# MSBA 635 Consulting Assignment 1

Report by **Nirosha Telu (Team 5)**

*As an analytics consultant, you are tasked with understanding the housing market for an area of Boston in the past. In particular, you are asked to build a model to predict house price given certain characteristics. Please use the data Housing.csv to complete the steps below—to access the data, go to Week 1 section then the Week 1 slides and data folder of Blackboard. The data dictionary is provided at the end of this document.*

*Clearly presenting results to clients or stakeholders is a critical part of your job. Therefore, please write in complete sentences for our client and clearly label and introduce your figures and tables. You will lose points not only for getting the answer incorrect but for incomplete sentences and not explaining figures. You can work together on the analysis as a group, but you will need to write your own findings individually. Each numbered section is worth 20 points for a potential total of 100 points.*

**Data Dictionary for Housing Dataset :**

Medv median value of owner-occupied homes in $1000

crim per capita crime rate by town

zn proportion of residential 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 nitrogen 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 mean of 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

lstat lower status of the population (percent)

1. **Purpose (1 sentence)** 
   1. **What is the business problem?**

The Boston Housing Consulting firm primarily focuses on the housing market and they wanted to predict the house prices basing on the data that they already have (historic data) and what features of the Boston area are influencing the costs of dwellings.

1. **What did you do? (1-2 sentences)** 
   1. **For example: In this report, initial exploratory data analysis has been performed and then a xx model has been applied to do what?**

In my report, to predict the influence of the different features (predictor variables) on the median value of the owner-occupied homes (response variable – medv), I wanted to initially understand the data and run an exploratory analysis using SAS.

The data exploratory analysis helped me to understand the insights, trends, and patterns of the data. With the graphical representation in my report, I wanted to do the deeper analysis in the correct direction.

After the data exploration is done, I jumped on to running a linear regression model on my data to understand the questions or concerns that the feature parameters have an influence on the pricing of the housing and correlations between each variable with the response variable.

1. **Data Contents (1-2 sentences)** 
   1. **What are the number of observations and predictor variables?**

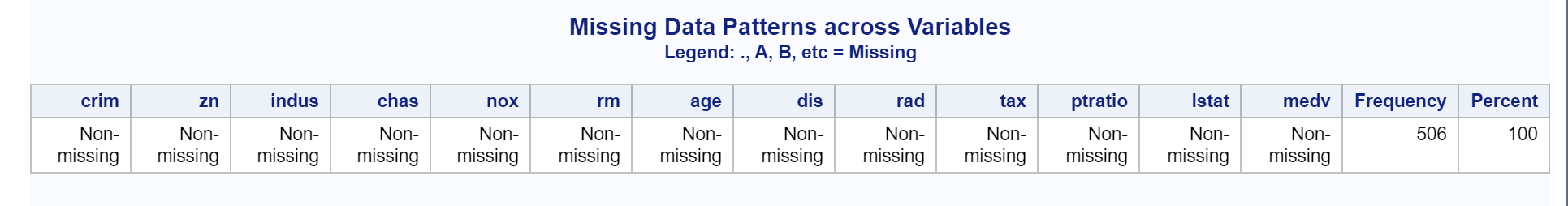
The total number of **observations** in this dataset is **506** and the feature variables are 13. Out of 13, one variable is the response variable and the rest of **12 variables are predictor variables.**

* 1. **Describe the predictor variables in general.**

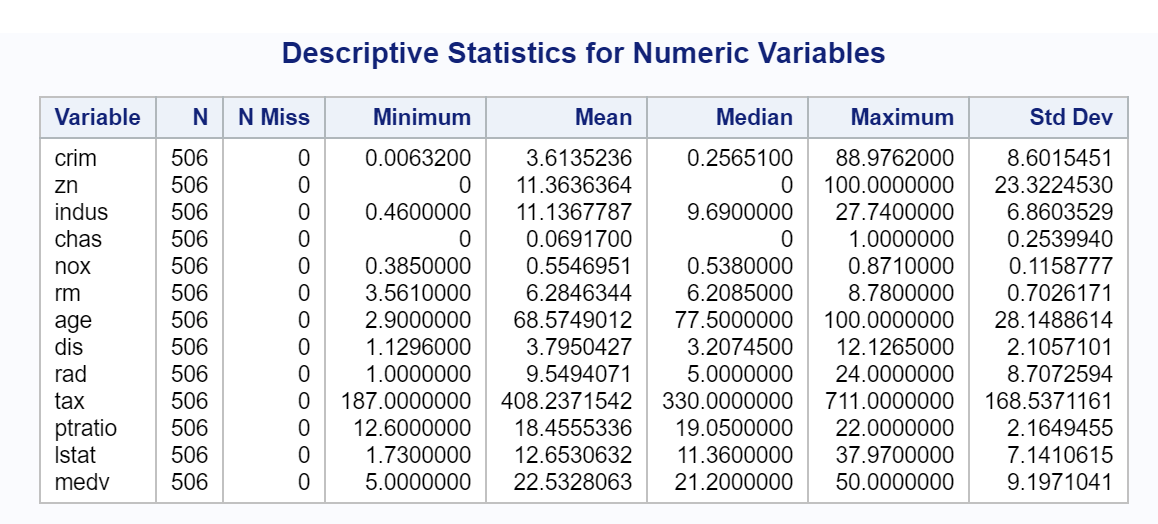
It is always crucial to understand the predictor variables and their behaviors with the response variable. In our housing dataset, we have 12 such variables and they individually have their features as in giving the data about the crime rate per town, rooms per house, is the property bound by Charles river or not, accessibility to radial highways, student-teacher ratio by town which gives the insight of the education in the area, distances to the Boston employment centers, any health hazardous locations nearby like the nitric oxide concentration, so on and so forth.

1. **Explain Exploratory Data Analysis findings (3-4 sentences)** 
   1. **Summary Statistics:** 
      1. **Are there missing values, outliers, or any inconsistencies?**

There are no missing values in the data. This is checked using the “Describe missing data” tab in “Data” task.

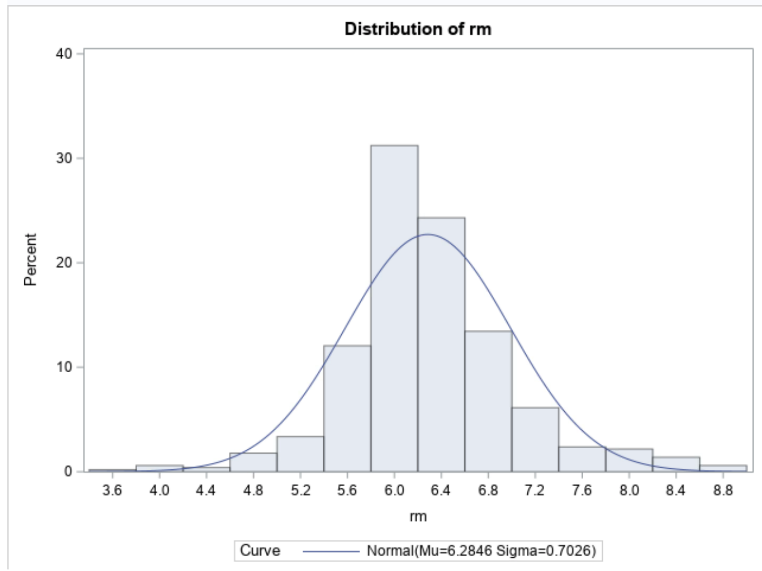


* + 1. **Provide a table of summary statistics for numerical variables**

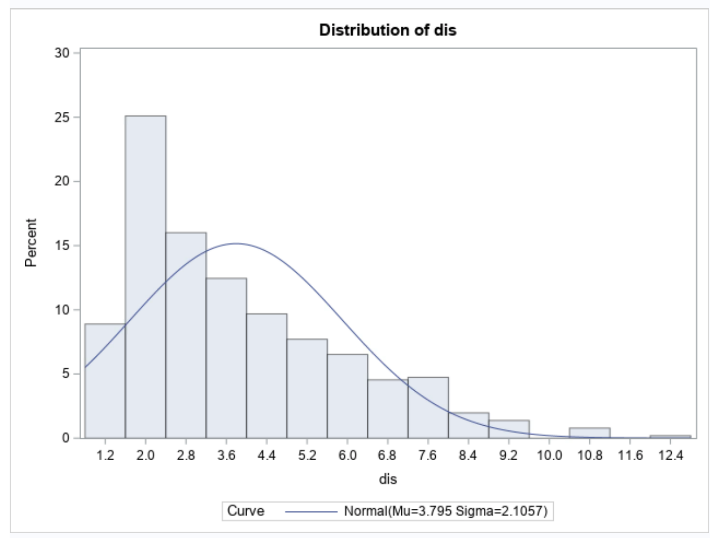


To understand more descriptive nature of the variables, the following plotting and graphs are used to understand the trends in data.

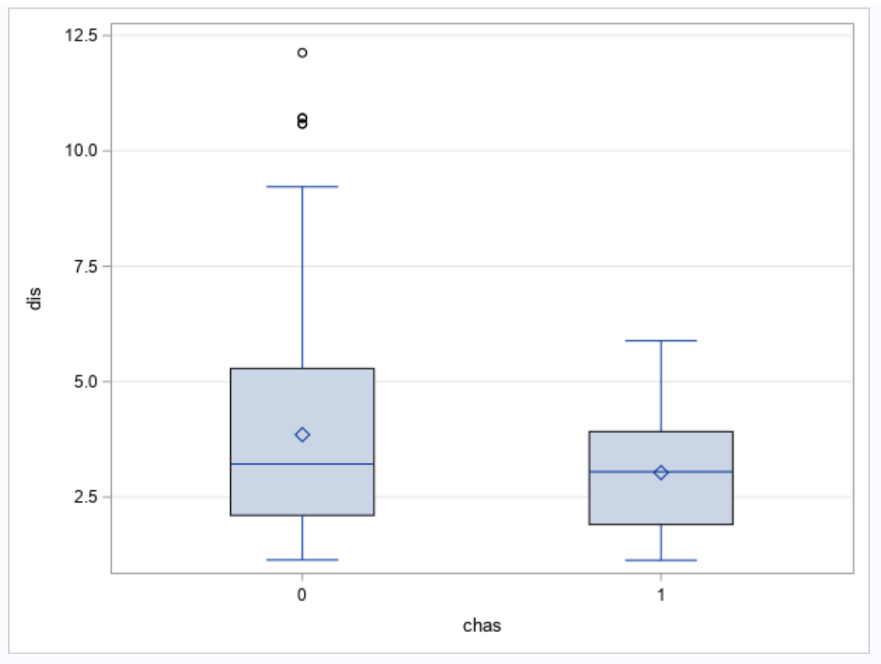
* + 1. **Histograms and density plots for a 1-2 continuous variables:**
    2. **Describe what you observe—normal distribution, skewed, etc.**

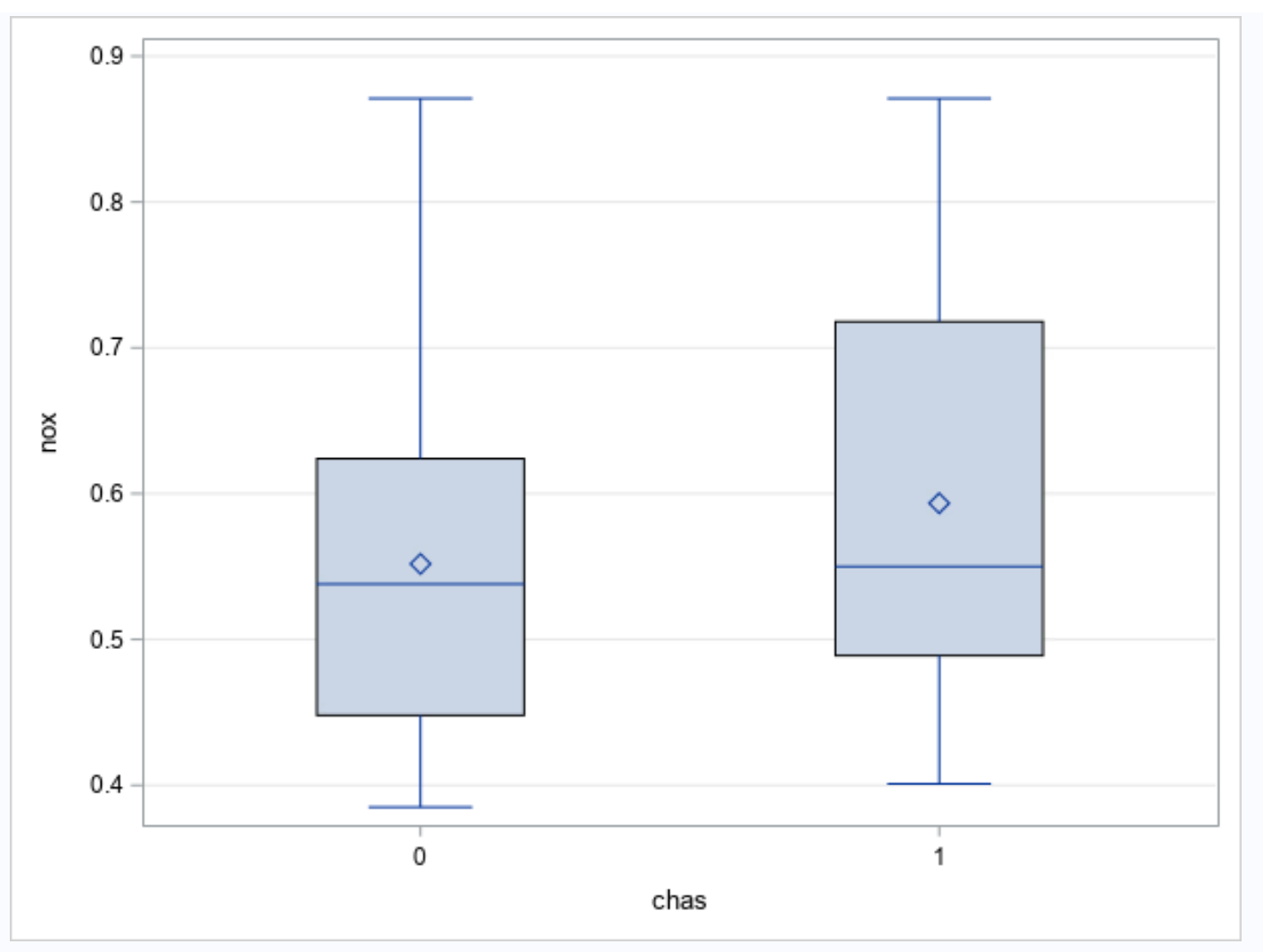


I plotted a histogram with density curve and observed a normal distribution in the number of rooms per dwelling (rm). The average number of rooms are 6.

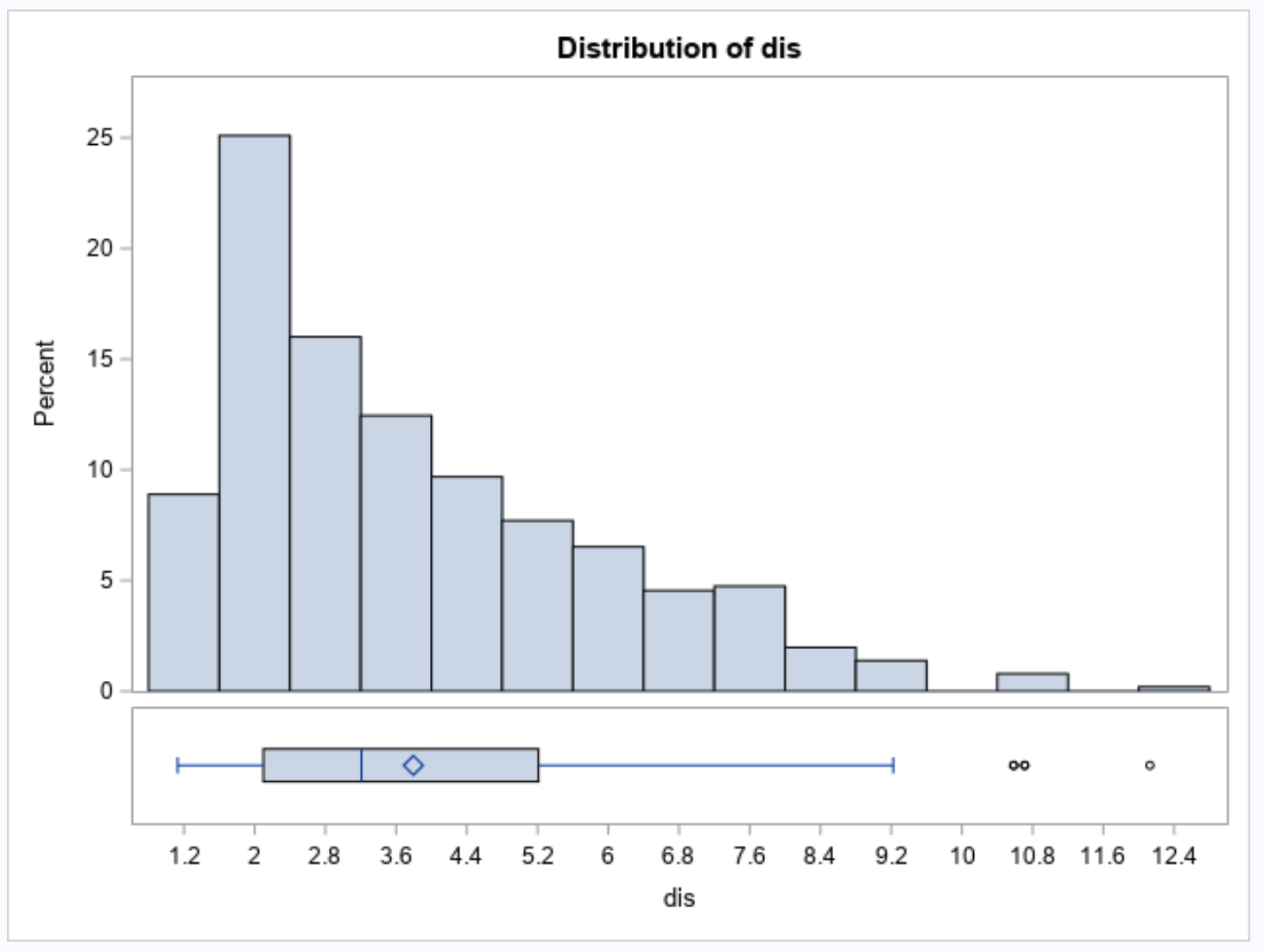
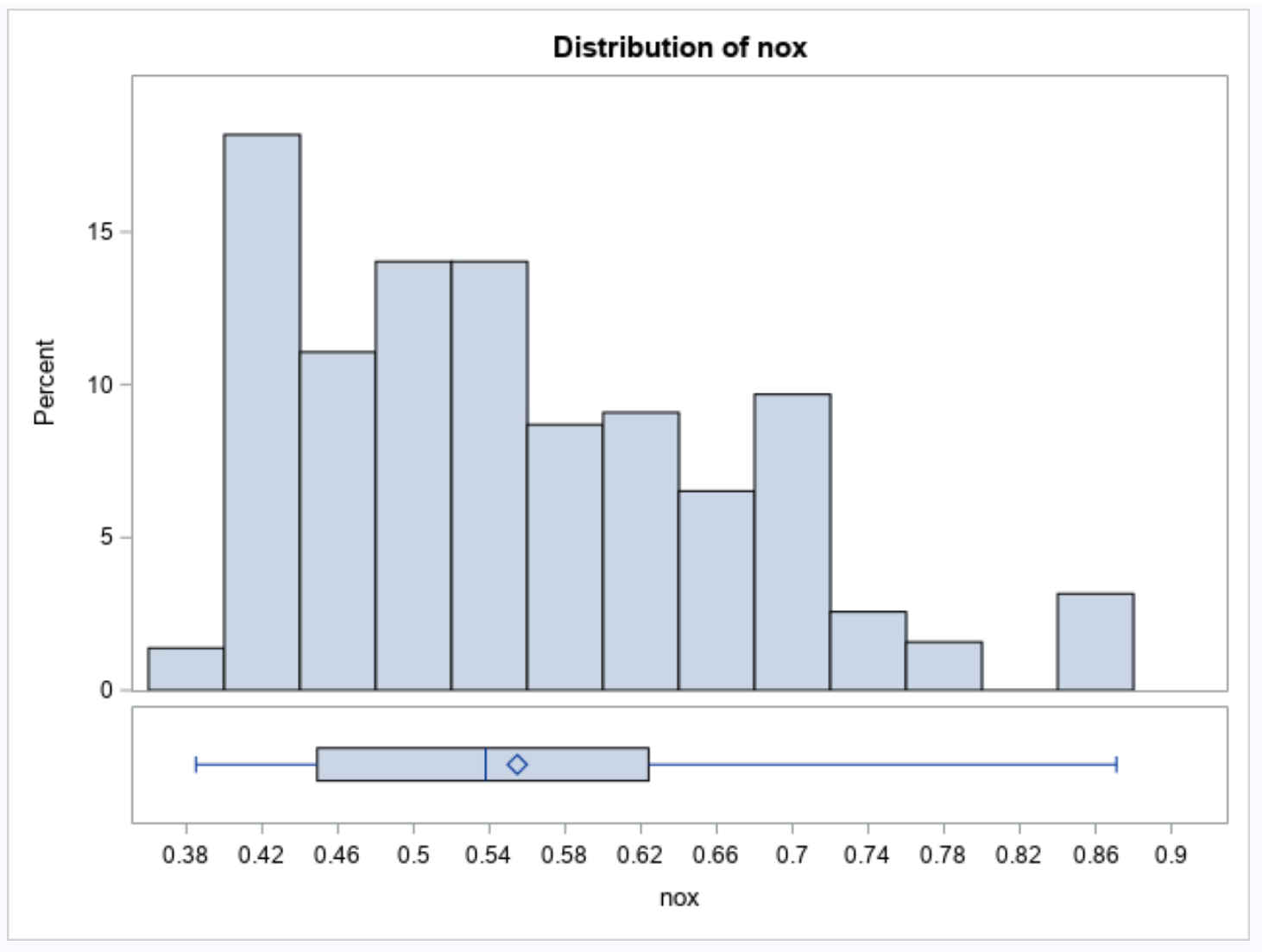
The distribution observed in the weighted distance to the Boston employment centers are slightly left-skewed and the major employment centers are within 2.0 miles.

* + 1. **If there are categorical variables, then build box plots for some variables by category.**

The only categorical variable in the data set is the chas – and this indicates if the housing tract is bound by river, it is 1 and otherwise, it is 0. According to the below box plot with the weighted distance (dis), most the properties that are close to the employment centers do not have the Charles river tract. There are outliers in the data as well and can be checked if they are causing any huge impact on the analysis. If they are causing, we can normalize the data according to the requirements. The variance in this plot is uneven between 0 and 1 and this can be addressed.

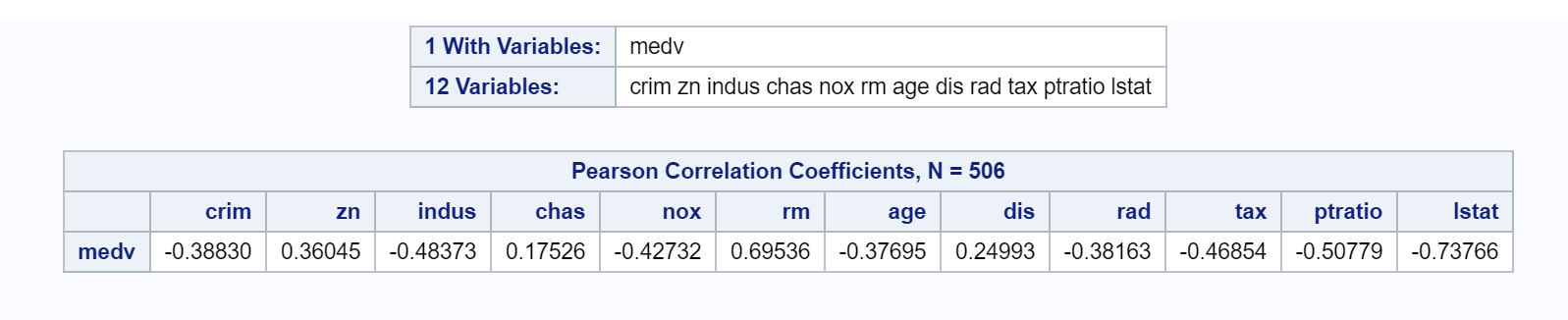
This box plot is built against the nitric oxide concentration and the plot visually depicts that there are no outliers, and the central tendency lies within the highlighted blue zone for both 0 and 1. In simple words, the tracts of the properties which are bounded by the river has a normally distributed and the properties which are not bound by the river is skewed downwards.

* + 1. **Provide the histogram, box plot etc.**



The above is a graphical distribution of both histogram and box plot using the data exploration in Statistics tab of task in SAS. Both the variables are left skewed, and the outliers are clearly noted in the “dis” variable plots.

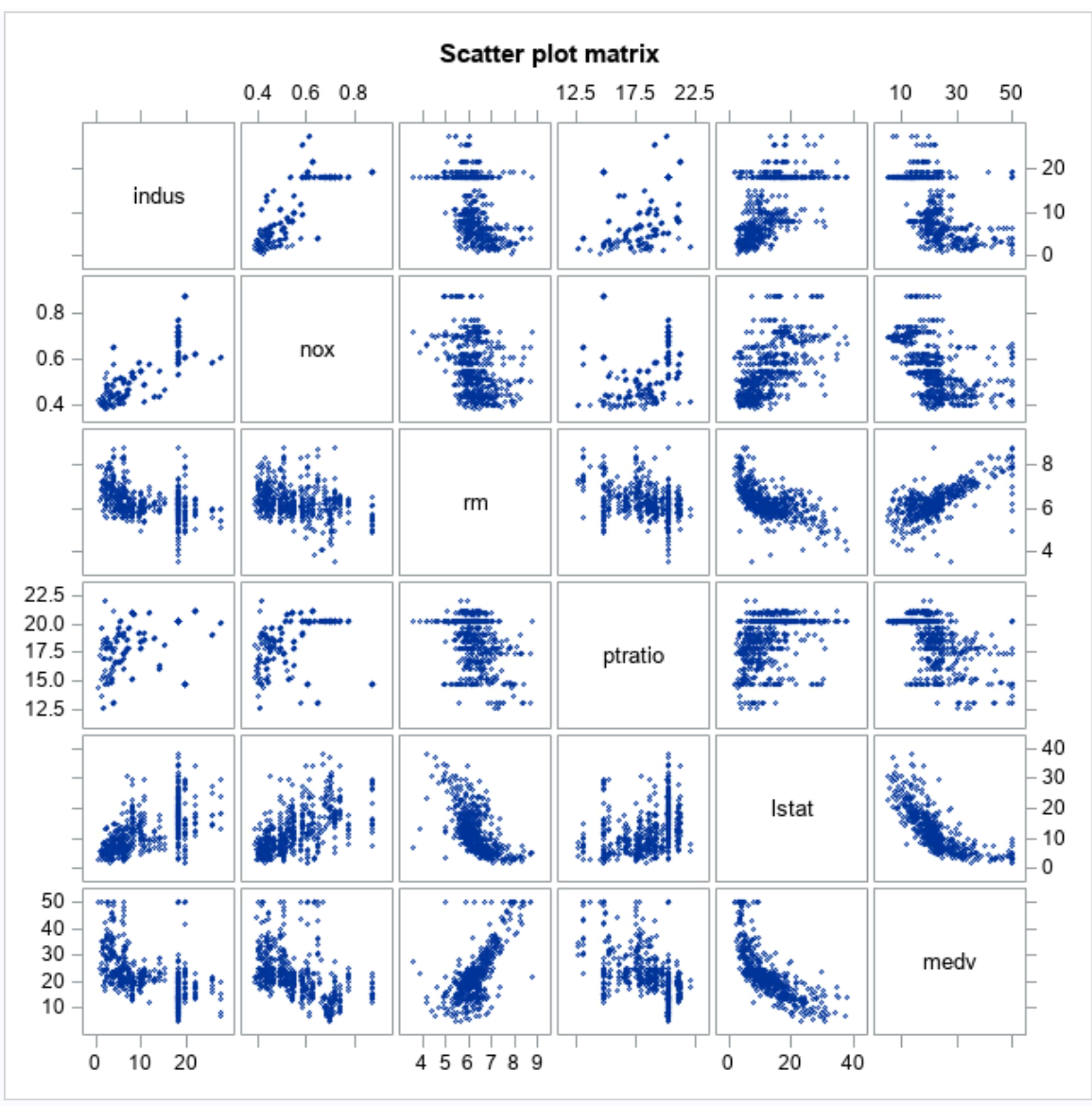
* 1. **Scatter plots:**
     1. **Describe any correlation you observe between the response and predictor variables.**



The correlation analysis is executed in SAS and by doing this, I could understand the relation between the response (medv) variable and the predictor variables. According to the coefficients in the above table, “rm” is positively correlated better when compared to other variables and the “lstat” is negatively correlated.

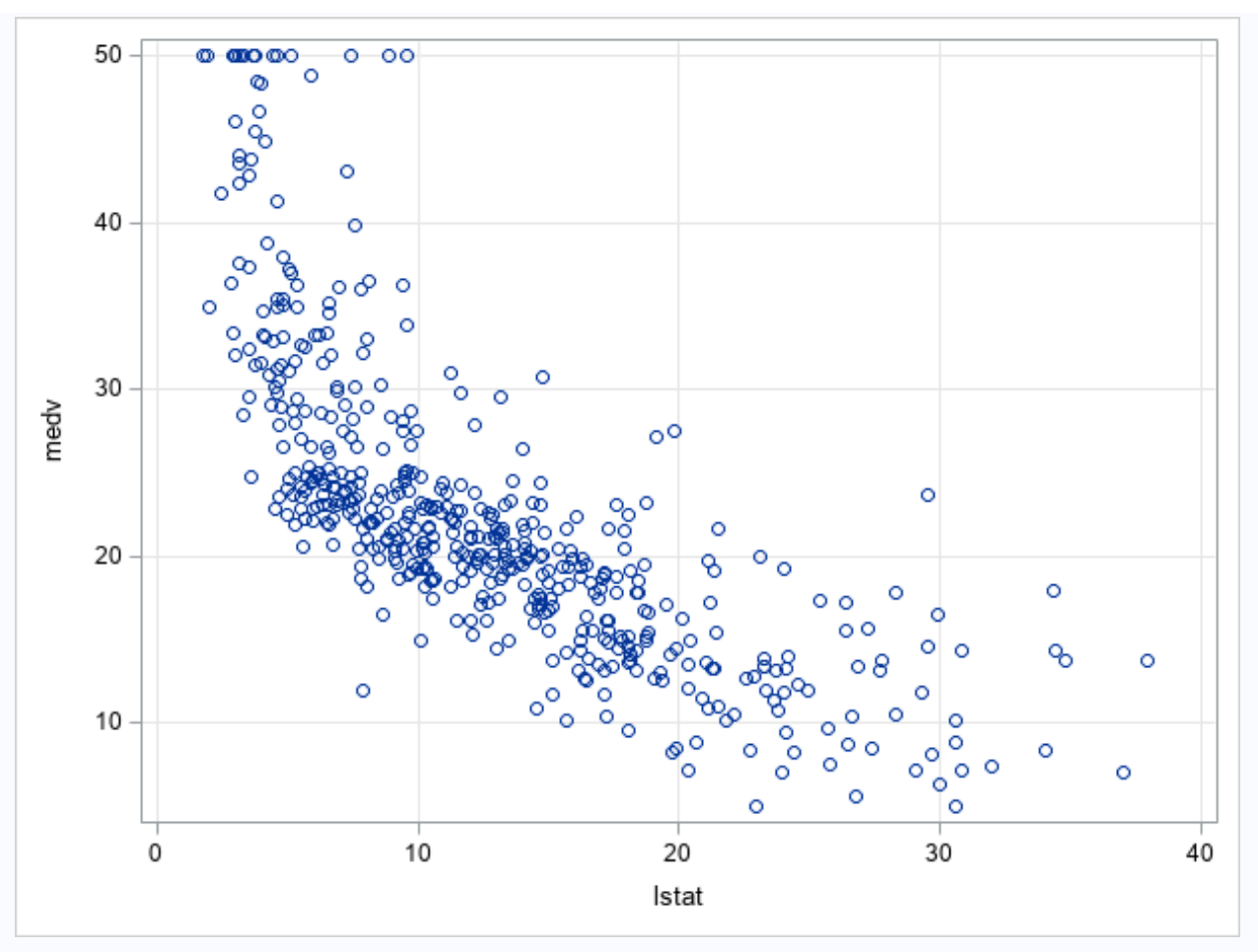
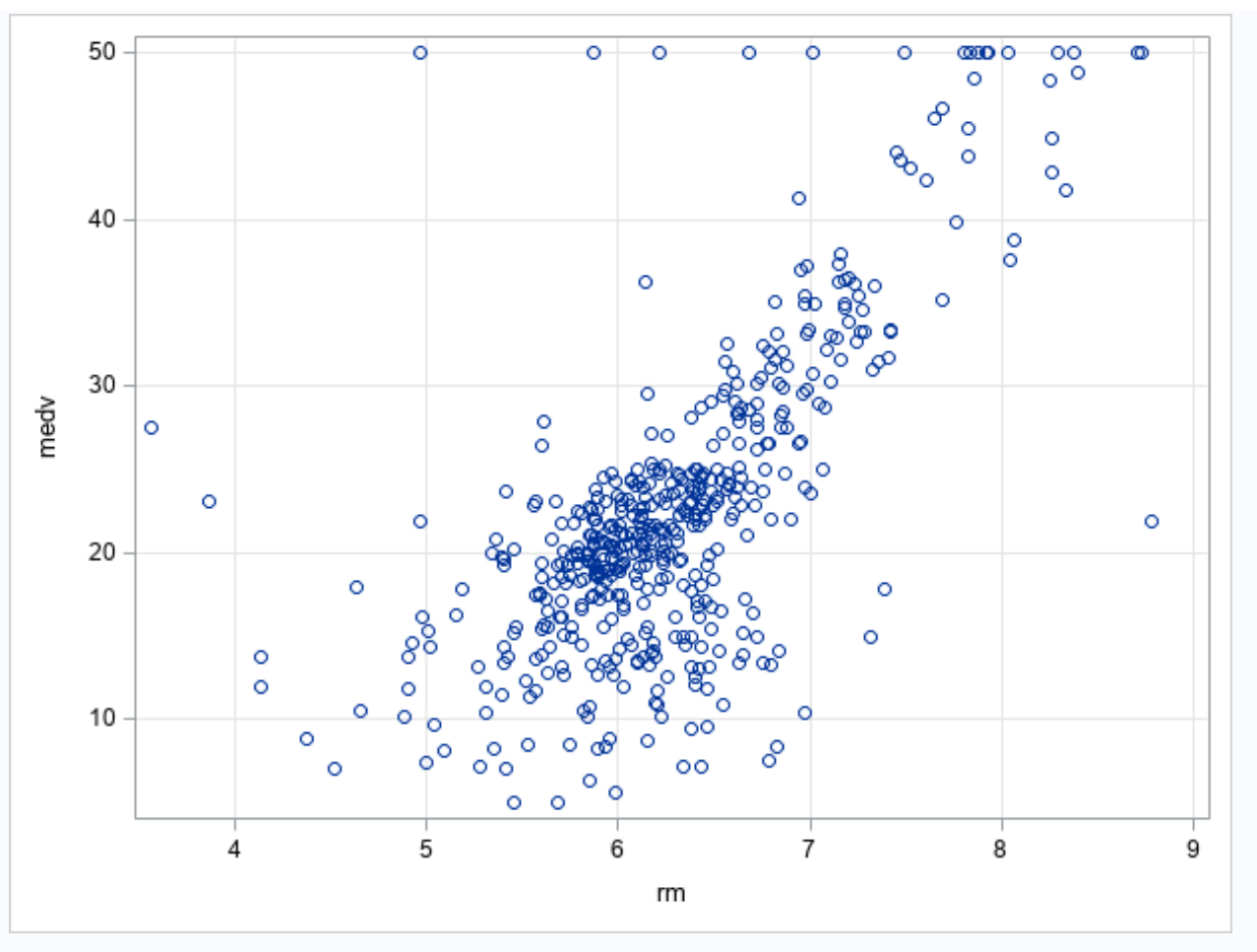
It is always crucial to understand in simple words, as the number of rooms in the house increases, the pricing increases and as the percentage of lower status people increases in the Boston area, the housing prices will drop which could be an advantage or a disadvantage to the property owners.

* + 1. **Provide the scatterplot between the response and predictor variables.**



I chose five variables that exhibit a decent correlation with the median house value for my scatter plot.

The correlation between “rm” and “lstat” are graphically depicted to see the trends using a scatter plot.

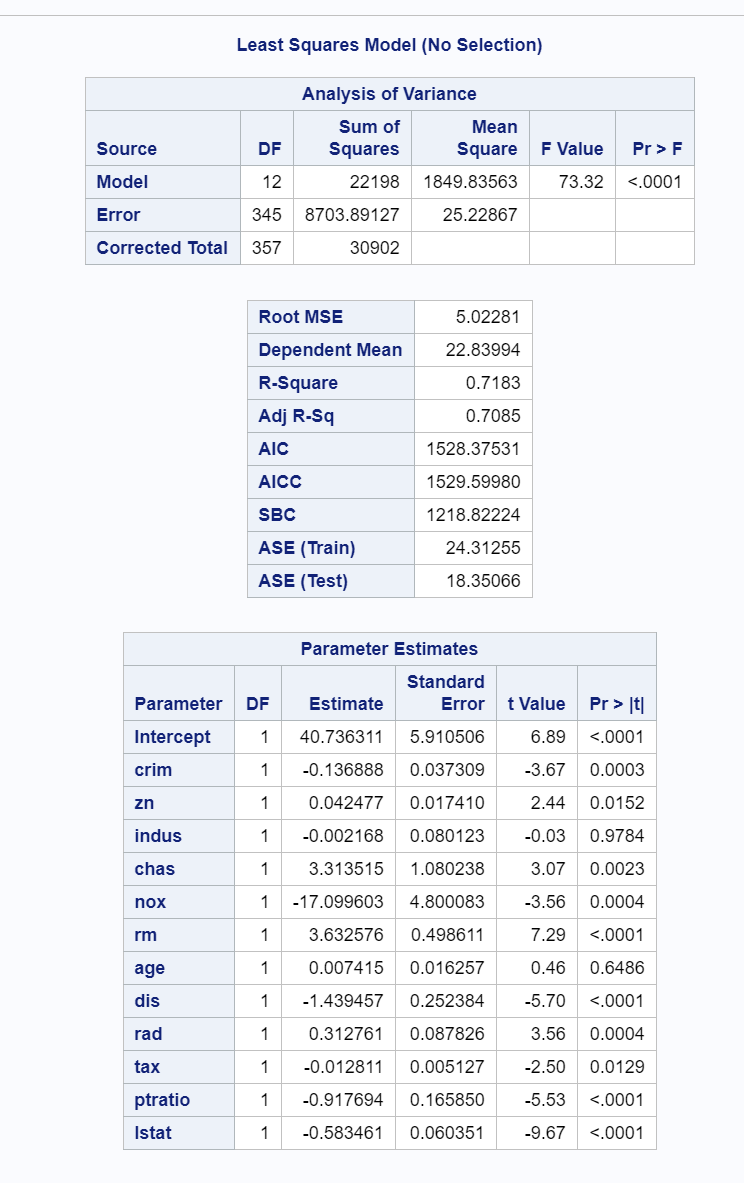
1. **Modeling and results**
   1. **Generalization Approach (1-2 sentences)**
      1. **How and why is training and testing data used for your model?**

To predict the unseen data, we usually split the data into training and testing data. We do this type of modelling in order to get the best fit of the model reducing the variance and the bias of the data to make predictions better or accurate for the business problem.

While doing linear regression with splitting the data, we are likely assuming that the predictor variables are linearly related to the response variable while the errors in the data are normalized. But in my model, I have not treated the data even though I have outliers to see how the raw data behaves according to the business problem.

* 1. **Model (1 sentence)**
     1. **What does your model do? (A linear regression model was used to…)**

Below are the analysis reports after splitting the data. I used 80-20 splitting with a random seed of 32. This linear regression model is used to predict the significance of each parameter with the model generated. With the ANOVA scores, the model is significant, and it has a good fit as the p-value is less than 0.05 significance level.

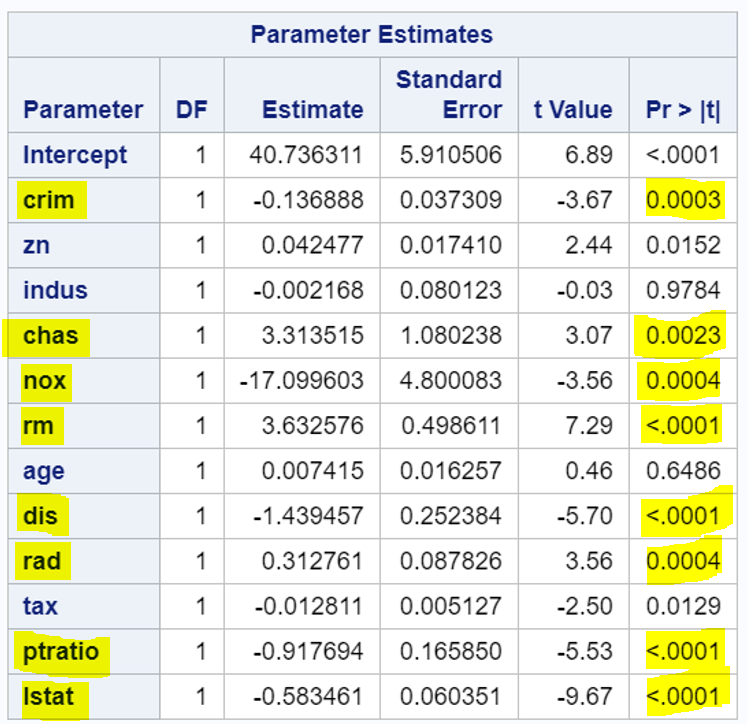


* 1. **Interpret results (3-4 sentences or more)**
     1. **Is there a relationship between the predictors and median house price?**

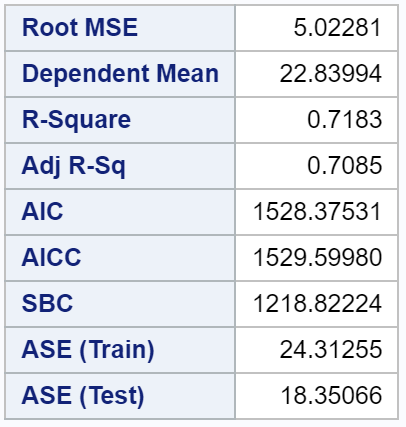
There are only few variables that are significant, related and might impact the pricing in the Boston Housing. The insignificant variables can be treated to normalize the data or can be avoided based upon the problem statement. The relation between the variables can be identified if their p value is <0.05.

* + 1. **What predictors have a statistically significant relationship to median house price?**

The highlighted predictor variables have a statistically significant relationship to the median house price.



* + 1. **What is the average squared error (ASE) of the test set?**



The ASE of the test set is **18.35066**

* + 1. **What are your recommendations to the client?**

As an analyst consultant, I would recommend the Housing firm to predict the future house pricing basing on the factors that are significant to the response variable according to the data and, I would not suggest considering the factors which are causing an increase in the housing cost that might lead to a loss in the business. For example, as the rooms are increasing in a property, the house cost would go up and we must minimize the rooms and be with the average (mean) number of rooms to have a budgeted home to the owners. Likewise, when the houses are being built in the areas where there are middle to low-income people, there is no point in investing high priced homes and that might lead to a loss to the Boston housing firms.

I feel this is the way we must consider every feature in the data and its impact to the pricing and make better decisions to invest in home.