CLUSTERING OUT-MIGRATION DESTINATION OF MILLENNIALS FROM NEW JERSEY

Nor Zam Azihan MH

10th October 2019

1.0 Introduction

1.1 Background

New Jersey located in Atlantic region of the United States, bounded by New York at the north and east; Atlantic Ocean at the east, southeast and south; Delaware River and Pennsylvania at the west; Delaware Bay and Delaware at the southwest. It is the most densely populated in the US with 9 million residents as of 2017. In addition, it is also populated one of the most diversify ethnic and religion in the US. Despite of that, as much as 30 municipalities in New Jersey are rated as the top 100 safest municipalities in the US. It is also housing the highest number of millionaires and as 2017, considered as the second wealthiest US state.

New Jersey economy is multifaceted, mostly cantered around ports, pharmaceutical, biotechnology, information technology, financial industry, tourism and agricultural outputs. In 2016, the estimated gross state product was $575 million.

1.2 Problem Description

The rising housing cost in New Jersey has resulted in out-migration of the millennials. The destination includes the Los Angeles, Miami, New York, Washington and others. It has become a problem for millennials to find new destination to migrate. Therefore, this project aims to clusters the migration destinations for the. This could help the millennials in choosing the best destination to settle down.

2.0 Data Description

Data used for this project came from two sources. Data on destination counties were retrieved from Table1. Population Density and Non-Driving Commuting in Counties That Disproportionately Attract New Jersey Millennials, from njfuture.org website.

Foursquare.com was used to gather data on list of common venue in the counties as well as their coordinates.

The coordinate for each destination country were retrieved using geocoder. The latitude and longitude were used to get the most common venues from the foursquare.com. List of 10 most common venues was created. KMeans clustering method were used to cluster the counties based on the common venues. Subsequently, the clusters were compared with the population density, commuting percentage and percentage of millennials migrants retrieved from destination counties table.

KMeans clustering also used to cluster the destination counties using population density, commuting percentage and percentage of millennials migrants in which were then compared to the initial cluster groups retrieved using common venues.

3.0 Methodology

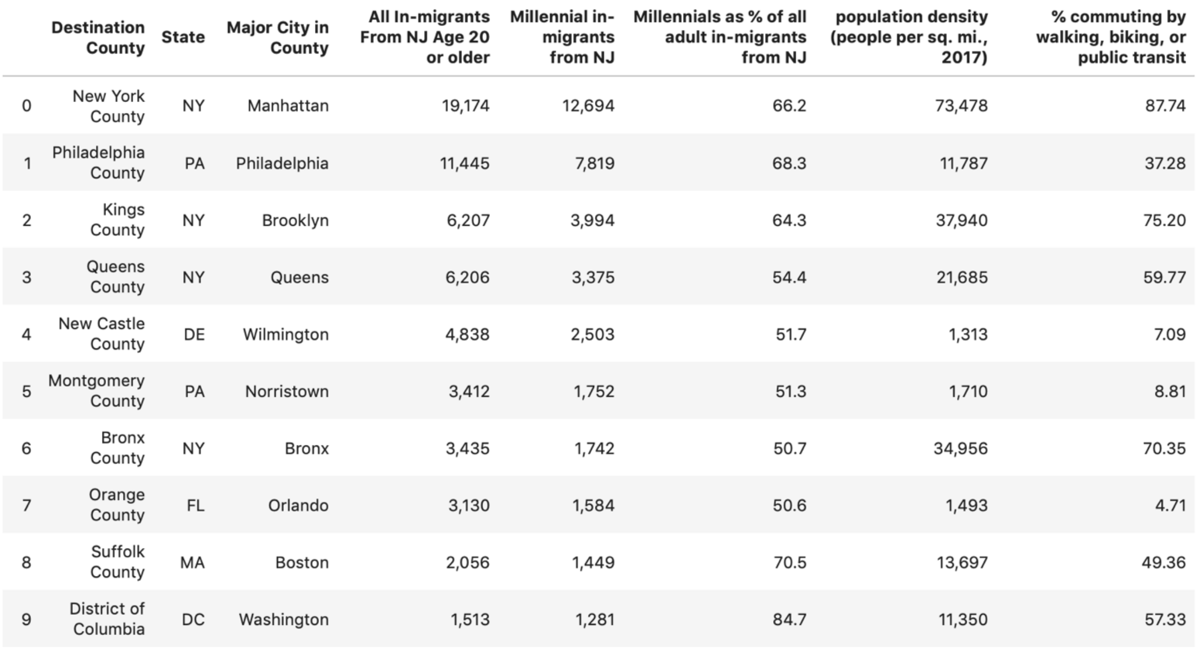
Data analysis was done using Python 3.6. Python packages involved were Matplotlib, Numpy, Pandas, Seaborn, Sci-kit learn, Folium, Geopy and Beautiful Soup 4. All analysis was done in JupyterLab 1.0.

3.1 Data Cleaning and Exploratory Data Analysis (EDA)

Firstly, table was extracted using Beautiful Soup 4 from njfuture.org website. Data cleaning was done after table was extracted. The State of each Destination county is separated into a different column to ease retrieval of coordinate later on. Type of data for each columns were determined and changed into integer and float type for further analysis. The cleaned extracted table is shown in **Table 1**.

Exploratory Data Analysis was done summarizes the data mean, median, standard deviation, max, mean and the quartiles (**Table 2**). Based on the EDA, all the variables were positively skewed. However, there were no missing values for the data.

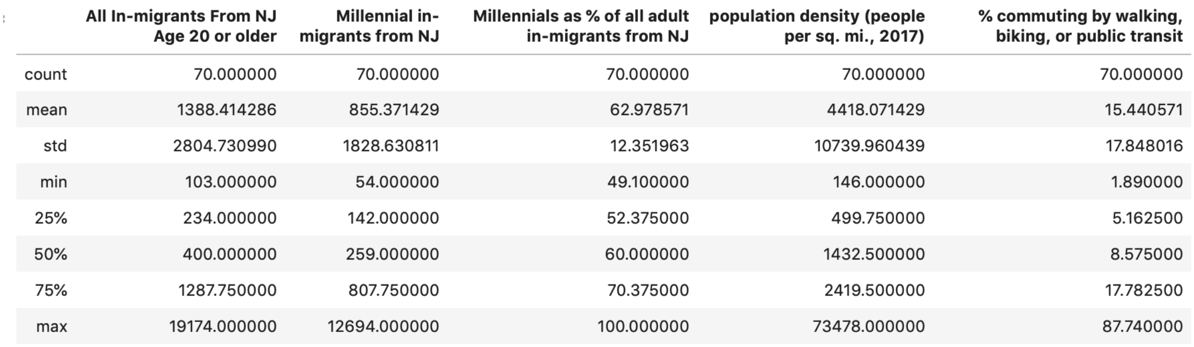
Table 1. The First 10 Rows of Population Density and Non-Driving Commuting in Counties That Disproportionately Attract New Jersey Millennials



	Destination County	State	Major City in County	All In-migrants From NJ Age 20 or older	Millennial in-migrants from NJ	Millennials as % of all adult in-migrants from NJ	population density (people per sq. mi., 2017)	% commuting by walking, biking, or public transit
0	New York County	NY	Manhattan	19,174	12,694	66.2	73,478	87.74
1	Philadelphia County	PA	Philadelphia	11,445	7,819	68.3	11,787	37.28
2	Kings County	NY	Brooklyn	6,207	3,994	64.3	37,940	75.20
3	Queens County	NY	Queens	6,206	3,375	54.4	21,685	59.77
4	New Castle County	DE	Wilmington	4,838	2,503	51.7	1,313	7.09
5	Montgomery County	PA	Norristown	3,412	1,752	51.3	1,710	8.81
6	Bronx County	NY	Bronx	3,435	1,742	50.7	34,956	70.35
7	Orange County	FL	Orlando	3,130	1,584	50.6	1,493	4.71
8	Suffolk County	MA	Boston	2,056	1,449	70.5	13,697	49.36
9	District of Columbia	DC	Washington	1,513	1,281	84.7	11,350	57.33

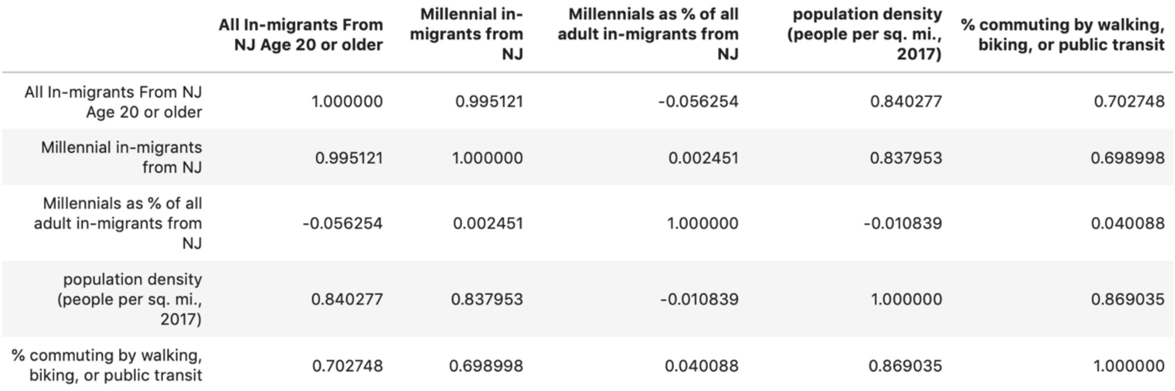
Each columns were also tested for correlation through regression plots (**Table 3**). Number of all in migrants from New Jersey age 20 years and above had a very strong correlation with the millennials in-migrant from New Jersey. On the other hand, correlation between percentage of commuting by walking, biking or public transport and millennials percentage of all adult in-migrants from New Jersey was the lowest compared to others.

Table 2. Mean, Median, Standard Deviation, Max, Mean And The Quartiles Of All Variables



	All In-migrants From NJ Age 20 or older	Millennial in-migrants from NJ	Millennials as % of all adult in-migrants from NJ	population density (people per sq. mi., 2017)	% commuting by walking, biking, or public transit
count	70.000000	70.000000	70.000000	70.000000	70.000000
mean	1388.414286	855.371429	62.978571	4418.071429	15.440571
std	2804.730990	1828.630811	12.351963	10739.960439	17.848016
min	103.000000	54.000000	49.100000	146.000000	1.890000
25%	234.000000	142.000000	52.375000	499.750000	5.162500
50%	400.000000	259.000000	60.000000	1432.500000	8.575000
75%	1287.750000	807.750000	70.375000	2419.500000	17.782500
max	19174.000000	12694.000000	100.000000	73478.000000	87.740000

Table 3. Correlation Between All Variables



	All In-migrants From NJ Age 20 or older	Millennial in-migrants from NJ	Millennials as % of all adult in-migrants from NJ	population density (people per sq. mi., 2017)	% commuting by walking, biking, or public transit
All In-migrants From NJ Age 20 or older	1.000000	0.995121	-0.056254	0.840277	0.702748
Millennial in-migrants from NJ	0.995121	1.000000	0.002451	0.837953	0.698998
Millennials as % of all adult in-migrants from NJ	-0.056254	0.002451	1.000000	-0.010839	0.040088
population density (people per sq. mi., 2017)	0.840277	0.837953	-0.010839	1.000000	0.869035
% commuting by walking, biking, or public transit	0.702748	0.698998	0.040088	0.869035	1.000000

4.0 Results

4.1 Clustering Major Cities Destination Using Most Common Venues

Number of optimal cluster (k) was determined using Elbow Method and Silhouette Method. Thus, based on **Figure 1** and **Figure 2**, the optimal number of clusters (k) were three.

Figure 1. The Silhouette Method For Optimal Number of Clusters (k) of Major Cities

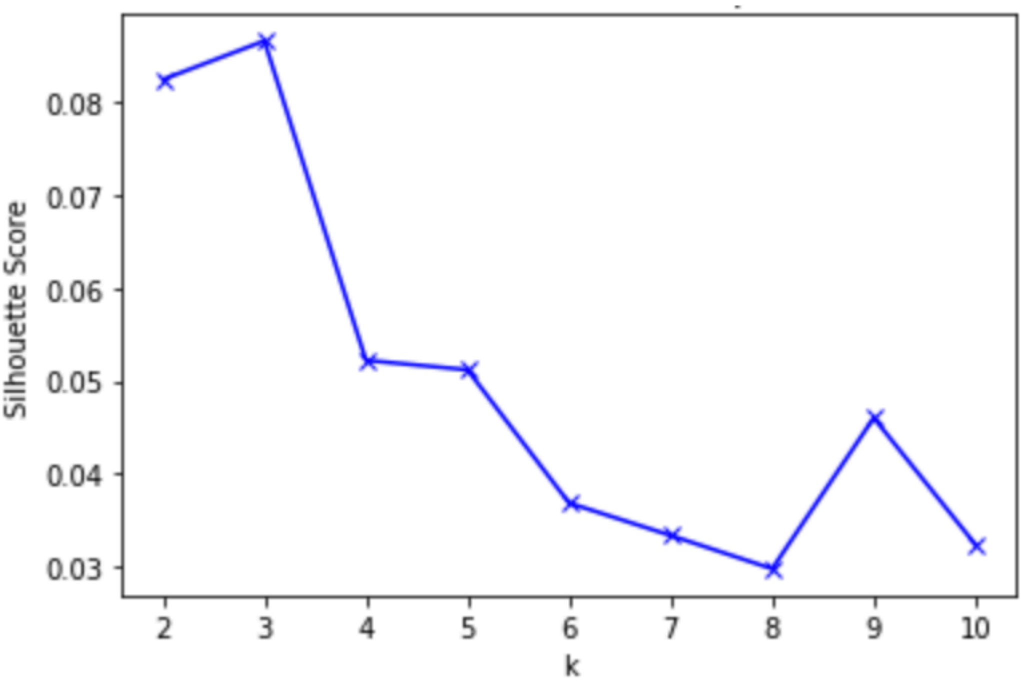
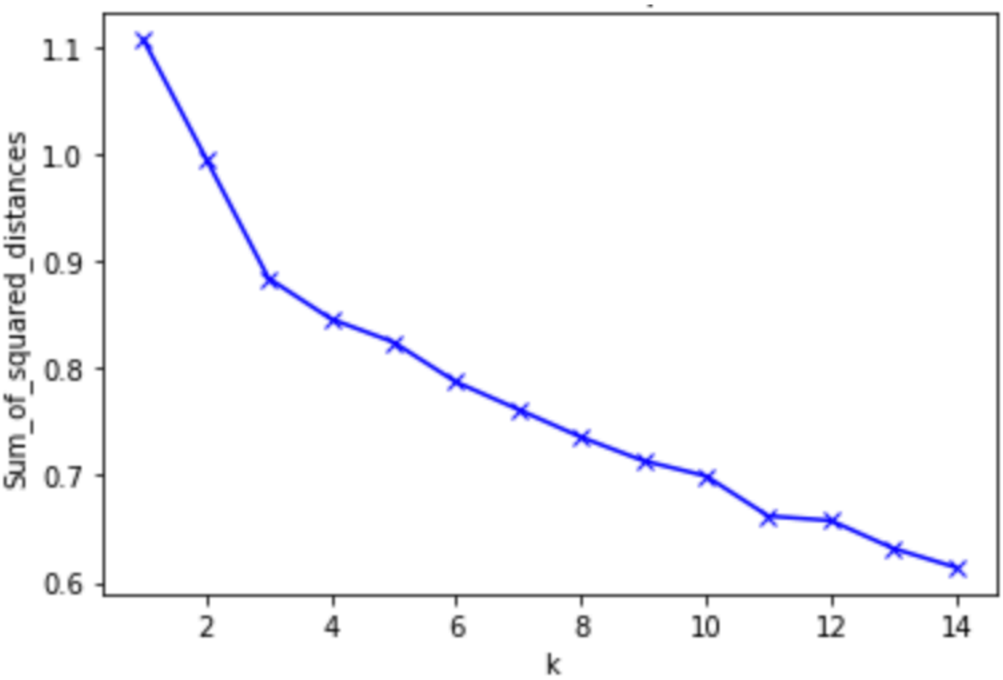


Figure 2. The Elbow Method For Optimal Number of Clusters (k) of Major Cities

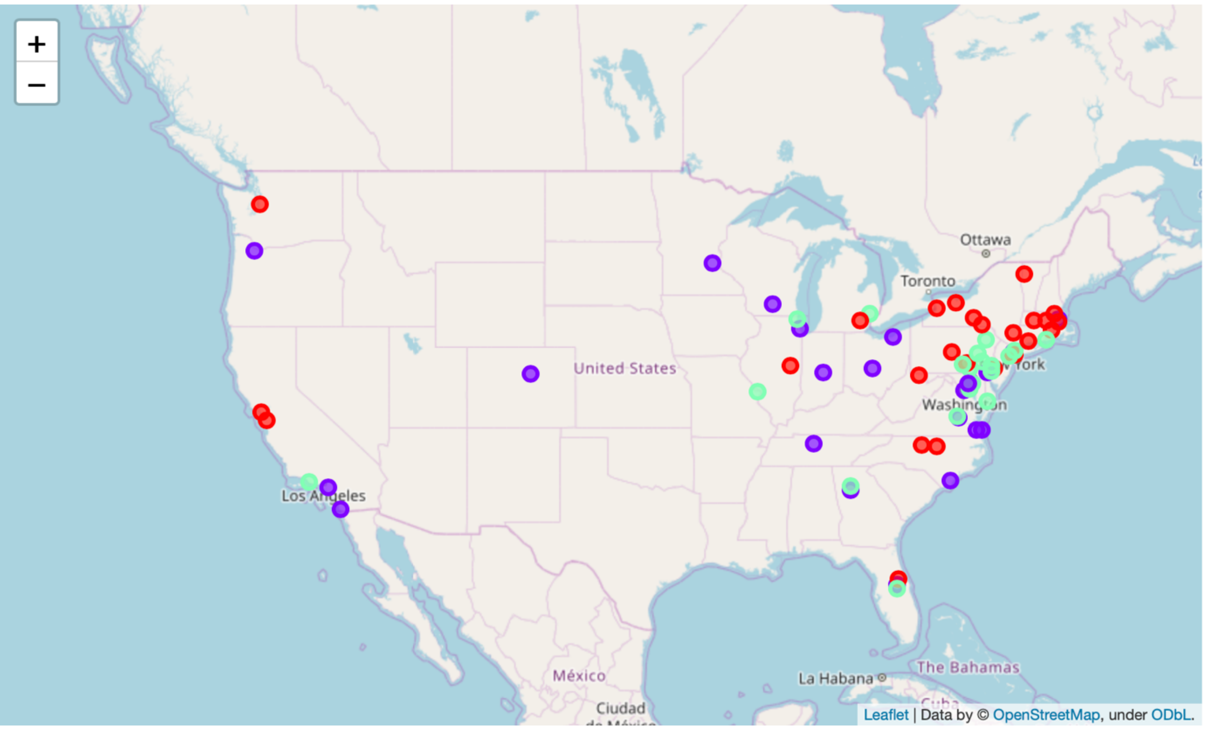


Each Major Cities were clustered into 3 clusters, which was the optimal number of clusters, using KMeans clustering methods. The Major Cities clusters characteristics are described in **Table 4**. Subsequently, each of the Major Cities were plotted into United State (US) map using Folium. The result is shown in **Figure 3** below.

Table 4. Description of Clusters for Major City Destination For Ney Jersey Millennials Based on Common Venues

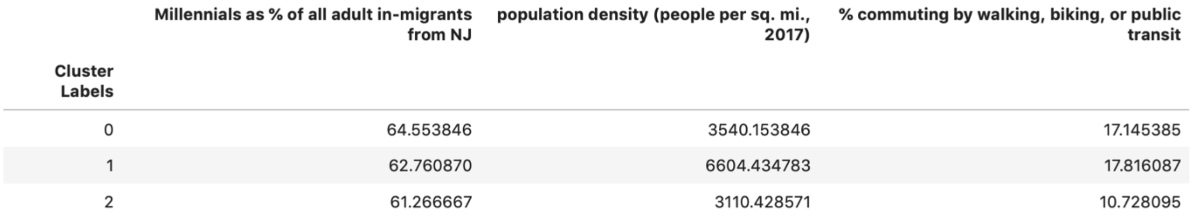
No.	Cluster Name	Label Color	Description
1.	Cluster 0	Red	Gastronomy and recreational
2.	Cluster 1	Purple	Entertainment and hobbies
3.	Cluster 2	Light Green	Light food and convenience store

Figure 3. Map Plot of Major City Destination For Ney Jersey Millennials Based on Common Venues Clustering



The mean of Millennials as percentage of all adult in-migrants from New Jersey, Population density and percentage commuting by walking, biking, or public transit were then calculated for each clusters as shown in **Table 5**. It showed that the average percentage of millennials were almost similar for all the clusters. Cluster 1, cities with better entertainments and hobbies had an average of higher density compared to the other clusters. Whereas, Cluster 2, with better light food and convenient store had less percentage of commuting by walking, biking an public transport.

Table 5. Mean of Millennials As Percentage Of All Adult In-Migrants From New Jersey, Population Density And Percentage Commuting By Walking, Biking, Or Public Transit For Major City Cluster of Common Venues



	Millennials as % of all adult in-migrants from NJ	population density (people per sq. mi., 2017)	% commuting by walking, biking, or public transit
Cluster Labels			
0	64.553846	3540.153846	17.145385
1	62.760870	6604.434783	17.816087
2	61.266667	3110.428571	10.728095

4.2 Clustering Destination Counties

Four variables were selected for clustering of destination counties for millennials. These variables were major city clusters, millennials as percentage of all adult in-migrants from new jersey, population density and percentage commuting by walking, biking, or public transit. KMeans clustering method was used for clustering purposes. Prior to that, all the variables selected were transformed into standard scaler.

Elbow method and Silhouette method were used for determining optimal number of clusters as shown in **Figure 4** and **Figure 5.** Thus, revealed the optimal number of clusters of four.

Subsequently, the average of Millennials as percentage of all adult in-migrants from New Jersey, Population density and percentage commuting by walking, biking, or public transit for each clusters were calculated and showed in **Table 6.**

Figure 4. The Silhouette Method For Optimal Number of Clusters (k) of Destination Counties

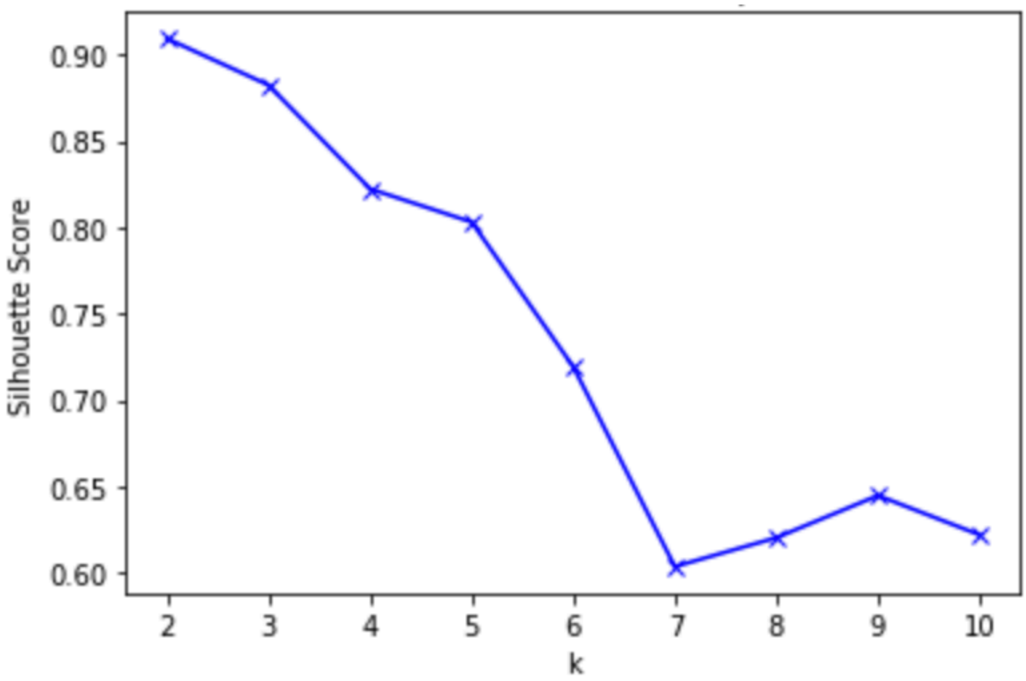


Figure 5. The Elbow Method For Optimal Number of Clusters (k) of Destination Counties

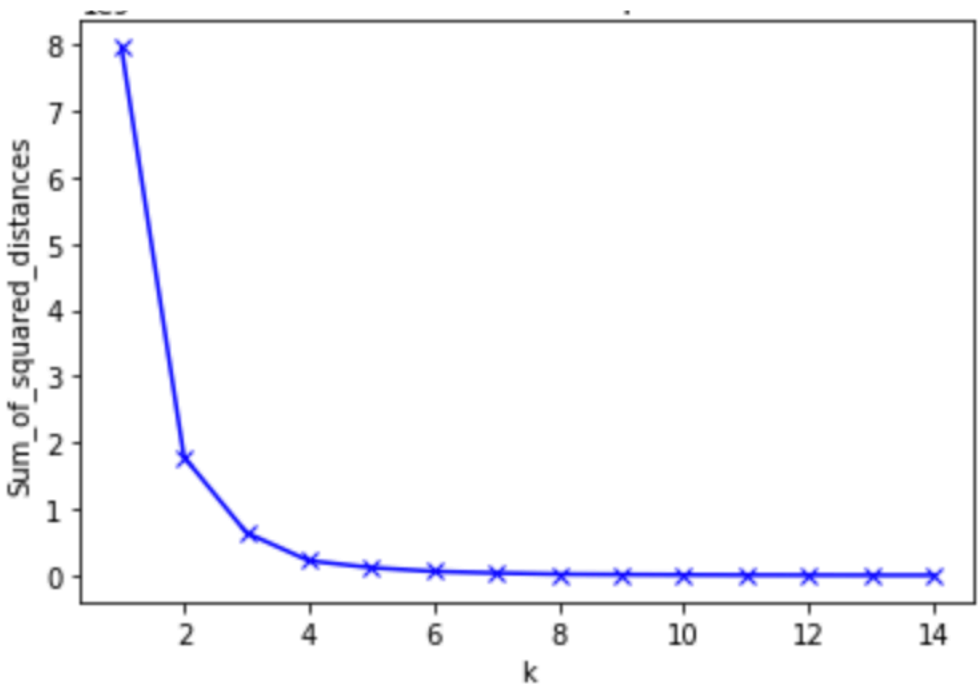
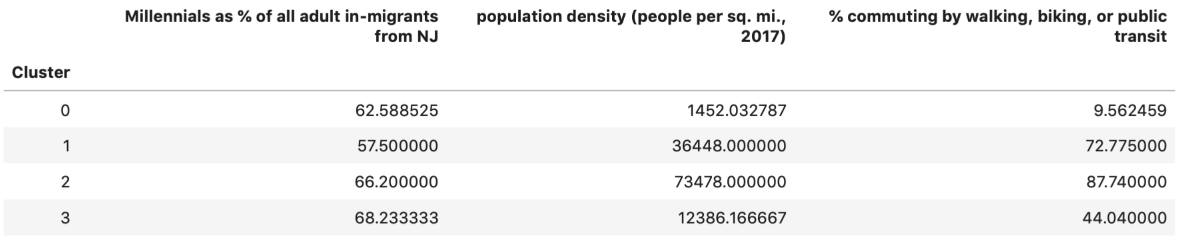


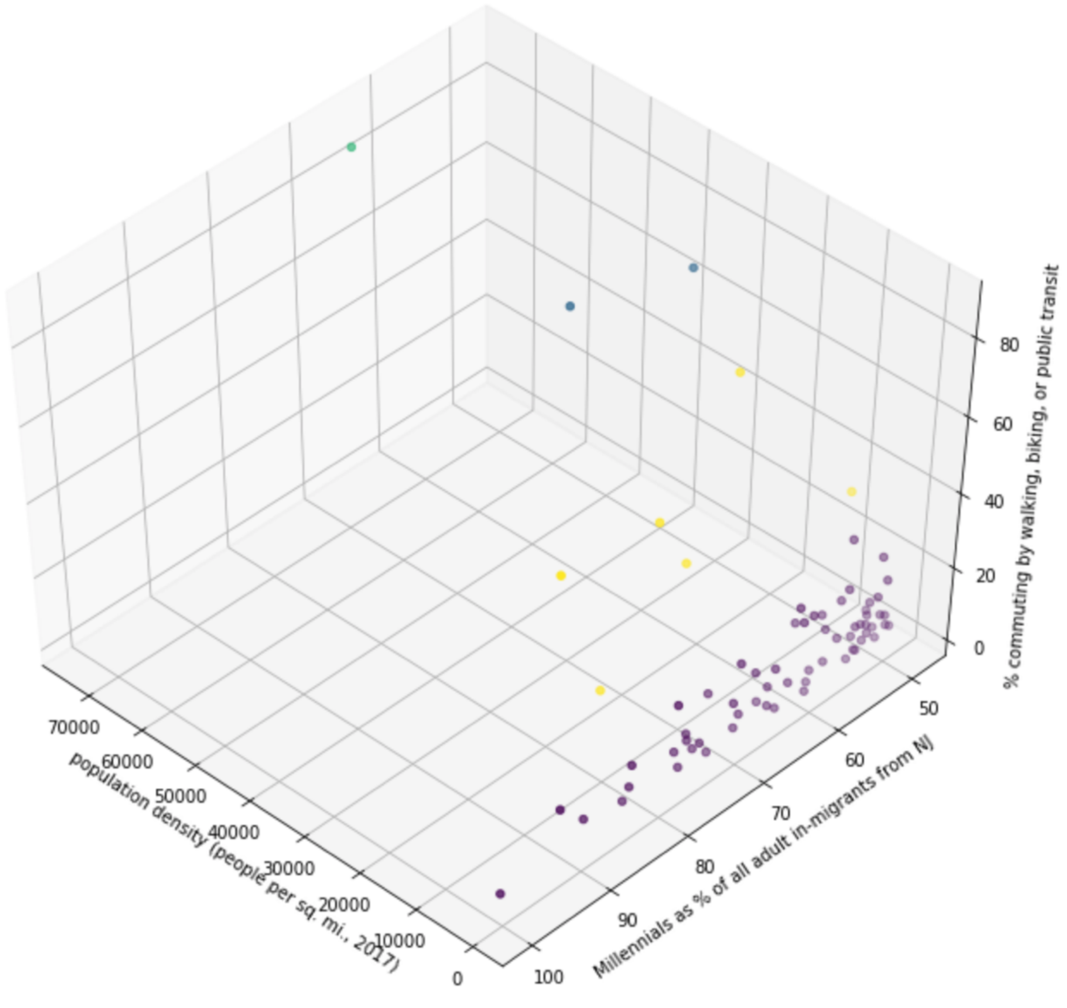
Table 6. Mean of Millennials As Percentage Of All Adult In-Migrants From New Jersey, Population Density And Percentage Commuting By Walking, Biking, Or Public Transit For Destination County Cluster



Cluster	Millennials as % of all adult in-migrants from NJ	population density (people per sq. mi., 2017)	% commuting by walking, biking, or public transit
0	62.588525	1452.032787	9.562459
1	57.500000	36448.000000	72.775000
2	66.200000	73478.000000	87.740000
3	68.233333	12386.166667	44.040000

The data were plotted into 3D graph to further explain characteristics of each cluster (**Figure 6**). The graph showed that all four clusters were clearly separated. Cluster 0 (purple) had the lowest population density and lowest percentage of commuting by walk, bike or public transport. Most of the destination counties were in this cluster. A wide range of millennials chose this county compared to the other adults, with percentage of 50% to almost 100%. There were few destination counties of Cluster 2 (blue) with low middle population density and having a higher percentage of commuting compared to Cluster 0. Cluster 2 (light green) only consisted of one county, which had a very high density and commuting percentage by walking, biking or public transport.

Figure 6. 3D Plot of Millennials As Percentage Of All Adult In-Migrants From New Jersey, Population Density And Percentage Commuting By Walking, Biking, Or Public Transit



5.0 Discussion

The results showed that the distribution of millennials in-migrants were mostly located on the eastern part of United States, near to New Jersey. Only few chose to migrate to the western or the centre of United States. Choosing the destination for migration may be difficult, but nearby county and states could be a better choice. Nonetheless, other contributing factors might play a role on this migration, for example, the housing price, job opportunities and etc, which were not covered in this study.

KMeans clustering method is one way of grouping the major cities or destination counties. In this study, common venues were used for clustering the major cities. This enable the major cities to be classified into 3 main clusters, namely a city with gastronomy and recreational, entertainment and hobbies, as well as light food and convenience store. All the clusters had almost similar preference among millennials. However, using the foursquare.com API for common venues always differs depending on the time when request is made since it always being updated. Thus, it raises the issue of results replicability. Despite of that, the use of foursquare API can be very useful if the latest and updated data is required.

The results also identified 4 clusters of destination counties based on the 4 variables, namely the Millennials As Percentage Of All Adult In-Migrants From New Jersey, Population Density And Percentage Commuting By Walking, Biking, Or Public Transit For Destination County Cluster and Major Cities clusters. These clusters can be explained clearly based on the level of population density, and percentage of commuting by walking, biking or public transport. Majority of the selected destination counties had a low population densities as well as low percentage of commuting by walking, biking or public transport, which is in Cluster 0. This characteristics may due to other factors such as housing price, traffics, and etc, which were not covered in this study. Only one destination county had high population density and high percentage of commuting by walking, biking or public transport. This destination may serve as different purpose for these group of millennials, e.g. for seeking a job. These gaps however, need to be further explored in future studies.

6.0 Conclusion

In conclusion, this study provide a better understanding of the destination for millennials in-migrant from New Jersey. Thus, giving much better insight for the millennials to choose a destination to settle on based on what their purpose for migration. Nonetheless, there are still gaps for future studies to look further into other factors in each clusters of destination county.