at the city level and does so for over 100 municipal jurisdictions across North America.

The Doing Business North America team collected data on 63 different regulatory and economic indicators across six different categories. The data collected came entirely from official and publicly-available sources.

## About the Data



This project manually collects data on primary regulatory burdens businesses face throughout the entire life-cycle of a business, ranging from Starting a Business to eventually Resolving Insolvency if the business were to shut down / go bankrupt. The report contains data on 63 regulatory indicators within the following six categories:

1. Starting a Business
2. Employing Workers
3. Getting Electricity
4. Paying Taxes
5. Land and Space Use
6. Resolving Insolvency

These six categories are then combined to create a catch-all value known as the 'Ease of Doing Business'. This is the value used in all analysis for this project.

#### Special Thanks to Arizona State University and the Center for the Study of Economic Liberty

**Research Question #1**

**Question: Does the *'Ease of Doing Business'* Score Correlate with Business Activity?**

In order to determine whether the *'Ease of Doing Business'* Score correlates with relative business activity for a given location, we must first determine what the *'Ease of Doing Business'* Score measures and how it is calculated.

The *'Ease of Doing Business'* Score is a metric created that focuses on the regulatory burdens a small- to medium-sized business would face from the birth of the business to the death of the business in cities across North America. It takes these regulatory burdens (over 60 included in this report) and creates a single number used to represent the regulatory climates in these different locations.

The *'Ease of Doing Business'* Score is calculated by using the raw data and applying a basic linear transformation equation of: ((W-C)/(W-B))\*10, where W is the Worst Regulatory performance for a given indicator, B is the Best Regulatory performance for a given indicator, and C is the performance for the current observation (the city being calculated). For example, if the lowest minimum wage is $7.25 across the U.S., the highest minimum wage is $15.00 across the U.S., and I want to know the value for Phoenix AZ (a $12.00 minimum wage), that would be: (($15.00-$12.00)/($15.00-$7.25))\*10 => ~3.87. A location with a $15 minimum wage would get a 0.00 and a location with a $7.25 minimum wage would get a 10.00, where 10 is granted to the "best regulatory performance", and 0 is granted to the "worst regulatory performance".

This process is repeated for all indicators in a category, and then across all categories to get a final *'Ease of Doing Business'* Score. The locations with the highest scores are granted the highest rank, and the locations with the lowest scores are granted the lowest rank.

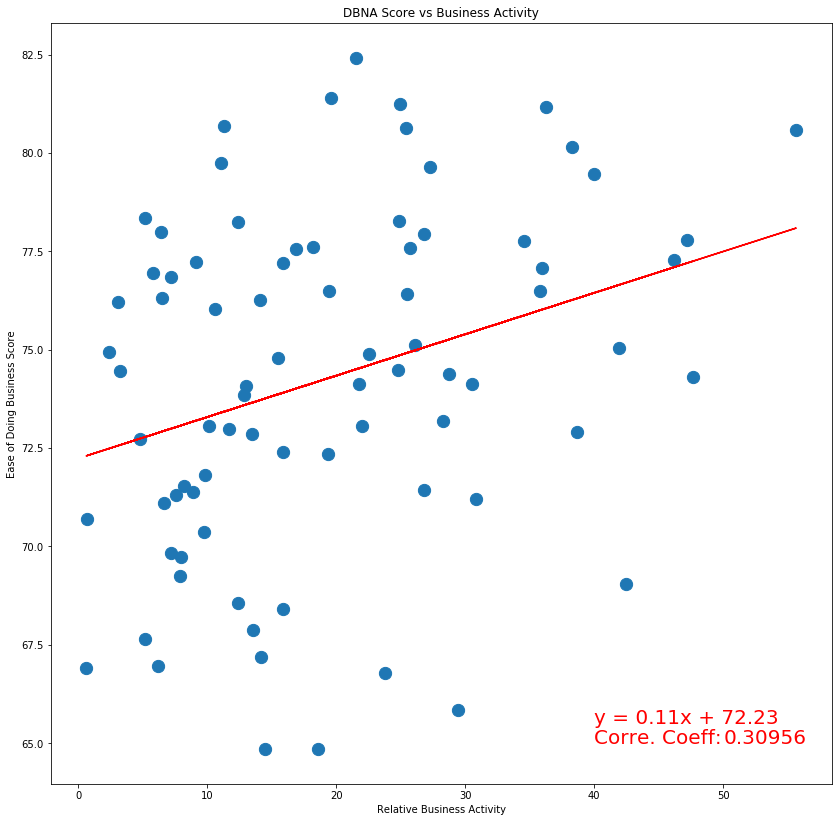
With the *'Ease of Doing Business'* Scores calculated, the top 10 cities are as follows:

1. Raleigh - North Carolina (82.42)
2. Jackson - Mississippi (81.39)
3. Tulsa - Oklahoma (81.25)
4. Sioux Falls - South Dakota (81.17)
5. Charleston - South Carolina (80.69)
6. Houston - Texas (80.64)
7. San Antonio - Texas (80.58)
8. Colorado Springs - Colorado (80.16)
9. Cincinnati - Ohio (79.75)
10. Cheyenne - Wyoming (79.65)

The assumption is that there should be relatively higher amounts of business activity in locations that have lower regulatory burdens, because lower burdens / barriers to entry incentivize taking higher risk and being involved in entrepreneurial activity. To test this theory, we need to find a metric of relative business activity.

Because the DBNA data specifically collects information geared towards small- to medium-sized businesses, data for this particular category of business type from the Census and other government websites was used. The Census provides the number of businesses with 25 or fewer employees at the city level, which is exactly what we need. The number of businesses is then divided by the city population, to get a number that reflects "the Number of Businesses relative to the local population". Cities with more businesses relative to population indicate that the population is, on average, more likely to be involved in entrepreneurial activity.

We can now begin to ask whether the DBNA *'Ease of Doing Business'* Score correlates with relative business activity. To do this, we use Numpy, Scipy and Matplotlib to produce the following results:



The results of our calculations tell us that: ***The R-Value between DBNA Scores and Relative Business Activity is 0.309564***, meaning the *'Ease of Doing Business'* Score and *'Relative Business Activity'* calculation have fairly strong, positive correlation at roughly 31 percent.

**Conclusion**

The results of this first research question were very satisfying. This data-set has never undergone any form of analysis before, so I was quite pessimistic in its ability to produce coherent results. Not only did it produce coherent results, but it also confirmed our initial hypothesis that regulatory burdens may indeed play a role in relative business activity. The magnitude of that correlation isn't perhaps as large as initially anticipated, but it does provide some context to where businesses start. It also allows us to now incorporate additional, non-regulatory indicators to see if we can complete a bigger picture on answering the question "What are the determinants when choosing where to open a business". The next steps forward are clear, and these are questions we plan to answer in time.

# Research Question #2

## Question: What New Information Can Clustering Provide?

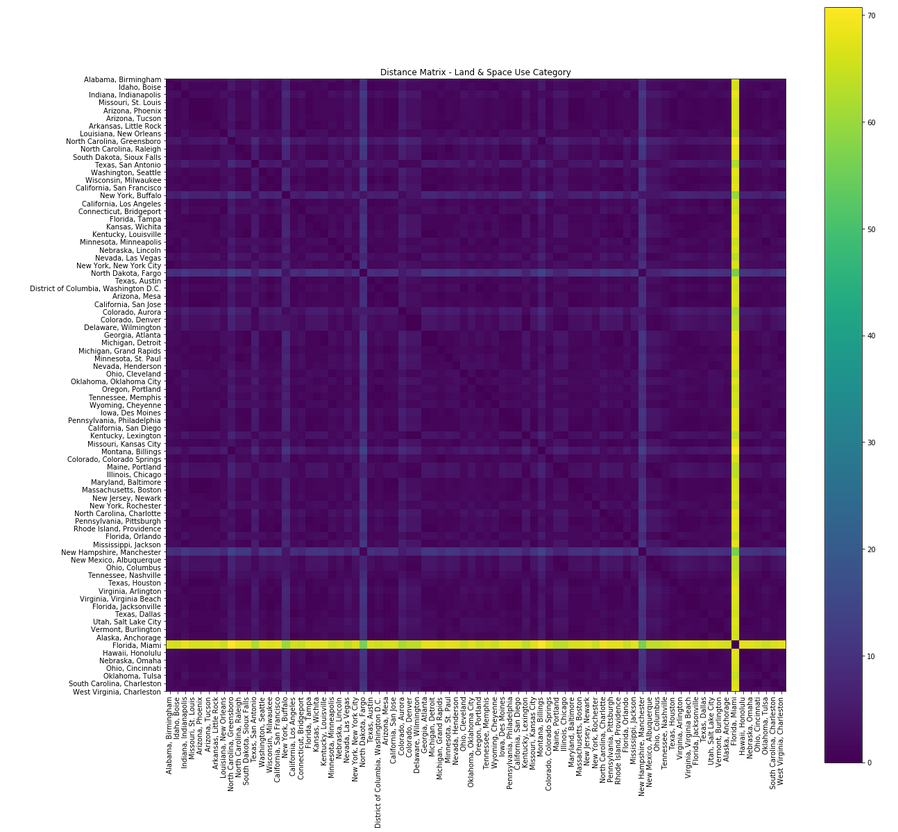
This question is intentionally open-ended, because it allows us to tackle multiple problems. First and foremost, does the DBNA Data cluster in a way that is coherent? Does clustering the data or clustering the scores (post-linear transformation) provide more interesting results? What clusters provide the most interesting results (by category, by all indicators, etc.)? If the data is able to be clustered, are clusters formed around the 'Ease of Doing Business' Scores/Ranks, or at least correlate with them? Do different clustering methods produce the same results, and if not, which clustering algorithm is best? There are many different questions that can be be answered, and my goal is to tackle as many of them as possible. Due to time constraints, we will focus on general outcomes and answering the questions proposed above.

### Clustering Method - KMeans vs Affinity Propagation

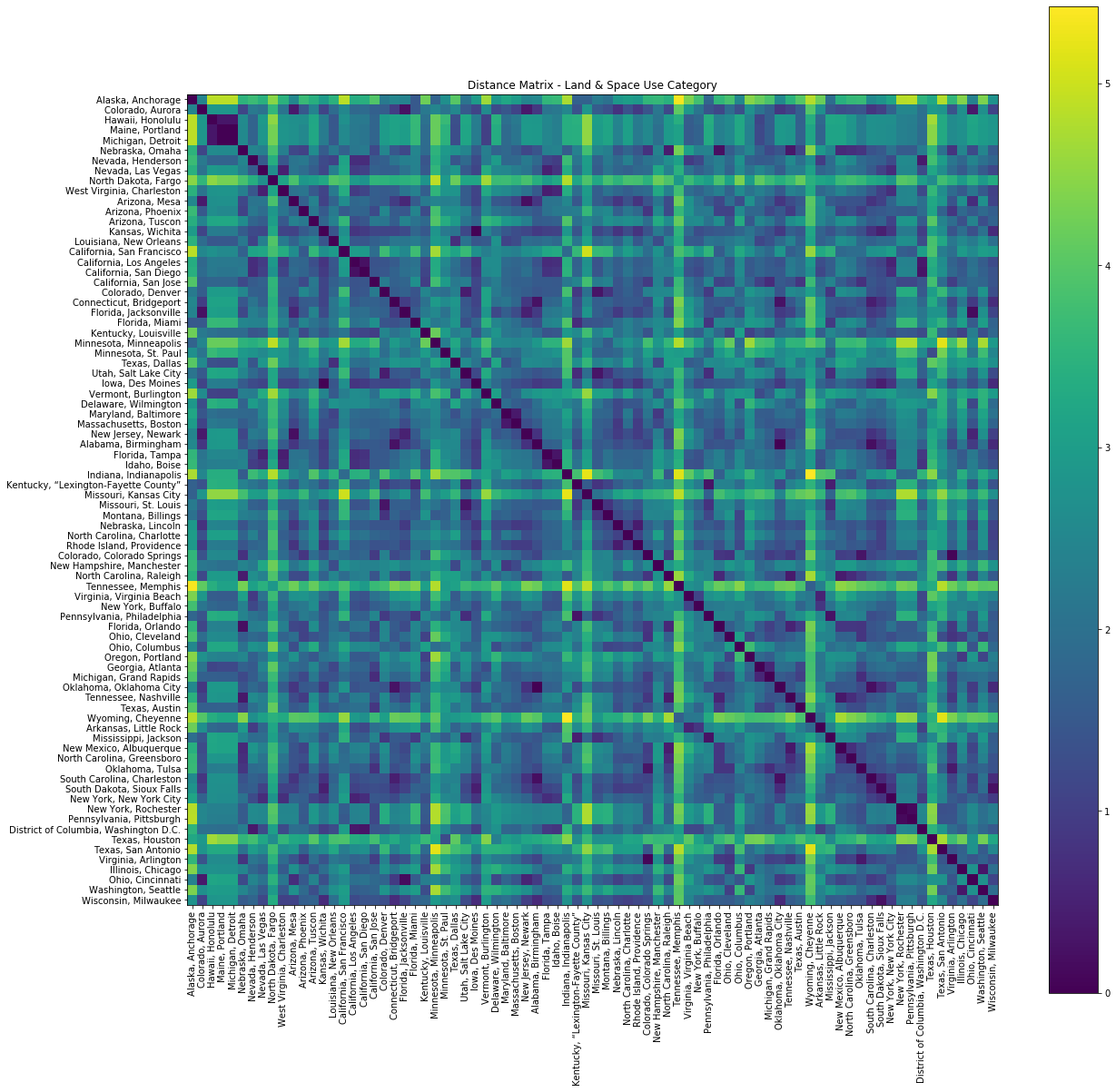
Before beginning a deep dive into the clustering, I wanted to test different algorithms to see which one produced the most accurate results. The two methods I was primarily interested in testing were KMeans clustering (a clustering method we've talked about in class) and Affinity Propagation (a different method of clustering that Scikit-learn can handle). KMeans is a nice introductory method of clustering because it simply calculates the Euclidean distances between different points. After assigning the number of clusters (the bulk of the "art" for this clustering method), you can get some interesting results. However, I believe that Affinity Propagation is a better clustering method for data that's as dynamic and varied as what's found in this particular data-set. Some indicators simply calculate number of procedures (ranging from values such as 3 to 13), while others are dollar values with massive distances ($10 to $100,000). Affinity Propagation easily allows for the inclusion of a weighting system, allowing data to be slightly more normalized. As a result, it is the method that I plan to use for all further clustering analysis.

### Clustering Raw Data vs Normalized Data

The Ease of Doing Business data is very dynamic, and as a result, even after switching to Affinity Propagation to use weighted values in clustering, there is still issues with the scales on some values. Take the 'Land and Space Use' category for example. In this category, there are some indicators with values that are averaged around 100 (with most values between 0 and 200). However, the Affinity Propagation method uses the Maximum Value when calculating the distance between points, and the distance matrix ends up looking something like this:

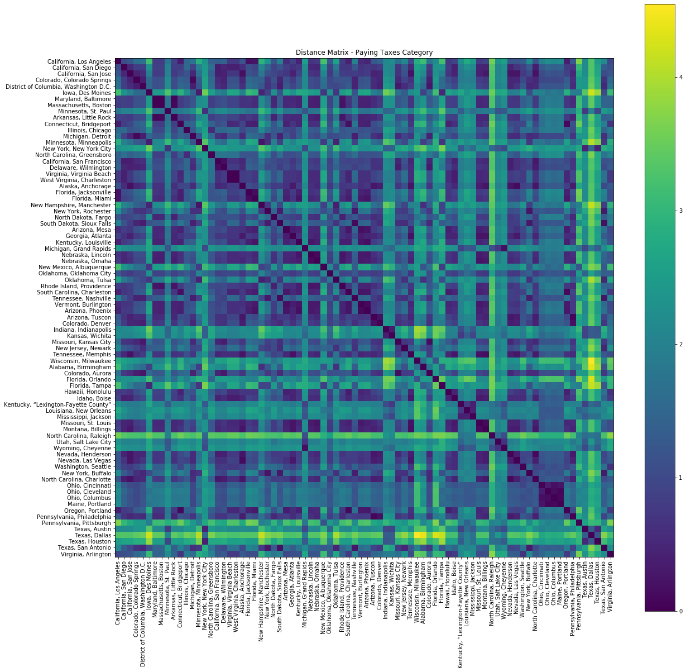
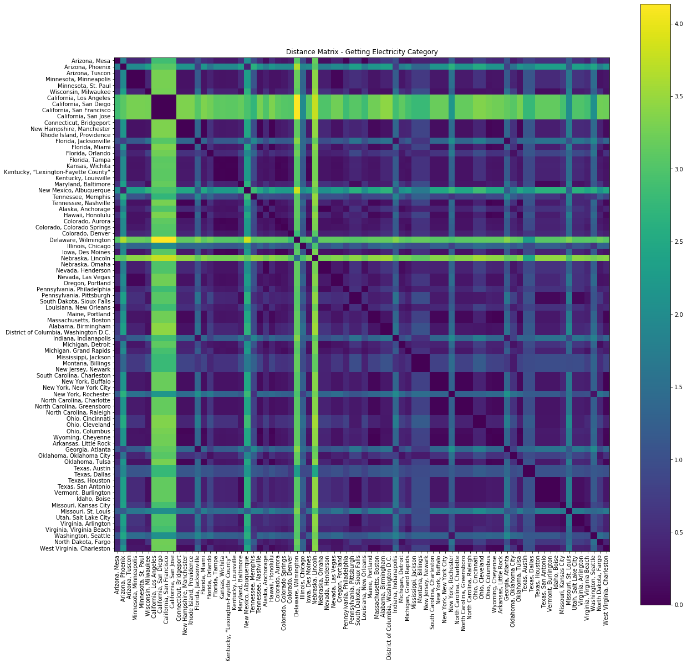
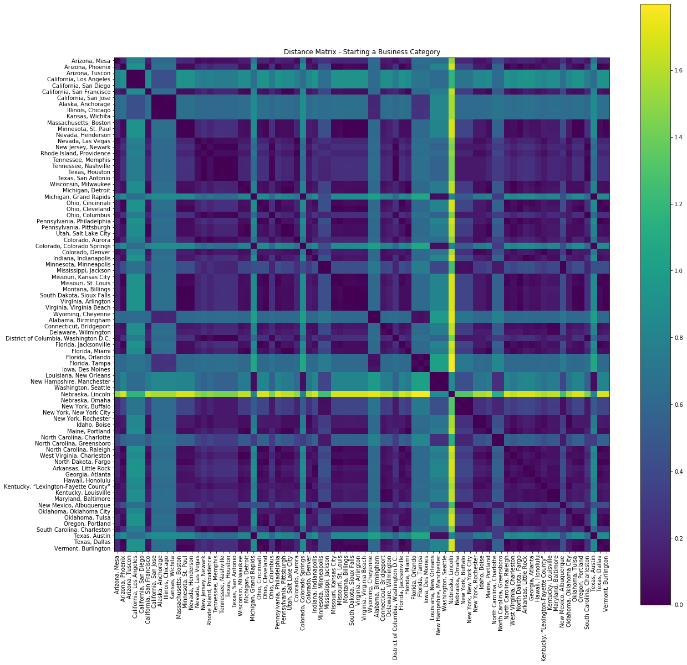


Using the Maximum Values has an obvious limitation in that outliers can have a dramatic impact on the scale in which the distance matrices are calculated. One of the most intuitive ways to fix this problem is to omit the outliers and adjust the scale in which Affinity Propagation is calculating values. Fortunately, we have already done this when answering Research Question #1. We have taken the data and applied a basic linear transformation to get the values to always be between 0 and 10 (where 0 is a bad outcome and 10 is a favorable outcome). When doing this linear transformation, we also removed outliers, so the data is not only more accurate but also easier to work with. When applying the same clustering technique on the **exact** same data (but after doing a simple transformation on the data), the results are a lot more interesting and digestible:



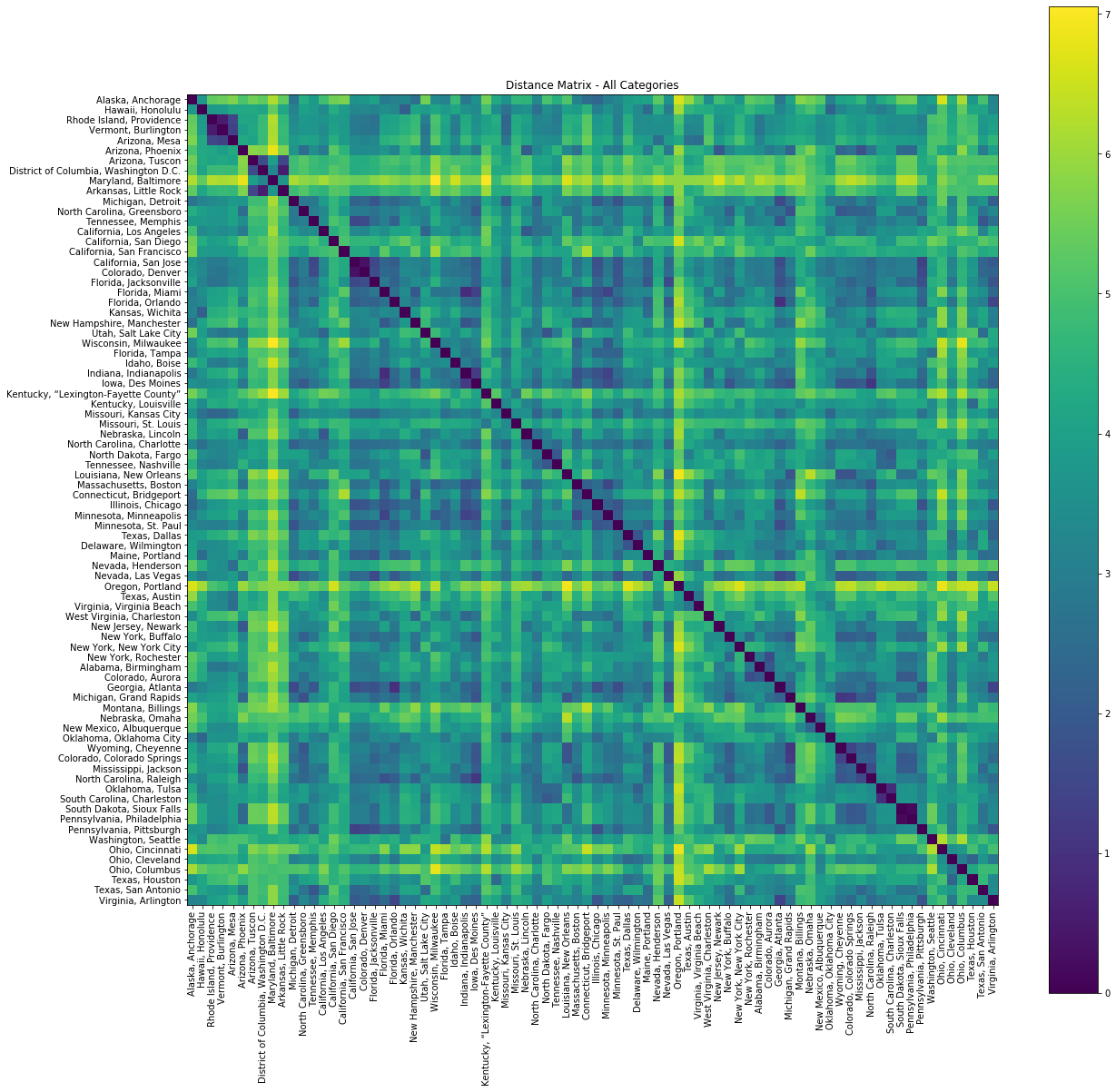
This graph produces results that are a lot more interesting. You can now begin to see some clusters forming within the distance matrix (areas with dark blue values), and begin to see which locations do not cluster well together at all (areas with yellow values). This process is then repeated for the other five categories (you can see those images [here](https://github.com/PaulBernert/DBNA/tree/master/images)).

##### Other Category Previews:

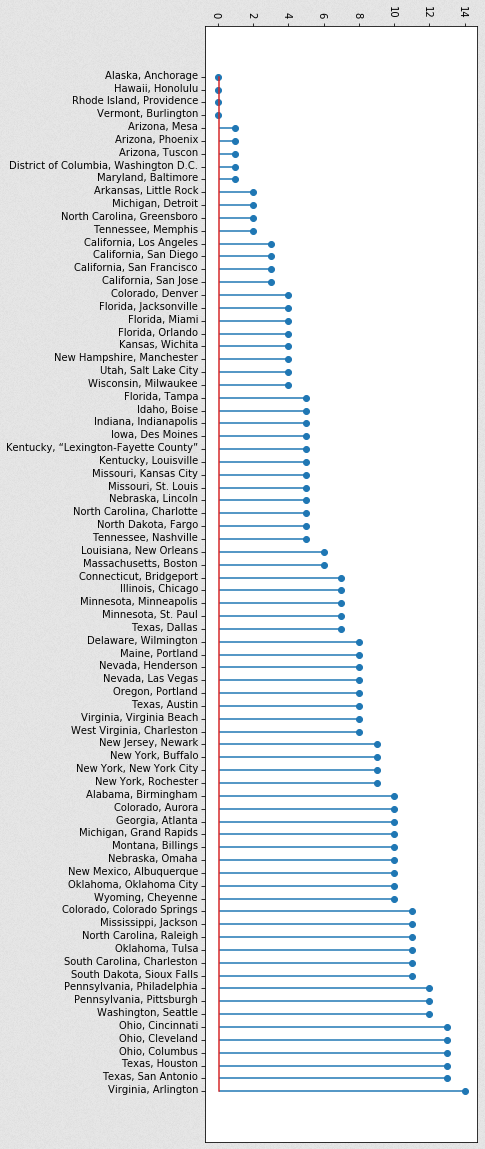


### Clustering Normalized Data Across ALL Categories

With every category individually clustered, the next test was to see whether clusters can be formed across all categories. The distances for each individual category generally had a small range (sometimes 0-1, sometimes 0-4), so my initial hypothesis was that the categories would be able to cluster (albeit some of the clustering would look forced due to high distances between points). However, after using Affinity Propagation to do clustering, these were the results:



I think the results are quite interesting. It's clear that clusters do indeed form, and the average distance between locales isn't as bad as initially predicted. The range of the distances goes from 0 to 7, which isn't absurd given the high variance for distances between some of the categories. Affinity Propagation produces the following clusters:

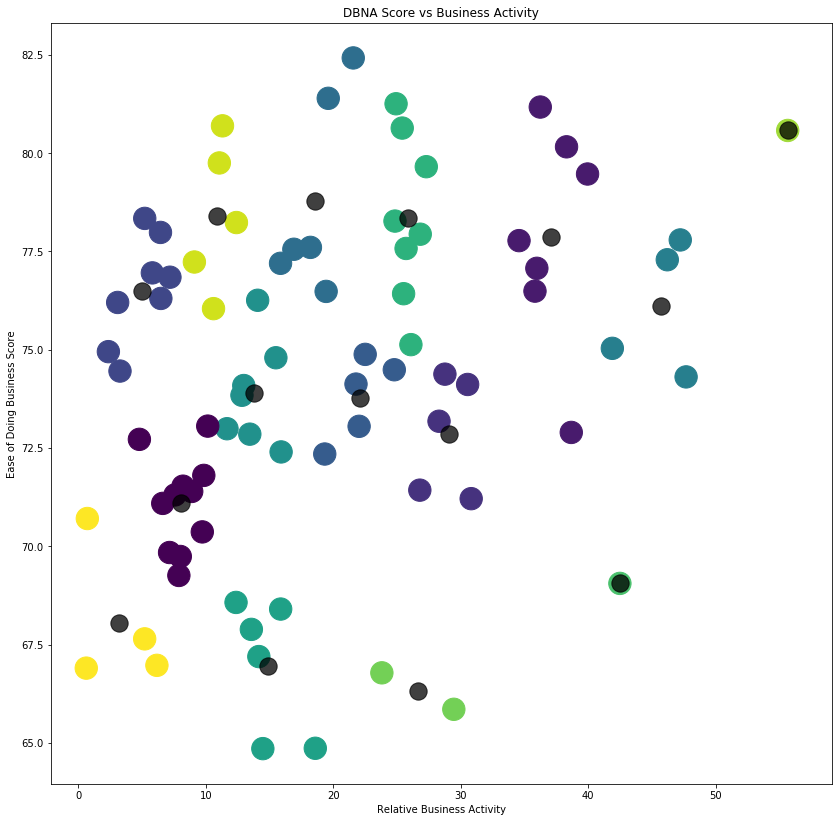


### Cluster Analysis

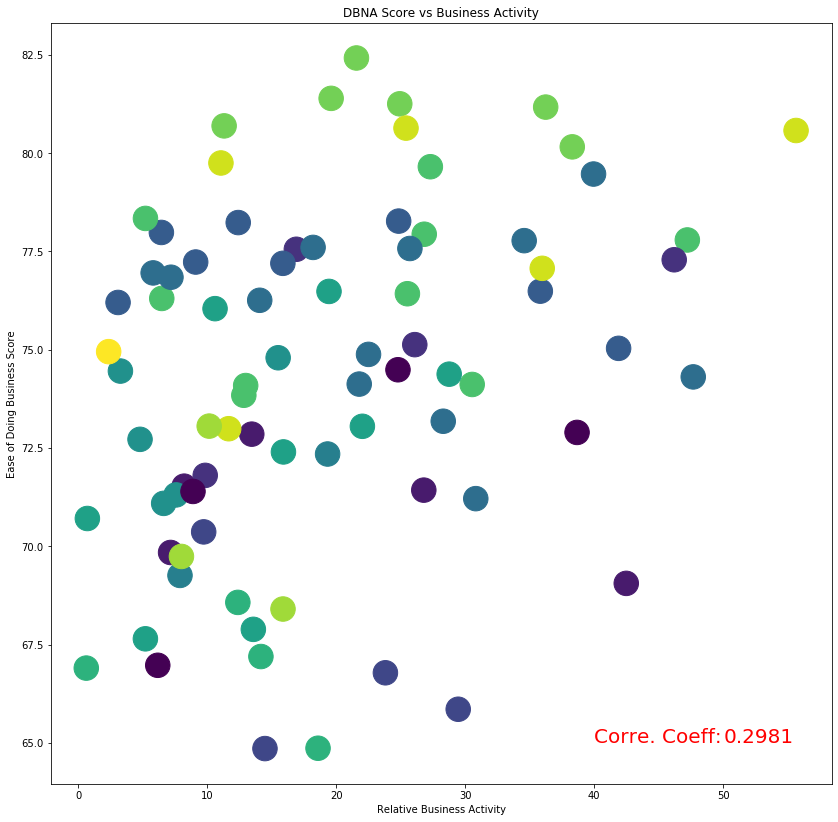
With the clusters assigned, one of the objectives was to solve whether there was a relationship between clusters and the 'Ease of Doing Business' Rank and Score. Ideally, the clusters are created around the 'Ease of Doing Business' Scores, where Cluster 1 contains cities with the highest Score, Cluster 2 the 2nd highest Scores, etc. Previous analysis confirmed that the DBNA data does indeed reflect where there is relatively higher amounts of business activity, so by transfer-ability, we should expect to see clusters form around the calculated Ranks and Scores. To test this, we need the following parameters:

1. Location (State, City)
2. 'Ease of Doing Business' Score
3. 'Ease of Doing Business' Rank
4. Cluster Number
5. Relative Business Activity

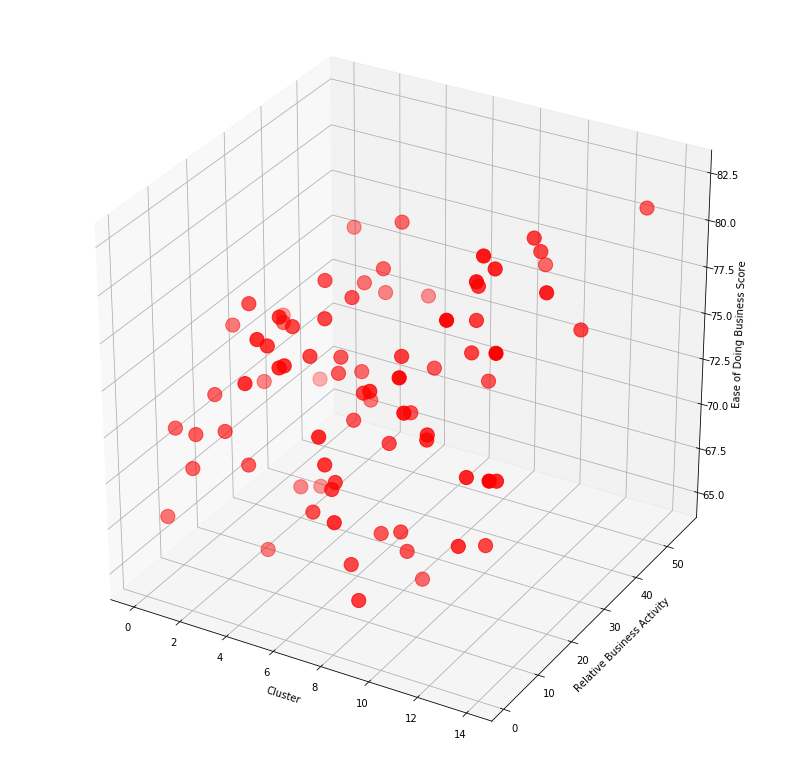
Using these five parameters, we now have everything to begin testing clustering methods. However, it is important to recognize that our clustering is around pre-determined groups based on the distance matrices calculated in the analysis above. It is **NOT** clustering around our previous research question, where we calculated points for the 'Ease of Doing Business' Score against Relative Business Activity. That clustering would produce these results:



While an interesting graph, it creates a new set of clusters that contradict and conflict with our previously-determined clustering. However, it does create the foundation for some analysis we can do using the correct clusters. We can continue to use our correct set of clusters to determine whether clusters form around ranks and scores like we initially anticipated. That results in the following visualization:



With this visualization, have reached the final results of analysis on this particular project. All further analysis is done using the results found here. Notice how it appears clusters are formed not around local dots, but across the horizontal spectrum (the top 10% of the graph is Green, the next 10% is Yellow, the following 10% is Dark Blue, etc.). This behavior indicates that clusters do indeed form around 'Ease of Doing Business' Scores. However, it is apparent that it's not a perfect correlation between Rank and Cluster. Observe the following two visualizations, using the exact same data but split into two different methods:



It's difficult to take three massive sections of data and assume that these things are the sole cause of one another, as there can simply be correlation without causation. Like with the previous research question, this is simply the beginning of the potential analysis that can be done with the Doing Business North America dataset, and I encourage the reader to take this dataset and conduct your own research to see what interesting results you can find!