ISEN 613-Spring 2023
Course Project I

 # YOUR NAME: Sidharth Thazhathedathu
To: President X
From: Data Consultant Y
Subject: Analysis of Boston Suburban Housing Values, Executive Summary'

This project is tasked to find trends on the data on the housing situation in the suburbs of Boston and make predictions based on it. There are many different variables that have to be taken into consideration during this analysis. In this project, we have taken the crime rate, proportion of residential land zoned for lots over 25000 sq.ft., the proportion of non-retail business acres per town and areas where the Charles River exists have been considered as part of the geological survey. We have also considered the factors that affect the residents' health like the amount of nitrogen oxides concentration. Furthermore, factors like the average number of rooms per residence have to be considered and the proportion of owner-occupied units built prior to 1940 are essential to gain structural knowledge of the building. There are daily concerns that people normally have, like the teacher-pupil ratio, the proportion of black population, the cost of public services in each community and the percent of lower status of population have been taken into consideration. This can show how friendly and diverse a neighborhood can be from this region. The distances to five employment centers and radial highways can show accessibility to radial highways for different regions.

The summary shows that the owner-occupied buildings average ages are around 68-69. The highest recorded age of a building is 100 and the lowest is 2 years. There is good accessibility to radial highways. Crime stats show that crime rate is low averaging at 3.6 but there is a region with high crime rate of 88.9, this region should be considered a red flag and we need to further contemplate before making an investment in this region. The concentration of nitrogen oxide is low which ranges between 0.385 and 0.871 parts per 10 million. The number of rooms per dwelling varies from 4 to 8, where the mean average is found to be 6.

There is a high relation between the number of proportions of non-retail business and nitrogen levels which is expected. As expected, crime rate and the number people living in that region are almost inversely dependent. Where there is a high crime rate, there are many people living in those regions. The number of rooms has been consistent irrespective of when a building was constructed. There is a relationship between the price of houses in Boston and the number of rooms as expected. The distance between the employment centers and the region with high concentration of nitrogen oxides. There is a high correlation between the lower status people and the median value of owner-occupied homes. There is a high correlation between the tax paid and the distance to highways. The homeowners with the top 5% cost houses generally have more rooms. The houses with high cost live close to highways and most of them are situated in regions where crime rate is very low. The distance to employment centers does not matter too much on the price of the house, they do maintain an average pupil-teacher ratio. The top 5% cost value ranges between $43000-$50000. The cost of houses is higher when it is near Chase River. It is possible to compromise on the cost if we buy buildings that are older.

If a person wants to live in a relatively new house with good schools with 6 rooms, good accessibility, a very low crime rate, the price value varies between $38,700 and $50,700. With 0.01 increase in parts per 10 million, the value of the decrease by it 8%, $1800.

With the model that was chosen we can see that for every unit increase in rooms the price of house increases by 10%, $2253 and if the house is near Charles River the cost increases by 12%, 2703.6 and we should choose regions where the nitrogen oxide levels, crime rate and distance to the employment

center are low. The other variables for this model show very little dependence to the cost of a house based on the model that is chosen.

Technical Report

For this project we have considered the Boston dataset, from the Mass package. The package contains a variety of independent variables that are important to find trends. Here, the dependent variable is the 'Median value of owner-occupied homes in $1000's' that is MEDV. There are 13 different variables, and there are 506 cases.
logmedv is taken as log(medv) to create a better model.

From, there were no Missing values in the dataset.

Now we are going to assess the pairs plot of the top 5 % MEDV to see the correlation between the variables.

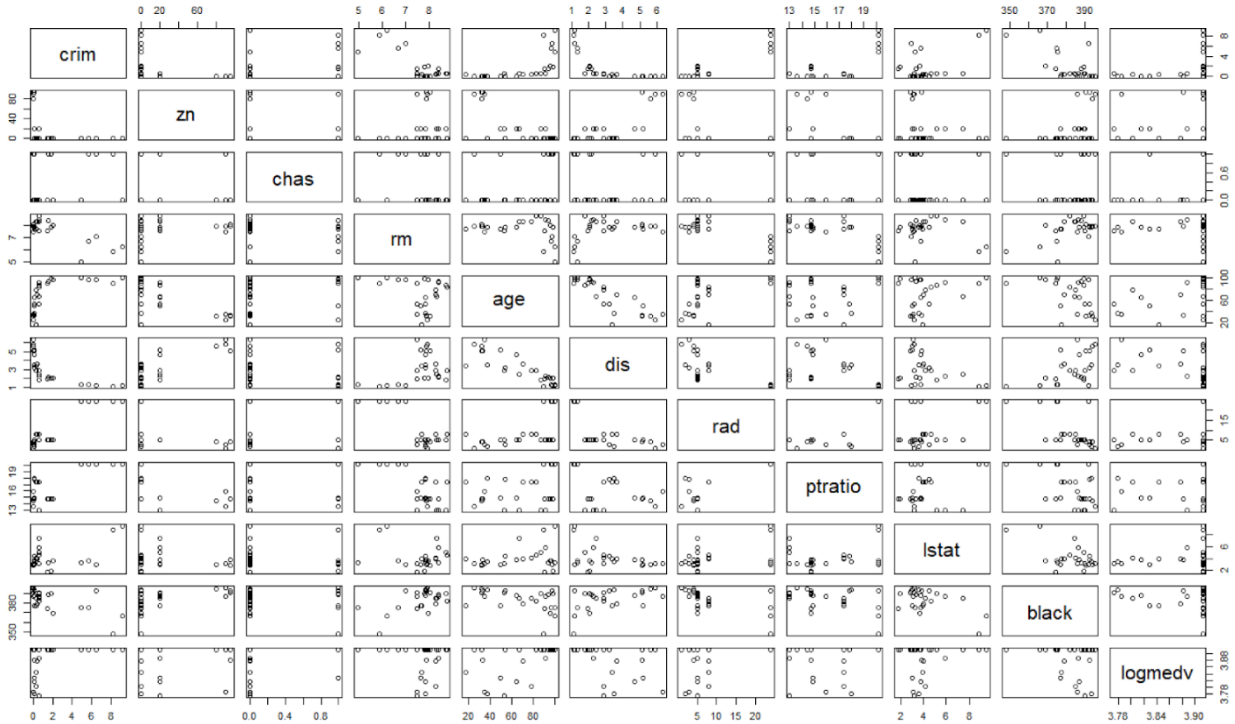

Figure 1: Pair plots for top 5%

From the pair plots of the top 5 % expensive homes, it is found that the crime rate is also very low averaging at 1.8. There are very few low statuses people who own expensive homes, averaging at 4.155 population percent. They generally have homes that have many rooms averaging at 7.631 rooms per dwelling. This shows that the houses are very big which is expected for expensive homes. The pupil-teacher ratio averages around 16.17 which is lower than the mean of the whole. We can see that the chas predictor has more points at '0' than at '1' suggesting that most top 5% people don't live close to the Charles River.

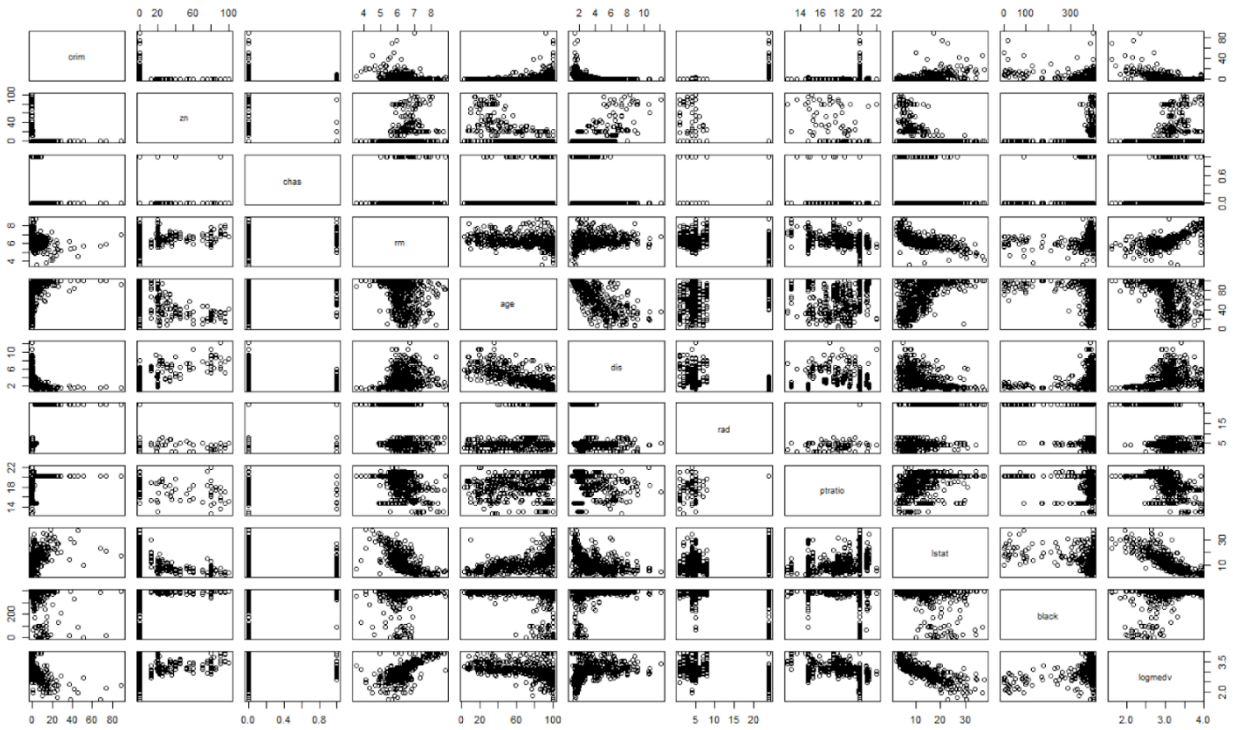

Figure 2: Pair plots for the all the data.

From the given plots we can see that the lower status generally prefers less expensive costs as we see in the pair plots a decreasing trend as the log(medv) increases. We can see that as the number of rooms increase the cost of the houses also increase. The crime rate is generally low when the house costs more.

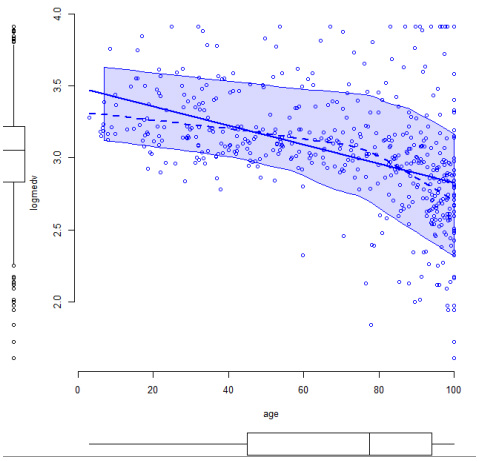

Figure 3: plot of logmedv and age

As age of the residence increases the cost of the dwelling decreases

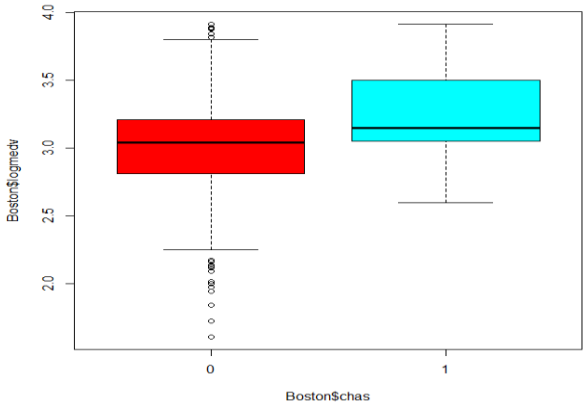

Figure 4: Boxplots of logmedv and dummy variable

The boxplot shows that people near the Charles River pay less compared to people who don't. The plot is between the Charles River and Log median value of the cost per 1000$ per town

The 8 predictors that were needed to be considered were included and the variables that shows the percentage of lower status population(lstat), black population and nitrogen oxide levels were high(nox) were these were included, because the significance of these predictors were found to be high and there was a significant increase in the performance in the model. So, a total of 11 variables have been chosen for the final model. The indus predictor shows very little significance comparatively. After calculating VIF (Variance Inflation Factor), the variable tax is not included because it shows high collinearity.
From the selected model, we have the slopes of the predictors show how much the predictor contributes to response(logmedv). Here, crime rate, age of the dwellings, distance to the employment centers, pupil-teacher ratio and percentage of lower status population have a negative contribution to the response. The predictors: proportion of residential land zoned higher than 25000 sq.ft, if the neighborhood is close to the Charles river, the distance to the highways, the number of rooms in a residence and proportion of black people per town. The price of the houses increases for decrease in nitrogen oxide values. The model is having an adjusted R sq. of 0.777 which shows how reliable the model is. For every unit increase in room it contributes to 1.103 increase in value of the median. The amount how much each variable contributes to median value of the cost of homes per 1000$ is given:

Table 1: The contribution to response variable: medv

crim	zn	chas	rm	age	dis	nox	rad	ptratio	lstat	black
0.989901	1.000006	1.12358	1.103081	1.000137	0.953043	0.41822	1.00540	0.960042	0.971969	1.00043

Null Hypothesis test: log(medv) does not have any relation with the variables

Alternate Hypothesis: There is a relationship with other variables.

Before finding significance

```
Call:
lm(formula = log(medv) ~ crim + zn + chas + rm + age + dis +
    nox + rad + ptratio + lstat + black, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73458 -0.10630 -0.01126  0.09403  0.87274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.0108415  0.2063234  19.440  < 2e-16 ***
crim        -0.0101504  0.0013352  -7.602 1.48e-13 ***
zn           0.0006674  0.0005442   1.226 0.220674
chas         0.1165197  0.0347309   3.355 0.000855 ***
rm           0.0981076  0.0167875   5.844 9.26e-09 ***
age          0.0001372  0.0005370   0.255 0.798499
dis         -0.0480955  0.0079024  -6.086 2.32e-09 ***
nox         -0.8717483  0.1462035  -5.963 4.73e-09 ***
rad          0.0053885  0.0016495   3.267 0.001163 **
ptratio     -0.0407786  0.0052176  -7.816 3.32e-14 ***
lstat       -0.0290491  0.0020564 -14.126  < 2e-16 ***
black        0.0004304  0.0001091   3.946 9.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.193 on 494 degrees of freedom
Multiple R-squared:  0.7819,    Adjusted R-squared:  0.777
F-statistic:   161 on 11 and 494 DF,  p-value: < 2.2e-16
```

After finding significance

```
Call:
lm(formula = logmedv ~ crim + chas + rm + dis + rad + lstat +
    black + nox + ptratio, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73252 -0.10612 -0.01410  0.09214  0.87773

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.0213020  0.2051665  19.600  < 2e-16 ***
crim        -0.0100087  0.0013294  -7.529 2.44e-13 ***
chas         0.1161515  0.0346669   3.351 0.000868 ***
rm           0.1014707  0.0162931   6.228 1.01e-09 ***
dis         -0.0442353  0.0065520  -6.751 4.10e-11 ***
rad          0.0056058  0.0016295   3.440 0.000630 ***
lstat       -0.0288434  0.0019305 -14.941  < 2e-16 ***
black        0.0004319  0.0001088   3.970 8.26e-05 ***
nox         -0.8712566  0.1395967  -6.241 9.32e-10 ***
ptratio     -0.0426888  0.0049175  -8.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1929 on 496 degrees of freedom
Multiple R-squared:  0.7812,    Adjusted R-squared:  0.7773
F-statistic: 196.8 on 9 and 496 DF,  p-value: < 2.2e-16
```

The standard error shows the how precise is the slope. The p- value shows the significance of the predictor. The variables age and zn show less significance. So, the new model has variables that are all significant. The F- statistic of the first model is 161 and for second model is 196.8.

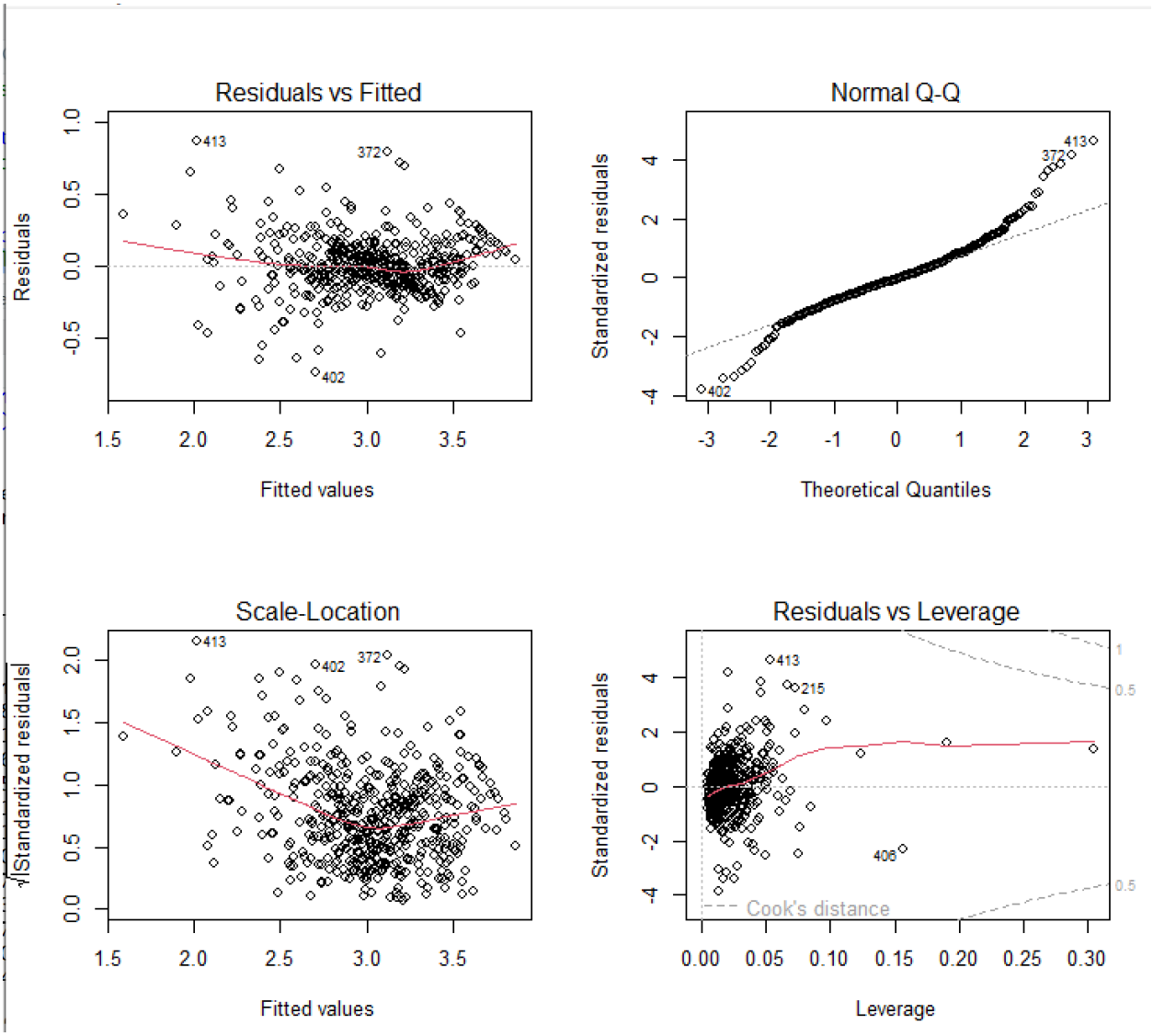

Figure 5: Diagnostic plots of first model

Diagnostic plots.

From **the residual vs leverage** points, we can see that there aren't overly influential points. 406,413 is close to Cook's distance. We can see if there is overfitting or underfitting of the model. There is no funnel shape so there is no heteroscedasticity.

The **Scale-Location** is used to check equal variance also known as homoscedasticity. The red line shown in the graph is not linear so we can see that the model does not have equal variance. It does not show homoscedasticity.

The **Q-Q plot** is linear between -2 and 2. This plot is used to check if the model follows a normal distribution.

In the **Residuals vs Fitted** values show that the red-line does not perfectly follow the horizontal line. But it does not deviate too much and shows the linearity of the model. The inclusion of interaction term coils increases the linearity of the model as well.

Final Formula: $Log(medv) = -0.0101504 * crim + 0.0006674 * zn + 0.1165197 * chas + 0.0981076 * rm + 0.0001972 * age - 0.0480955 * dis - 0.8717485 * nox + 0.0053885 * rad - 0.0407786 * ptratio - 0.0290491 * lstat + 0.0004305 * black$