# Clustering Analysis on the Boston Housing Dataset

Andrew Roberts
3/15/2018

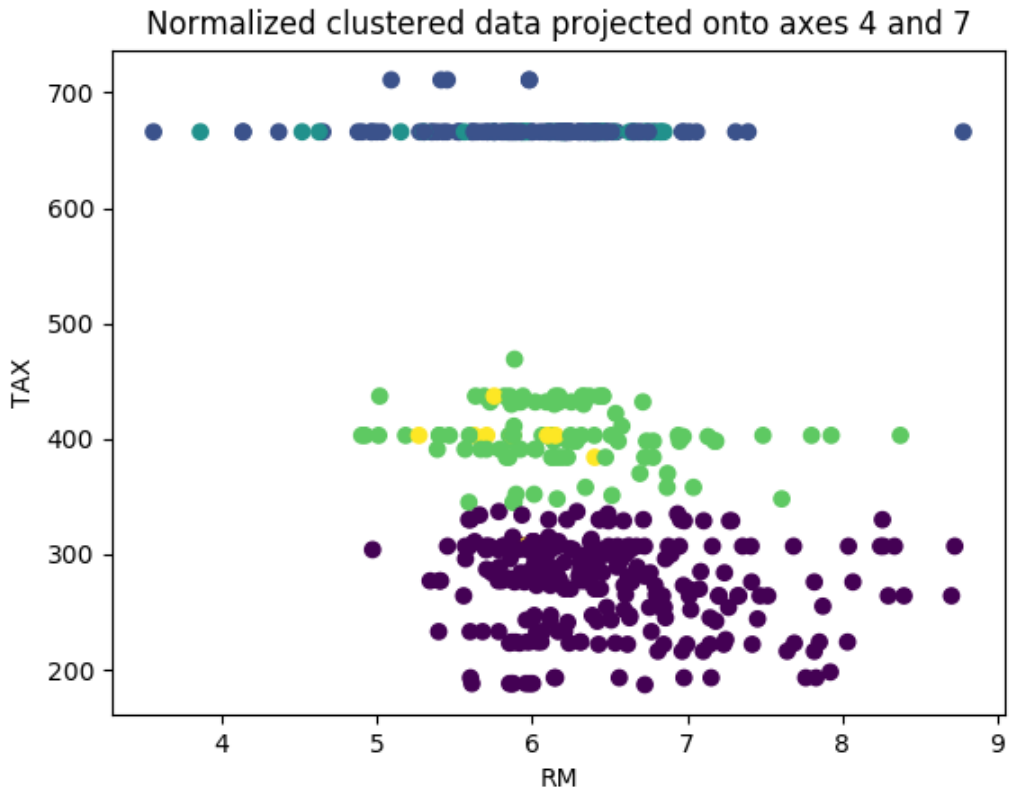
## Introduction

In this paper I perform an unsupervised clustering analysis of the Boston Housing Dataset (BHD), utilizing k-means and Principal Component Analysis (PCA). The Boston Housing Dataset contains housing data for 506 census tracts of Boston taken from the 1970 census. Although it is often analyzed using supervised learning techniques, I focus entirely on unsupervised clustering and therefore remove the target variable (median house price) from my analysis. Additionally, I remove the two categorical features in the data because k-means relies on Euclidean distance, which has little meaning for non-continuous data. The remaining features in the 506 x 11 dataset are given below:

| Feature | Description |
|---|---|
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq. ft. |
| INDUS | proportion of non-retail business acres per town |
| NOX | nitrogen oxides concentration (parts per 10 million) |
| RM | average number of rooms per dwelling |
| AGE | proportion of owner-occupied units built prior to 1940 |
| DIS | weighted mean of distances to five Boston employment centres |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |
| BLACK | 1000(Bk-.63)^2 where Bk is the proportion of black individuals by town |
| LSTAT | lower status of the population (percent) |

## Naïve Analysis

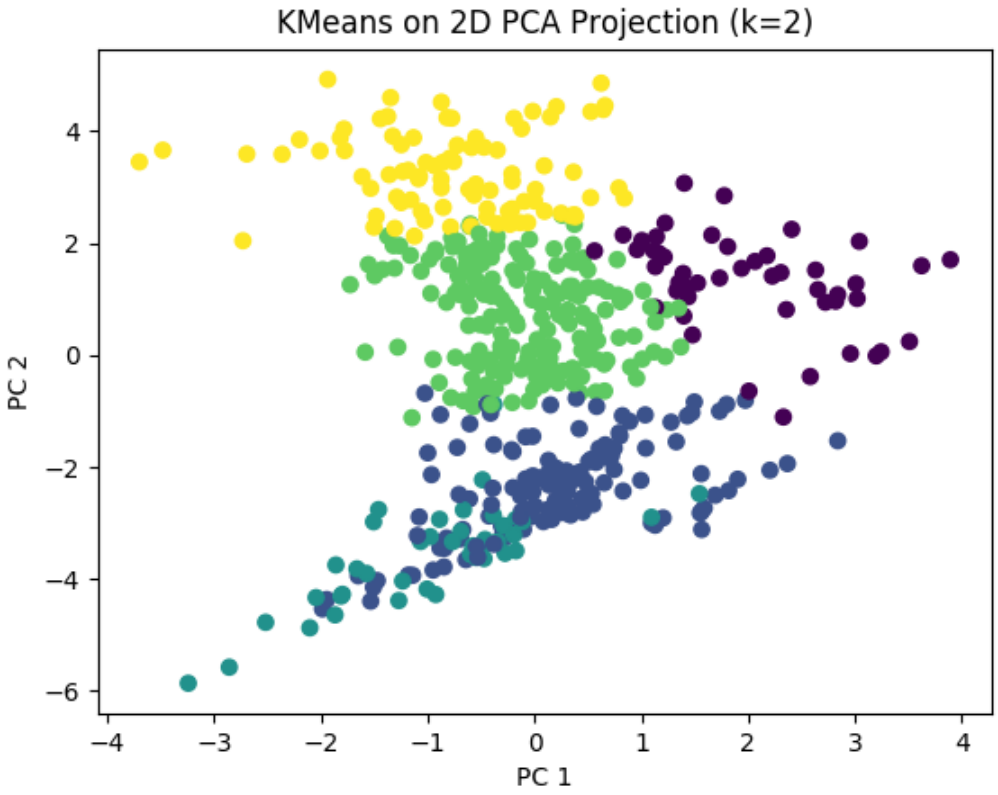
I begin by projecting the 11-dimensional observations onto arbitrary 2-dimensional subspaces to attempt to discern any obvious clustered structure in the data. I also begin by arbitrarily choosing the number of clusters, k, to be 5. After z-normalizing the data, the results projected onto the axes associated with RM and TAX look like this (here I cluster on the normalized data but display the original data because the clusters are more distinguishable):



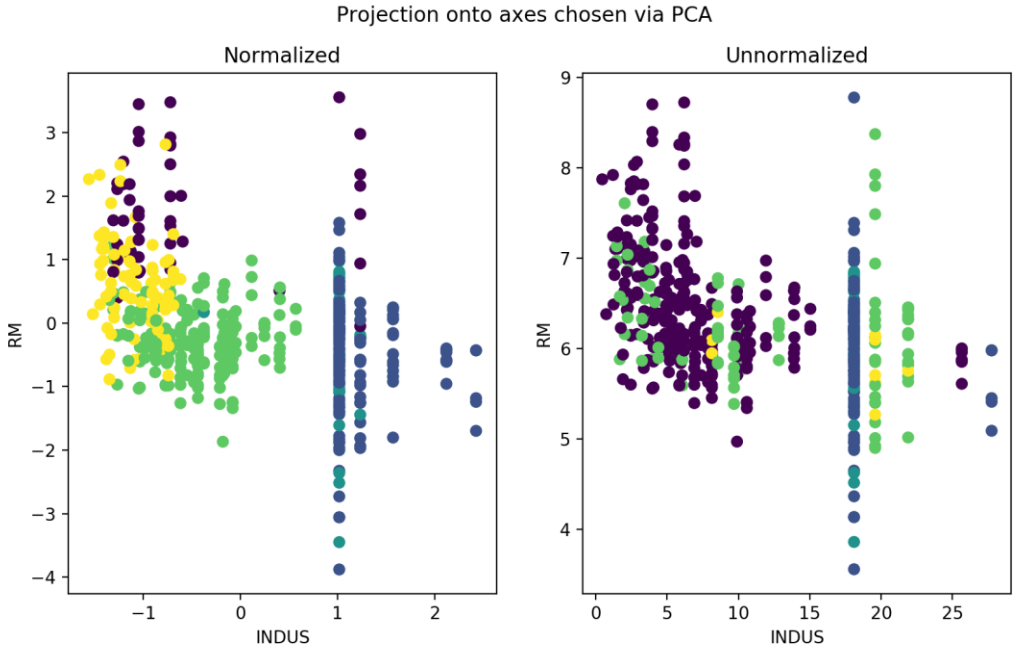

For now, I continue my analysis with k=5; later in the paper I turn to empirical methods to find the optimal value of k. In the image above, the clustering structure is certainly not completely clear, but shows some promising results. To determine if there exists a 2-D projection to better reveal the cluster structure in the data, I find the optimal 2-D linear subspace that best approximates the data using PCA. Operating under the idea that clusters may be more clearly defined in a projection onto this subspace, I run k-means with k =5 and then project all observations onto the subspace spanned by the first two principal directions (plot on next page).

This perspective provides a better-defined view of the cluster structure in the data. PC1 and PC2 are linear combinations of the original features, so I calculate the Pearson correlation coefficient between each of these and the original features in the BHD. A high correlation between a certain feature and principal component indicates that the feature is weighted heavily in the linear combination. I calculate thirteen correlation coefficients for each principal component but only list the top three for each below.

| Var1 | Var2 | Corr(Var1, Var2) |
|---|---|---|
| INDUS | PC1 | -0.87 |
| NOX | PC1 | -0.85 |
| DIS | PC1 | 0.82 |
| RM | PC2 | 0.51 |
| PTRATIO | PC2 | -0.49 |
| DIS | PC2 | -0.44 |




I provide intuition as to why I believe these are the dominant correlations later in this paper, but for now I simply take the features corresponding to the highest correlation with PC1 (INDUS) and PC2 (RM) to form the 2-D plane onto which I project the data. The result is shown below for both normalized and unnormalized data:



While not as clearly defined as those in 2-D principal component space, the clusters still look fairly reasonable, especially when the data is normalized. The two methods discussed in this section are intended to provide a general grasp of the data and its potential for cluster analysis. It should be emphasized that k was arbitrarily chosen, highlighting the need for more careful analysis.

## Z-Normalization

In the preceding section, I plot cluster results for both normalized and unnormalized data. I do this to illustrate the impact that scaling can have on k-means results, but will confine the rest of my analysis to only the normalized observations. The ability of k-means to cluster data relies on Euclidean distance, which is clearly sensitive to scale. For example, in the BHD CRIM is typically a small proportion on the order of .001 while INDUS can take values on the order of 10. The difference in these orders of magnitude can exert great influence on the determination of clusters. I have no reason to believe certain variables should be given more influence than others in the cluster determination, so I normalize each feature to mean zero and variance one.

## Finding the Optimal k

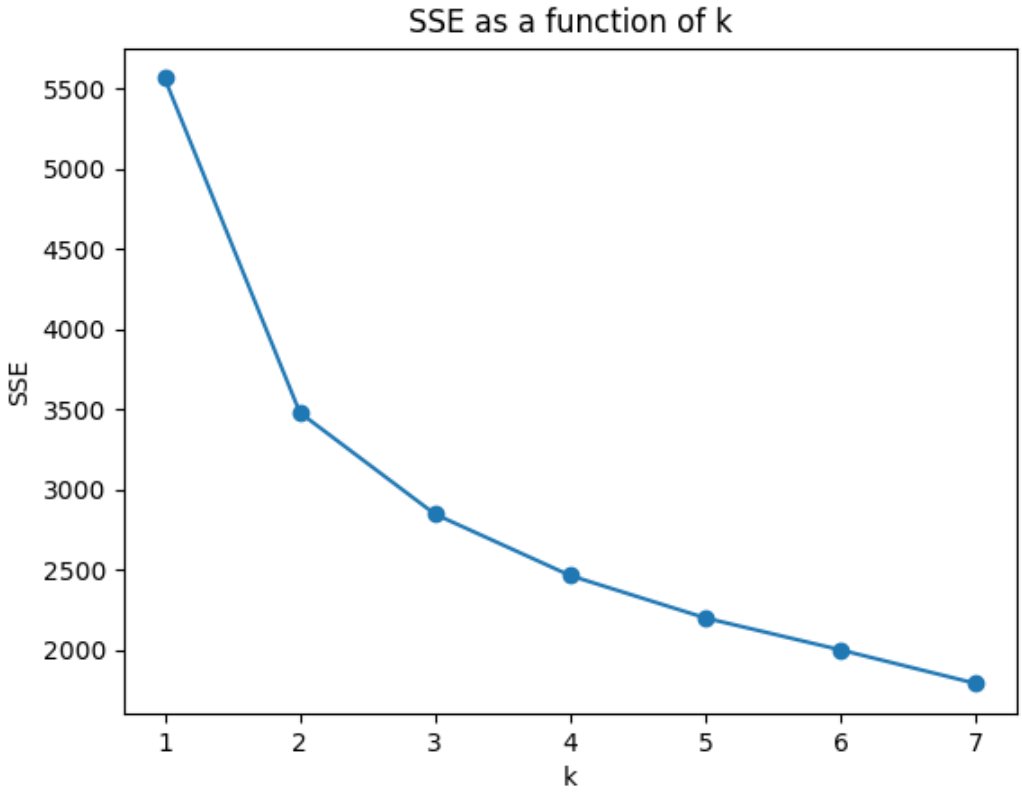
The choice of k to be 5 was completely arbitrary, so I employ empirical techniques in this section to determine the optimal number of clusters. One idea is to look at the 2-D PCA projection to see if clear clusters reveal themselves on this subspace. Observing the image shown above, it appears that the choice of k to be 5 segments the data too much, breaking up what appear to be large clusters into smaller sub-divisions. At a glance, k=2 might be a reasonable pick; the mass of points at the top of the plot seem to form one cluster, while the approximately linear spread of observations at the bottom may form another. I test this hypothesis empirically using two methods.

### *The elbow method*

One common metric to quantify the quality of a clustering is the sum of squared errors:

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

Here, the outer summation sums over the $k$ clusters and the inner calculates the sum of the squared Euclidean distance of each point in cluster $C_i$ to its centroid $c_i$. As a function of $k$, SSE will be minimized when the number of clusters equals the number of observations. Clearly, this is not useful but the SSE can be plotted for different values of $k$ and the number of clusters can be chosen based on the "elbow" of the graph, in other words where an increase in $k$ does not cause a significant decrease in SSE. I plot this curve for the normalized dataset below.



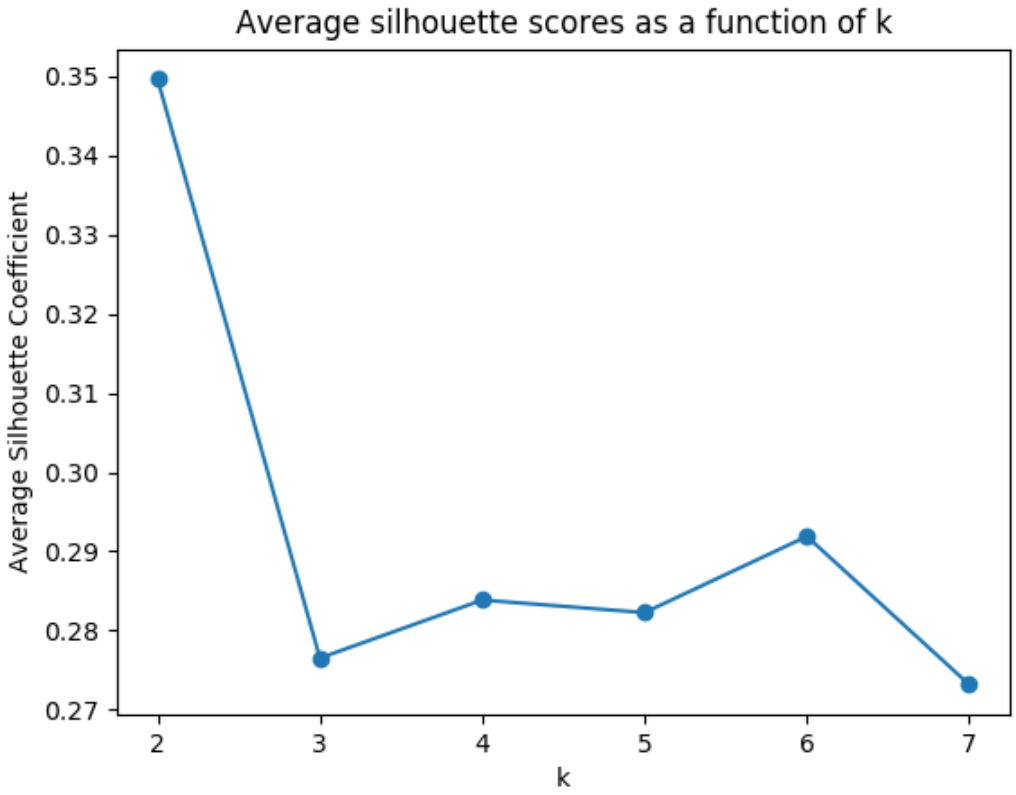

Potential choices for k based on this plot are 2 and 3, aligning closely with my previous determination using PCA.

### *The Silhouette Method*

Another method to evaluate the quality of a clustering relies on *Silhouette Scores*, a measure of how well each point lies within its cluster that ranges between -1 and 1. For each observation *i,* a silhouette score is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

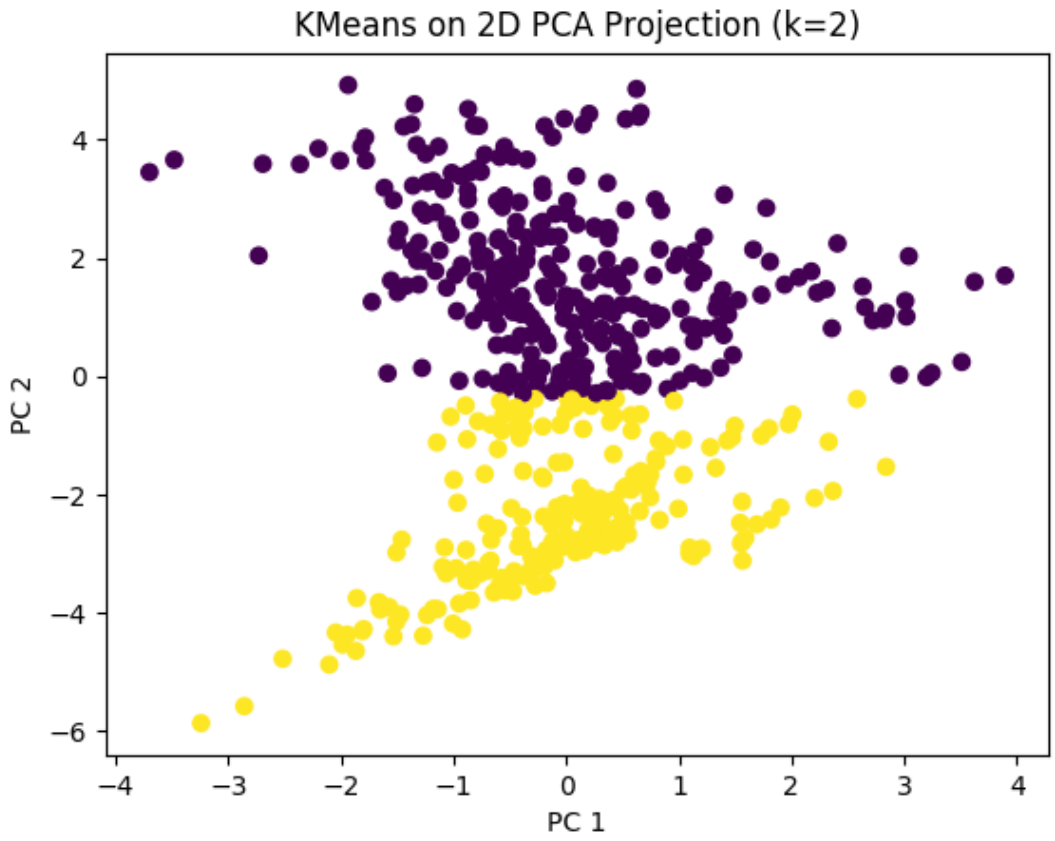
I define *a(i)* as the average Euclidean distance between point *i* and all other points within the same cluster, while *b(i)* is the minimum average Euclidean distance from *i* to all of the points in any other cluster. Intuitively, a good clustering would imply that *a(i)* is small and *b(i)* is large or, in other words, that S(i) is close to 1. I calculate the average silhouette score across all observations for different values of k and plot the results below.



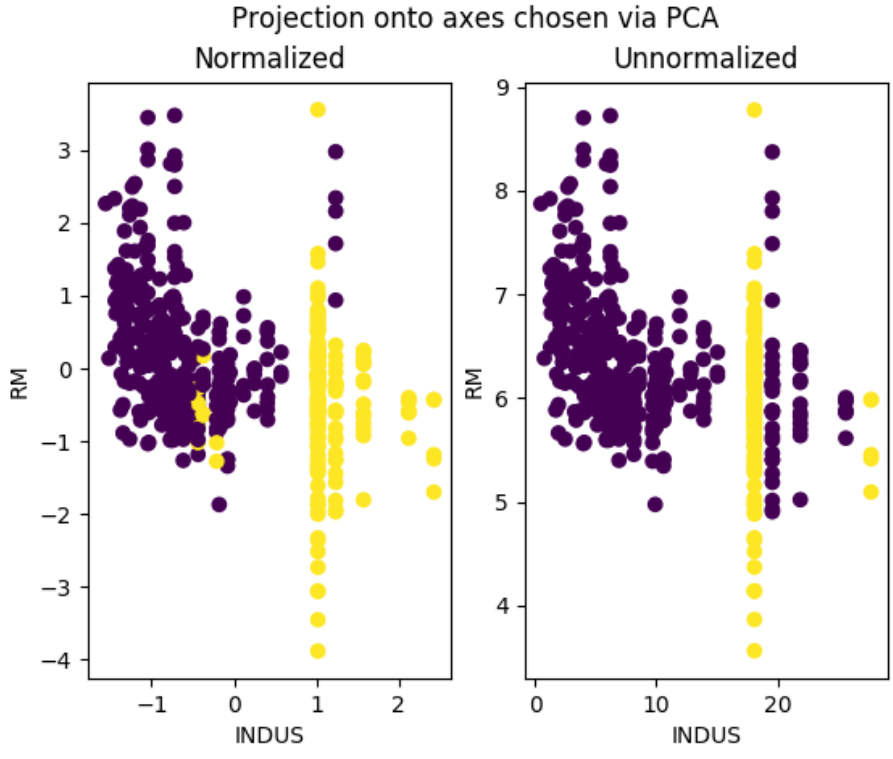

The value of k that maximizes the average silhouette score is 2, again supporting the previous findings.

### Clustering with k = 2

The empirical results from the last section imply that 2 is a good choice for k for this particular dataset so I proceed to run k-means on the BHD with k=2 in an attempt to uncover the reasoning behind the formation of the clusters. The below plot shows k-means applied to the 2-D PCA projection shown previously.




The clustering structure resembles what I hypothesized previously. For the sake of comparison, I also present the identical graph to the one shown in the initial analysis except with k=2 instead of 5.



At a simple glance, the clusters certainly appear to be more reasonable relative to when k was set to 5. But what do these clusters actually represent? This question requires one to relate the mathematics with an intuitive understanding of the features in a real-world context. In an attempt to accomplish this, I calculate the correlation of each feature with the cluster labels themselves (which is simply a binary variable).

| Feature | Corr(Feature, Labels) |
|---|---|
| INDUS | 0.858 |
| TAX | 0.816 |
| NOX | 0.786 |
| DIS | -0.653 |
| AGE | 0.637 |
| LSTAT | 0.630 |
| CRIM | 0.493 |
| B | -0.445 |
| PTRATIO | 0.389 |
| ZN | -0.388 |
| RM | -0.363 |

The largest correlations (in absolute value) provide valuable insight into the nature of the clusters found by k-means. INDUS, TAX, NOX, and DIS all act as proxy variables representing the proximity of houses to Boston. For example, the proportion of industry and air pollution (Nitrogen Oxide) intuitively should both increase as one moves away from the suburbs and nears the city. Digging deeper into the data, it becomes clear that the one-labeled observations are those of closer proximity to Boston while the zero-labeled data is farther from the city-center. The mean values for some of the key features help to illustrate this.

| | Mean Value (Label = 0) | Mean Value (Label = 1) |
|---|---|---|
| **INDUS** | 6.46 | 18.53 |
| **CRIM** | 0.23 | 8.91 |
| **DIS** | 4.89 | 2.07 |
| **NOX** | 0.48 | 0.67 |

In summary, through the utilization of vector quantization and dimensionality reduction techniques I have been able to determine that the observations in the BHD fall generally into two clusters, divided primarily on their proximity to highly populous and industrial locations (i.e., Boston).