Finding the Next NBA City

Daniel Haas

September 2019

1. Introduction

The sports entertainment industry is a multi-billion dollar industry. Comprised of four major sports leagues (the NFL, the NBA, MLB and the NHL) each with roughly 30 teams, the impact of these leagues is felt around the country. Yet, in spite of the seeming national saturation of sports teams and leagues, there is continual talk of adding additional teams to new cities to further expand their respective sports.

A league that has seen growing popularity as of late is the National Basketball Association (NBA). In light of this, talk has been swirling of adding an expansion team to a new market. To further fuel these rumors, the NBA has relatively fewer teams than its two Fall/Winter sports league rivals; 30 teams vs 32 for the NFL and (soon to be) 32 for the NHL.

With this as a backdrop, the NBA is looking to expand its team base and, therefore, fan and revenue base via expansion. The challenge is finding a city that is suitable for expansion and will create a corresponding increase in revenue without diluting the current product.

The goal of this study is to help the NBA narrow down its search for suitable expansion cities. This will be done by looking at both bulk city data (population, TV market size, income) and the interests of the locals via popular venues and locales within the respective cities.

2. Data Acquisition and Cleaning

Data needed to help find the next location for an NBA team starts with a list of US cities ranked by population size. Additional details needed for each city will be population, location, density, per-capita income information, TV market size and details about popular venues in the area. Cities that currently have an NBA team need to be identified. Finally, characteristics of the cities interests will be assessed using FourSquare venue data.

Using the argument that the NBA is booming and its teams are successful, a clustering algorithm will be applied to this data set to find non-NBA cities that are similar to current NBA cities. These cities should be setup for similar success in the NBA and will become the ‘candidate’ cities for NBA expansion and further exploration of their capability to accommodate an NBA team.

Data for this study is going to come from five sources. One of these sources is the FourSquare locale data available through the the FourSquare developer API. The other four sources are:

1. City Population Data via Wikipedia

2. TV Market Data via Wikipedia
3. Metro Area income Data via Wikipedia
4. List of current NBA teams via Basketball-Reference.com

The first three data sources in the list above are located in Wikipedia and will need to be scraped. With the last data source, it is possible to copy the data to csv file and save locally. This step was done manually and then imported into a pandas dataframe via the csv_read function.

Once scraped, the data needs to be cleaned and prepared for the analysis to follow. The four city data sets listed above need to be merged into a single dataframe. A column labeled 'NBA' will be coded 1 for cities that already have an NBA team and 0 for cities that don't.

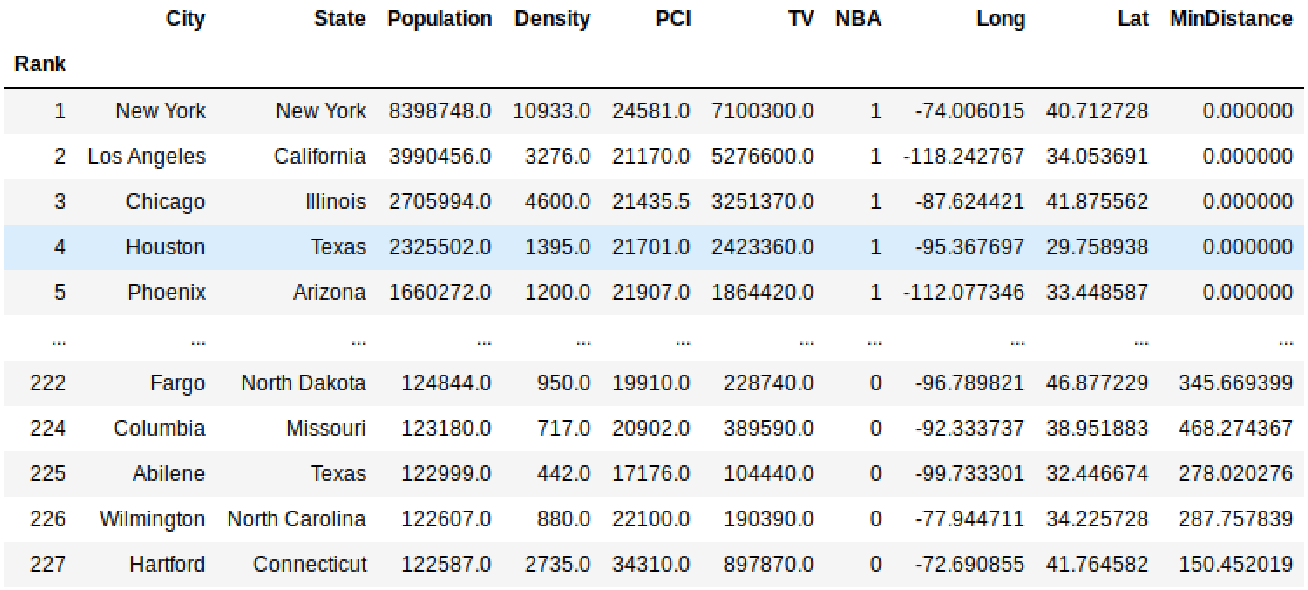
Irrelevant data from these sources can get dropped. The features carried forward will be City, State, Population, Population Density, Per Capita Income, TV Market Size and NBA Team.

Both the TV market data and the per-capita income are provided for a region (multiple cities per row). To merge these data sets, a check is made whether the city is listed in the row and the corresponding per-capita income or TV market size is returned. All numerical data is then converted to type float.

After merging these data sets, some of the cities are missing either per capita income or TV market size. To handle this missing data, the data is sorted by population and the per capita income and TV market size are interpolated to fill the NaNs.

With the data sets cleaned and merged, we end up with information on 314 cities. To further reduce the size of the data set, two actions were taken. First, the longitude and latitude of each city was pulled from the python module geopy and the distance from each city to the nearest NBA city was calculated. Cities less than 150 km from a current NBA city were excluded due to a regionally close NBA city already existing. Second, only the 120 largest cities of the remaining set were carried forward (27 of which are NBA cities). This is a reasonable restriction since a sizable city is needed to support an NBA franchise.

The cleaned and merged dataframe looks like that presented in figure 1.

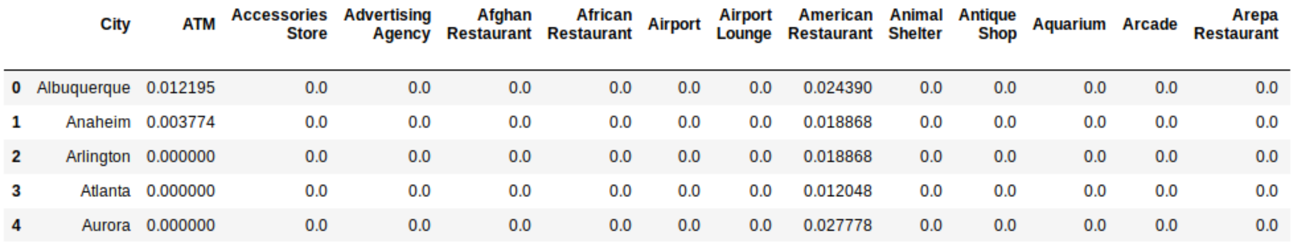


	City	State	Population	Density	PCI	TV	NBA	Long	Lat	MinDistance
Rank										
1	New York	New York	8398748.0	10933.0	24581.0	7100300.0	1	-74.006015	40.712728	0.000000
2	Los Angeles	California	3990456.0	3276.0	21170.0	5276600.0	1	-118.242767	34.053691	0.000000
3	Chicago	Illinois	2705994.0	4600.0	21435.5	3251370.0	1	-87.624421	41.875562	0.000000
4	Houston	Texas	2325502.0	1395.0	21701.0	2423360.0	1	-95.367697	29.758938	0.000000
5	Phoenix	Arizona	1660272.0	1200.0	21907.0	1864420.0	1	-112.077346	33.448587	0.000000
...	...	...	...	...	...	...	...	...	...	...
222	Fargo	North Dakota	124844.0	950.0	19910.0	228740.0	0	-96.789821	46.877229	345.669399
224	Columbia	Missouri	123180.0	717.0	20902.0	389590.0	0	-92.333737	38.951883	468.274367
225	Abilene	Texas	122999.0	442.0	17176.0	104440.0	0	-99.733301	32.446674	278.020276
226	Wilmington	North Carolina	122607.0	880.0	22100.0	190390.0	0	-77.944711	34.225728	287.757839
227	Hartford	Connecticut	122587.0	2735.0	34310.0	897870.0	0	-72.690855	41.764582	150.452019

Figure 1. Cleaned and merged dataframe ready for importation of FourSquare data.

The final step is adding the FourSquare city locale data. Since the FourSquare API limits the number of venues returned to 100, the explore venues call was made 9 times for each city. Once at the longitude and latitude of the city plus 8 more time at the different permutations of longitude and latitude +/- 0.2. This also helps to capture the broader region when assessing the cities capability to accommodate an NBA city.

Venues by locale are one-hot encoded as venue category and the mean for each venue category in a given city is merged with the dataset shown in figure 1. Example of the normalized city-venue data is shown in figure 2.



	City	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant
0	Albuquerque	0.012195	0.0	0.0	0.0	0.0	0.0	0.0	0.024390	0.0	0.0	0.0	0.0	0.0
1	Anaheim	0.003774	0.0	0.0	0.0	0.0	0.0	0.0	0.018868	0.0	0.0	0.0	0.0	0.0
2	Arlington	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.018868	0.0	0.0	0.0	0.0	0.0
3	Atlanta	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.012048	0.0	0.0	0.0	0.0	0.0
4	Aurora	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.027778	0.0	0.0	0.0	0.0	0.0

Figure 2. Normalized city venue data.

3. Methodology

Exploratory Data Analysis

To guide decisions in the data analysis phase, some exploratory analysis was conducted first to better understand the data collected.

Cities in the data set are shown on the map in Figure 3. Cities with an NBA franchise are colored Red, cities without a franchise are colored Blue. The size of the city is proportional to the size of the marker. Initial observation of the map indicates regions where there are few NBA franchises. For example, the mid-west region between Tennessee and Nebraska is a region relatively bare of NBA franchises. The northwest is also generally lacking for teams, with only 1 team in the region.

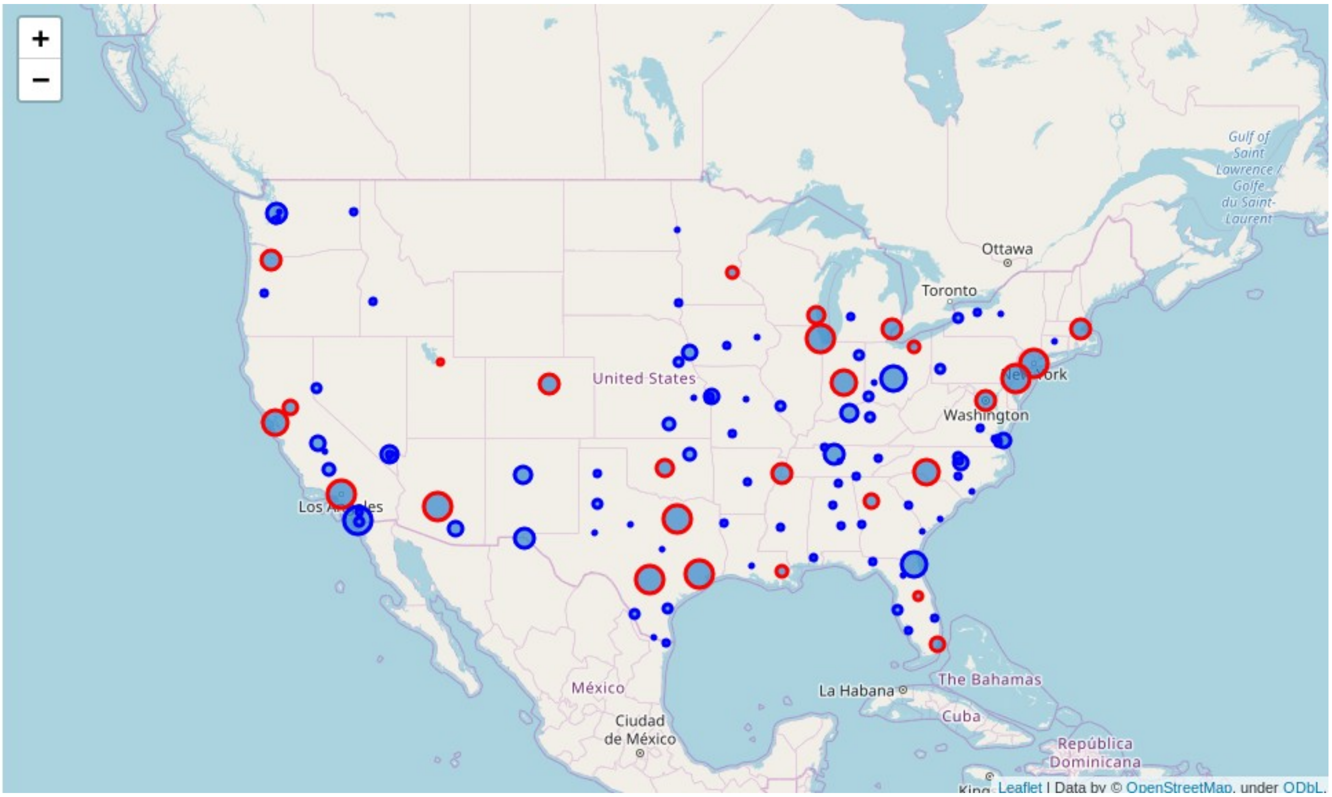
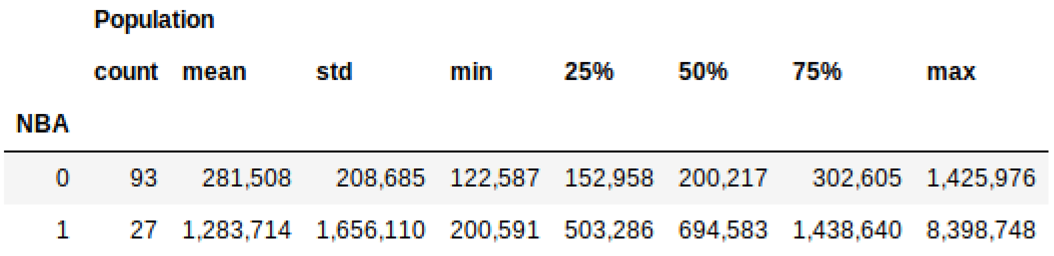
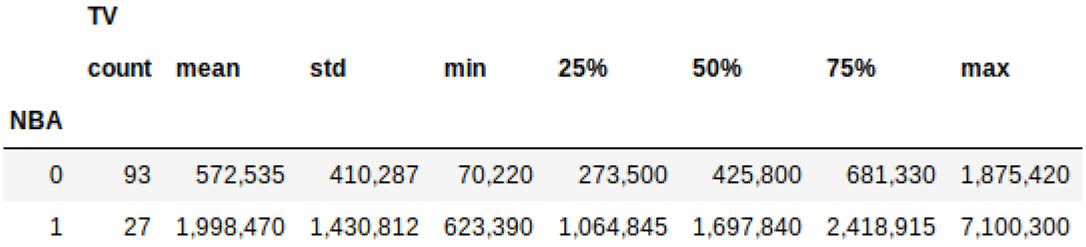


Figure 3. Map of US depicting cities already with an NBA franchise (RED) and those without (BLUE). Size of the marker is proportional to the size of the city.

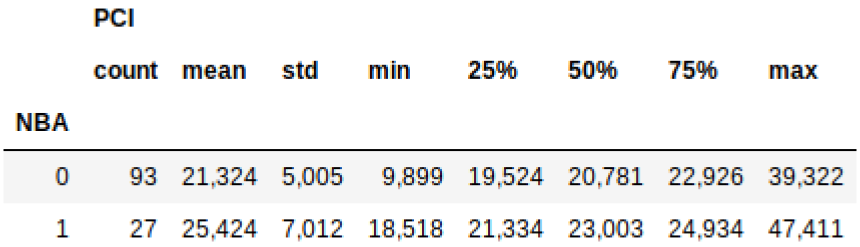
As a first look at how the NBA cities and non-NBA cities compare, the cities were grouped by city type (NBA vs non-NBA) and descriptive statistics were accumulated as depicted in Table 1.



	Population							
	count	mean	std	min	25%	50%	75%	max
NBA								
0	93	281,508	208,685	122,587	152,958	200,217	302,605	1,425,976
1	27	1,283,714	1,656,110	200,591	503,286	694,583	1,438,640	8,398,748



	TV							
	count	mean	std	min	25%	50%	75%	max
NBA								
0	93	572,535	410,287	70,220	273,500	425,800	681,330	1,875,420
1	27	1,998,470	1,430,812	623,390	1,064,845	1,697,840	2,418,915	7,100,300



	PCI							
	count	mean	std	min	25%	50%	75%	max
NBA								
0	93	21,324	5,005	9,899	19,524	20,781	22,926	39,322
1	27	25,424	7,012	18,518	21,334	23,003	24,934	47,411

Table 1. Summary statistics comparing average NBA and non-NBA cities.

Not surprisingly, the mean NBA city population and TV market are larger than non-NBA cities. That being said, NBA cities did not make up the 27 most populous cities (30 teams but 1 team is in Canada and Los Angeles and New York each have 2 teams) as gleaned from the overlapping min and max for each data set. This indicates there may be additional markets with the potential to house an NBA team.

Distributions between NBA and non-NBA cities can be more clearly seen in the box-plots of figures 4-6, reaffirming the observation made above.

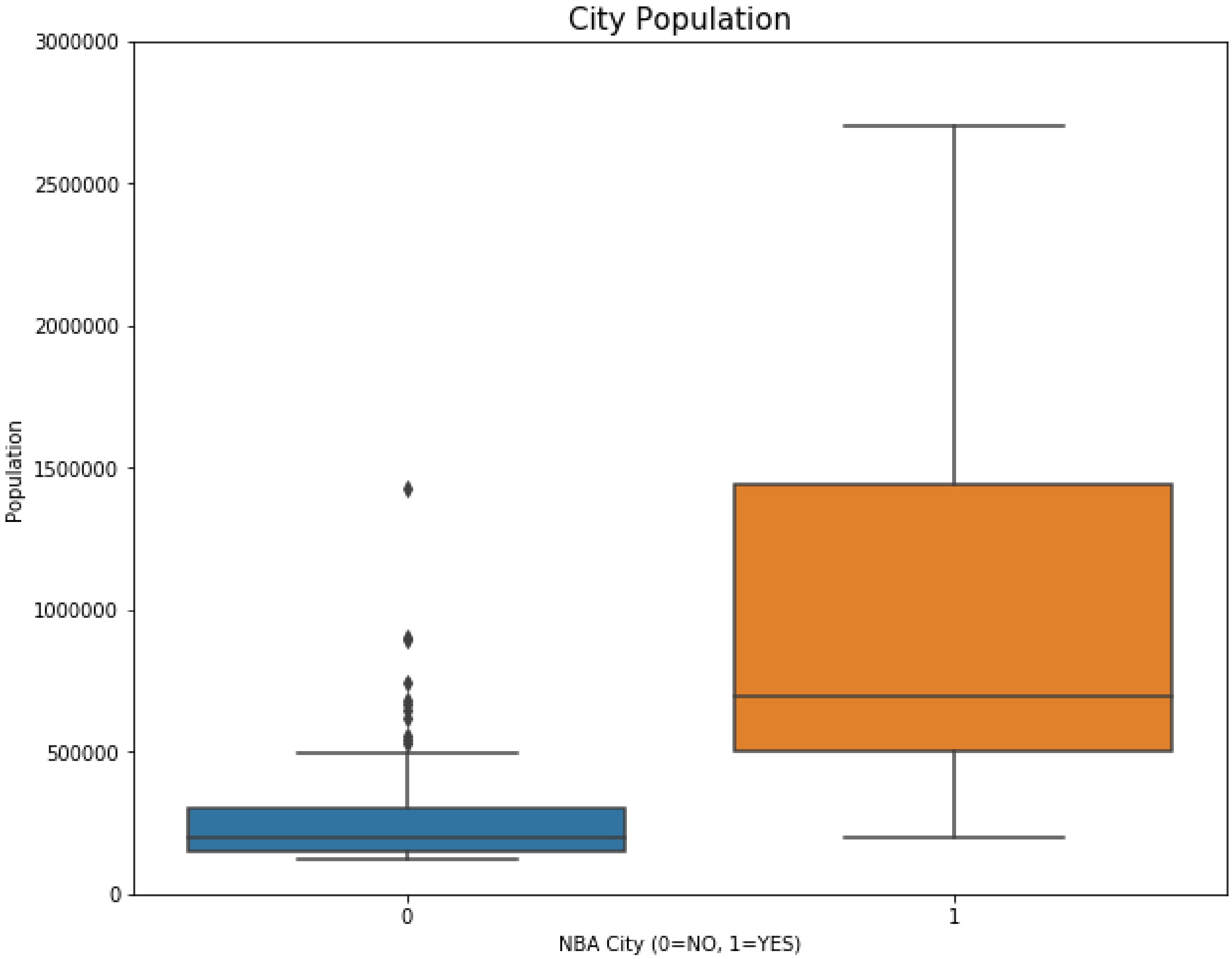


Figure 4. Box plot of population data comparing NBA cities vs non-NBA cities. There are numerous non-NBA cities with populations greater than those of cities with NBA teams.

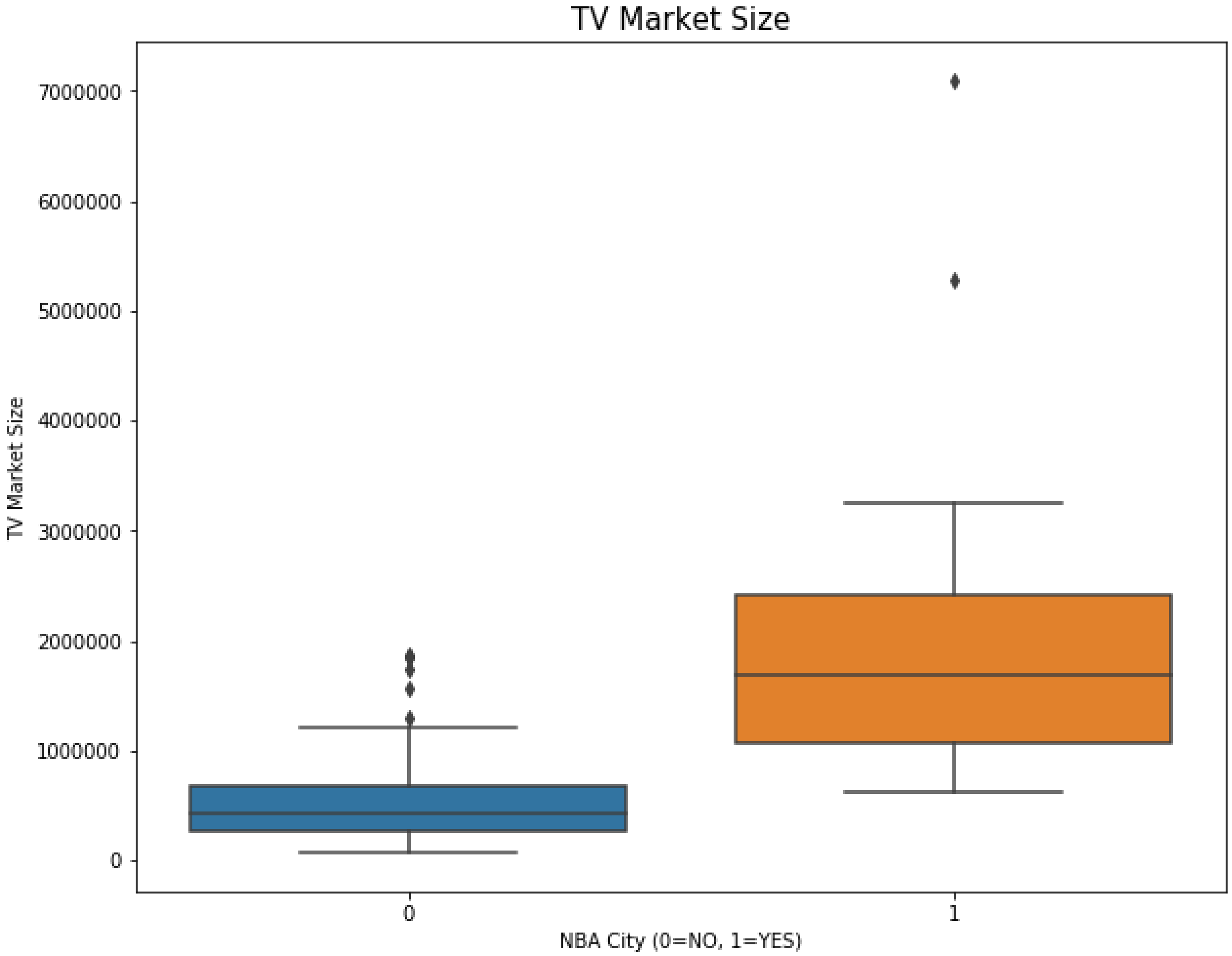


Figure 5. Box plot of TV market size comparing NBA cities vs non-NBA cities. There are numerous non-NBA cities with TV market sizes greater than those of cities with NBA teams.

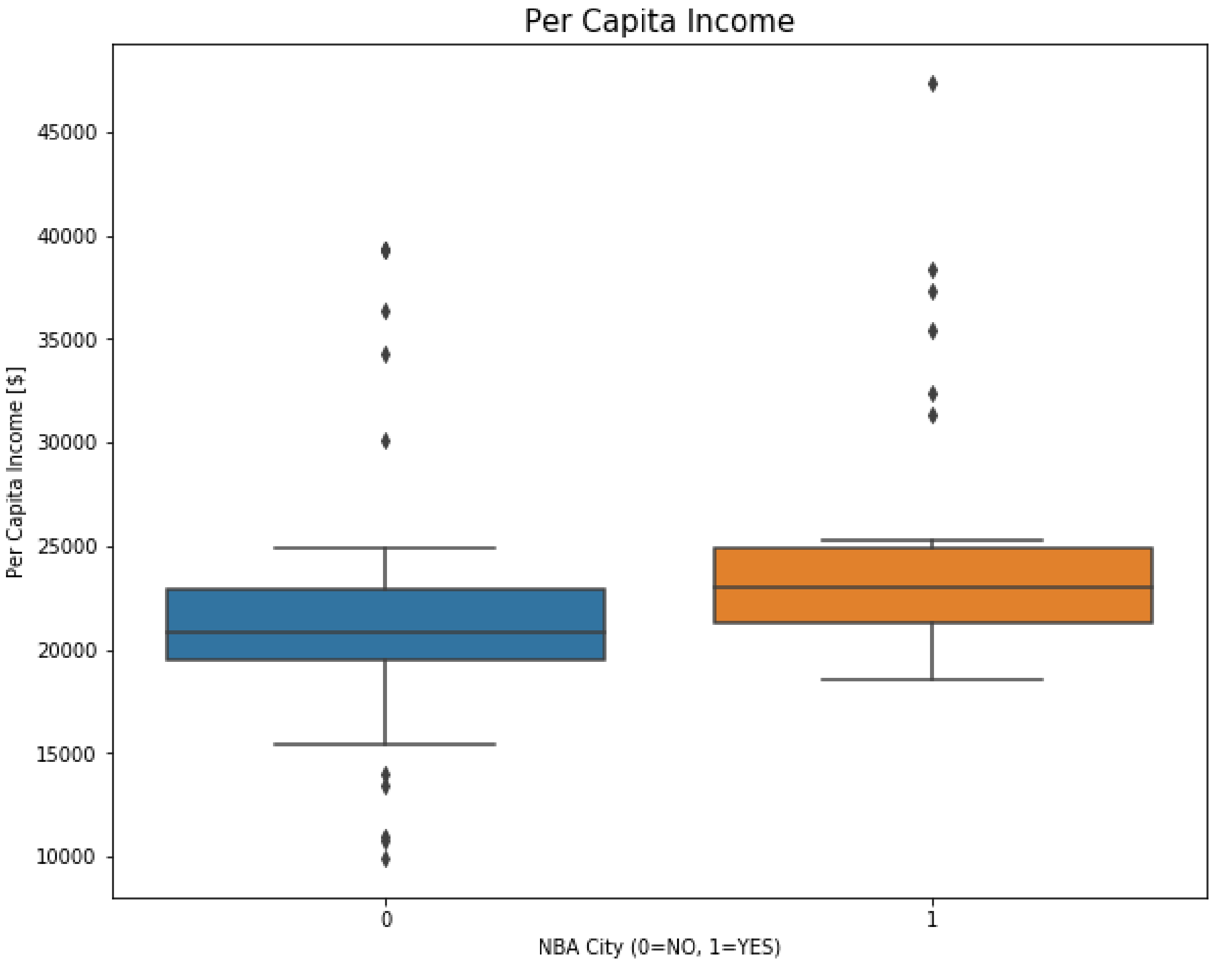


Figure 6. Box plot of Per-Capita Inceom comparing NBA cities vs non-NBA cities.

A scatter plot of TV market size vs population is provided in figure 7 with yellow symbols representing cities with an NBA team and purple symbols representing cities without an NBA team. This plot confirms the sentiment stated earlier that there are plenty of non-NBA cities with a large enough population and a large enough TV market to accommodate an NBA team.

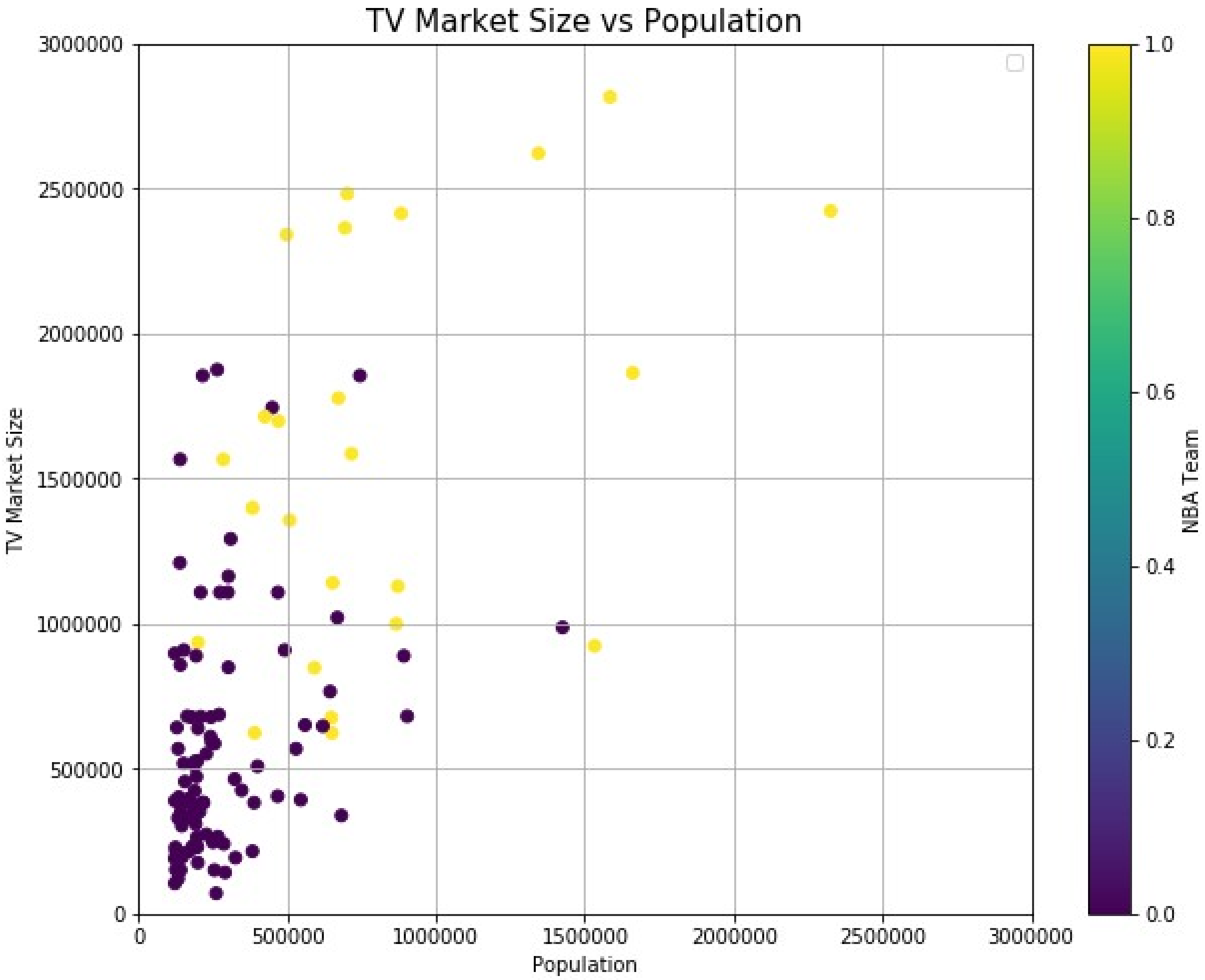


Figure 7. Scatter plot of TV market size vs Population. Non-NBA cities in purple. NBA cities in Yellow.

To visualize the FourSquare locale data, the cities were grouped by whether or not they had an NBA team and the mean locale information was taken. A bar chart showing how the mean percentage of total venues by category compared between these means is provided in Figure 8.

To simplify the data in figure 8, a delta of the two means was created and a bar chart showing cases where the delta was greater than 0.0015 was created as shown in figure 9. Delta is NBA city minus non-NBA city implying positive bars are more often found in NBA cities than non-NBA cities.

Though mostly similar across cities, these comparisons indicate that some venues are more characteristic of NBA cities than non-NBA cities. For example, Gyms and fitness center tend to be more typical of NBA cities where as fast food eateries are more common in non-NBA cities.

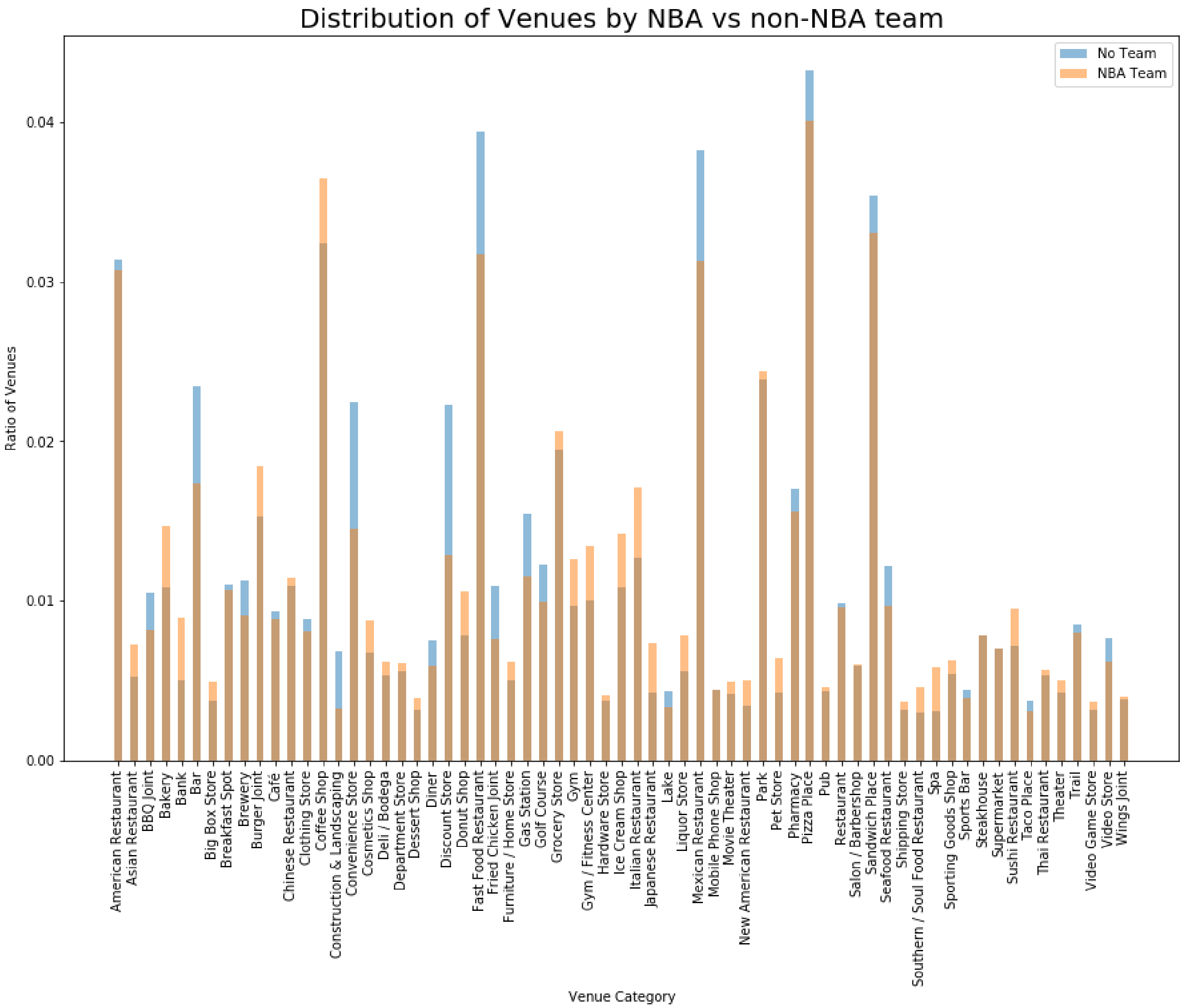


Figure 8. Comparison of percentage of total venues by category.

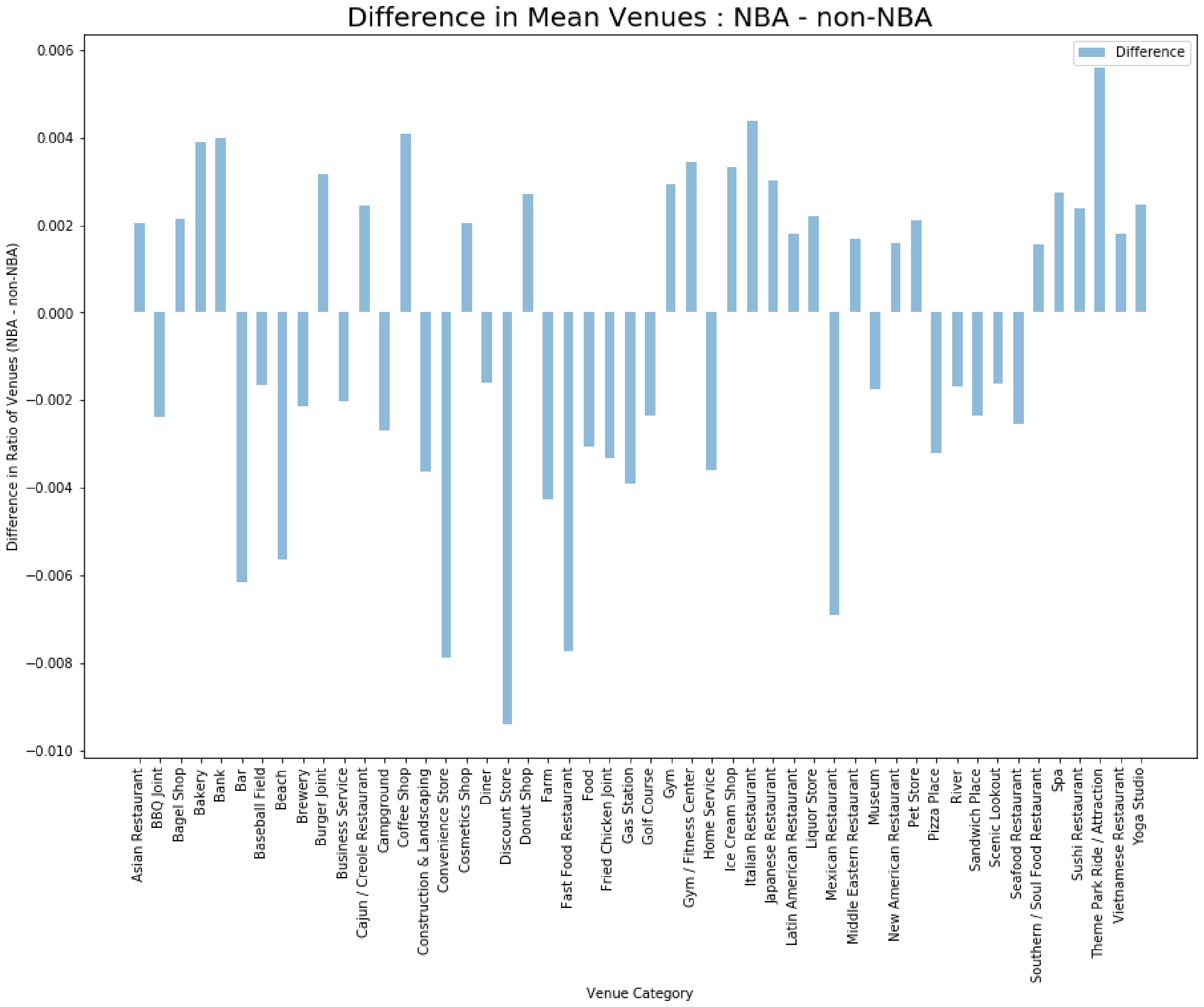


Figure 9. Delta comparison of percentage of total venues by category. Delta is NBA city minus non-NBA city.

Modeling

The modeling approach used to find the next NBA city is clustering. Two clustering algorithms were applied, each yielding similar results.

A few additional preparation steps were needed to get the data ready for clustering. First, the NBA cities were divided into 3 groups, the 10 most populous, the 10 next most populous and 7 least populous. The mean of each of these groups was then taken to represent an average ‘large’ NBA city, an average ‘Mid-size’ NBA city and a ‘small’ NBA city. These are the targets around which the candidate non-NBA cities should cluster. This action left the remaining data set size at 96, 93 non-NBA cities and 3 different sized average NBA cities.

Finally, the text (City and State) and location (Long, Lat and NBA) data were dropped from the dataframe leaving the locale, population, TV market and per capita income data as the inputs to the clustering model.

Both a KMeans model (with k=7 and init='k-means++') and an Agglomerative Hierarchical clustering (with k= 4 and linkage = 'complete') were used to cluster the cities. The conclusions drawn from each method of clustering are similar so the results and discussion sections that follow focus on only the KMeans clustering approach.

The clustering labels generated by KMeans were added to the dataset for further analysis.

4. Results

The KMeans algorithm returned seven clusters of cities. The value counts of these clusters are depicted in table 2.

Cluster Number	Number of Cities	NBA City?
0	14	Avg. Small
1	20	--
2	1	Avg. Large
3	6	Avg. Medium
4	5	--
5	9	--
6	41	--

Table 2. Cluster groups generated by KMeans clustering on city data.

Of the seven clusters returned, three contained one of the average NBA city types segmented earlier. The average Large NBA city is in a class by itself, the average Medium NBA city is in cluster with 5 other cities, and the average Small NBA city is in a cluster with 13 other cities.

Figure 10 depicts these clusters on a map of the US. The NBA city clusters are shown in the Pacific Ocean with the size of the marker indicative of the size of the sample. From Table 2 and Figure 10, one can conclude that:

- Cluster 0 = Avg Small NBA City = **RED** in Figure 10
- Cluster 3 = Avg Medium NBA City = **GREEN** in Figure 10
- Cluster 2 = Avg Large NBA City = **BLACK** in Figure 10

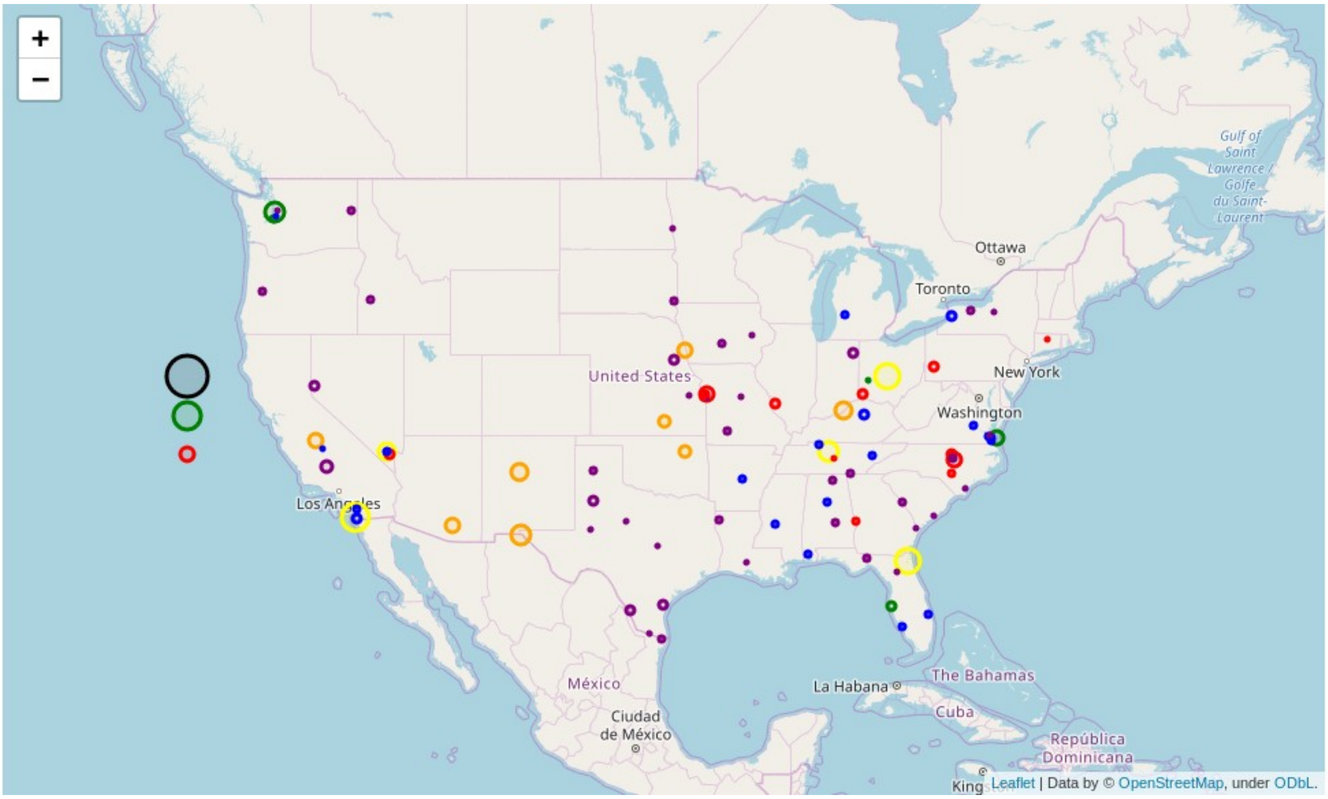
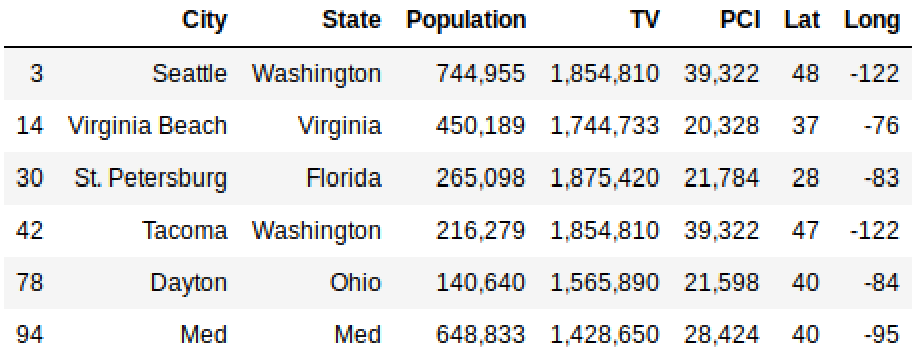


Figure 10. Clusters returned by KMeans algorithm. NBA cities clusters are shown in the Pacific Ocean with the size of the marker indicative of the size of the sample.

Of the three clusters containing representative (i.e. Avg) NBA cities, focus should be placed on the Medium size NBA city cluster (Cluster 3, GREEN symbols in figure 10). The large NBA city cluster only contains itself. The small NBA city cluster contains 14 cities and markets that would likely be less accommodating to an NBA city than the Medium sized group. Table 3 lists out the Cities and details contained in this cluster.



	City	State	Population	TV	PCI	Lat	Long
3	Seattle	Washington	744,955	1,854,810	39,322	48	-122
14	Virginia Beach	Virginia	450,189	1,744,733	20,328	37	-76
30	St. Petersburg	Florida	265,098	1,875,420	21,784	28	-83
42	Tacoma	Washington	216,279	1,854,810	39,322	47	-122
78	Dayton	Ohio	140,640	1,565,890	21,598	40	-84
94	Med	Med	648,833	1,428,650	28,424	40	-95

Table 3. Cities in the Avg Medium sized NBA city cluster (cluster 3, GREEN symbols in Figure 10).

5. Discussion

As the NBA considers expansion, the next NBA owner and the league itself will need to decide where to place the new team. Given the recent success of the NBA with its increased TV ratings and revenue, the assumption is that the current grouping of NBA cities is setup for success. Finding additional cities that fit this template would provide candidates for expansion.

During this study, it was observed that though many of the most populous cities in the US had an NBA team, it wasn't just the top 27 most populous cities that had an NBA team. This indicated there are US cities with populations and TV markets large enough to sustain an NBA team.

To narrow down the list of candidate cities for expansion, the cities were grouped by KMeans clustering on population and city venue data to see which cities were most like the existing NBA cities. Of the 93 non-NBA cities in our dataset, 5 were considered representative of our average Medium sized NBA city. As shown in Table 3, these were Seattle and Tacoma, WA, Virginia Beach, VA, St. Petersburg, FL and Dayton, OH.

Based on the information in this study and report, the recommendation would be to **select Seattle** as the location for the next NBA city. Not only does it have the largest population amongst our candidate cities, but another city on the list, Tacoma, is in the same state and would likely support interest in a new Seattle NBA team.

6. Conclusion

This study looked at cities suitable for an NBA franchise as the NBA considers expansion. The information in this report suggested Seattle, WA is likely the best candidate for expansion based on population, TV market and city venue makeup.

That being said, there are additional features that could be added to the data set which may change the answer. A couple of things that could be added in future work include:

- Are sports-related venues rated higher in NBA cities than non-NBA cities?
- Conduct polls in the cities to get overall fan interest in a new team and add that to the data set
- Are college sports teams filling the need for an NBA city?

Furthermore, there are additional considerations that need to be investigated that can't be reflected in the data before making any final decisions. To name a few:

- Is the city willing to fund or build a stadium?
- Is there an area within the city to build a new stadium or otherwise house the team?
- Is there ample public transit and/or parking nearby to get people to the stadium?

These notes aside, however, the model and data presented here provide a good starting point for narrowing down a list of US cities for possible expansion.