Purple State Plunge: A Deep Dive into Pennsylvania's Cities

Joseph J. Kamerdze III
Coursera Capstone Project

1 Introduction

1.1 Background

The state of Pennsylvania is the 5th most populous state in the United States. Although the state's two largest cities lie on its east and west coasts, the state itself is dotted with smaller cities of varying sizes. Many of these cities were founded on the steel and coal industries, and because of the decline in these industries, have seen declines in their populations over the years. Others have managed to survive and even thrive as they have transitioned to meet the demands of the 21st century.

1.2 Business Problem

With the advent of the increased availability of location data and machine learning, someone who is interested in starting a business, investing in, or determining how to allocate funding for new business development in Pennsylvania can leverage these new capabilities to aid in the decision-making process. The purpose of this project is to explore the population growth trends among the cities of Pennsylvania between the years of 2010 and 2018 and leverage the data from Foursquare to cluster these cities together based on the cities' venues. This clustering can then be used to characterize growing and declining cities based on the most common venues in those cities, which can be used to determine where and how to direct business investments.

2 Data

2.1 Data Sources

The data used for this project consists of census data from a website hosted by the state of Pennsylvania, a database of location coordinates for each city in Pennsylvania, and Foursquare location data. The census data was pulled from the website, "data.pa.gov", and consists of population data for all municipalities in Pennsylvania from the 2010 census. This database also contains population estimates of each municipality from 2010 to 2018. The data from this source, which contains population data for 2,201 municipalities, will be filtered to focus only on the state's 57 cities. This data will be used to calculate each city's population growth over 2 and 9 years, the overall growth rate, and the growth rate for each year.

The data for the city coordinates exists in an Excel spreadsheet and was not readily available in database format online. The website, "data.pa.gov", provides a .json file that contains shape files for each of the state's municipalities, but that file only contains coordinate data for the municipality boundaries, rather than coordinates for the city itself. As a result, the database was required to be built manually by inputting the coordinate values onto an Excel spreadsheet. That spreadsheet was then uploaded to GitHub.

Finally, location data from Foursquare was used to pull the most common venues for each city. A 5-mile radius around each city's location coordinates was used to search for venues, and the top 10 most common venues were returned for each city.

2.2 Data Cleaning

The 2010 census database was scraped from the pa.gov website, and it contained census data for each municipality in Pennsylvania. This included boroughs, townships, counties, and cities, as well as the overall state. Conveniently, each of the 57 cities had the word "city" in its name, so the dataframe used for this analysis, named "dfCity", could be easily created for all records whose name contained the text string "city." The population numbers for the state of Pennsylvania overall were also reserved as a variable for use in subsequent data comparisons.

2.3 Calculations

The database only contained the estimated values for the years 2010-2018, so the growth rates needed to be calculated. Columns were added to the dataframe to show the calculated percentages of growth after each year for each city. There was also a column added to show the growth rate for the previous 2 years (2016-2018), as well as the growth rate for the entire 9-year period (2010-2018). After the columns were added for the growth rates for each city, the dataframe was ready to be merged with the spreadsheet of location coordinates.

3 Methodology

3.1 Exploratory Data Mapping

To get an idea of how the 57 cities are distributed across the state of Pennsylvania, a Folium map was created using the cities' names and coordinates (see Figure 1 below).

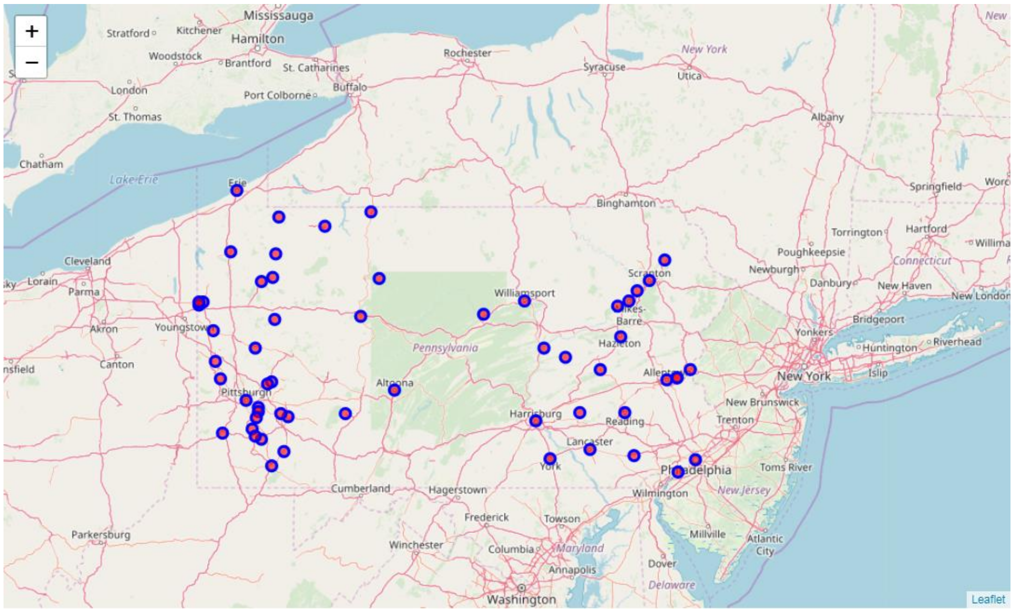


Figure 1: Pennsylvania's City Locations

The map in Figure 1 shows that most of the cities are located on the east and west sides of the state, with only about 10 of the 57 cities scattered across the center. A couple questions that this could raise for subsequent data analysis would be, “are the cities on either end of the state growing at a higher rate than the other end?”, or “is there a particular area of the state that is growing at a higher rate than the rest?”

3.2 Graphical Representation of City Population Growth vs Overall State Growth

Calculations based on the dataframe found that the population of the entire state of Pennsylvania grew by about 0.75% from 2010 to 2018. A histogram was created to show the distribution of population growth among the cities in the dataframe. The vertical red line on the histogram in Figure 2 below represents this 0.75% mark to more easily visualize the number of cities that grew at a higher rate than the state as a whole.

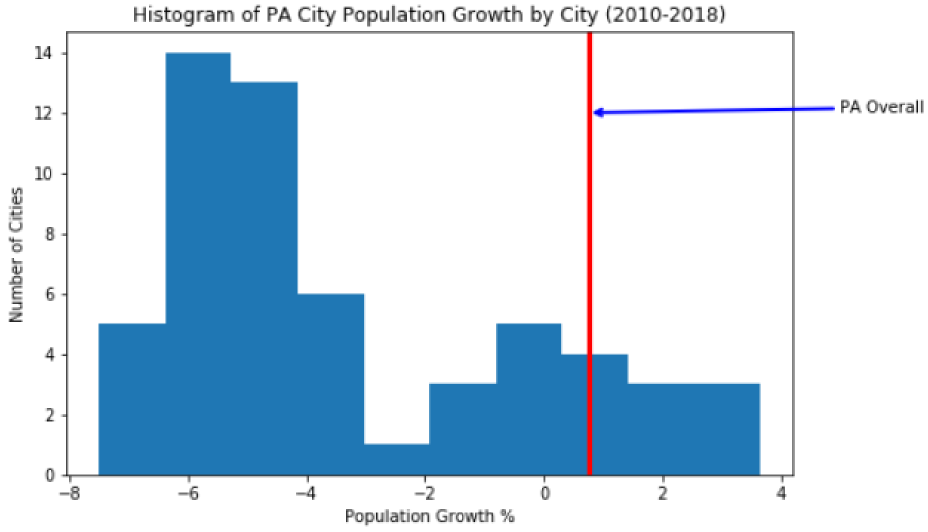
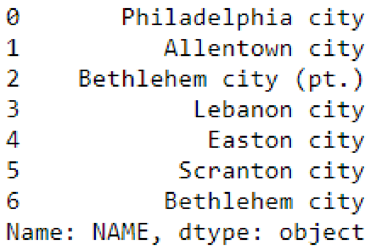


Figure 2: Histogram of PA City Population Growth by City (2010-2018)

The creation of a dataframe of cities that grew at a higher rate than the entire state found six cities. These six cities are listed in Figure 3 below and shown on the map in Figure 4 below.



```
0        Philadelphia city
1           Allentown city
2     Bethlehem city (pt.)
3             Lebanon city
4              Easton city
5            Scranton city
6           Bethlehem city
Name: NAME, dtype: object
```

Figure 3: PA Cities with higher growth rates than PA as a whole

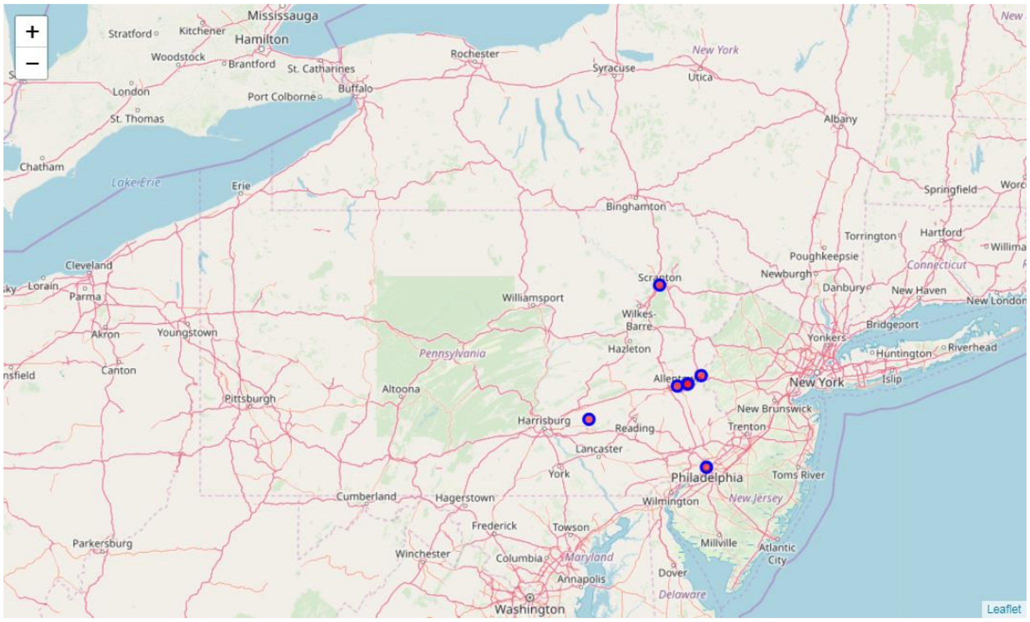


Figure 4: Map of PA cities with higher growth rates than PA as a whole

While the analysis performed over the past 9 years reveals which cities have grown the fastest from 2010-2018, it would be of higher interest to a potential investor which cities are actually on the rise population-wise. A list of cities that have grown faster than the state over the last 2 years (Figure 5) and a histogram (Figure 6) of this data was created to show the distribution of population growth among the cities over the last 2 years from 2016-2018. The map is shown in Figure 7.

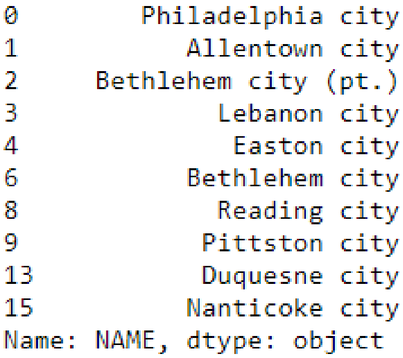

```
0          Philadelphia city
1             Allentown city
2       Bethlehem city (pt.)
3               Lebanon city
4                Easton city
6             Bethlehem city
8               Reading city
9              Pittston city
13             Duquesne city
15            Nanticoke city
Name: NAME, dtype: object
```

Figure 5: List of cities with higher growth rate than PA as a whole

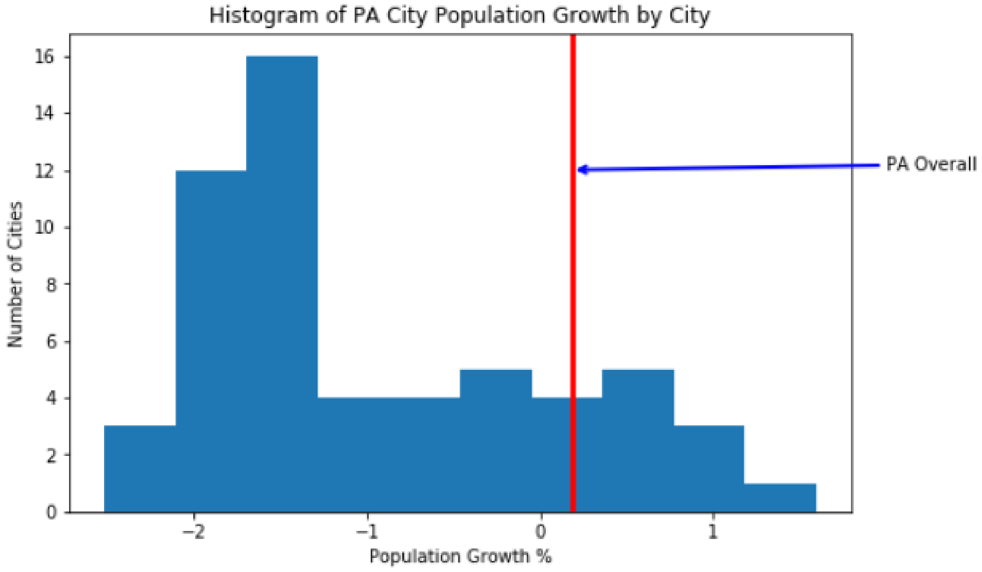


Figure 6: Histogram of PA City Population Growth by City (2016-2018)

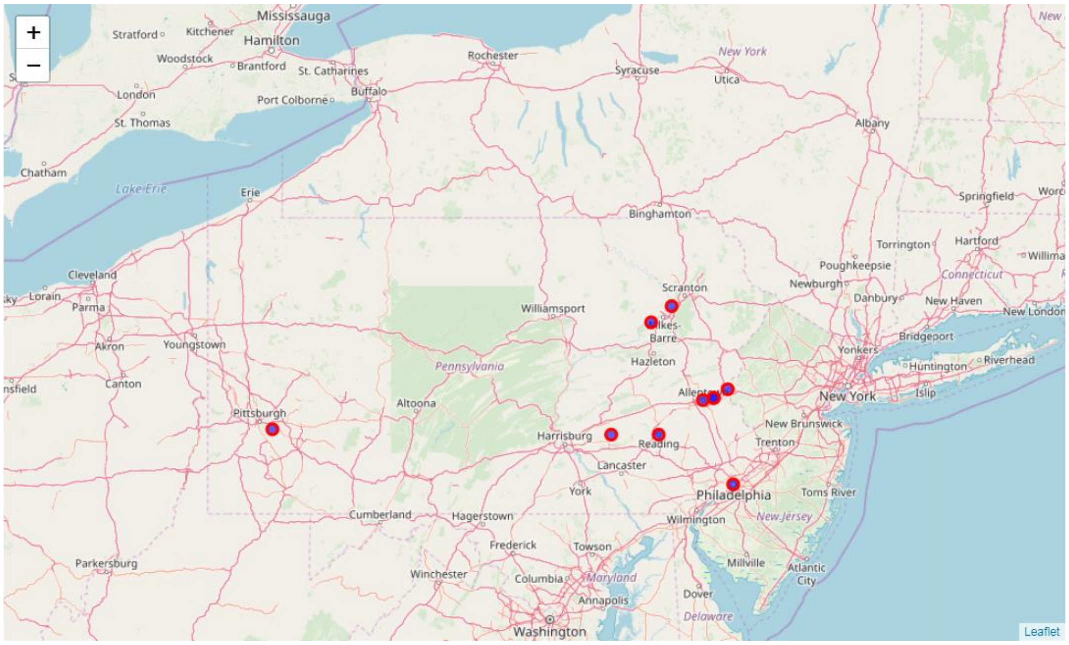
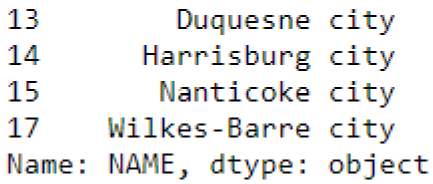


Figure 7: PA cities with higher growth rate than PA as a whole (2016-2018)

As the list and the associated map show, several of the cities that have grown over the period 2010-2018 have continued to grow from 2016-2018 showing sustained population growth.

Potential investors would also be interested to find any cities that might have been on the decline over the past 9 years, but are experiencing a turnaround in growth over the past 2 years. This would create the opportunity to get in on the “ground floor” to invest in a city that is on the verge of an economic turnaround. To determine which cities fall into this category, a subset of the dataframe was created for all cities that have had negative population growth from 2010-2018. Then, a subset of that dataframe of shrinking cities was created for any cities that have experienced positive growth over the period of 2016-2018. Those “hopeful cities” are listed below in Figure 8 and shown on a map in Figure 9.



```
13         Duquesne city
14       Harrisburg city
15        Nanticoke city
17     Wilkes-Barre city
Name: NAME, dtype: object
```

Figure 8: "Hopeful Cities" have experienced population decline from 2010-2018, but have grown from 2016-2018

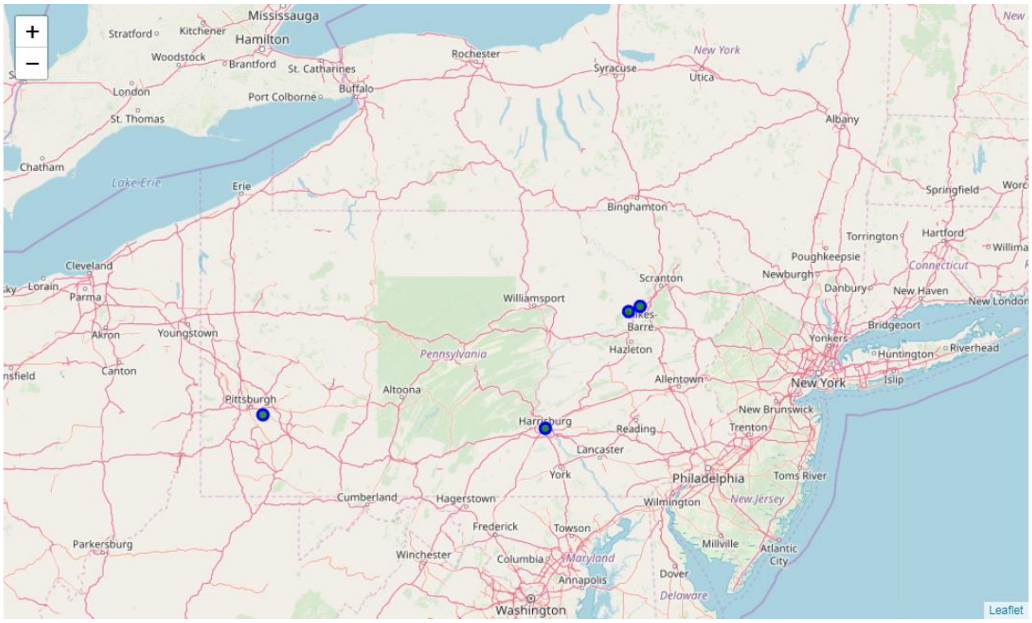
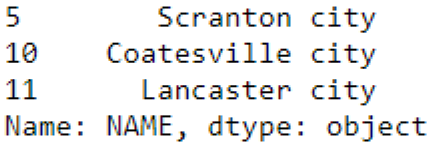


Figure 9 Map - "Hopeful Cities" have experienced population decline from 2010-2018, but have grown from 2016-2018

On the other hand, it would also be useful to know If there are any cities that have experienced overall population growth from 2010-2018, but those cities have seen declining populations in the last 2 years. This would indicate that a city might be on the verge of an economic decline. The analysis found 3 such cities, and they are listed in Figure 10 and shown on the map in Figure 11.



```
5          Scranton city
10      Coatesville city
11        Lancaster city
Name: NAME, dtype: object
```

Figure 10: "Slowing Cities" - have experienced population growth from 2010-2018, but have seen population decline from 2016-2018

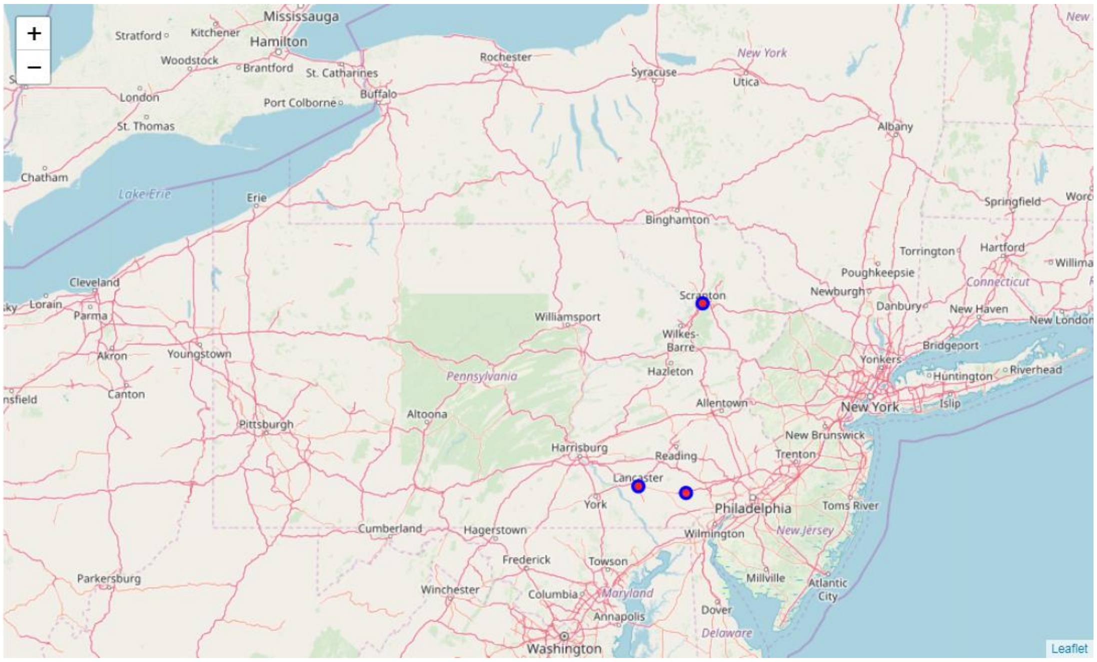


Figure 11:Map - "Slowing Cities" - have experienced population growth from 2010-2018, but have seen population decline from 2016-2018

3.3 Foursquare API

The Foursquare API was used to find the top 10 most common venues in each of Pennsylvania's 57 cities. The coordinates of each city were used as the location data, and a 5-mile (8000m) radius around these coordinates was used as the search area. The cities were then clustered using k-means clustering in 5 distinct clusters based on the top 10 most common venues in each city. A discussion of these clusters is included in the Results section of this report.

4 Results

4.1 Clustering

The Foursquare API was used to find the top 10 most common venues for each of the 57 cities in Pennsylvania. These cities were then divided into 5 different clusters using k-means clustering based on the venue data from Foursquare. A map of the cities and their respective color-coded clusters is shown in Figure 12 below.

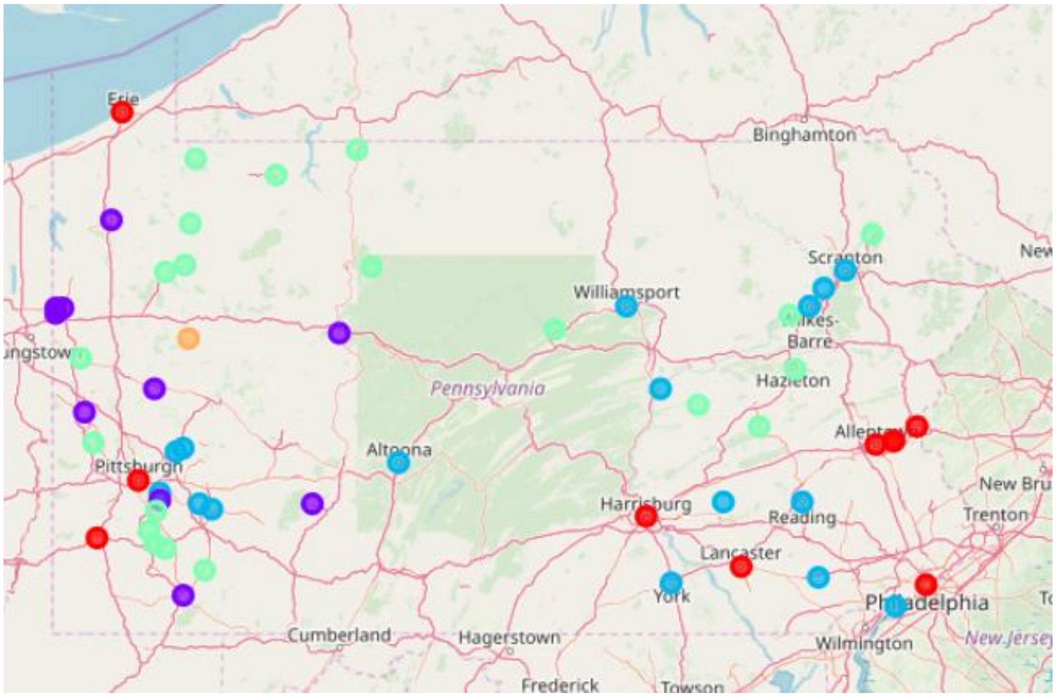


Figure 12: Map of all 57 cities in PA. Each color represents a different cluster, with 5 clusters total.

Based on the geographic locations of the cities and their relative population sizes, each cluster was given a name.

4.1.1 Cluster 1 – Big Cities (Color=Red)

This cluster is characterized by the relatively large population size of its cities. It contains the largest cities in Pennsylvania, including its two largest – Philadelphia and Pittsburgh. These cities are primarily located on the east and west coasts of the state, with the exception being Harrisburg, the state capitol. The majority (6) of these 10 cities have seen population growth from 2010-2018, but a some of them have seen population declines. See Figure 13 and 14 below for a histogram of city populations and growth rates for this cluster. Note: Philadelphia and Pittsburgh were removed from the histogram. Their populations so greatly outweigh the other cities that it creates a highly skewed histogram.

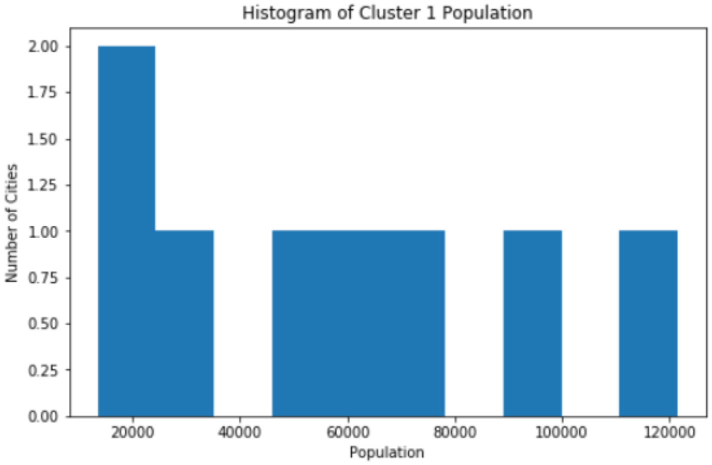


Figure 13: Histogram of Cluster 1 city populations

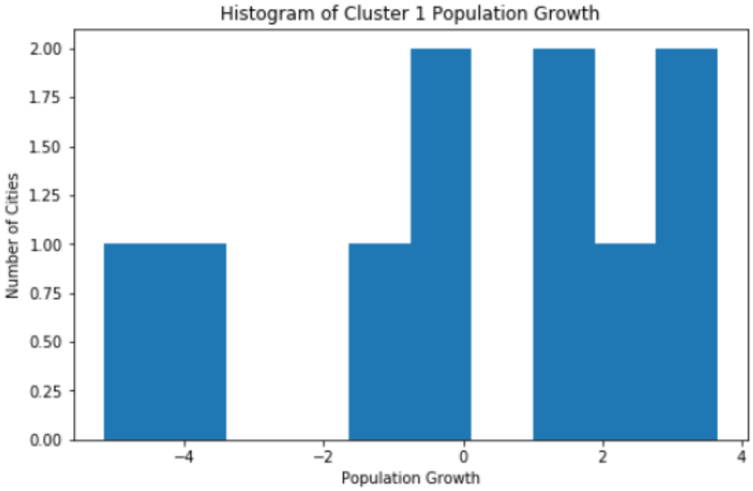


Figure 14: Histogram of Cluster 1 city population growth rates

4.1.2 Cluster 2 – Small Western Cities (Color = Purple)

This cluster contains mostly small cities with a population of less than 20,000, all of which reside on the western half of the state. All 10 of the cities in this cluster have experienced population declines from 2010-2018, as well as in the most recent two years.

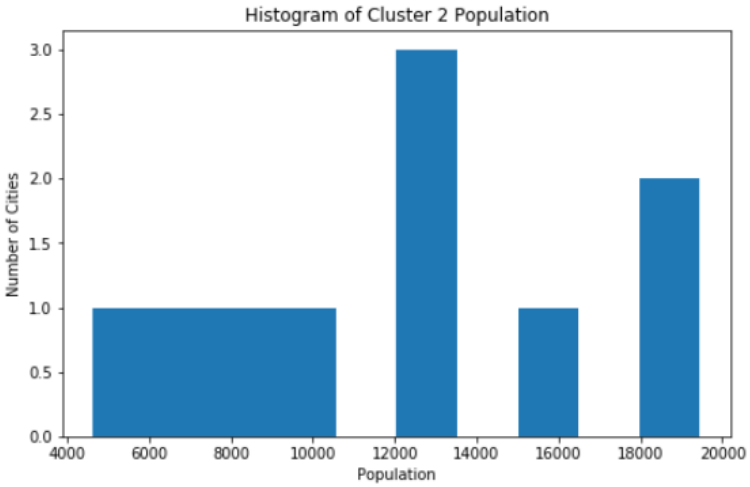


Figure 15: Histogram of Cluster 2 city populations

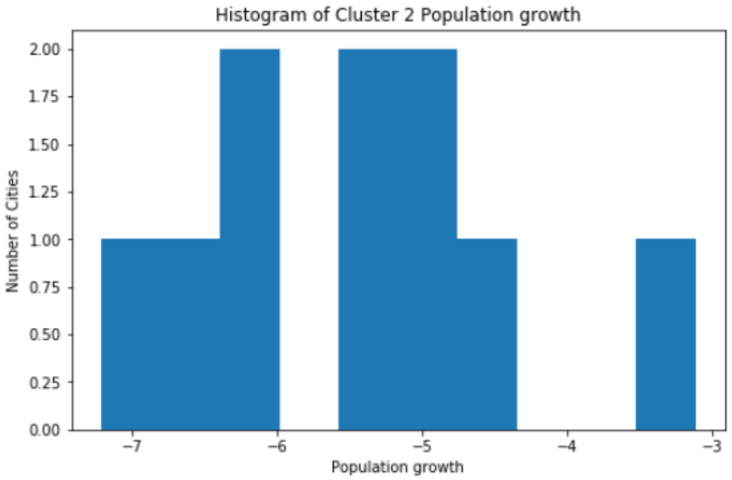


Figure 16: Histogram of Cluster 2 city population growth rates

4.1.3 Cluster 3 – "Purple State PA" – The Representative Sample (Color = Blue)

This cluster could be referred to as a representative sample of the entire state. The cities are dispersed across the entire state with a wide range of populations. The phrase "Purple State" refers to the state's political leanings. Pennsylvania is classified as a "swing state" in presidential elections because it swings between voting conservative (red) and liberal (blue), hence the combination of red and blue = purple. The phrase is being used here to describe the combination of east / west coasts, high / low populations, as well as growing / declining populations.

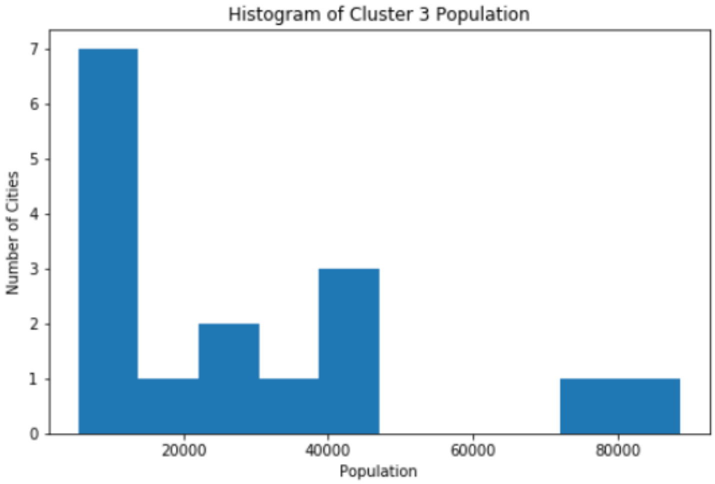


Figure 17: Histogram of Cluster 3 city populations

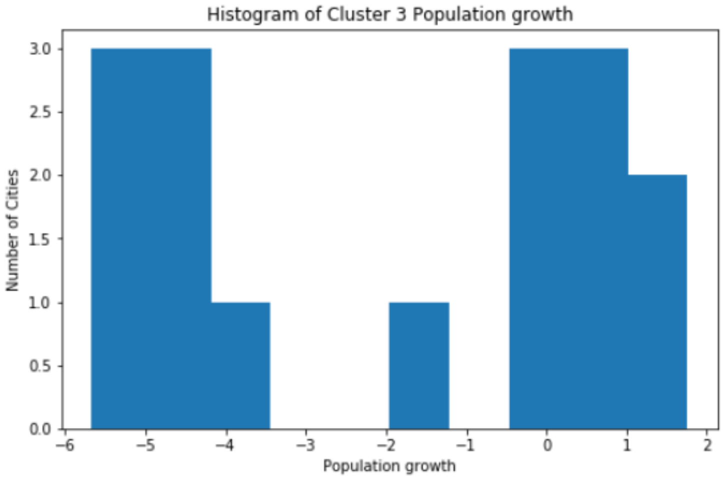


Figure 18: Histogram of Cluster 3 city population growth rates

4.1.4 Cluster 4 – Small, Declining Cities (Color = Green)

Of the 20 cities in this cluster, 12 have populations with fewer than 10,000 people. All 20 of them have experienced population declines from 2010-2018. However, one of this cities, Nanticoke City, is listed on our list of “Hopeful Cities”, which means that, even though its population has declined over the past 9 years, its has seen an increase in the last 2.

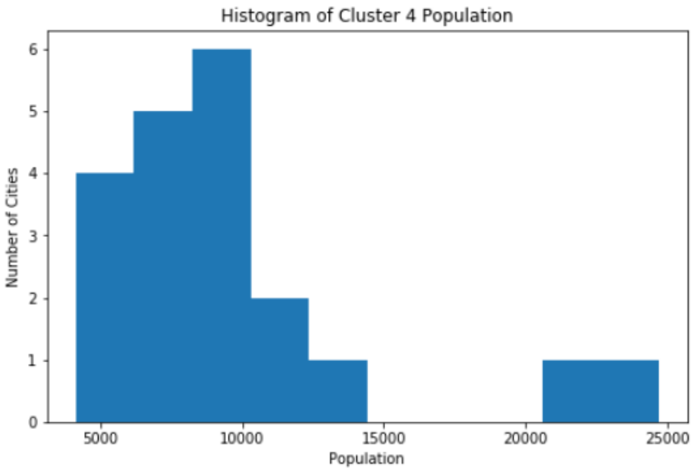


Figure 19: Histogram of Cluster 4 city populations

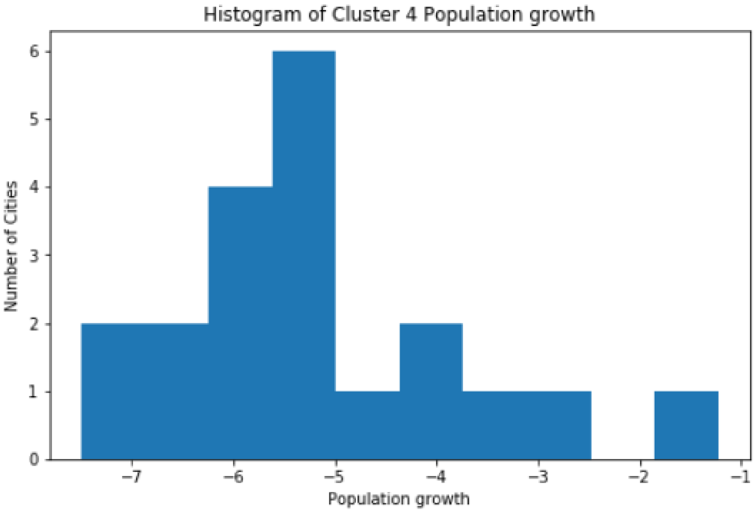


Figure 20: Histogram of Cluster 4 city population growth rates

4.1.5 Cluster 5 – Parker City: The Smallest City in the US (Color = Orange)

With a population of around 800 people, Parker City is in a cluster of its own, and this is for good reason. According to Wikipedia.com, Parker City is the commonly referred to as the country’s smallest city. It is located on the western half of the state and has experienced a steady population decline from 2010-2018.

5 Discussion of Results

5.1 Context

In order to better understand the results of this analysis, it is important to understand the overall picture of the population growth of Pennsylvania. According to the United States Census Bureau, the population of the United States grew by almost 6% between the years 2010-2018 (https://www.census.gov/library/visualizations/interactive/population-increase-2018.html). In contrast, the analysis performed in this report found that the state of Pennsylvania only grew by 0.75% over the same time period. By comparison to the country, the state of PA experienced very low relative growth. Our analysis found that only 6 of the 57 cities in the state grew at a higher rate than the state overall, and over half of this growth occurred in the state's largest city, Philadelphia (around 55,000 people, or 3.65% from 2010-2018).

5.2 Effects of Geography and Population Size on Clustering

While the majority of the city population growth occurred in cities on the eastern side of the state, a review of the population data in each of the clusters and visual observations of the cluster map in Figure 12 show that geography and population had modest-to-strong effects on how the cities were clustered together. Because the cities were clustered based on most common venues, and because PA is such a large state, it's likely that variations in local culture play a part in the proliferation of particular venues. Population size may play a similar role, as "bigger" and "smaller" cities tend to attract people with different preferences.

5.3 Discussion of Clustering

The purpose of this analysis was to characterize different types of cities based on their most common venues. This characterization would then inform a potential business owner or investor how direct their funds and efforts. One of the challenges to solving this problem would be – With all the combinations of the different kinds of cities (east/west, big/small, growing/declining), how do we know what a "model" city in each of these categories looks like? Fortunately, 3 out of the 5 clusters generated by the k-means clustering in this analysis contains a combination of growing and declining cities.

5.3.1 Discussion of Cluster 1

Cluster 1 contains the largest cities in PA. These cities are split between declining/growing populations, and east/west coasts. Of particular interest is that this cluster also contains 2 of the "Hopeful Cities", or cities that have experienced population declines over the 9-year period but have seen growth in the past 2 years (Harrisburg and Wilkes-Barre). If one were looking to open a business in Wilkes-Barre, for example, they may not want to open another bar (the most common venue in the city), but rather look to a growing city like Bethlehem as a model and opt to open a coffee shop (Bethlehem's third most common venue, and the 9th most common venue in Wilkes-Barre).

5.3.2 Discussion of Cluster 3

Cluster 3 can be seen as a cross-section of the entire state of PA. The venues “Pizza Place” and “American Restaurant” look to be the most common venues across the cities in this cluster. If one were looking to open a business in a mid-size Pennsylvania city, the data in this cluster would inform the decision-making process in a more general way. Perhaps, instead of doubling down on pizza and burgers, one might explore opening a more unique venue that has shown success in growing cities from other clusters, like a Mexican Restaurant or a Sandwich Place.

5.3.3 Discussion of Cluster 4

Cluster 4 consists of small cities with population declines. However, one of those cities, Nanticoke city, shows the potential for a rejuvenation with population increases in the past 2 years. If an investor were looking to open a business in a small, Western PA city, an Italian Restaurant in Nanticoke serving comfort food and good-quality coffee might be the best bet.

6 Conclusion

This study analyzed the 57 cities in the state of Pennsylvania in terms of population growth and geographical location, and sought to characterize these cities based on their most common venues using K-Means Clustering. The analysis found a relationship between population size, geographic location, and most common venues which could likely be attributed to local cultures and demographic preferences. The results of this analysis could be used to inform decisions by potential entrepreneurs looking to start a business in Pennsylvania. This model of analysis could also be used for other states with similar population demographics. Further analytical opportunities include the further exploration of a city’s population growth/decline based on job market growth/declines, but this analysis is beyond the scope of this report.