1. Objective and Overview

The objective of this project is to use Principal Component Analysis technique to analyze the Boston Housing dataset in order to understand the relationship between the independent variables and median home values.

2. Data

The Boston Housing dataset was originally published by Harrison, D. and Rubinfeld, D.L. in, '*Hedonic prices and Demand for Clean Air*', J. Environ. Economics Management, vol. 5, 81-102, 1978., as part of an investigation on the models and methodology used to measure willingness to pay for clean air. It contains data collected by the U.S. Census Service in 1970 for 506 census tracts in the Boston, Massachusetts metropolitan area. Within it are 14 variables, 13 of which are continuous and 1 that is categorical. The data set is often used to baseline machine learning algorithms that predict the median home value for each tract based on the other variables. A full description of each variable is below:

1. Crime Rate - per capital crime rate by town
2. Zoning - proportion of residential land zoned for lots over 25,000 sq.ft.
3. Industry - proportion of non-retail business acres per town
4. Charles River - 1 if tract bounds river, 0 otherwise
5. Nitric Oxide - nitric oxides concentration (parts per 10 million)
6. Average Rooms - average number of rooms per dwelling
7. Units Older 1940 - proportion of owner-occupied units built prior to 1940
8. Employment Centers - weighted distance to five Boston employment centers
9. Highways - index of accessibility to radial highways
10. Tax Rate - full-value property tax rate per $ 10,000
11. Pupil Teacher Ratio - pupil-teacher ratio by town
12. African American - $1000(af - 0.63)^2$ where af is the proportion of African Americans by town
13. Lower Status - percent lower status of the population
14. Median Value - median value of owner-occupied homes in $ 1000s

Note that I questioned whether to include the African American feature. Race may have been an appropriate feature to include in 1978, but I do not think it is today. However, after seeing the biplot, I decided to retain the variable as it led to some interesting conclusions.

The dataset does have missing values. There are a total of six columns with missing values, impacting 112 rows. Upon review, I was comfortable making the assumption that the missing values in the Zoning and Charles River columns could be replaced with 0 due to the distribution of the data. After making that update, 79 rows with NAs remained, which were excluded from the analysis.

Given that the objective of this exercise is to visualize the correlation between the median home value and the other features, I added a Category feature to the dataset, which will be incorporated into the biplot. The categories are below:

1. 1 = below $ 10K
2. 2 = from $ 10K to $ 20K
3. 3 = from $ 20K to $ 30K
4. 4 = from $ 30K to $ 40K
5. 5 = from $ 40K to $ 50K

The histograms in Figure 1 visualize the distribution of the data for each variable. The continuous variables were binned in order to best visualize their general distribution. I also compared the plots before and after removing the rows with NAs. The omission did not impact the distribution of any of the variables.

3. Principal Component Analysis

The PCA analysis was run on the 13 independent variables (all but the Median Value feature). The scree plot in Figure 2 shows the percent variance explained by each of the 13 dimensions. As you can see, the first dimension explains 50.7% of the variance and the second explains 11.0%. In fact, the dataset only has three true dimensions as determined by performing a permutation test. The permutation test showed that eigenvalues greater that 0.712 are significant. This is also in line with the results if we take an average percentage, which finds that the dimensions with a percent of explained variance greater than 7.69% are significant.

The PCA biplot for the Boston Housing data is displayed in Figure 3. Upon review, we can draw the following conclusions:

1. The data points form two clear clusters, with the cluster on the right corresponding to tracts with a lower median property value (presumably the less desirable neighborhoods) and the cluster on the left corresponding to the tracts with a higher median property value (the more desirable neighborhoods).

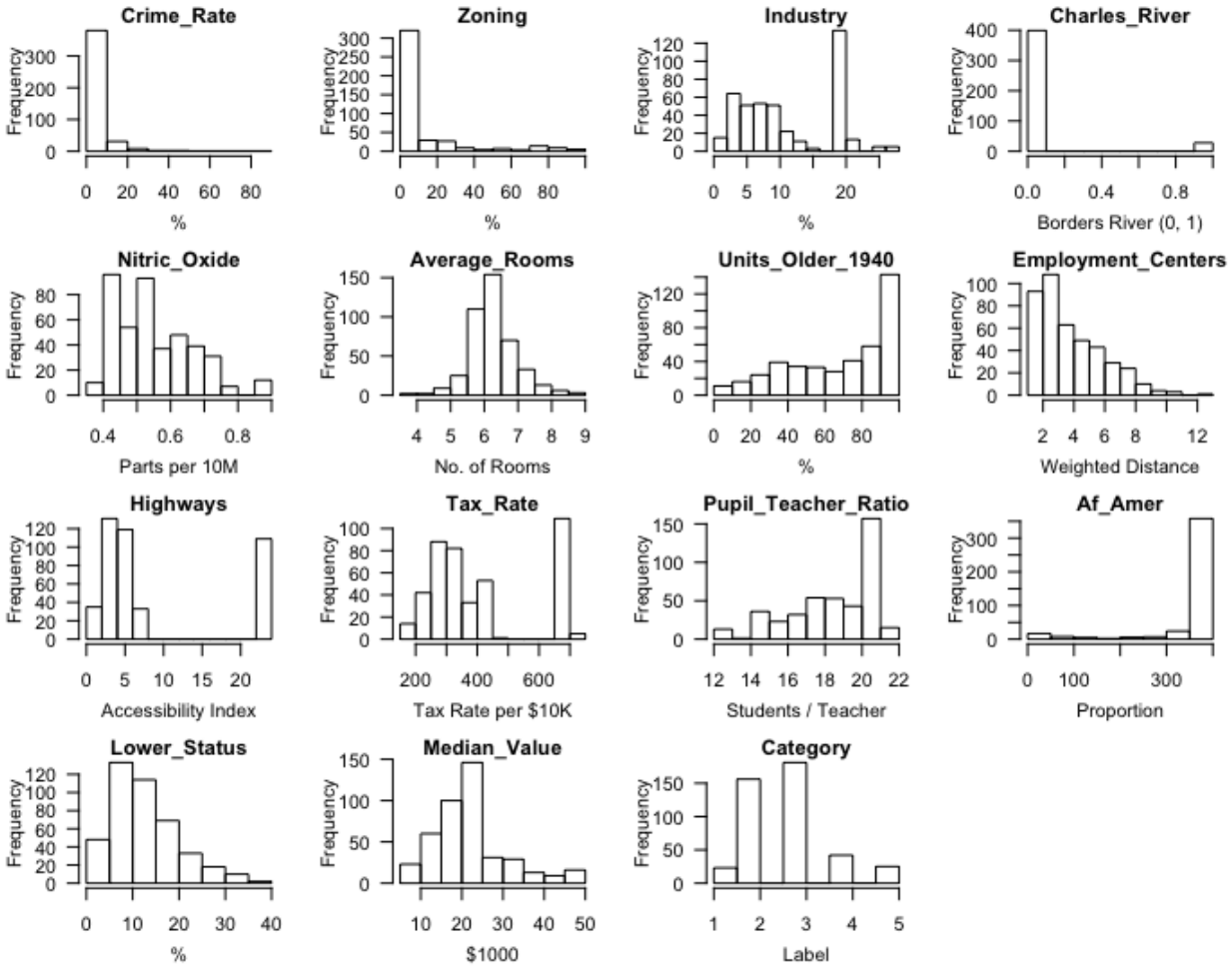

Figure 1: Data distribution histograms for each feature, including the manually added Category feature

2. Examples of features with a larger influence on the first dimension are the Tax Rate and Employment Centers.

3. Examples of features with a larger impact on the second dimension are African Americans and Zoning.

4. There is positive correlation between some variables. For example, Crime Rate, Highways, Tax Rate and Pupil-Teacher Ratio are correlated, as are Employment Centers are Zoning.

5. There are also examples of variables that are negatively correlated. For example, the percentage of African Americans in a neighborhood has an inverse correlation to the

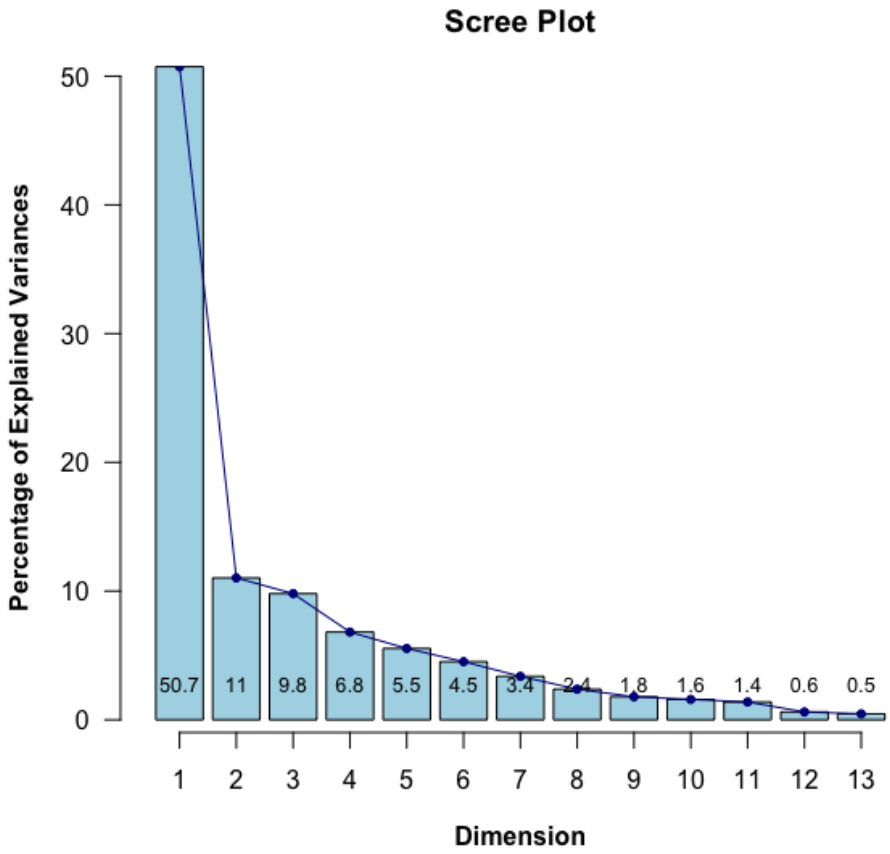

Figure 2: Sree plot illustrating that approximately 50% of the variance is explained by the first dimension and 11% by the second.

crime rate. Also, the number of Employment Centers and Zoning are negatively correlated to the number of units older than 1940.

4. Conclusions

The biplot enables strong conclusions to be drawn from the data. Due to the two clusters, it is clear that neighborhoods with a higher crime rate, are closer to a highway, have a higher percentage of older homes and have a higher tax rate are indicative of lower median home values, where those that have a lower crime rate, are further from highways, are closer to employment centers, have a higher percentage of residential land zoned for over 25,000 sq. ft. and have more rooms on average are indicative of neighborhoods with a higher median value. Unfortunately the tract location (or latitude / longitude) was not provided in the dataset, so we can not see where each tract is located, but we could infer that the tracts

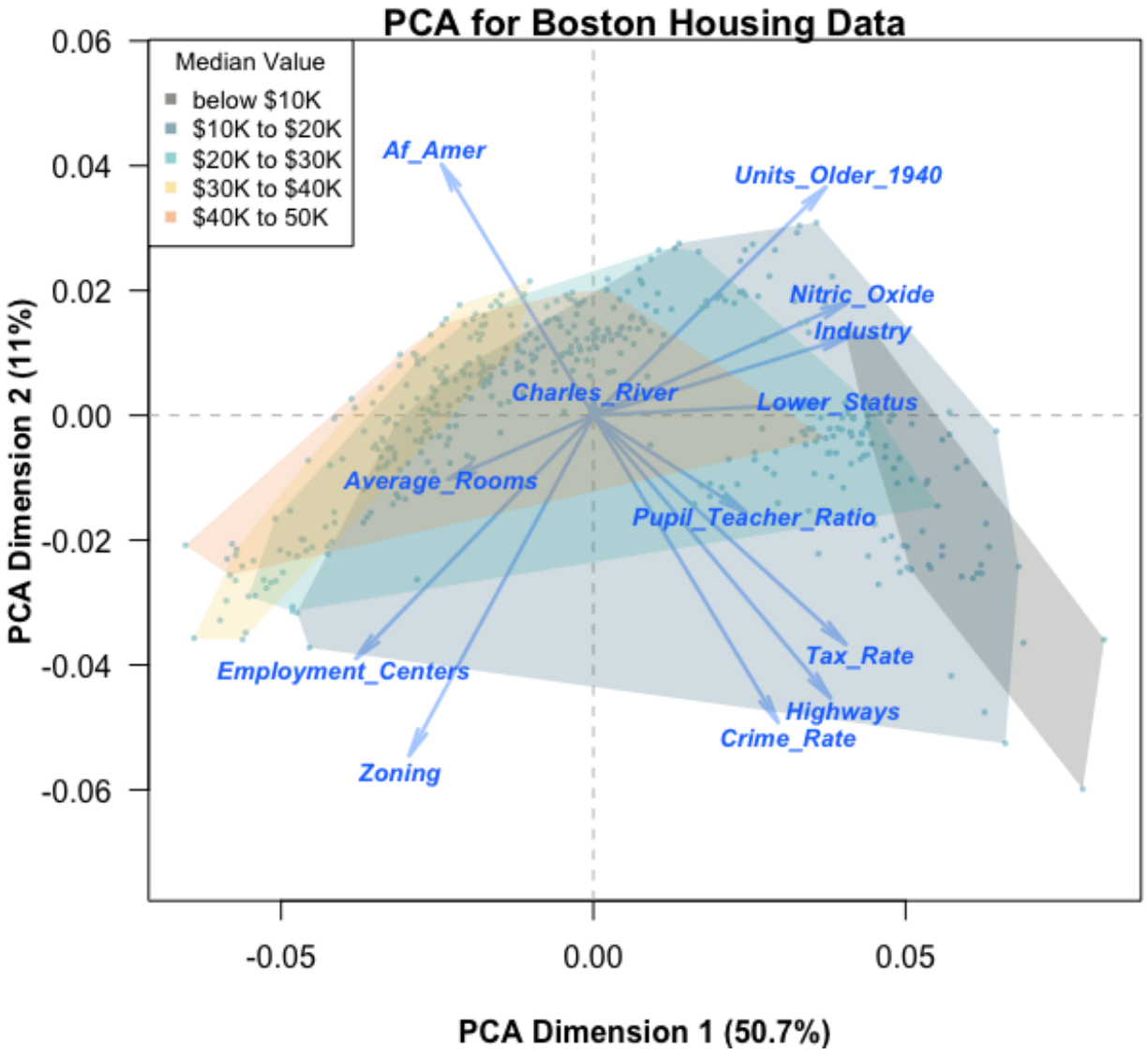

Figure 3: PCA biplot highlighting how the median value is impacted by the other features

with lower median values are likely located in more congested neighborhoods / towns with smaller homes, than those with higher median values, which are likely located in the more affluent suburbs.

As mentioned earlier in the document, I hesitated to include the African American feature in the analysis, as I would prefer to use variables representative of the type of neighborhood (crime rate, location, quality of education, etc.) than the race of the residents. In the end, I kept the variable because I found the conclusion unexpected, which made me question my own biases. I do not know why the variable was originally included in the dataset, but perhaps

in 1978, when the paper was written, there was a belief that race impacted home prices. I was actually happy to see that the percentage of African Americans in a neighborhood was negatively correlated to the crime rate and increased as median home values went up.