**Optimal Locations to Build Additional Hospitals in New Jersey**

IBM Data Science Capstone Project Report

By: Michael Manian

## Introduction

### Background

Hospitals are an integral part to society. Hospitals, also referred to as medical centers, are health care institution that provide patient treatments with specialized medical, nursing staff and medical equipment. They serve to provide treatment and help cure illnesses all around the world. Over the years, technology and science have discovered over 10,000 diseases that could affect humans and most of those diseases do not have a cure. In order to survive global pandemics, there must be enough medical supplies and hospitals to take care of those affected. I have decided to do my project on this topic as it is an important real-life scenario and has an impact considering all that is going on across the world with the current COVID-19 pandemic.

### Problem

COVID-19 has proved to be one of the worst pandemics the world has faced. The United States especially got hit the hardest with this virus compared to other countries. Back in March, when the virus first hit the US, New Jersey was one of the many states that were affected the most. It got to the point where there were not enough hospital beds for those infected across the tristate. Hospitals were jam packed and nurses and doctors were overwhelmed and running out of resources. Taking into account the number of municipalities, population, income, and number of hospitals, this project aims to show where an addition of new hospitals would be most beneficial in the state of New Jersey. This not only will help in future pandemics, but also help on a regular basis in terms of distributing patients across different hospitals closest to them to reduce hospital wait time etc.

### Interest

Adding new hospitals in New Jersey would be of interest to the Department of Health and Human Services, considering they are the ones in charge of enhancing the health and well-being of all Americans through providing effective health and human services. It would also be of interest to any fresh doctors and nurses looking for a job.

## Data

### Data sources

I used the following data sources:

* Basic data about New Jersey, including counties and municipalities was scrapped from [**Wikipedia**](https://en.wikipedia.org/wiki/List_of_municipalities_in_New_Jersey).
* Population data per county as of early 2020 and median income data per county as of 2018 was also scrapped from [**new jersey demographics**](https://www.newjersey-demographics.com/counties_by_population), and [**Indexmundi**](https://www.indexmundi.com/facts/united-states/quick-facts/new-jersey/median-household-income#table), respectively.
* **Python’s Geopy library** was used to locate the geographical coordinates of each county. Geopy was not able to find two of the coordinates so I used **Google Maps** to manually get those.
* **Foursquare API** was used to gather all the hospitals located near each county.

### Data usage/cleaning

Data that was extracted from all three websites was scrapped and combined into one table for readability. This includes counties, municipalities, population, and income. The [Wikipedia](https://en.wikipedia.org/wiki/List_of_municipalities_in_New_Jersey) website had lots of excess information on its table such as municipal type, form of government, community established, and year incorporated. I dropped all these columns since they provided no value to solve the problem. The website also contained population data from 2010 and 2017, which I dropped and replaced with the 2020 data scrapped from the [new jersey demographics](https://www.newjersey-demographics.com/counties_by_population) website. The geographical coordinates for each New Jersey county was gathered from the Geopy library and appended to the final table. Two locations were not found so I manually had to append the latitude and longitude through searching on Google Maps. Next, hospital data was gathered from Foursquare. This included hospital latitude and longitude as well as the hospital name. Foursquare ended up gathering hospitals from New York, Pennsylvania, and Delaware since they are relatively close to New Jersey. Since I am only focused on New Jersey, I dropped the hospitals that were from a different state. Once all the New Jersey hospitals were gathered, I was able to get the number of hospitals per county to see which counties have more hospitals and which have less. Out of all this, only the number of counties, population, income, and the number of hospitals per county were used to solve the problem. Municipality names and hospital names are categorical variables and do not impact this analysis, so they were omitted. Data was then normalized and grouped into clusters using machine learning algorithms to determine a correlation between each other. The goal of this is to locate counties that did not have many hospitals and correlate that to the population and income in that area, and the overall density of how many towns reside in the county. Finally, this data will be used to inference a solution to the problem stated and figure out the ideal locations for building additional hospitals.

## Methodology

### Basic data analysis

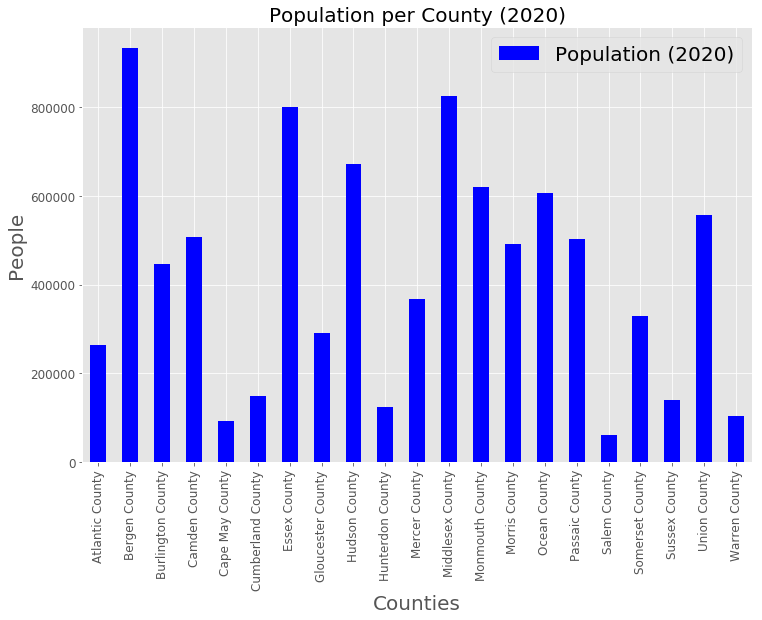
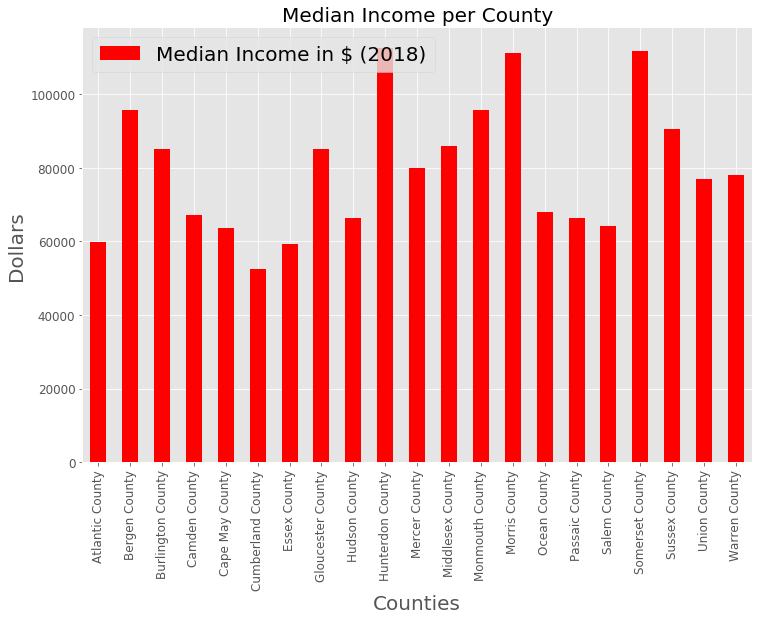
Before delving into all the advanced machine learning techniques used, I made four basic bar charts to depict each variable per county. Figure 1 shows the population as of early 2020 for each county. Figure 2 shows the median income per county based on the 2018 Census. Figure 3 shows the number of hospitals per county based on Foursquare data. And finally, Figure 4 shows the number of municipalities (towns) per county.

Figure 1

Figure 2

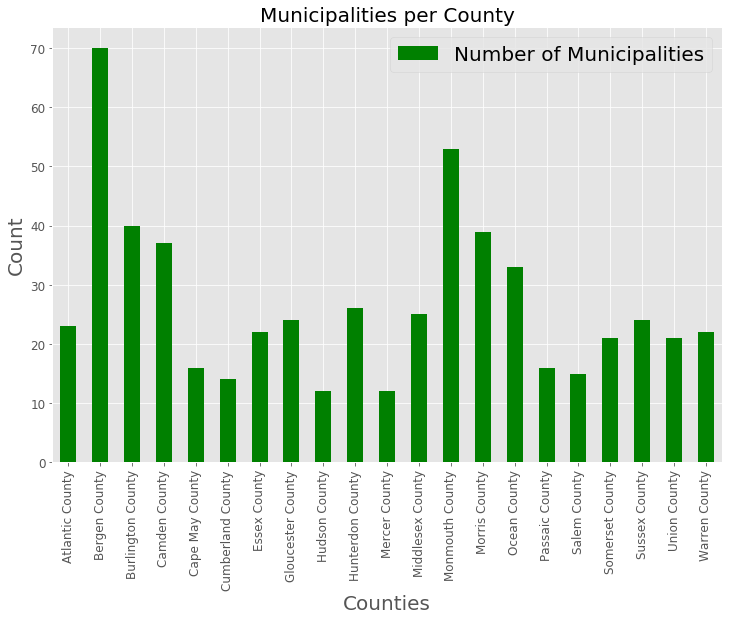
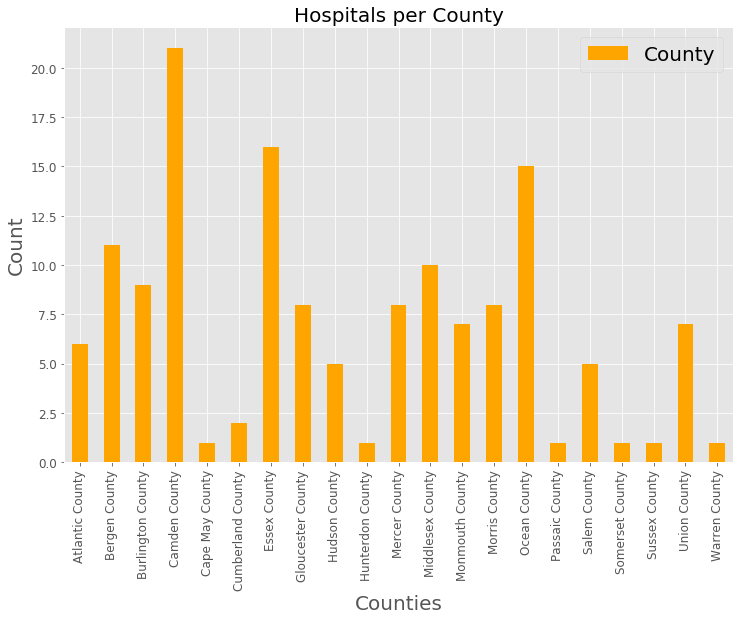
 Foursquare API returned many hospitals some of which were also not in New Jersey. It also returned many duplicates since the towns in NJ are relatively nearby. Duplicates and out of state hospitals were dropped from my table. To figure out in fact which hospital belongs to which county, I used a distance metric to calculate the distance between each hospital and every county in NJ. A simple one-line SQL command was used to cross join (Cartesian product) the Dataframe with the 21 counties and their geographical coordinates and the Dataframe with the 144 hospitals and their geographical coordinates returned by Foursquare. Then I was able to calculate the distance and then sort the Dataframe to pick only the shortest distance to each hospital to a county to determine which hospital belongs to which county.

Figure 4

Figure 3

Now that I have gathered the four pieces of critical data (Population, Income, Number of Municipalities, Number of Hospitals), I can proceed with advanced data analysis. I decided to segment the data into pairs to see the correlation between each variable. Based on that I can inference a solution to the problem.

### Machine learning techniques

For analyzing the data, I had to use machine learning tools. All the variables in the data (population, income, number of municipalities, number of hospitals) are all quantitative variables, in other words they are numerical. Since these features are relatively inconsistent with each other, I had to normalize the data. Normalizing makes model training less sensitive to the scale of features and in turn leads to a more accurate model. I used Python’s sklearn preprocessing and StandardScalar libraries to do this. The data is standardized through scaling, fitting, and transforming the data segments. The result is a normalized dataset which could then be used for classification and clustering. To cluster the data into groups, K-means algorithm was used. This is an unsupervised machine learning model that is a method of vector quantization which partitions n observations into k clusters. To determine the optimal number of k clusters, the silhouette score method was used. Determining the optimal k is important because it will lead to more accurate results. The silhouette score measures how similar a point is to its cluster versus the next closest cluster. The higher the score, the better.

### Correlation between Population and # of Municipalities

I hypothesized that the lower the population the less towns there are in the county and vice versa. Figure 5 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=4, which means 4 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 4 groups and display the clusters in the form of a scatter plot as shown in Figure 6. Based on the plot we can see that my hypothesis is correct since there is a strong positive relationship between the two variables. The cluster in yellow seems to be an outlier group.

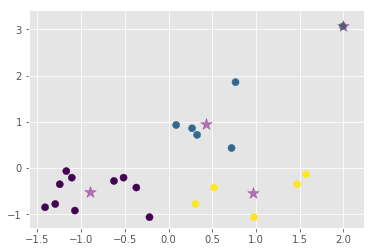


Figure 6

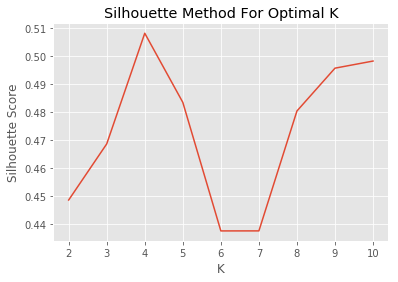


Figure 5

### Correlation between Median Income and # of Municipalities

I hypothesized that there would not be much of a correlation between the income and number of towns. Figure 7 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=5, which means 5 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 5 groups and display the clusters in the form of a scatter plot as shown in Figure 8. It turns out there is a slight linear trend to my surprise, with the cluster in light blue as the outlier group. It appears that in counties with less municipalities, resides people with low income.

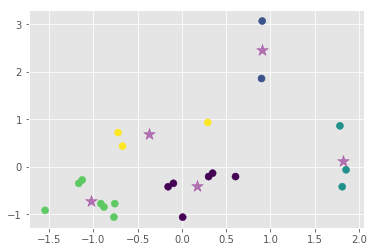


Figure 8

### Correlation between # of Hospitals and # of Municipalities

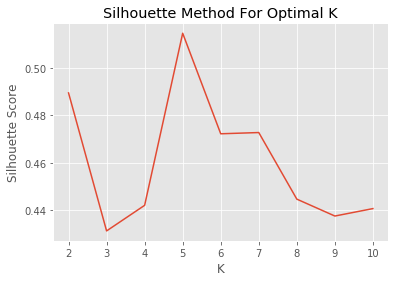


Figure 7

I hypothesize that there is a strong correlation between number of hospitals and number of towns in a county. The counties with many towns would likely have much more hospitals than those without. Figure 9 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=3, which means 3 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 3 groups and display the clusters in the form of a scatter plot as shown in Figure 10. Based on the plot we see my hypothesis is correct as there is a linear relationship. The cluster in purple seems to be an outlier group.

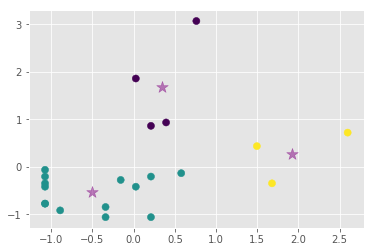


Figure 10

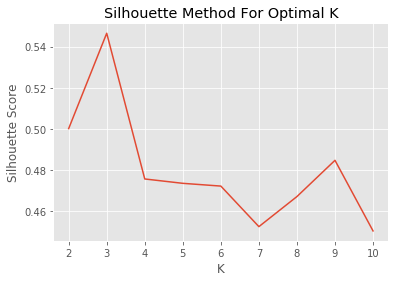


Figure 9

### Correlation between Population and # of Hospitals

I hypothesize that the higher the population the greater number of hospitals and vice versa. It makes sense since a county with lot of people would need access to more medical supplies in case of a pandemic for example. Figure 11 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=6, which means 6 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 6 groups and display the clusters in the form of a scatter plot as shown in Figure 12. Based on the plot there seems to be a slight positive linear relationship, but it is hard to notice at first.

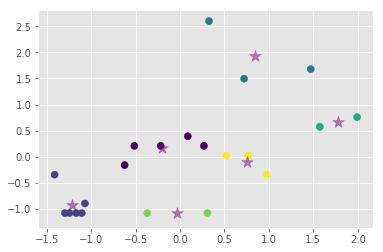


Figure 12

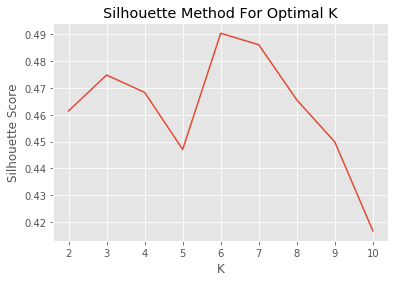


Figure 11

### Correlation between Median Income and # of Hospitals

I hypothesize that there is no correlation between income and number of hospitals. Figure 13 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=5, which means 5 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 5 groups and display the clusters in the form of a scatter plot as shown in Figure 14. Based on the plot, we can see that there is not really any relationship. If anything, there seems to be a neutral horizontal relationship with the light blue cluster being the outlier group.

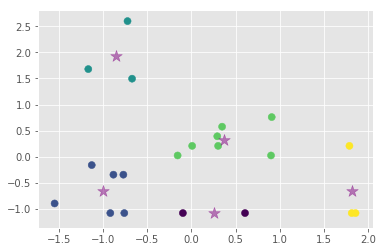


Figure 14

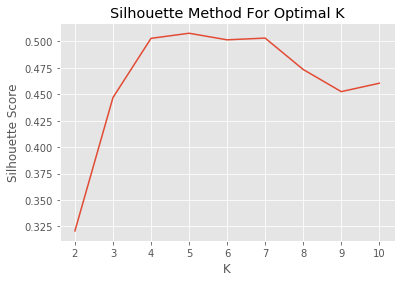


Figure 13

### Correlation between Population and Median Income

I hypothesize that there is no correlation between population and income. Figure 15 shows the graphical representation of the silhouette score so we can see the appropriate value for k. As we can see, the highest score is at k=5, which means 5 clusters is optimal for this dataset. Finally, we use K-means to cluster the dataset into 5 groups and display the clusters in the form of a scatter plot as shown in Figure 16. Based on the plot my hypothesis is correct. The plot is hard to read but it looks like the yellow and purple clusters are outlier groups. Or the data could very well be unrelated which is why the points looks scattered.

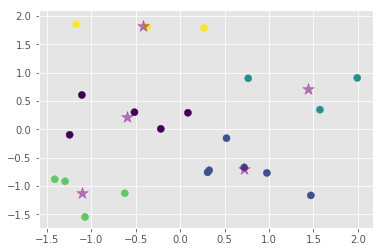


Figure 16

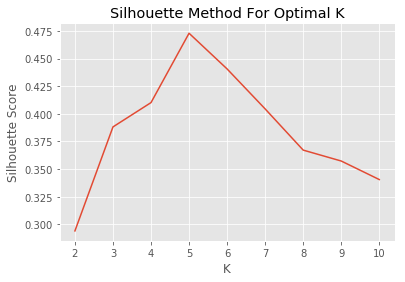


Figure 15

## Results

### Analysis of Population and # of Municipalities

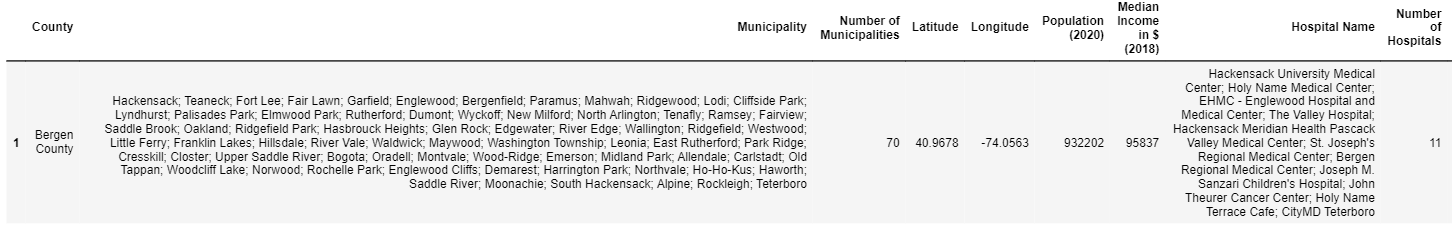
As we saw this combination had 4 clusters:



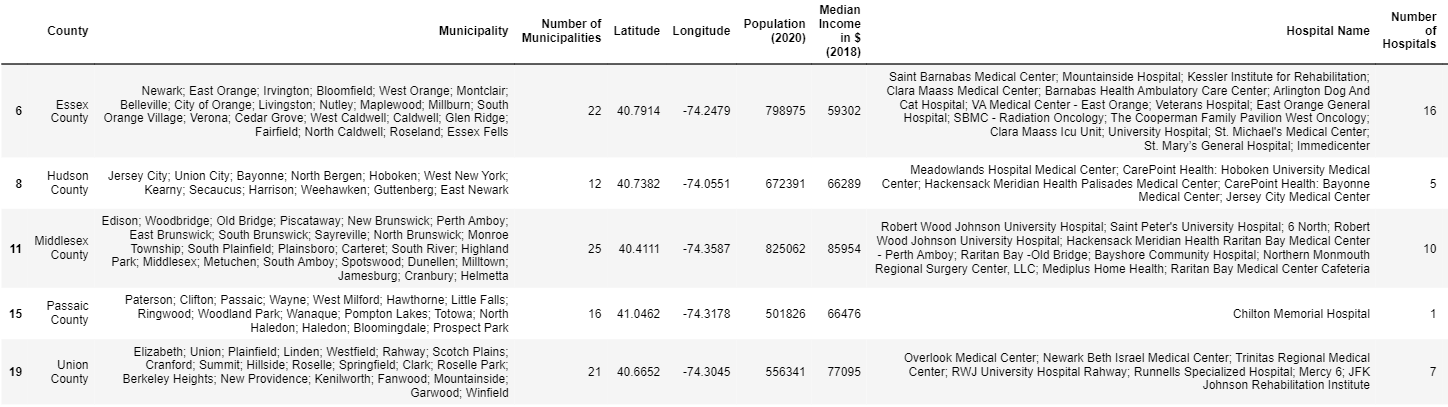
Cluster 1 above contains 10 counties. The counties with the least number of municipalities and the least population were grouped together.



Cluster 2 above contains 5 counties. The counties with an average amount of municipalities and population were grouped together. Not the most but also not the least.



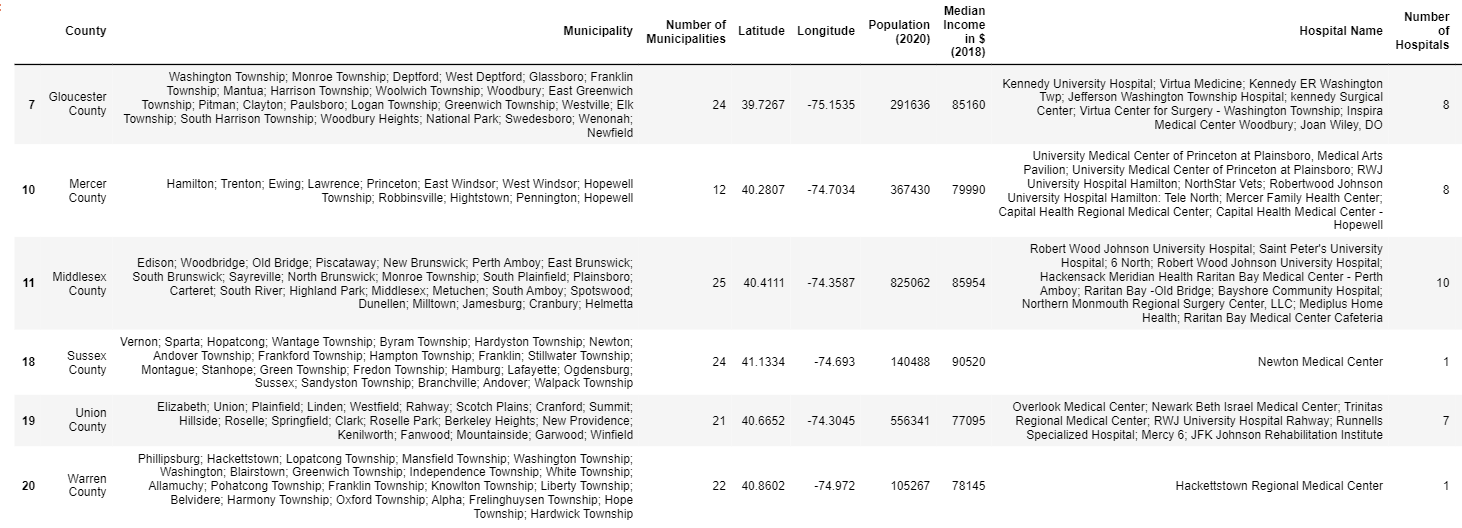
Cluster 3 above contains 1 county. This county has the most municipalities in New Jersey and the highest population.



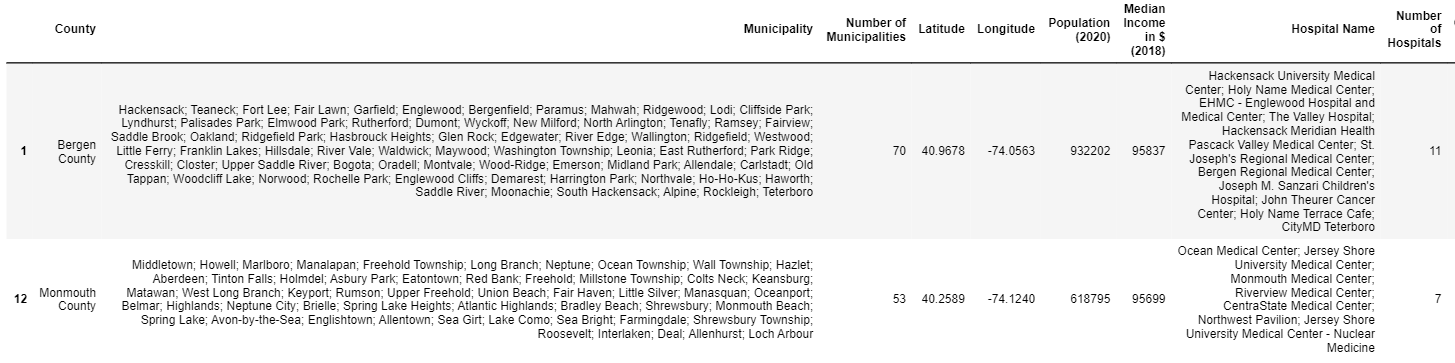
Cluster 4 above contains 5 counties. This appears to be an outlier cluster. The counties here have a lot of population but not a lot of municipalities.

### Analysis of Median Income and # of Municipalities

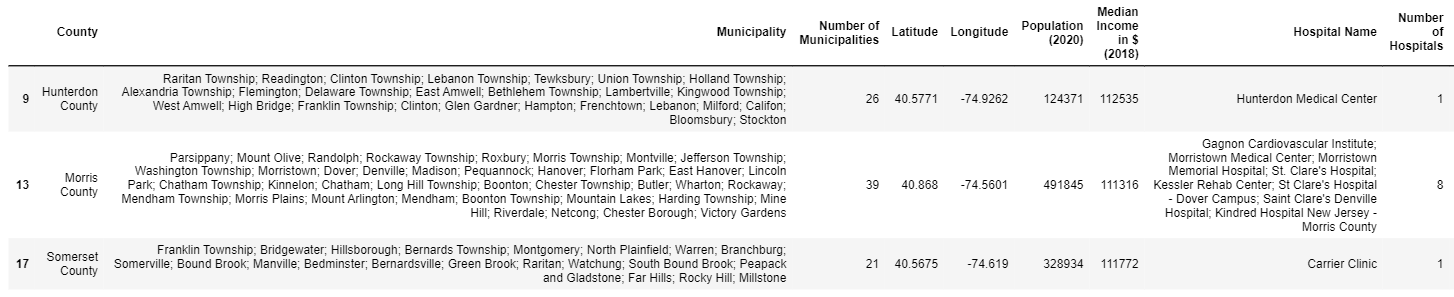
As we saw this combination had 5 clusters:



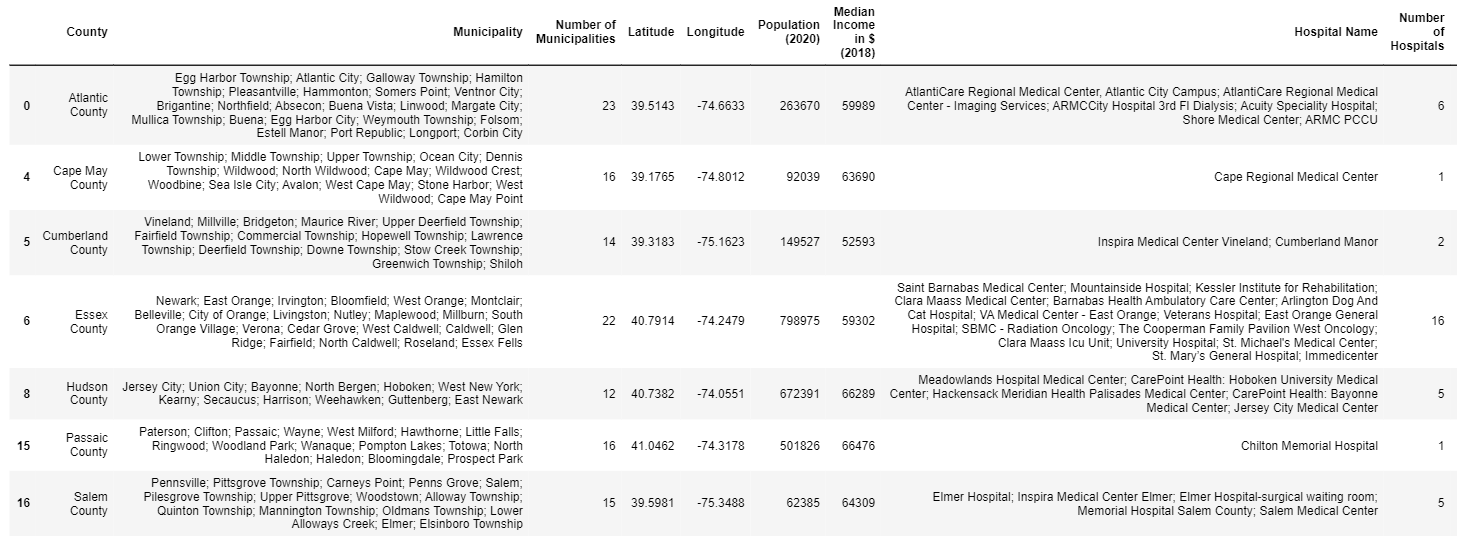
Cluster 1 above contains 6 counties. The counties with a high income and average number of municipalities were grouped together.



Cluster 2 above contains 2 counties. The counties with the highest number of municipalities and a high income were grouped together.



Cluster 3 above contains 3 counties. The counties with the highest income and an average number of municipalities were grouped together. This is the outlier group.



Cluster 4 above contains 7 counties. The counties with the lowest income and lowest number of municipalities were grouped together.



Cluster 5 above contains 3 counties. The counties have an average amount of municipalities and an average amount of income. Not the most but also not the least.

### Analysis of # of Hospitals and # of Municipalities

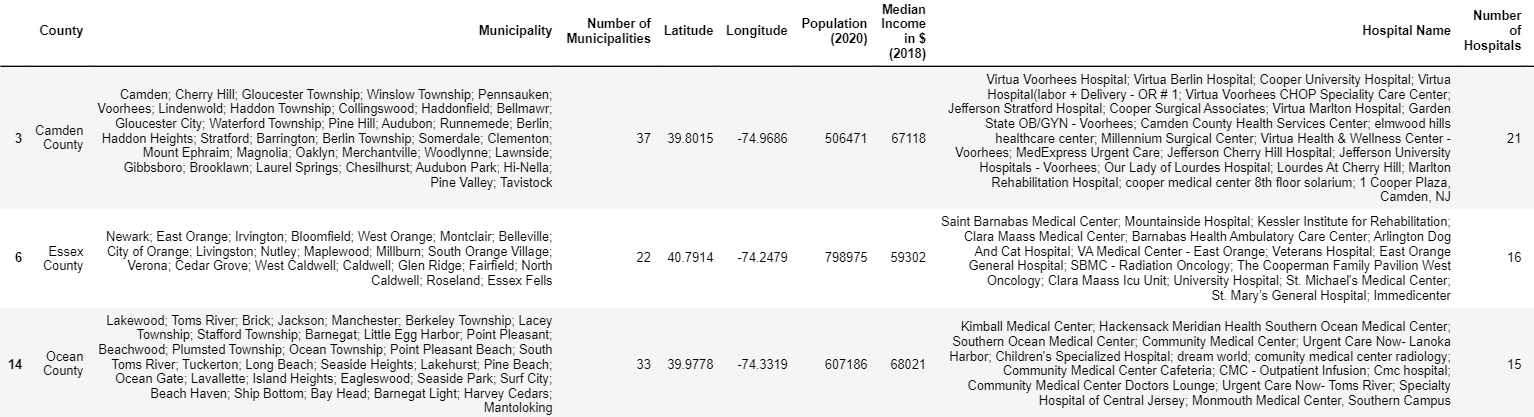
As we saw this combination had 3 clusters:



Cluster 1 above contains 4 counties. The counties with the greatest number of municipalities and a lot of hospitals were grouped together. This could be an outlier group because the municipality to hospital ratio does not add up.



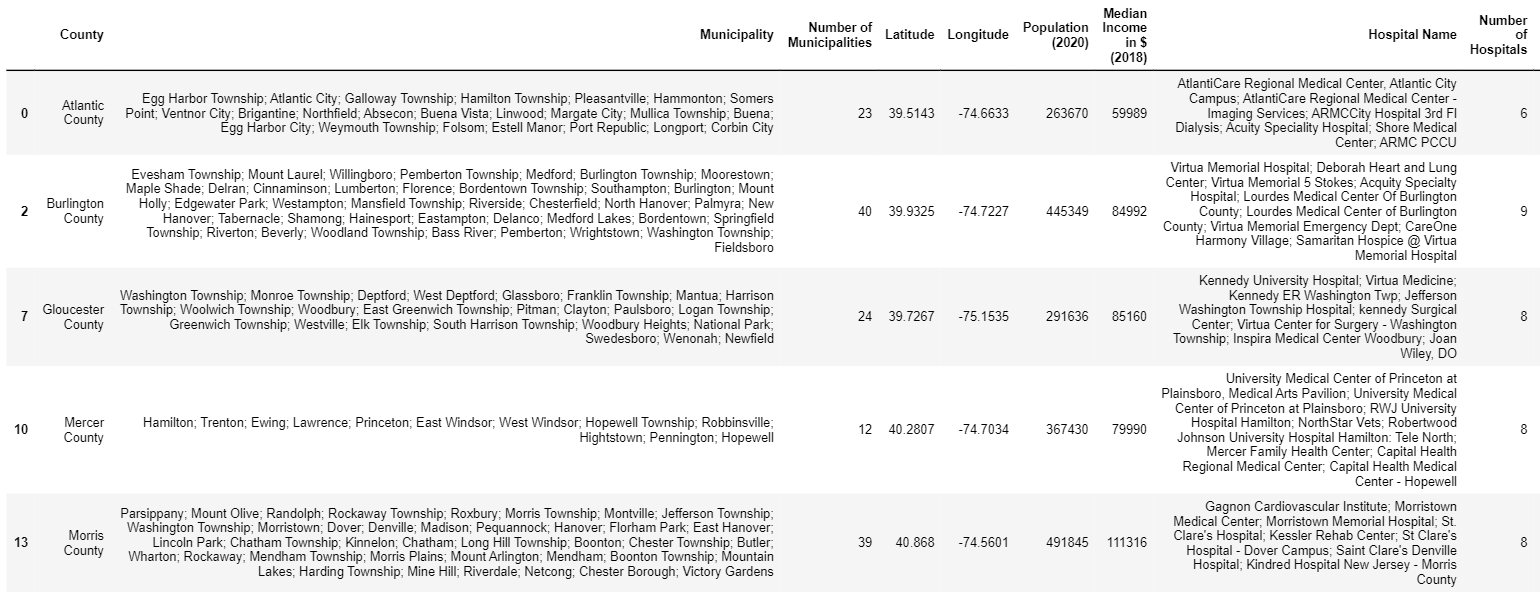
Cluster 2 above contains 14 counties. The counties with the least number of municipalities and the least number of hospitals were grouped together.



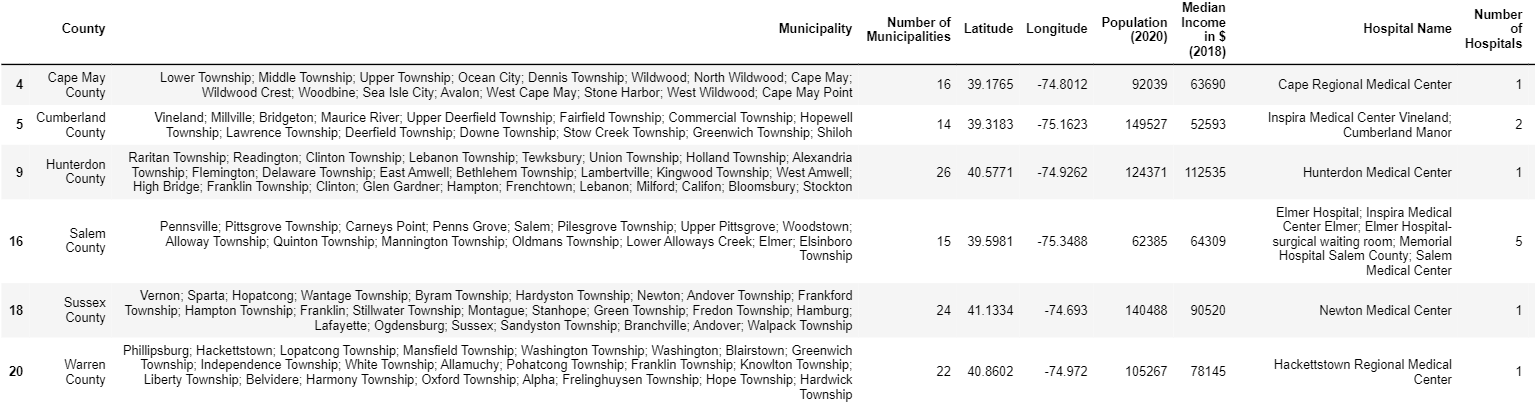
Cluster 3 above contains 3 counties. The counties with an average number of municipalities and the greatest number of hospitals were grouped together.

### Analysis of Population and # of Hospitals

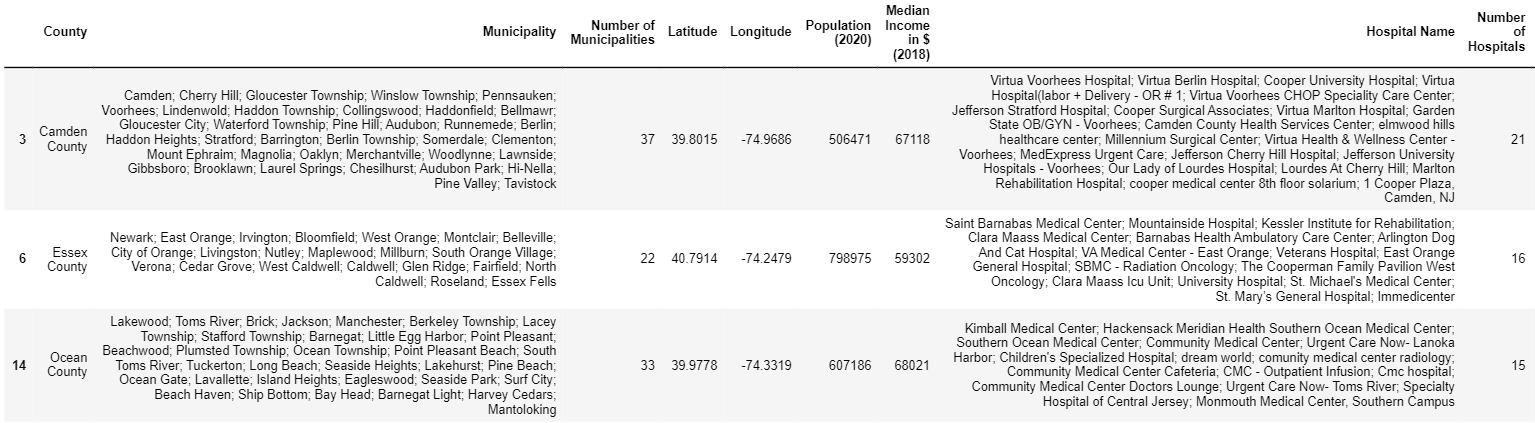
As we saw this combination had 6 clusters:



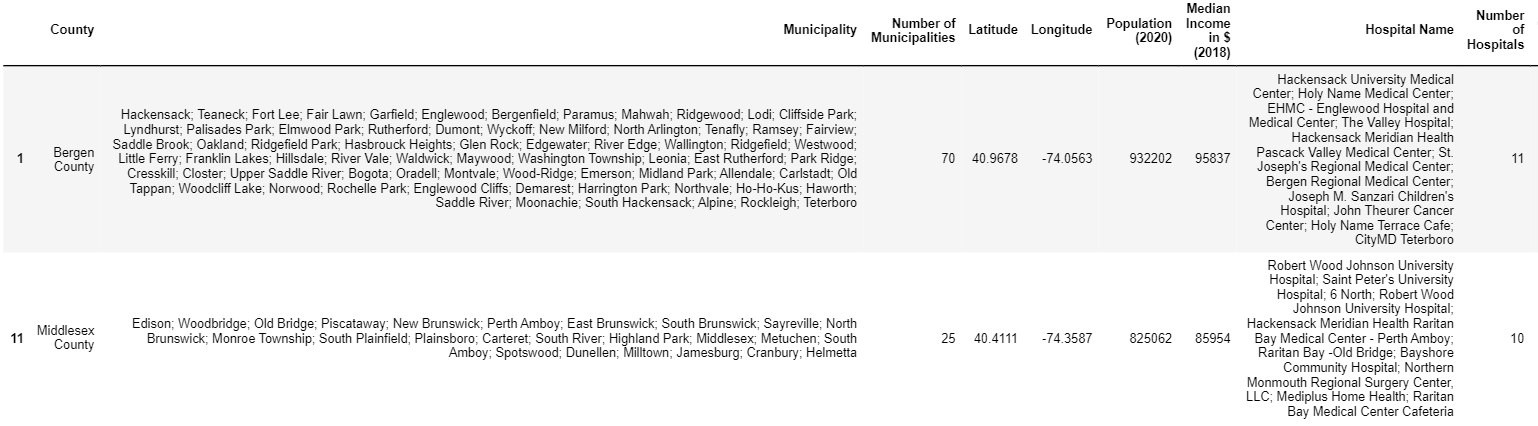
Cluster 1 above contains 5 counties. The counties with an average population and an average number of hospitals were grouped together. Not the most but also not the least.



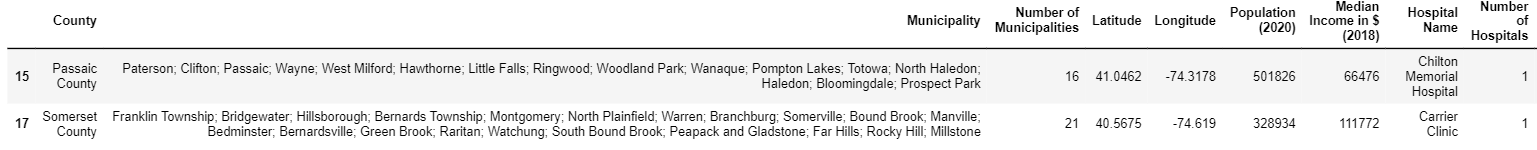
Cluster 2 above contains 6 counties. The counties with the lowest population and the least number of hospitals were grouped together.



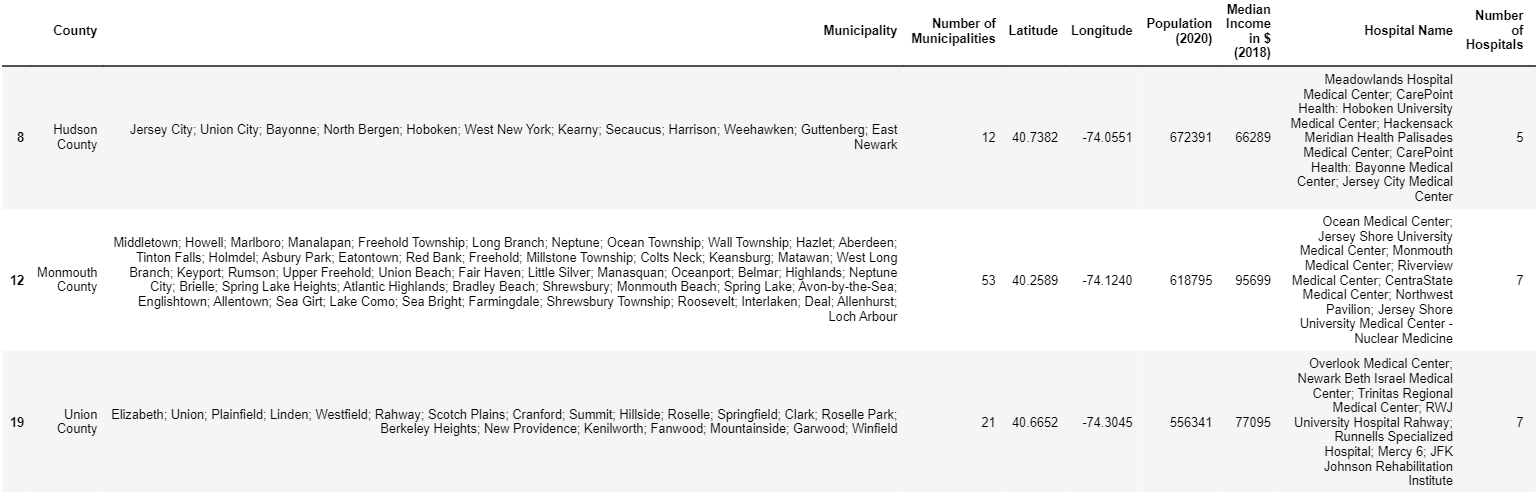
Cluster 3 above contains 3 counties. The counties with a lot of population and the greatest number of hospitals were grouped together.



Cluster 4 above contains 2 counties. The counties with the most population and a lot of hospitals were grouped together.

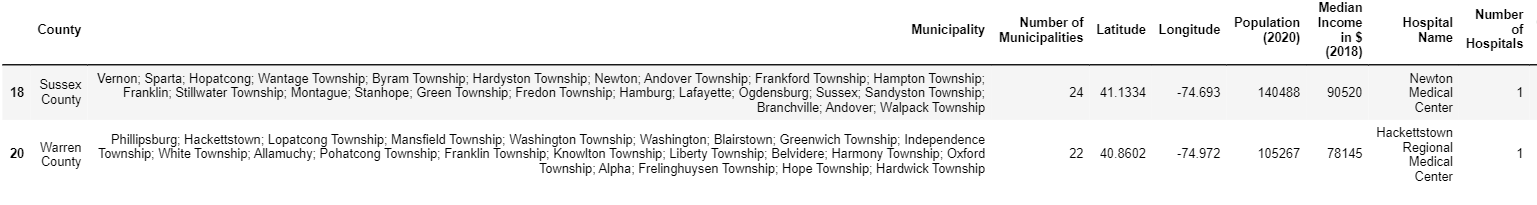


Cluster 5 above contains 2 counties. The counties with an average population and the lowest number of hospitals were grouped together. This could be a possible area where additional hospitals could be built.

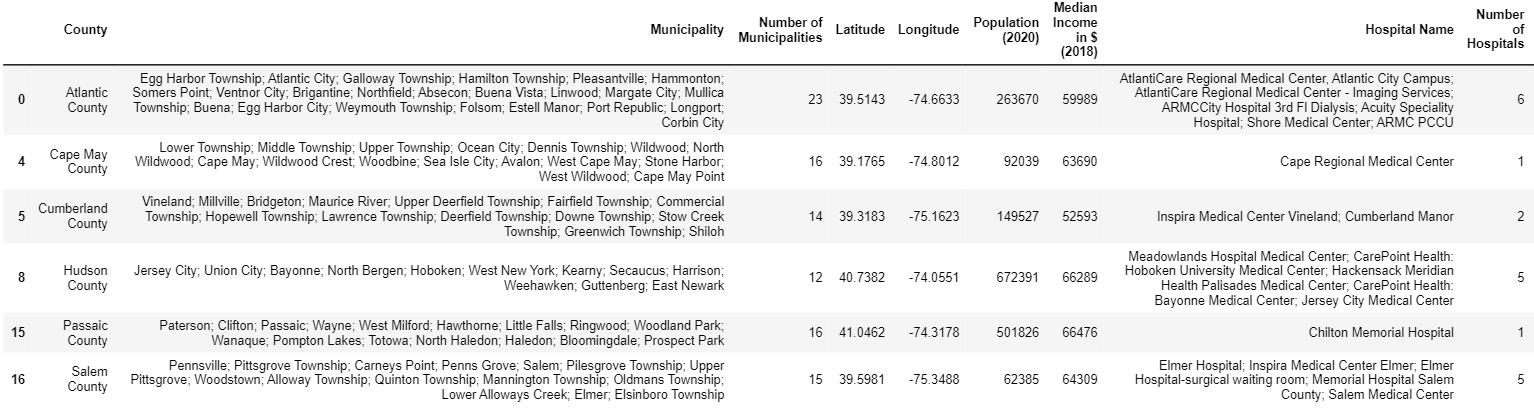


Cluster 6 above contains 3 counties. The counties with a lot of population and an average number of hospitals were grouped together.

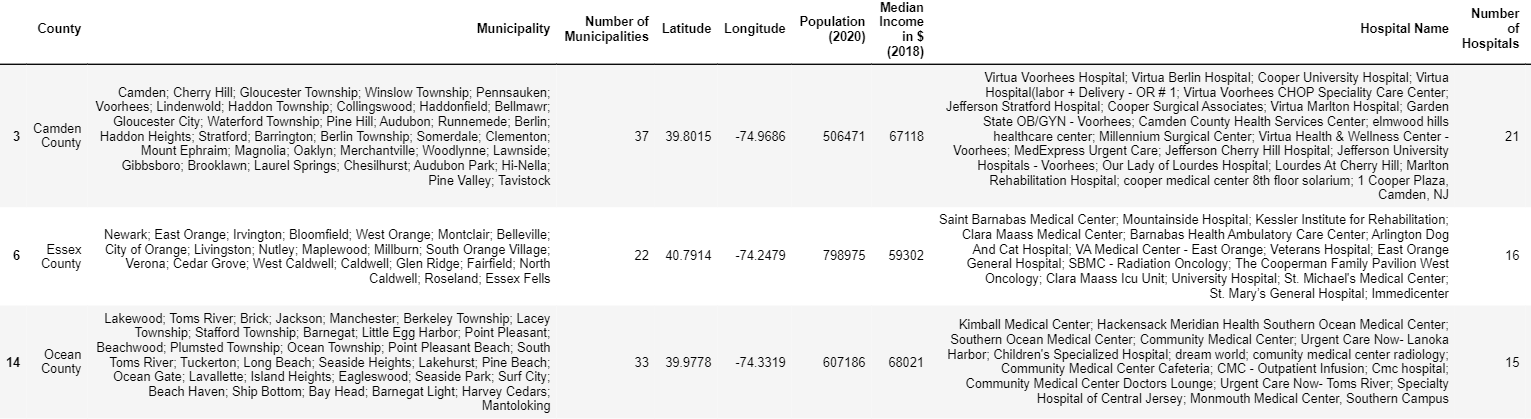
### Analysis of Median Income and # of Hospitals

As we saw this combination had 5 clusters:   


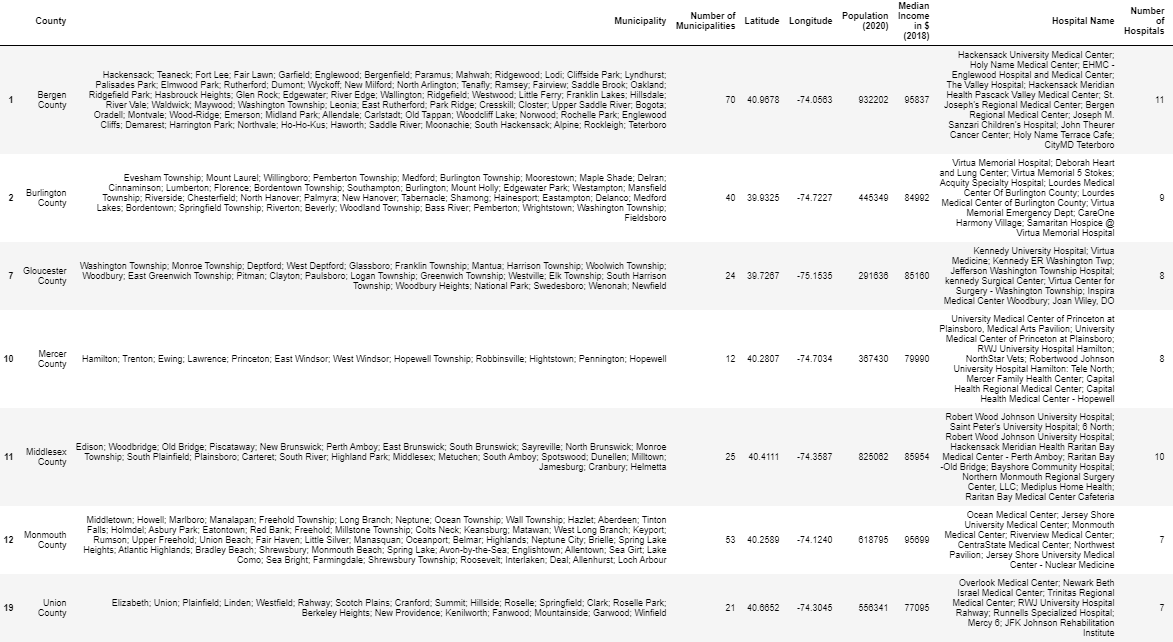
Cluster 1 above contains 2 counties. The counties with an average amount of income and the least number of hospitals were grouped together.



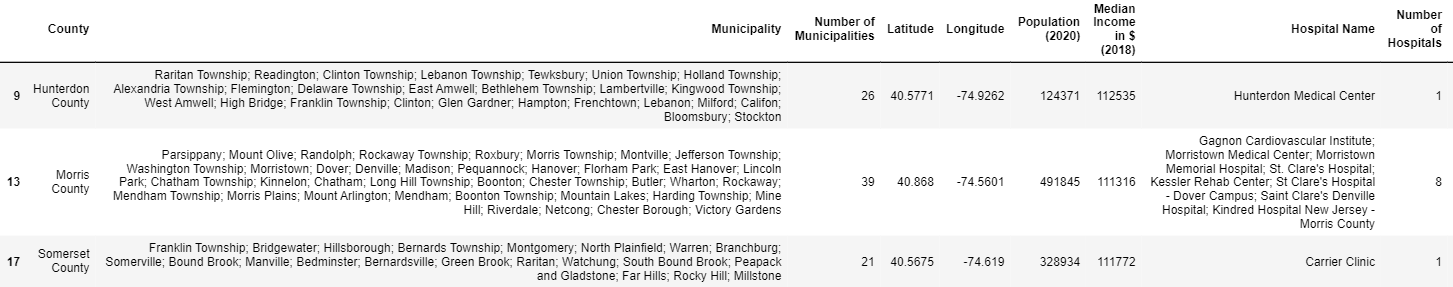
Cluster 2 above contains 6 counties. The counties with an average income and a low number of hospitals were grouped together.



Cluster 3 above contains 3 counties. The counties with a low income and the greatest number of hospitals were grouped together. This is an outlier group.



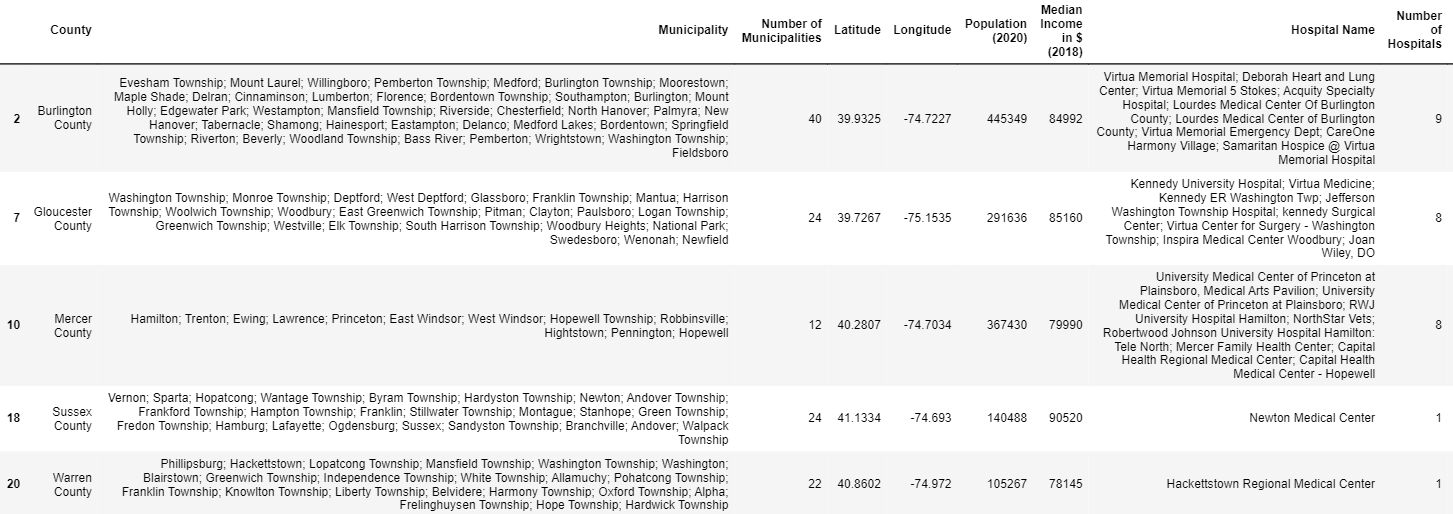
Cluster 4 above contains 7 counties. The counties with a lot of income and a lot of hospitals were grouped together.



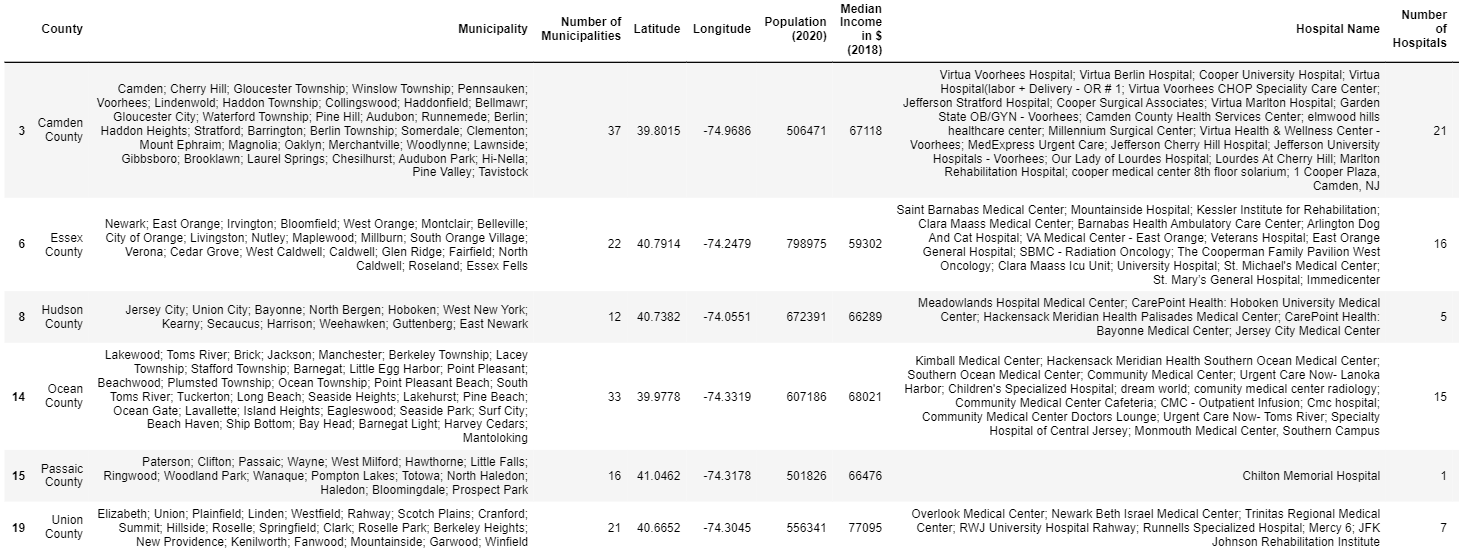
Cluster 5 above contains 3 counties. The counties that have the most income and least number of hospitals were grouped together. This could also be an outlier group.

### Analysis of Population and Median Income

As we saw this combination had 5 clusters:



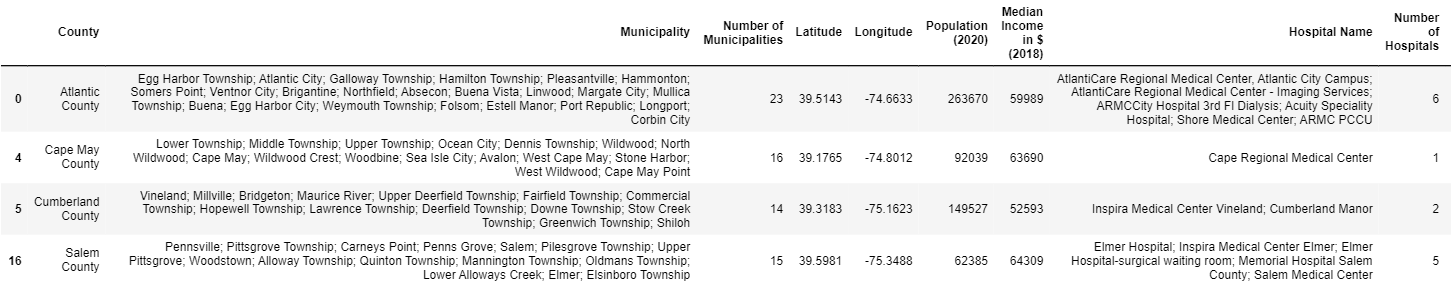
Cluster 1 above contains 5 counties. The counties have an average to low population and high income. This could be an outlier group.



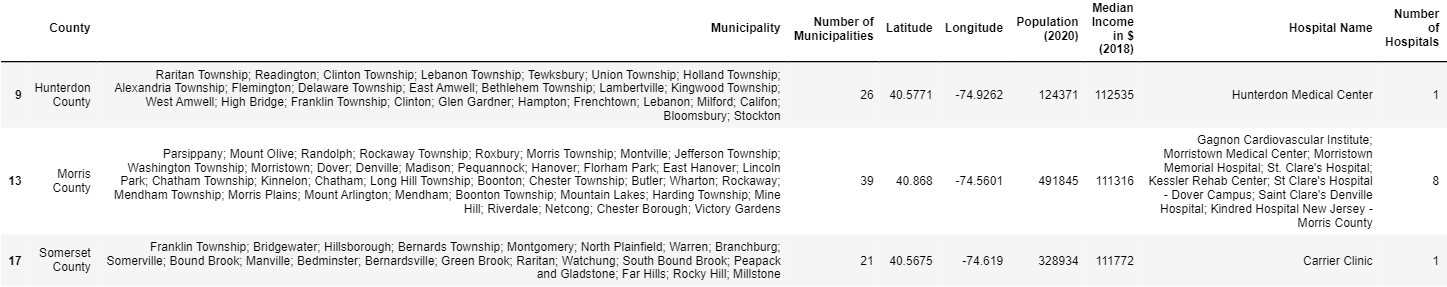
Cluster 2 above contains 6 counties. The counties with an average population and an average income were grouped together.



Cluster 3 above contains 3 counties. The counties with a high population and a high income were grouped together.



Cluster 4 above contains 4 counties. The counties with the lowest population and the lowest income were grouped together.



Cluster 5 above contains 3 counties. The counties with a low population and the most income were grouped together. This could be an outlier group.

## Discussion

Looking back at the observations, we can see that the higher the population the higher the number of municipalities and vice versa. We also saw that lower income families tend to live in counties with not as many towns and middle-class families live in counties with many towns. Also, there are more hospitals in counties with more towns. This makes perfect sense since there would be a higher population in those towns too. I also noticed that the lower incomed counties had the most hospitals which I found interesting because hospital and medical bills are very expensive. But at the same time, I also found that higher incomed counties also have a lot of hospitals, granted not as much as the other counties. This made me realize the one big limitation in this project, that is the Foursquare API. Foursquare most likely did not locate all the hospitals in New Jersey simply because sufficient data was not entered into its systems. We can also conclude that population and income have no correlation in this situation as proven by the scatter plot and the data cluster tables. That brings us now to the main question in this project, that is which locations would benefit from additional hospitals. After analyzing the data and all the outliers the charts and tables produced, I have come up with solution. I compared the outliers for each different combination to come up with the top 5 counties. The recommended counties that I would suggest would greatly benefit from an addition of a new hospital is Hunterdon County, Morris County, Somerset County, Passaic County, and Middlesex County. Middlesex County only has 10 hospitals, yet it is the 2nd most populated county in New Jersey. It also consists of 25 municipalities which is more than average. Passaic County only has 1 hospital and 15 municipalities. It also has over 500,000 people living in the area which is more than average. The median income for people living in this county is on the lower side so I would recommend building a charity hospital here to help aid those who cannot afford the cost. Somerset County only has 1 hospital and over 300,000 people living in the area. It also has 21 municipalities. That is a lot of people and towns for only having one hospital. Morris County has 39 municipalities, around 500,000 people living in the area yet only has 8 hospitals. Finally, Hunterdon County has only 1 hospital and has 26 municipalities. Although its population is roughly 130,000, one hospital is not enough to share amongst 26 different towns.

## Conclusion

All in all, in this study I was able to analyze data from various sources and compare the individual relationship with each other to determine any correlations to come up with ideal locations to build additional hospitals. Machine learning techniques such as clustering and normalization were used to help achieve the results. This model can greatly help New Jersey prepare for future pandemics and even daily! This model can be improved by implementing some socioeconomic data as well as data from the COVID-19 pandemic, but they were more difficult to extract. Also analyzing the towns rather than the counties would have provided a better model however there are 565 towns in New Jersey which would have been difficult to implement especially with the limitations of Foursquare and Geopy. With this additional data and more accurate hospital readings/coordinates, we would be able to visualize a regression line on the scatter plots much easier since the points would be much closer together.