***Executive Report***

In this report, I look into trying to predict the price, or the median value, of real estate. I have been provided a data set that includes many variables that make up the price of real estate, such as distance to radial highways, crime rate, and much more than can be found in the appendix and at the bottom of page 2.

To start, I wanted to use this data to find how the crime rate per capita affects the real estate value of a given neighborhood. The results of my findings are found at the top of page 4, but I found that in general, areas with a lower crime rate per capita generally have higher median values for real estate. Additionally, I wanted to look at all the variables that affect real estate value and find which ones are the most important. The results of these findings are found on the bottom of page 4 and top of page 5. I found that there are about 4 variables that affect price the most, which are listed from most important to least, LSTAT, NOX, RM, and CRIM. These findings are also backed up by the decision tree created on at the bottom of page 4 and by the variable importance table on page 5. Lastly I wanted to investigate the relationship between the percentage of the lower status of the population and the crime per capita, and the relationship between the proportion of non-retail business acres per town and the accessibility to radial highways. I found that From looking at the coefficients table at the bottom of page 5, we can say that for two areas or neighborhoods, that differ in LSTAT by 1, are expected to differ in CRIM by 0.8905 with the area that has a larger LSTAT having a larger CRIM. Lastly, from looking at the coefficients table at the top of page 6, we can say that for two areas or neighborhoods, that differ in RAD by 1, are expected to differ in INDUS by 0.6905 with the area that has a larger RAD, or more access to radial highways, having a larger INDUS.

While the findings of the report can be found throughout this report, I have also included a description of the data on page 2, what methods I used and why I used them on page 3, and an appendix on the last page for more information on any terminology used in this report.

**Problem statements**

- One of the main things I am trying to predict from this data, is the price of real estate based on many different indicators, such as crime rate of the area, the student-teacher ration of the town, and much more. For this report, I want to specifically look into the following:

* How crime per capita affects the price
* What most affects the price

I also want to look into:

* The relationship between the percentage of the lower status of the population and the crime per capita
* The relationship between the proportion of non-retail business acres per town and the accessibility to radial highways

**Description of the data**

Text

Description automatically generated with medium confidence- The data I am working with for this project all relates to real estate price and any factors associated with it. A quick overview of the data is included to the left. Moving on, once the data was downloaded, there were missing values in the RM, or the average number of rooms, column. To fix this, I replaced the empty columns with the column mean. I choose to do this since I had nothing else to replace it with and it seemed to be the best representation of the data. Next, I created histograms for all the columns and looked to see if they were normal or not and if I needed to transform any of the data. Some columns were already normal such as the RM column, but most of the others were not. To change this, I transformed each of them by logging them. Some of the data needed the transformation and doing so caused the data generally smaller. Additionally, since a few of the columns lowest values were zero or were there were zeros mixed in, I had to take the log plus one + the lowest value. So, if the lowest value was zero, I took the log of the data + 1.

**Methods considered, and the best methods**

**-** For the first question on how the crime rate influences the median value of homes, I choose to use rPart to create a decision tree.

- I also chose the same method to answer my second question about what most affect the price.

When choosing what methods to use, I considered the vanilla regression model, the regularized linear regression model, the vanilla regression model tuned over the values of cp, or the rPart method, and the random forest model. To find which model was the best to use, I ran all of them and compared their RMSE. What I found is that three of the four models were all pretty good and close to each other. From there I used my own discretion to choose which models to use to make my figures and gather my metrics.

- For analyzing the relationship between the percentage of the lower status of the population and the crime per capita, I choose to do a simple linear regression. I choose to do this specific method because I thought it would be the best and easiest way to find the relationship, the p-value, and the slope and intercept between the two specific variables.

I also chose the same method for analyzing the relationship between the proportion of non-retail business acres per town and the accessibility to radial highways.

**Summary of results**

Diagram

Description automatically generated

Figure X

**-** For the first question on how the crime rate influences the median value of homes, I choose to use rPart to create a decision tree. This created the decision tree shown to the right. From looking at the tree, we can see that a higher crime rate, as CRIM, leads to a lower median value. With the highest median value, 1.47, being between 0.26 and 0.16 rate of crime. There is also an outlier, or an interesting result with a seemingly high median value, 1.14, being related to the crime rate being equal to or less than 1.3. In general, from the model the lower the CRIM, the higher the median price.

- To find what indicators most affect the median value of real estate, I created a decision tree to show all the most important indicators. This tree is shown below.

Diagram

Description automatically generated

From looking at the tree to the above on page 4, it breaks down the most important variables that make up the median value of real estate. The tree includes the percentage of lower status population as LSTAT, the crime rate, which I covered on page 4, the average number of rooms per house as RM, proportion of owner-occupied units built prior to 1940 as AGE, the nitric oxides concentration as NOX, and finally, the weighted distances to five Boston employment centers as DIS. An interesting finding from the tree, is that the highest median value, 1.64, is found if the LSTAT is equal to or below 0.99 and if the amount of rooms in the house are above 7.4. Another find is that the lowest median value is found if the LSTAT is equal to or above 0.99, the CRIM rate is equal to or above 0.85, the LSTAT is also equal to or above 1.3, and if the NOX is equal to or above 0.31.

|  |  |
| --- | --- |
| Variable | Overall |
| LSTAT | 100.0 |
| NOX | 75.44 |
| RM | 73.19 |
| CRIM | 70.68 |
| DIS | 43.48 |
| TAX | 26.87 |
| AGE | 26.69 |
| B | 24.93 |
| INDUS | 15.00 |
| PTRATIO | 13.07 |
| RAD | 4.44 |
| CHAS | 0.00 |
| ZN | 0.00 |

- Also created using the rPart method, was the variable importance table found below. The table shows the most important variables that affect the median price. From looking at it, we can conclude that LSTAT is the most important variable that affects the price. Continuing, NOX, RM, and CRIM are also very important at all around 70-75% of importance. We can further prove this and get a more in depth look into how these variables affect the price in the decision second tree at the bottom of page 4.

- Another thing I wanted to look at was the relationship between the percentage of the lower status of the population and the crime per capita. To investigate this, I ran sample linear regression. I found that from this relationship, we have an incredibly small p-value of less than 2e-16. This means that this relationship is statistically significant meaning that there is something going on here. From looking at the coefficients table below, we can say that for two areas or neighborhoods, that differ in LSTAT by 1, are expected to differ in CRIM by 0.8905 with the area that has a larger LSTAT having a larger CRIM. The model also typically makes an error of 0.374 when predicting the CRIM rate. Essentially, areas that have a higher population of the lower status of the population have a higher crime rate per capita.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -0.5673 | 0.0667 | -8.5 | < 2e-16 |
| LSTAT | 0.8905 | 0.0625 | 14.2 | < 2e-16 |

Residual standard error: 0.374 on 509 degrees of freedom

Multiple R-squared: 0.285, Adjusted R-squared: 0.284

F-statistic: 203 on 1 and 509 DF, p-value: <2e-16

Additionally, if we look at the QQ plot below, we can see that this relationship is about normal, with maybe being a little right skewed.

Chart, line chart, histogram

Description automatically generated

**-** Lastly, I wanted to investigate and analyzing the relationship between the proportion of non-retail business acres per town and the accessibility to radial highways. This relationship also seems to be statistically significant due to the low p-value. From looking at the coefficients table below, we can say that for two areas or neighborhoods, that differ in RAD by 1, are expected to differ in INDUS by 0.6905 with the area that has a larger RAD, or more access to radial highways, having a larger INDUS. The model also typically makes an error of 0.221 when predicting the INDUS rate. At a higher level, it looks like the areas with access to more radial highways have a larger proportion of non-retail business acres.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.6905 | 0.0230 | 30.1 | < 2e-16 |
| RAD | 0.4142 | 0.0257 | 16.1 | < 2e-16 |

Residual standard error: 0.221 on 509 degrees of freedom

Multiple R-squared: 0.337, Adjusted R-squared: 0.336

F-statistic: 259 on 1 and 509 DF, p-value: <2e-16

Additionally, if we look at the QQ plot below, we can see that this relationship is not very normal, could be bimodal, and maybe a little bit right skewed.

Chart, scatter chart

Description automatically generated

**Appendix**

CRIM: per capita crime rate by town.

ZN: proportion of presidential land zoned for lots over 25,000 sq.ft.

INDUS: proportion of non-retail business acres per town.

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise.

NOX: nitric 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 distances to five Boston employment centers.

RAD: index of accessibility to radial highways.

TAX: full-value property-tax rate per $10,000.

PTRATIO: pupil-teacher ratio by town.

B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

LSTAT: % lower status of the population.

MEDV: Median value of owner-occupied homes in $1000's.

rPart: Recursive Partitioning and Regression Trees.

Simple linear regression: A method that allows us to compare the average or expected value of y between variables.

QQ Plot: A probability plot for comparing two probability distributions.

Residual standard error (RMSE): The amount the model typically misses by when predicting something.

P-Value: A number describing how likely it is that your data would have occurred by random chance. A value of less than 0.05 is statistically significant while a value above is not.

Decision Tree: A graphical representation of possible solutions to a decision based on certain conditions.