Sean McLean

ALY 6040

Module 1: Technique Practice

Introduction

The focus of this report is to dissect and analyze a dataset that looks at information of homes in the Boston area. The dataset consists of 14 attributes instrumental in the pricing of each home in the dataset, which amount to 506. Using summary statistics, correlations, and data visualizations, the exploration of the data will seek to show patterns and trends that can explain how homes are priced.

Analysis

The 14 attributes look at certain characteristics of the homes themselves and some of the features of the surrounding area that might impact whether someone will buy a home and at what price. Some of the home attributes include the age of units being sold, the amount of rooms per house, and property tax rates. Outside characteristics of the homes include pupil-teacher ratio per town, distances to radial highways and employment centers, and crime rates per town. A look at the summary statistics from all the columns in the dataset shows that all the attributes contain integer or numeric values. Implementing the View() function in R provides a glimpse into how the dataset looks and shows all the numeric ranges per attribute.

Using the head() function in R, the first six homes in the dataset are shown and gives a small sample size of the values of each attribute per home.

```
> head(boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

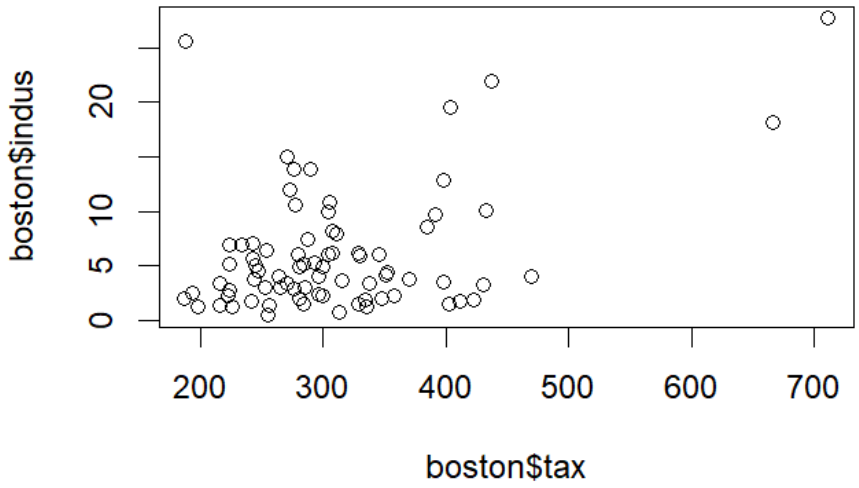
At first glance it appears that many of the values are close per unit with some variation in the age of the homes, the averages of the rooms per unit, and the percentage of the lower status of the population (lstat). The unit in the first row has the lowest crime rates (crim) in the entire dataset but has the highest nitric oxide concentrations (nox) and the shortest distances to the Boston employment offices (dis).

The dataset was well organized and cleaned up, with no duplicates or suspicious data located. An evaluation of the ranges per attribute didn't show any outliers that really stood out from the rest of the group. This would indicate that a more thorough look at correlations and visualizations of the data could show some potential outliers from this angle. Boxplots were created on some of the variables that showed outliers with crime rates, proportions of black people living in the city, and proportion of residential land zoned for lots over 25,000 square feet (about four times the

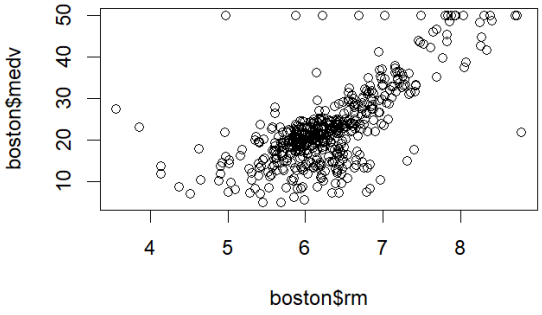
area of a basketball court) contained noticeable outliers when presented visually. These can have an impact on their mean and median values and be misleading when showing their relationships with other attributes in the dataset.

The cov() and cor() functions were used to show the strength of the relationships between each attribute. Looking at the covariances between the variables, the tax attribute showed very high values or very low values in its relationships. This indicates that taxes on the units are heavily affected by other attributes of the houses being sold. One variable that had values close to zero when compared to most of the other variables was the 'chas' attribute which pertains to whether the unit property borders the Charles River. This quality in a home surprisingly does not impact other features including the taxes on the house or any other housing feature.

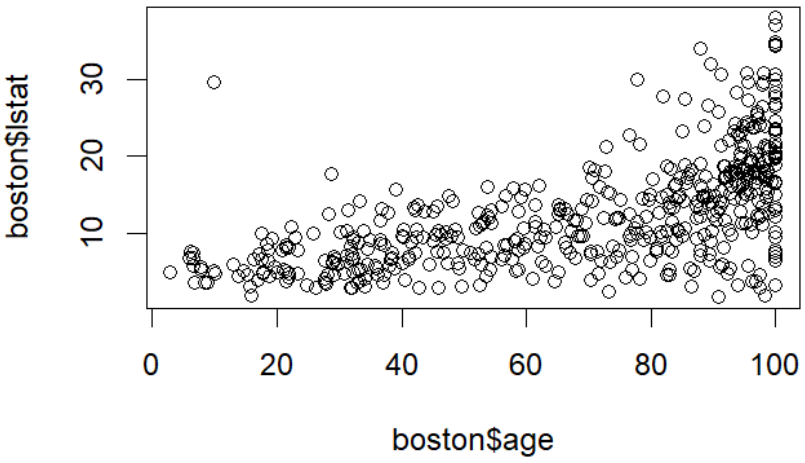
The cor() function was run to analyze the relationships between the variables with several showing strong correlations. To verify this, I used scatterplots to show the linearity between the attributes. A few examples of this include the connection between the variables 'indus' and 'tax'. The plot below shows that the taxes of units increase as the proportion of non-retail business acres per town also increases. This could potentially mean that houses have more value when there are fewer retail businesses in the town where the home is located.



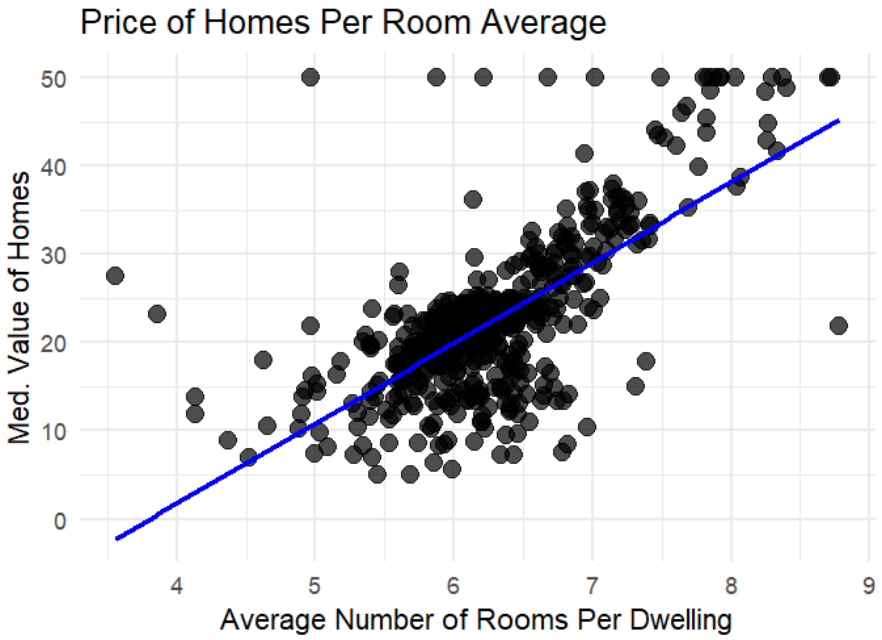
The plot below shows the correlation between the median values of owner-occupied homes and the property tax rates of homes. An analysis of the data visualization shows that the tax rate increases as the value of homes goes up in value. This would most likely be a strong relationship in that taxes are usually higher when the value of a home is also high.



The plot below shows the relationship between the age of the units versus the percentage of the lower status of the population. An evaluation of these data points shows that the age of housing is rising as the percentages of the lower status of the population increase. The percentages could be measured by total household income or another metric I am unsure of, but if it pertains to income this could indicate that households that are in the smallest percentile live in older units by what the scatterplot is showing. This number increases more as the house gets above a century old in age.



A ggplot was implemented to show the strong correlation between the value of homes per the average total of rooms per unit. While there are some outliers in the graph, there is a large cluster in the middle of the plot where most of the units contain five to seven rooms per unit. There does seem to be a linear relationship though as the more rooms a home has the higher value the unit will be. The ggplot provides a more in-depth look at the points and the darker the cluster gets with more data points being clumped together.



Conclusion

An overall assessment of the data shows that concerns like socioeconomic status could play a role in the values of the units and the surrounding areas' features. This could create issues where areas could be gentrified in the future. My next step would be to look at the strong relationships among the attributes as well as the trends and the patterns of the data for a better understanding of how they can impact the city and its real estate market. This would hopefully provide the opportunity to weed out any potential issues that could negatively affect the community. From there I would make recommendations to stakeholders based off those insights and continue to monitor how the market fluctuates with pricing and if the relationships between the attributes change over time.

References

Kabacoff, R. (2015). *R In Action. Data analysis and graphics with R. Second Edition.* (pp. 155-158, 416-417, 439-440, 543). Manning.