# ISEN 613-Spring 2023

# Course Project I

**Deadline: March 6 (Monday) @11:59pm**

**# YOUR NAME: Sidharth Thazhathedathu**

**# 427009745**

**# RULES:**

**#**

**# 1) You can discuss the homework with each other in general terms,**

**# but you must write your own solutions and not copy from anyone.**

**#**

**# 2) Edit your answers into this file following each problem.**

**#**

**# 3) Post your solutions to Canvas**

**#**

**# 4) 0.5 points per hour are deducted for late submissions.**

**# The filename should have this format: LastName-FirstName-project01.doc**

**# An example would be: Eksin-Ceyhun-project01.doc**

It is 1970 and you are hired by a real estate company in Boston as a data consultant. The company would like to make price estimates based on different preferences that a customer can have, so that they can assess if a customer’s budget is realistic or not. Your task is to build a model and write an executive summary that outlines important features that affect the price of a house.

Attach “Boston” data included in the MASS library. The data contains information from 504 geographic areas. There are 14 attributes in each area of the dataset. You may find the description of each attribute using help(Boston) code in R.

Your model should have the log transformation median house value (log(medv)) as the output variable. When you are searching for a meaningful regression equation, consider the following attributes:

1. Structural properties of a house (age, number of rooms, lot size allowed by zoning laws in the area)
2. Accessibility (distance to major employment centers and closeness to highways)
3. Neighborhood (crime rate, education quality, whether it is by the Charles River or not)

You may include other terms in your model, but you must discuss a model that contains the features above. You can use model diagnostics or other model selection methods to guide you in determining a final model. Decide on one model and use it to write an executive summary.

**Executive Summary**

Summarize your findings to the president of the real estate company at a level that a non-technical person (president or a realtor) can understand. The executive summary is **at most one pages**, **single spaced and in 12-point type.** Do not include any graphs or statistical concepts. Use the following page to write your executive summary.

In the executive summary,

1. When you discuss the effect of an attribute on price, you need to describe the meaning of the attribute and the ranges of values associated with the attribute. For example, if there is an effect of crime rate on house prices, it would be necessary to report not only the coefficient associated with the age but what is the worst crime rate and best crime rate across different areas.
2. Discuss important attributes that significantly affect the value of a house.
3. Introduce a common baseline scenario such as relatively young houses near decent schools with great accessibility and discuss its estimated price and price range.
4. Describe the attributes that yield highest house values (houses that are in top 5% in value)
5. Discuss certain bargains, such as if you sacrifice a certain neighborhood attribute you can afford to live at a bigger house for a budget that is at the mean of house values.
6. Use rounded numbers that are memorable.

**Technical Summary**

In the technical summary you speak to your peers. You show them that you performed a reasonable analysis and that you interpreted the results competently. This should not be a step-by-step report of what you did in R, but a summary of the most important steps in logical, not chronological, order. Even in a technical summary it is not of interest to hear, for example, how you used R; it is simply assumed that you know how to execute the available software.

The technical narrative should explain what values were used in the executive summary and how they were rounded. In addition, it should explain what contributions to the model were neglected because their effect on house value is too small.

The technical summary should, among other things, explain the final fitted model, term by term and estimate by estimate. It should mention model diagnostics that were performed and their outcomes, possibly accompanied by plots. Report data points (geographical areas) that you may have removed, if any, and why you did so.

Technical summary should at most be **four pages including any figures/tables you choose to include.** It should follow your executive summary page.

**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.

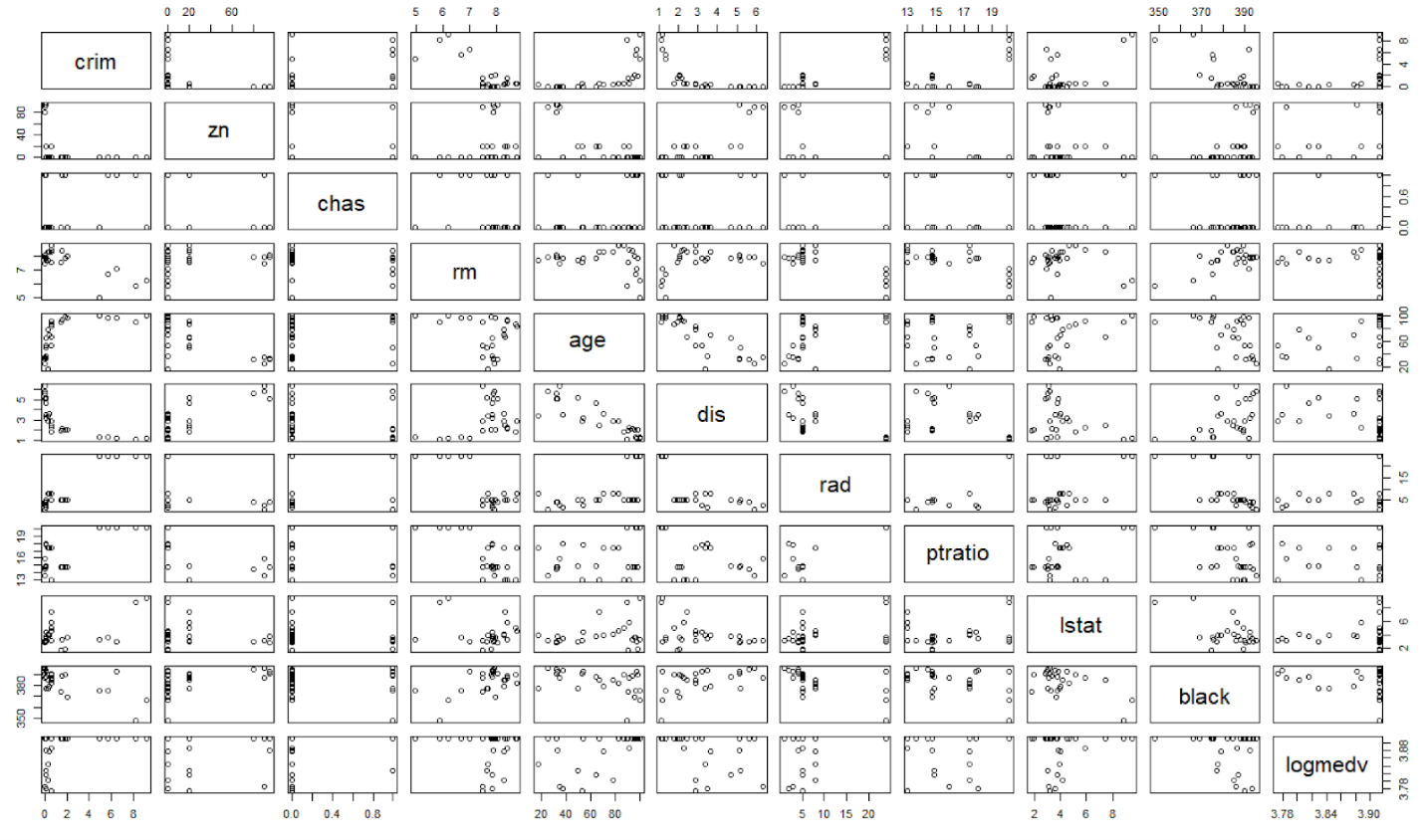
**Technical Report**

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.

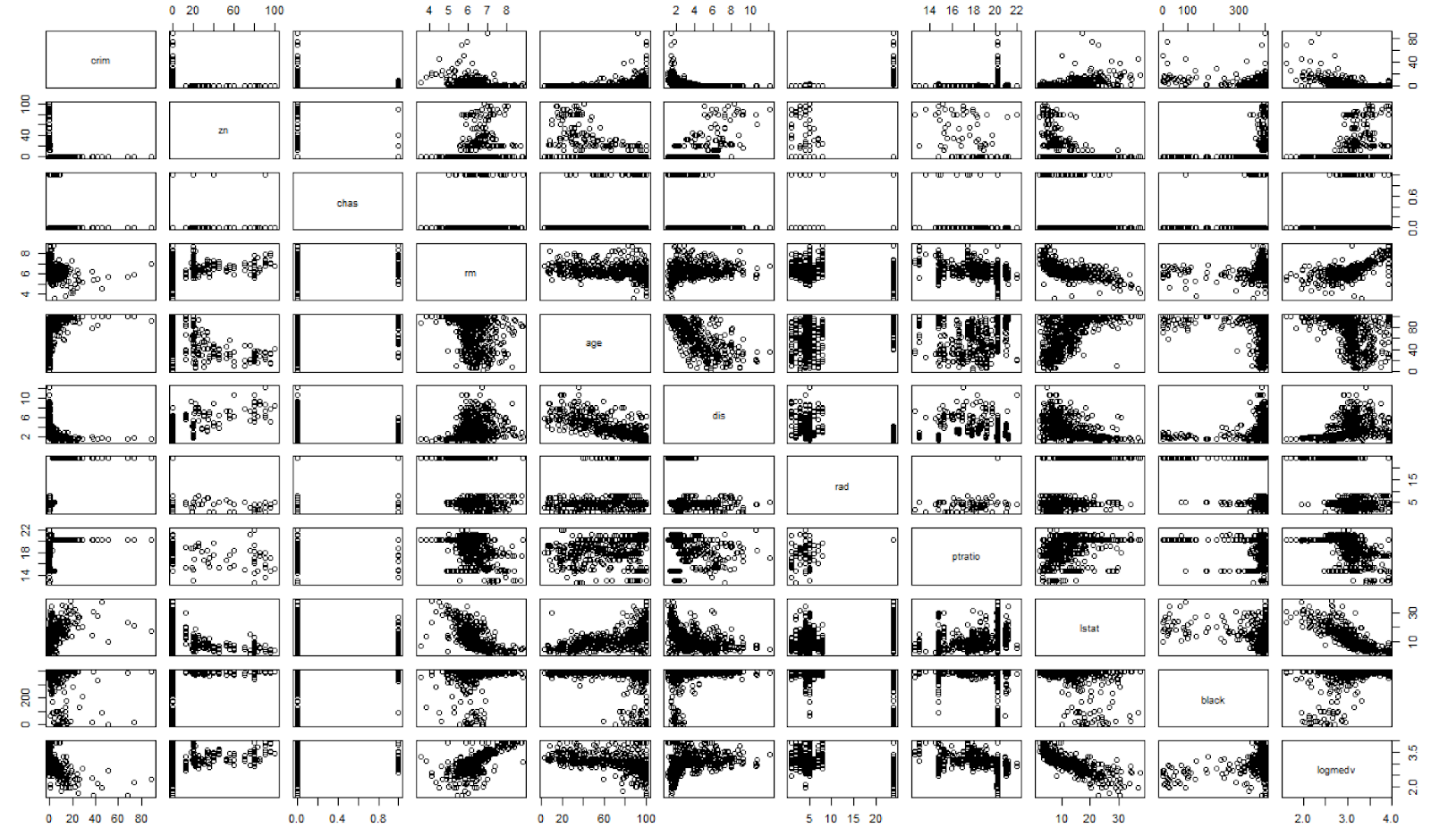
From, there were no Missing values in the dataset.

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** | **Figure 4: Boxplots of logmedv and dummy variable** |
| As age of the residence increases the cost of the dwelling decreases | 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**

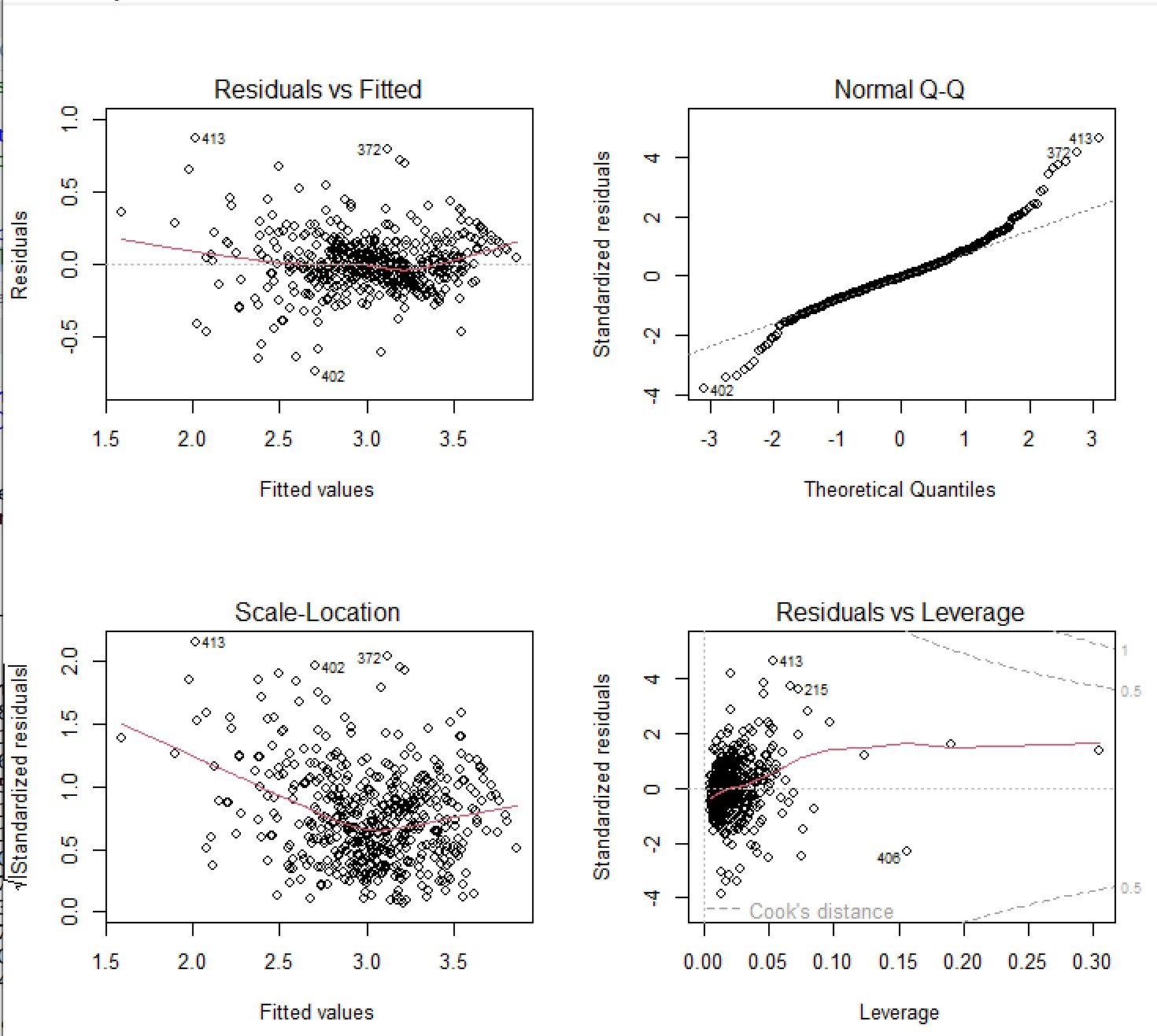
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| crim | zn | chas | rm | age | dis | nox | rad | ptratio | lstat | black |  |
| 0.989901 | 1.000006 | 1.12358 | 1.103081 | 1.000137 | 0.953043 | 0.41822 | 1.00540 | 0.960042 | 0.971969 | 1.00043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

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 | After finding significance |
|  |  |

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: