**Homework 1. Boston Housing Dataset**

Author: Rohit V. Akole

University of Connecticut

OPIM5604 Predictive Modeling

Professor: Pankaj Prakash

Abstract

This homework is about learning predictive modeling using SAS JMP software (JMP). Using this software, I have built the graphs i.e., Boxplot, Histogram, Scattered Plot. Demonstrating the correlation between different variables of the Boston Housing Dataset.

**Definitions** of the datasets of the housing in Boston Massachusetts are as follows:

**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** - 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 centres

**RAD** - index of accessibility to radial highways

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

**PTRATIO** - pupil-teacher ratio by town

**LSTAT** - % of lower economic status population

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

**CAT.MEDV** - Derived variable.Median value of owner-occupied homes is greater than $30,000 when CAT.MEDV = 1 and less than $30,000 when CAT.MEDV = 0

Table of Contents

[Question 1 4](#_Toc144457654)

[MEDV vs. CHAS Bar Chart 4](#_Toc144457655)

[Question 2 5](#_Toc144457656)

[MEDV Histogram and Box Plot 5](#_Toc144457657)

[Question 3 6](#_Toc144457658)

[MEDV vs. LSTAT Linear Regression Scatter Plot 6](#_Toc144457659)

[MEDV vs. LSTAT Smoother Curved Scatter Plot 7](#_Toc144457660)

[Question 4 8](#_Toc144457661)

[TAX vs. MEDV Linear Regression Scatter Plot 8](#_Toc144457662)

[TAX vs. MEDV Smoother Curved Scatter Plot 9](#_Toc144457663)

[Question 5 10](#_Toc144457664)

[RM vs. MEDV Linear Regression Scatter Plot 10](#_Toc144457665)

[RM vs. MEDV Smoother Curved Scatter Plot 11](#_Toc144457666)

# Question 1

Q1. Recreate the bar chart comparing median home values for homes (MEDV) that are not near the Charles River (CHAS = 0) versus those that are (CHAS = 1) as shown on slide 17 of the Lecture 1 presentation? Provide screenshots of your work? What’s the takeaway?

## MEDV vs. CHAS Bar Chart

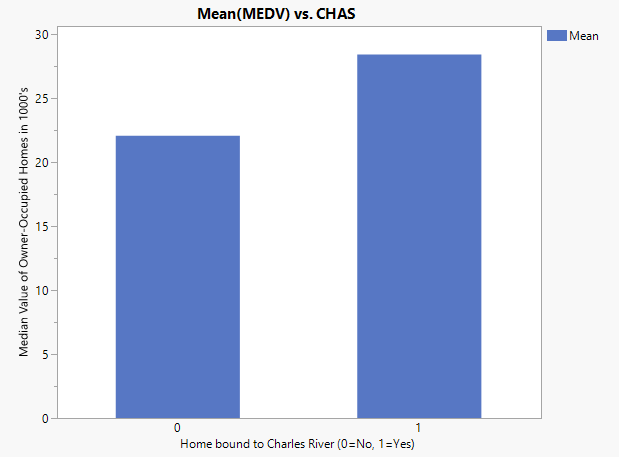


Figure 1. MEDV vs. CHAS Bar Chart

This bar chart demonstrates the Median value of owner-occupied homes in $1000's (MEDV) tract bound to Charles River. If yes then all those go under 1, if not then it goes under 0. Looking at this bar chart, we can see that the mean of MEDV for houses that are not bound to Charles River is 22.1 and 28.4 for those that are bound to Charles River. This bar chart also suggests that, if the house is bound to the Charles River, then it is more likely to be expensive as compared to those which are not.

# Question 2

Q2. Recreate the histogram showing the distribution of the median house value (MEDV) as shown on slide 20 of the Lecture 1 presentation? Provide screenshots of your work? Create a Box Plot showing the distribution of the median house value (MEDV)? Provide screenshots of your work? What is the IQR for MEDV? How would you characterize the distribution of MEDV based on the Histogram and the Box Plot?

## MEDV Histogram and Box Plot

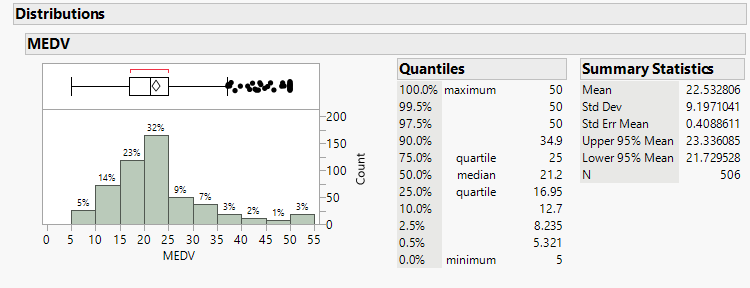


Figure 2. MEDV Histogram and Box Plot

The above histogram and box plot shows the data is skewed right (positive). The above histogram shows the maximum count (32%) at 20-25 and lowest (1%) at 45-50. On the box plot, 3% which are above 50 (maximum quartile) are the outliers. There are no outliers below the minimum quartile. We can see the 1st quartile at 16.95 and 3rd quartile at 25. With this information, we can get the IQR for the MEDV:

3rd Quartile – 1st Quartile = IQR

25 - 16.95 = **8.05**

# Question 3

Q3. Recreate the scatter plot between median house value (MEDV) and % lower status of the population (LSTAT) as shown on slide 21 of Lecture 1 presentation? Provide screenshots of your work? Describe the relationship between the two variables (+ or upward, - or downward, or scattered), and their form (linear, curved, scattered, etc.). Does the relationship make sense based on what you about the data based on the data dictionary?

## MEDV vs. LSTAT Linear Regression Scatter Plot

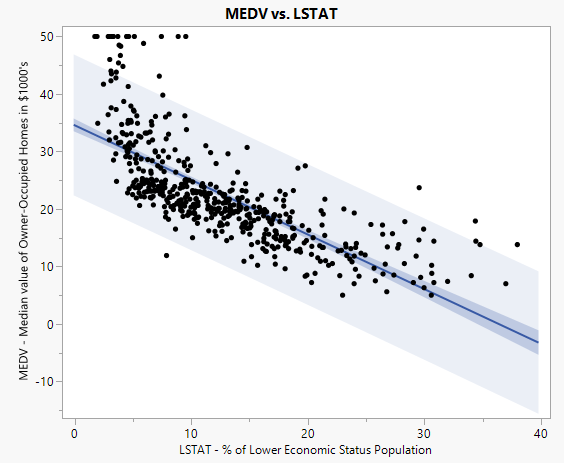


Figure 3. MEDV vs. LSTAT Linear Regression Scatter Plot with Prediction

As the % of Lower Economic Status Population increases, the MEDV decreases for them. Using the prediction, we can see as the LSTAT will increase or decrease, it is most likely to fall under near the linear regression in the prediction (light blue) area. LSTAT is more saturated in the below 20 LSTAT and below 30 MEDV. More MEDV falls below linear regression. The outliers are increasing as the LSTAT decreases. There are few outliers above 30 LSTAT. Maximum outliers are below 10 LSTAT. The Linear regression goes below 0 after 40 LSTAT.

## MEDV vs. LSTAT Smoother Curved Scatter Plot

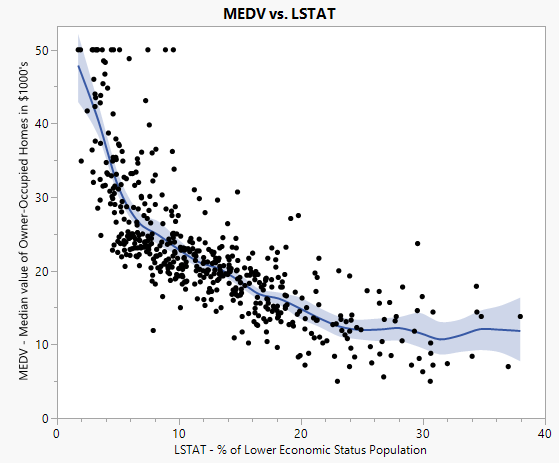


Figure 4. MEDV vs. LSTAT with Smoother Curved Scatter Plot

The smoother curved scatter plot shows the decline in MEDV is very sharp till the LSTAT reaches 10. As the LSTAT increases, the curve moves more horizontal rather than continuous decline. Here again we can see clearly that the points are more saturated around MEDV 10 - 30 and LSTAT 2 – 20. As the LSTAT increases the plot is more horizontally scattered and below 10 LSTAT the graph is vertically scattered. More of the LSTAT population are below the curved line.

# Question 4

Q4. Create the scatter plot between median house value (MEDV) and full-value property-tax rate per $10,000 (TAX)? Provide screenshots of your work? Describe the relationship between the two variables (+ or upward, - or downward, or scattered), and their form (linear, curved, scattered, etc.). Does the relationship make sense based on what you know about the data based on the data dictionary?

## TAX vs. MEDV Linear Regression Scatter Plot

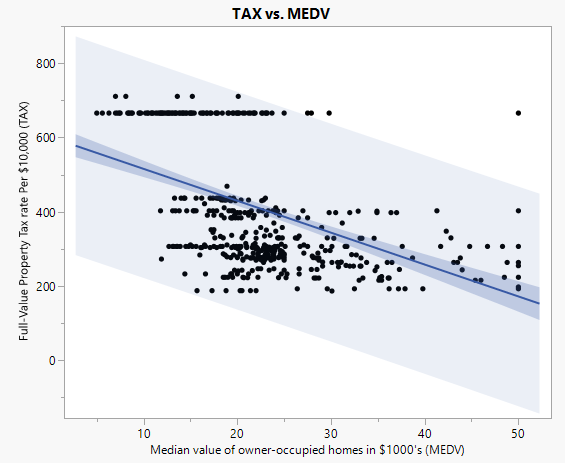


Figure 5. TAX vs. MEDV with Linear Regression Scatter Plot with Prediction

In this Scatter plot, we can establish that for a lot of properties, the tax rate was constant at 666 no matter what the MEDV is. The rest of the points are scattered horizontally around 200-500 range and a few are a little over 700 regardless of the MEDV. Using the linear regression line and prediction cloud we can certainly say that the more the MEDV, the less TAX is decreasing. So, the higher the MEDV the less Tax per $10,000 is accrued. One of the major outliers is at MEDV 50 and still, the TAX is at 666.

## TAX vs. MEDV Smoother Curved Scatter Plot

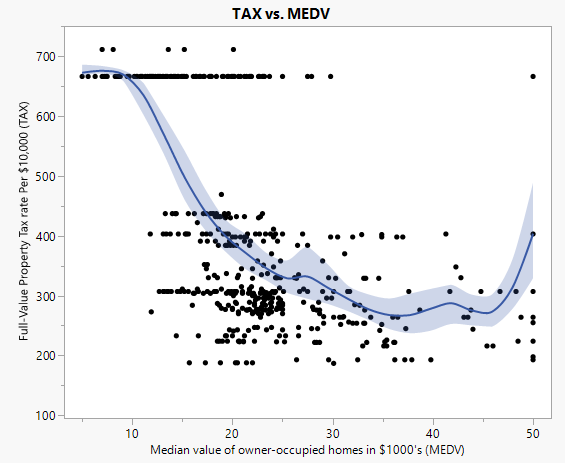


Figure 6. TAX vs. MEDV with Smoother Curved Scatter plot

Using the curved scatter plot, we can say that the TAX is decreasing as the MEDV increases. But again towards 45, it starts increasing. For some of them, no matter what the MEDV is, TAX is constant at a little over 700. In MEDV 30-50, there are no outliers or there are no points for which the tax is 666. Overall, the TAX vs. MEDV is horizontally scattered and the curve is decreasing but again increasing at the end of the plot.

# Question 5

Q5. Create the scatter plot between median house value (MEDV) and average number of rooms per dwelling (RM)? Provide screenshots of your work? Describe the relationship between the two variables (+ or upward, - or downward, or scattered), and their form (linear, curved, scattered, etc.). Does the relationship make sense based on what you know about the data based on the data dictionary?

## RM vs. MEDV Linear Regression Scatter Plot

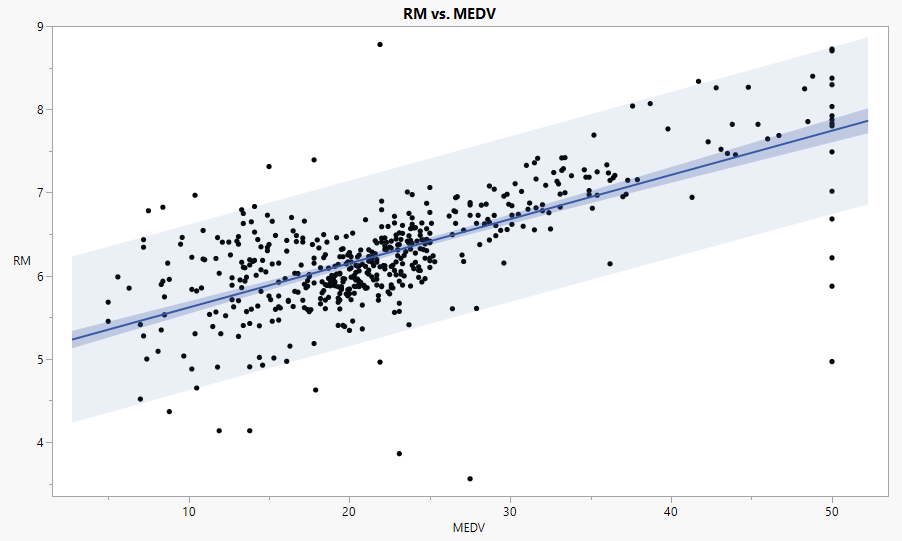


Figure 7. RM vs. MEDV with Linear Regression Scatter Plot with Prediction

There is a positive correlation between average number of rooms per dwelling and median value of owner-occupied homes in $1000’s. For some reason, there are data points which are constantly on MEDV 50 no matter how high or low the RM is. The data points are scattered across the plot. However overall linear regression suggests the upward relationship between RM and MEDV. There are a few significant outliers in this plot in between 20-30 MEDV above 8 and below 4 and on 50 MEDV as well whereas mentioned earlier, MEDV 50 has a vertical scatter from 4.5 to 9 RM.

## RM vs. MEDV Smoother Curved Scatter Plot

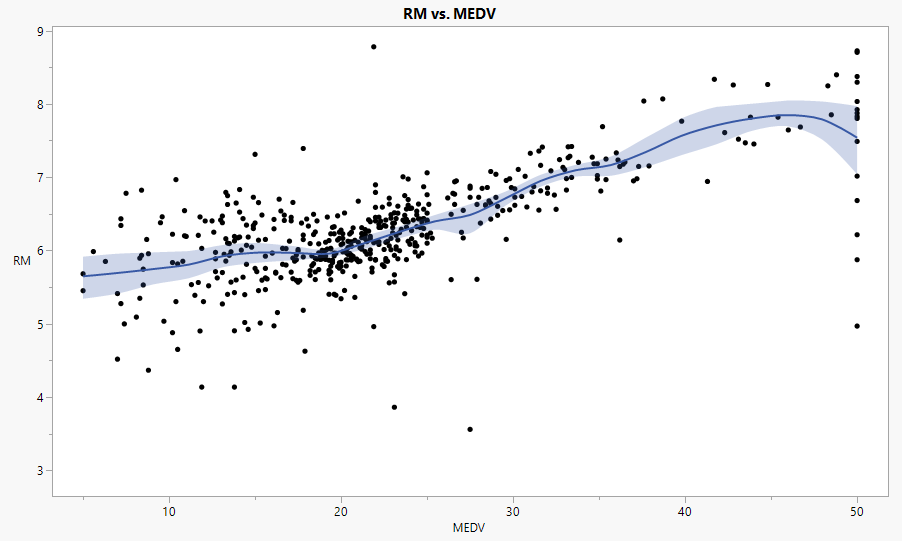


Figure 8. RM vs. MEDV with Smoother Curved Scatter plot

Using curved scatter plot, the curved line suggests positive inclination in the data as the MEDV increases, the RM is also increasing. At the MEDV 50 we can see that the curve is going towards a slight decline but that could be because the MEDV at 50 has different RM. The major group of data points are around 5-7 RM and 20-25 MEDV.