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, correlation and clustering algorithms will be applied to this dataset 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 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 datasets, 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 get fill the NaNs.

With the datasets cleaned and merged, we end up with information on 314 cities. To limit the size of the data when incorporating the FourSquare venue data, the data set is limited to the 100 largest 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.

Rank	City	State	Population	Density	PCI	TV	NBA
1	New York	New York	8398748.0	10933.0	24581.000000	7.100300e+06	1
2	Los Angeles	California	3990456.0	3276.0	21170.000000	5.276600e+06	1
3	Chicago	Illinois	2705994.0	4600.0	21435.500000	3.251370e+06	1
4	Houston	Texas	2325502.0	1395.0	21701.000000	2.423360e+06	1
5	Phoenix	Arizona	1660272.0	1200.0	21907.000000	1.864420e+06	1
6	Philadelphia	Pennsylvania	1584138.0	4511.0	22874.000000	2.816850e+06	1
7	San Antonio	Texas	1532233.0	1250.0	18518.000000	9.239900e+05	1
8	San Diego	California	1425976.0	1670.0	22926.000000	9.877600e+05	0
9	Dallas	Texas	1345047.0	1493.0	23616.000000	2.622070e+06	1
10	San Jose	California	1030119.0	2231.0	40392.000000	2.414470e+06	0

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

The final step is adding the FourSquare city locale data. 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 a 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.