Electronic Assignment Cover Sheet

Student Name	Paulius Zaicev
Student Number	10362242
Course Title	Science in Data Analytics - Part-Time
Lecturer Name	Terri Hoare
Module/Subject Title	B8IT103 Statistics for Data Analytics
Assignment Title	Applied Statistics - Regression Modelling.
No. of Words	6373

TABLE OF CONTENTS

TABLE OF FIGURES

INTRODUCTION

The Boston housing data set was originally published by Harrison, D. and Rubinfeld, D.L. in the J. Environ. Economics & Management, vol.5, 81-102 journal in 1978. The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 independent variables for homes from various suburbs in Boston, Massachusetts.

This report seeks to examine the multiple neighborhood attributes to predict the per capita crime rate by town (*CRIM*), in an attempt to find the suitable explanatory variables. Definitions for each variable are provided in the *Table 1*.

Table 1 The Boston data set variables
Source: http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

No.	Variable	Definition
1.	*CRIM*	per capita crime rate by town
2.	*ZN*	Proportion of residential land zoned for lots over 25,000 sq.ft.
3.	*INDUS*	Proportion of non-retail business acres per town
4.	*CHAS*	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5.	*NOX*	Nitric oxides concentration (parts per 10 million)
6.	*RM*	Average number of rooms per dwelling
7.	*AGE*	Proportion of owner-occupied units built prior to 1940
8.	*DIS*	Weighted distances to five Boston employment centres
9.	*RAD*	Index of accessibility to radial highways
10.	*TAX*	Full-value property-tax rate per $10,000
11.	*PT*	Pupil-teacher ratio by town
12.	*B*	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. (**proportion of IC2 and IC6 race people by town**)
13.	*LSTAT*	% lower status of the population
14.	*MV*	Median value of owner-occupied homes in $1000's

The excel data analysis tool pack and box plots are used to visualize, explore and summarize findings from the Boston data set. The descriptive statistic summary of each variable provides us with the following information:

- Mean – the sum of all samples divided by the number of values.
- Standard Error – a measure of the statistical accuracy of an estimate, equal to the standard deviation of the theoretical distribution of a large population of such estimates.

- Median – the median of a quantitative data set is the middle number when the measurements are arranged in ascending (or descending) order. If n is odd, the value of x for which half of the remaining values are larger and half are smaller. If n is even, the average of the two values in the middle.
- Mode – the most frequently occurring value, if any.
- Standard Deviation – the standard deviation is a measure of how widely values are dispersed from the average value (the mean).
- Sample variance – square of the standard deviation.
- Kurtosis – kurtosis characterizes the relative peak or flatness of a distribution compared with the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution. The kurtosis of a sample is consistent with a normal distribution if it is near value 3.
- Skewness – characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values. The skewness of a sample is consistent with a normal distribution for a population if it's absolute value is small, e.g. less than 0.3.
- Range – maximum value minus minimum value.
- Minimum – minimum value.
- Maximum – maximum value.
- Sum – sum of all values.
- Count – number of values, n. (Harvey, 2015)

1. DESCRIPTIVE STATISTICS

In the provided data set there are 506 observations with 14 variables. For better visualization the box plot for all variables is illustrated in the *Figure 1*. To visualize all variables in the one figure proportion for each variable was counted.

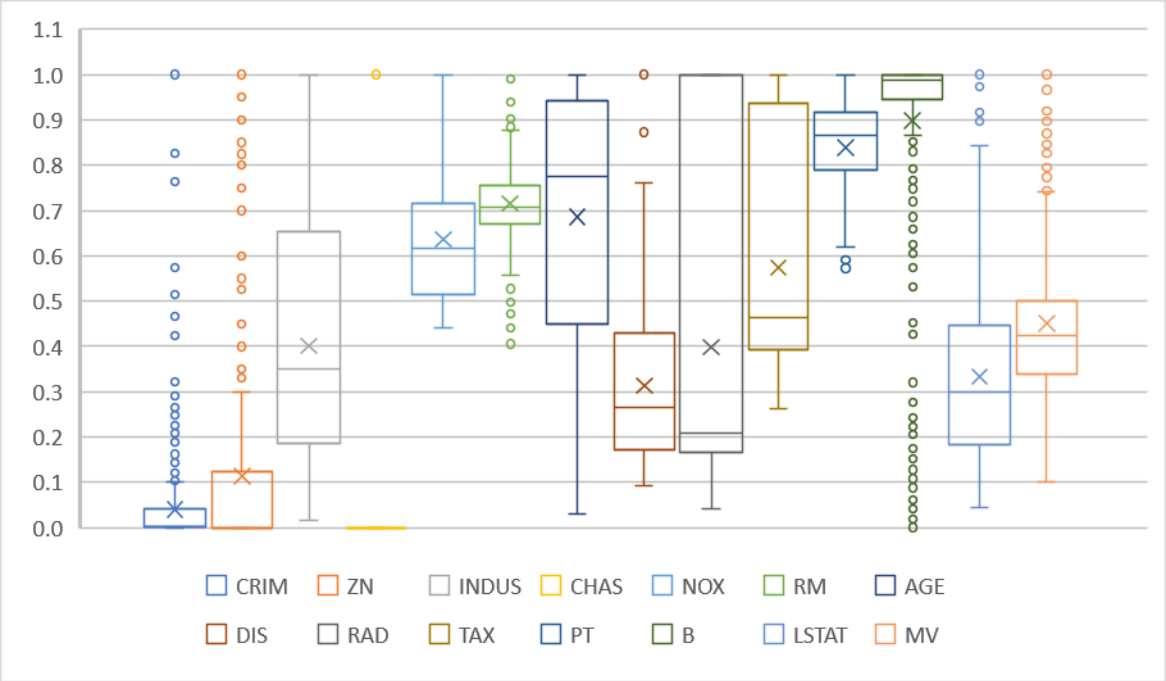


Figure 1 The box plots of all variables

The main goal of this analysis is to examine the multiple neighborhood attributes to give information to the regression model that will intern explain the variation in the per capita crime rate by town (CRIM). For this reason the CRIM is dependent variable in this data set, however, specific neighborhood attributes such as the weighted distance to five Boston employment centres (DIS), pupil-teacher ratio by town (PT), proportion of IC2 and IC6 race people by town (B) and lower status of the population (LSTAT) will be described and analyzed (Steele, Syndercombe Court, Balding, 2014). These four variable were selected due to expectation for most likely collinearity with the variable CRIM. Observation for these variable are listed below:

- DIS – the weighted distance to five Boston employment centres. Assumption can be made that if weighted distance are longer from the Boston employment centres, this area can be more dangerous and per capita crime rate can be higher.
- PT - pupil-teacher ratio by town is the ratio of students to teachers in primary and secondary schools in the neighborhood. First assumptions have been made that wealthier schools have higher budgets for salary so they have lower ratio of students

to teachers. Second, is that schools with a smaller number of pupils are more demanding and are more expensive, for this reason schools with less pupils might be located in a nicer area. So assumption can be made, that the higher ratio of students to teachers, the bigger per capita crime rate by town.

- B – the proportion of the IC2 and IC6 race people by town. Many people may believe that an area with many IC2 and IC6 race people will be not safe. This analysis will help to investigate if higher per capita crime rate have a positive correlation to the weighted proportion of IC2 and IC6 race people by town.
- LSTAT – the lower status of the population is the percentage of homeowners in the neighborhood considered as "lower class". Percentage of homeowners in the neighborhood considered as "lower class" refers to the number of working poor people among all people in the neighborhood. This analysis will help to investigated if this variable will have a positive correlation with per capita crime rate by town values.

Five variables – CRIM, DIS, PT, B and LSTAT are described below. The descriptive statistic for variables ZN, INDUS, CHAS, NOX, RM, AGE, RAD, TAX, PT and MV can be found in the Appendix A. The information will help to familiarize all variables and will help justify the results shown.

First variable, the CRIM (per capita crime rate by town) was analyzed by using the excel analysis tool pack (*Table 2*). Upon analysis, it was investigated that the smallest (minimum) per capita crime rate by town is 0.006 while the biggest (maximum) is 88.976. This generates a range of 88.970.

Table 2 The CRIM descriptive statistics summary by using the Excel analysis tool pack

CRIM			
Mean	3.614	*Kurtosis*	37.131
Standard Error	0.382	*Skewness*	5.223
Median	0.257	*Range*	88.970
Mode	0.015	*Minimum*	0.006
Standard Deviation	8.602	*Maximum*	88.976
Sample Variance	73.987	*Sum*	1828.443
Count	506.000		

According to the *Table 2* we can see that the most frequently repeated value is 0.015 (mode), while the average per capita crime rate by town is at 3.614 (mean) and at the same time value 0.257 (median) divides the values in two halves. The big variance between the mean and the

median show that the CRIM variable has a big skewness which consist of 5.223 and kurtosis consist of 37.131. This shows that the variable CRIM doesn't have a normal distribution. Furthermore, we can see that the sample variance of the CRIM is 73.987 and the standard deviation is 8.60.

To better visualize the distribution of the variable the box plot is provided below (*Figure 2*). The chart provides information on the lower quartile (Q_L=0.082) and the upper quartile (Q_U=3.682). It is interesting to see that 50% of the values are placed between the Q_L and the Q_U range. Additionally, the mean of 3.614 and the median of 0. 257 are illustrated in the *Figure 2*.

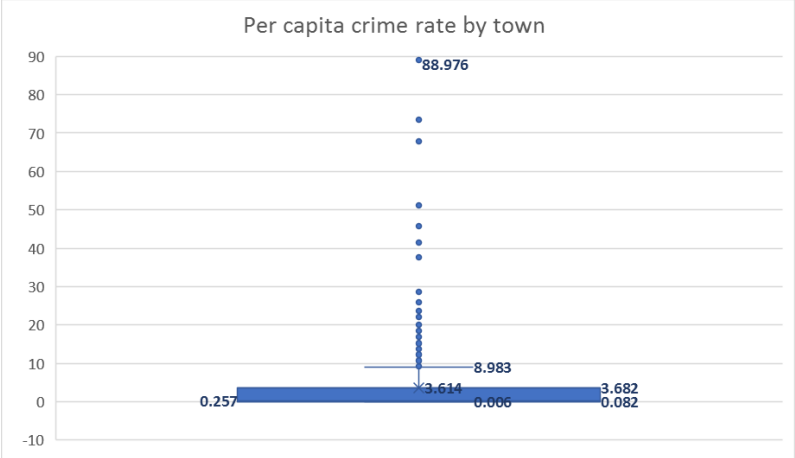


Figure 2 The box plot for per capita crime rate by town

According to the *Figure 2* the lower inner fence is placed at 0.006 while the upper inner fence is located at 8.983. *"Values that are beyond the inner fences are deemed potential outliers, because they are extreme values that represent relatively rare occurrences"* (McClave, Sinich, 2012). The outer fences, are defined at a distance 3(IQR) from each end of the box and in this box plot outer fences were calculated and can be located beyond the lower outer fence -10.717 and the upper outer fence of 14.481. Interesting to see that in the CRIM variable 37 (7.3%) observations fall between the upper inner fence and the upper outer fence, however, 30 (5.9%) observations fall beyond the upper outer fence. It is reasonable to have outliers in the CRIM variable, as there may be some dangerous and unsafe areas leading to this results.

The variable DIS, which represents weighted distances to five Boston employment centres was analyzed and a summary is presented in the *Table 3*. In the variable DIS the smallest (minimum) weighted distance to five Boston employment centres is 1.130 while the biggest proportion is (maximum) – 12.127. These proportions create the range of – 1.997.

According to the summary we can see that the most frequently repeated value is 3.495 (mode), while average weighted distance – 3.795 (mean) and the value 3.207 (median) divides values from the DIS variable in two halves.

Table 3 The DIS descriptive statistics summary by using the Excel analysis tool pack.

DIS			
Mean	3.795	*Kurtosis*	0.488
Standard Error	0.094	*Skewness*	1.012
Median	3.207	*Range*	10.997
Mode	3.495	*Minimum*	1.130
Standard Deviation	2.106	*Maximum*	12.127
Sample Variance	4.434	*Sum*	1920.292
Count	506.000		

The variance between the mean and the median show that the DIS variable skews by 1.012 and has a kurtosis of 0.488, which shows that variable doesn't have a normal distribution. Additionally, the sample variance of the DIS is 4.434 and the standard deviation – 2.106.

The box plot chart is provided below (*Figure 3*) for the DIS variable observation. The chart provides us with the information of the lower quartile 2.097 (Q_L) and the upper quartile 5.213 (Q_U). The graph shows that 50% of the values are placed between the Q_L and the Q_U range.

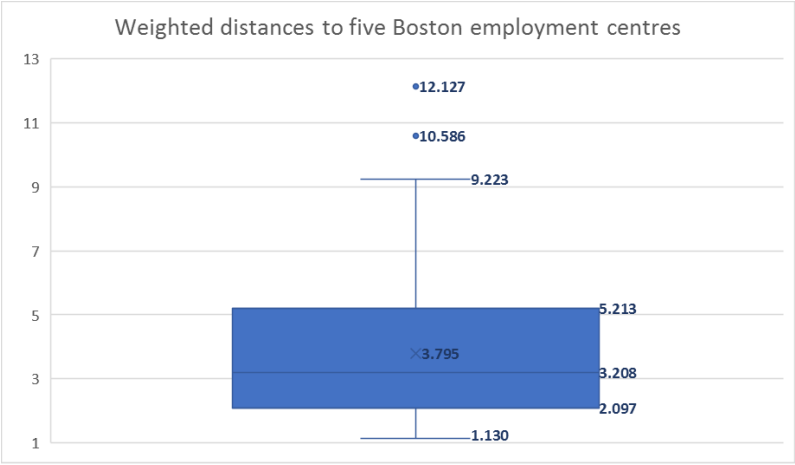


Figure 3 The box plot for weighted distance to five Boston employment centres

In the *Figure 3* the lower inner fence is 1.130 while the upper inner fence is located at 9.223. The values of lower and upper outer fences was computed and in this box plot no values are placed beyond the lower or upper outer fences, however, 6 (1.2%) observation are placed outside upper inner fence. The *Figure 4* illustrates how the DIS variable's observations are separated into 12

different intervals. The pareto line shows that 501 values are placed between inner fences which represents approximately 98.8% of values.

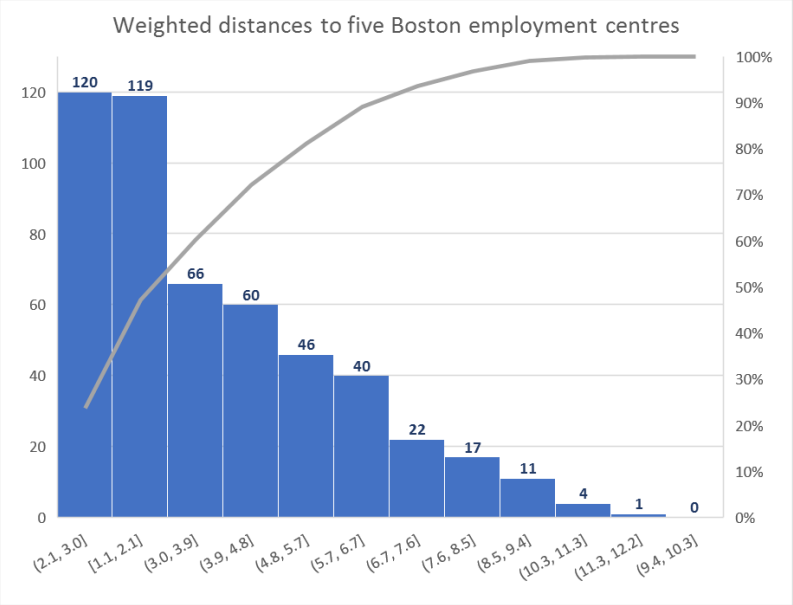


Figure 4 The histogram chart for proportion of weighted distance to five Boston employment centres

The chart illustrates that 6 observations are placed between the upper inner fence (9.223) and the upper outer fence (14.550) which confirm that approximately 1.2% of values from the variable are potential outliers, however, it is normal to have outliers in the DIS variable because some observed areas can have a bigger weighted distance to the employment centres.

The variable PT, which represents pupil-teacher ratio by town was analyzed and a summary is presented in the *Table 4*. In the PT variable the smallest pupil -teacher ratio (minimum) is 12.6 while the biggest proportion (maximum) – 22.0 and this represents the difference of ratio equal to 9.4. According, to the *Table 4* the most frequently repeated ratio is 20.2 (mode), while the average ratio is 18.456 (mean) and the ratio of 19,050 (median) divides observations in two halves.

Table 4 The PT descriptive statistics summary by using the Excel analysis tool pack.

PT			
Mean	18.456	*Kurtosis*	-0.285
Standard Error	0.096	*Skewness*	-0.802
Median	19.050	*Range*	9.400
Mode	20.200	*Minimum*	12.600
Standard Deviation	2.165	*Maximum*	22.000
Sample Variance	4.687	*Sum*	9338.500
Count	*506.000*		

The variance between the mean and the median show that the PT variable has the negative skewness of -0.802 and kurtosis of -0.285. Additionally, the sample variance of this variable is 4.687 and the standard deviation – 2.165. The box plot illustration (*Figure 5*) for the PT variable provides us with the information of the lower 17.375 (Q_L) and the upper quartile 20.200 (Q_U).

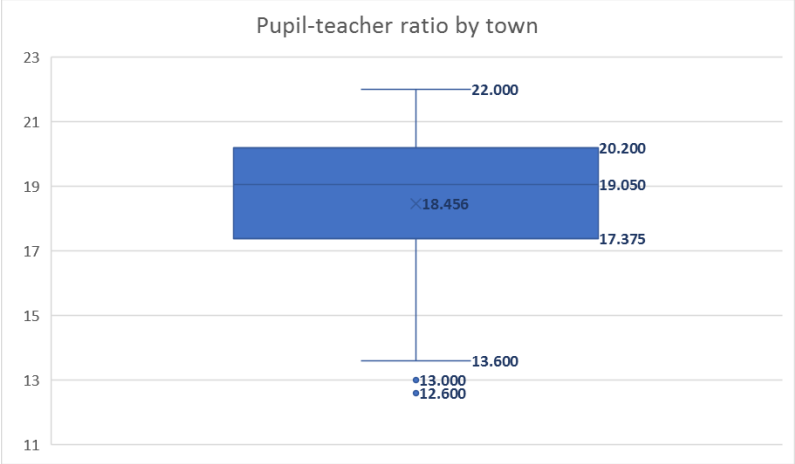


Figure 5 The box plot for pupil-teacher ratio by town

In the *Figure 5* the lower inner fence is located at 13.60 while the upper inner fence is located at 22.00. From the PT variable 491 (97%) observations are located between inner fences. No values are located beyond upper inner fence, however, 15 (3%) observations are placed below the lower inner fence. The *Figure 6* illustrates how the PT observations are separated into 10 different intervals. The Pareto line shows how many observations are stored in each interval. The interval between 19.22 and 20.2 stores the highest ratio, 161 observations and it shows that the pupil-teacher ratio between these intervals is the most common.

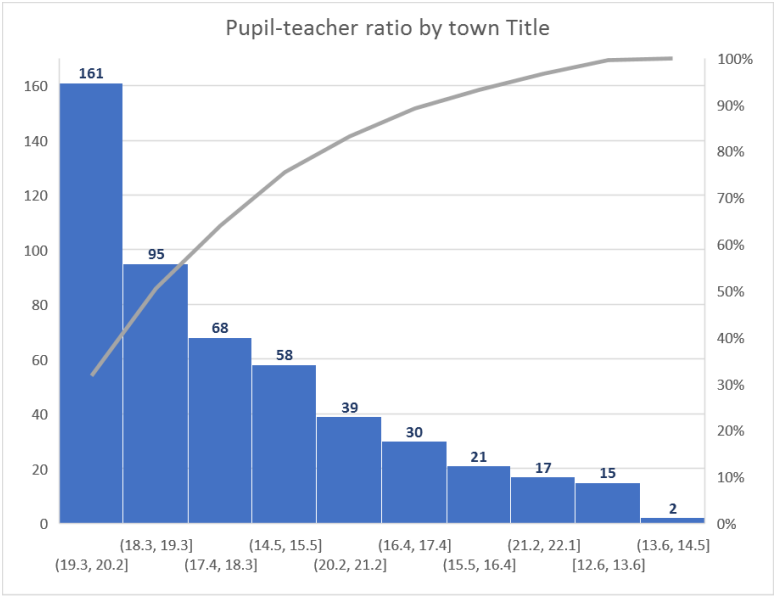


Figure 6 The histogram chart for pupil-teacher ratio by town

The chart illustrates that 15 observations are placed below the lower inner fence (13,60) which shows that approximately 3% of values are potential outliers, however, it is normal to have outliers in the PT variable because wealthier schools can have higher budgets for salary or schools with the smaller number of pupils and are more expensive for this reasons. These outliers are reasonable.

The variable B, represents the proportion of coloured people by town which was counted by the formula $1000(B-0,63)^2$. Further in the analysis coloured people will be described as IC2 and IC6 race people. The descriptive statistic summary for IC2 and IC6 race people proportion by town is illustrated in the *Table 5.* In the variable B the smallest (minimum) observed proportion of IC2 and IC6 in the area is equal to 0.320 while in the biggest proportion (maximum) is 396.900. This creates the difference of 399.580 between proportions. The *Table 5* illustrates that the most frequently repeated proportion is 3.900 (mode), while average proportion is 356.674 (mean) and the proportion of 391.440 (median) divides values into two equal halves.

Table 5 The B descriptive statistics summary by using the Excel analysis tool pack.

B			
Mean	356.674	*Kurtosis*	7.227
Standard Error	4.059	*Skewness*	-2.890
Median	391.440	*Range*	396.580
Mode	396.900	*Minimum*	0.320
Standard Deviation	91.295	*Maximum*	396.900
Sample Variance	8334.752	*Sum*	180477.059
Count	506.000		

The variance between the mean and the median show that the B variable has a negative skew equal to -2.890 and kurtosis equal to 7.227. Furthermore, the sample variance of the B is 8334.752 and the standard deviation is 91.295.

The box plot chart is provided below (*Figure 7*) for the B variable observations. Illustrated box plot for the B variable is different from previously explained ones. The box plot perfectly illustrates that this variable has a significant amount of outliers below lower inner fence. The chart provides us with the position of the lower quartile at 375.300 (Q_L) and the upper quartile 396.233 (Q_U).

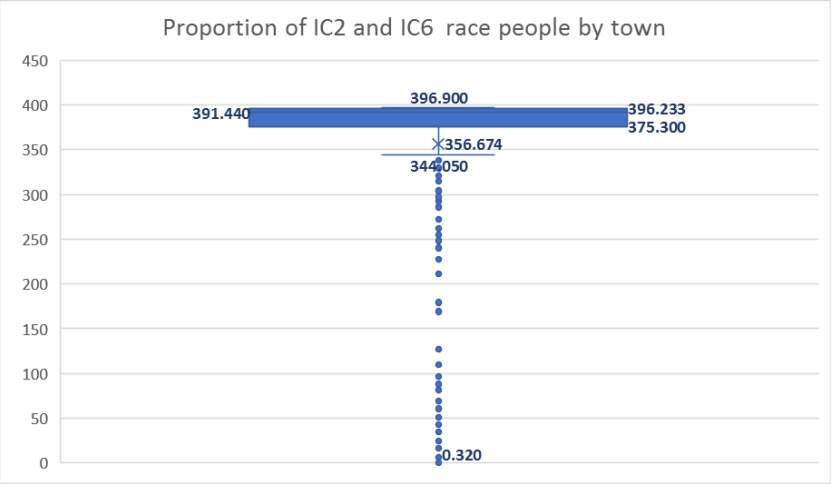


Figure 7 The box plot for proportion of IC2 and IC6 race people by town

In the *Figure 7* the lower inner fence is located at 344.05 while the upper inner fence is located at 396.900. Interesting to see that upper inner fence is a maximum observation in the B variable and that 85% of variables are placed between inner fences. The value of lower outer fence was computed and can be located at 312.503.

The *Figure 8* shows how the B variable's observations are separated into 10 different intervals and illustrates that 18 (3.5%) observation are placed between lower inner fence and lower outer fence, while 58 (11.5%) observations are placed below lower outer fence.

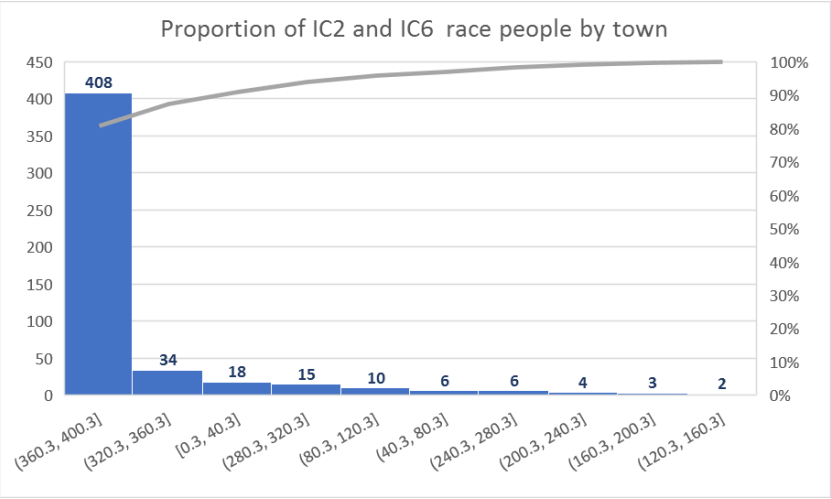


Figure 8 The histogram chart for proportion of IC2 and IC6 race by town

The chart illustrates that 15 % of observations are potential outliers, however, it is normal to have outliers in the B variable as some areas can have a larger proportion of IC2 and IC6 race people living in the area. It will be interesting to investigate this variable as assumption have been made that it should have a positive correlation with the CRIM variable.

The variable LSTAT, represents the lower status of the population in a percentage of homeowners in the neighborhood considered as "lower class". The percentage of homeowners in the neighborhood considered as "lower class" refers to the number of working poor people among all people in the neighborhood. The descriptive statistic summary for lower class people is presented in the *Table 6.* The smallest (minimum) percentage of working people in the area from the observed values is 1.730 while the biggest (maximum) percentage is 37.97. These variables create the difference of 36.240 between the poorest and richest area from the observations. The *Table 6* presents the most frequently repeated percentage in the LSTAT variable is 8.05 (mode), while average percentage is 12.653 (mean) and percentage of 11.360 (median) divides observation from the LSTAT in two equal halves.

Table 6 The LSTAT descriptive statistics summary by using the Excel analysis tool pack.

LSTAT			
Mean	12.653	*Kurtosis*	0.493
Standard Error	0.317	*Skewness*	0.906
Median	11.360	*Range*	36.240
Mode	8.050	*Minimum*	1.730
Standard Deviation	7.141	*Maximum*	37.970
Sample Variance	50.995	*Sum*	6402.450
Count	506.000		

The variance between the mean and the median show that the LSTAT variable has the positive skewness equal to 0.906 and kurtosis equal to 0.493. The sample variance is equal to 50.995 and the standard deviation is 7.141. The box plot chart is provided below (*Figure 9*) for the LSTAT variable observations. The box plot perfectly illustrates that this variable have 6 outliers. The chart provides us with the position of the lower quartile at 6,928 (Q_L) and the upper quartile 16,992 (Q_U).

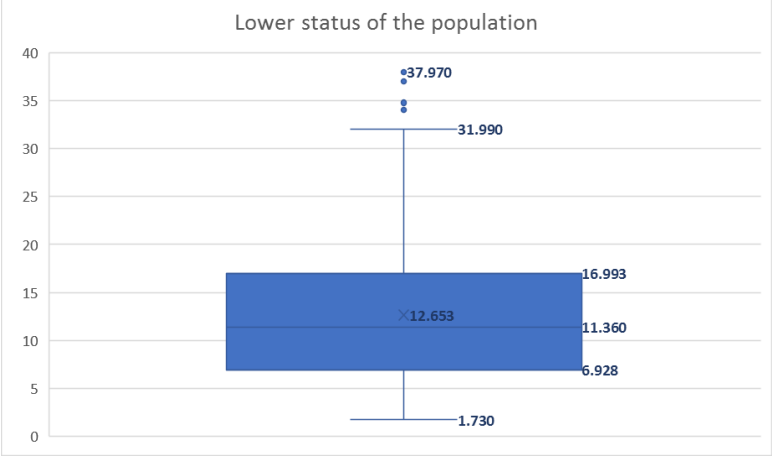


Figure 9 The box plot for percentage of lower status of the population in the area

In the *Figure 9* the lower inner fence is located at 1.730 while the upper inner fence is located at 31.990. The 98.8 % (500) of observations are placed between the inner fences. The chart illustrates that 1.2 % of observations are potential outliers, however, it is normal to have outliers in the LSTAT variable as some areas can have a larger percentage of poor working people and it will be interesting to investigate if this variable have a positive correlation with the CRIM variable.

The descriptive statistics summaries for variables ZN, INDUS, CHAS, NOX, RM, AGE, RAD, TAX, PT and MV are presented in the Appendix A. According to the Appendix A, all variables have skewness not equal to 0 and kurtosis not equal to 3, however, the variable RM is skewed 0.404 and kurtosis of 1.892 which is close to the normal distribution requirements. The box plot illustrations and quartile numerical values for each variable are presented in the Appendix B.

2. REGRESSION MODELS

"Simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variables. The adjective simple refers to the fact that the outcome variable is related to a single predictor." (Seltman, 2015)

In this analysis the CRIM is a dependent variable and simple linear regression models will be fitted between all 13 variables (Table 1) using R Studio software. However, the main goal of this analysis is to analyze specific neighborhood attributes such as weighted distance to five Boston employment centres (DIS), pupil-teacher ratio by town (PT), proportion of IC2 and IC6 race people by town (B) and lower status of the population (LSTAT) and find which variable has a strongest statistical relationship with dependent variable CRIM.

To create a correlation matrix plot between analyzed parameters pairs function is used in the R Studio software. (*Figure 10)* The pairs matrix help us to visualize possible relation between variables.

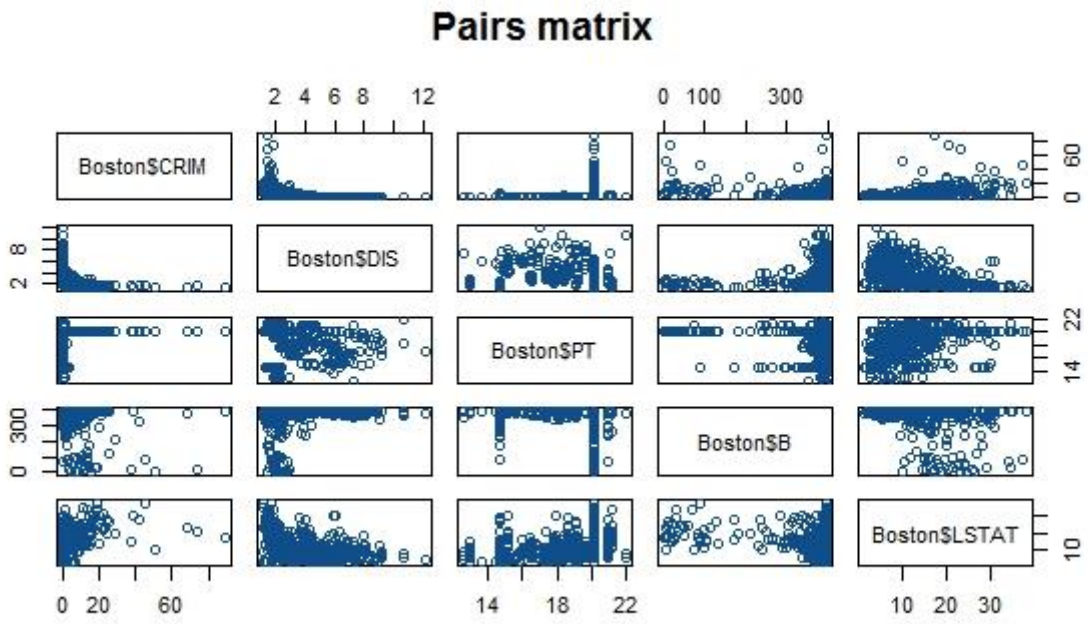


Figure 10 The pairs matrix between analyzed variables

In this particular case is very had to see relationship between CRIM and DIS, PT, B, LSTAT variables. For this reason correlation for each variable will be analyzed separately. Additionally, Appendix C illustrates pairs matrix between all variables which are listed in *Table 1*.

To find the strongest relationships between the dependent variable CRIM and independent variables the correlation coefficient will be analyzed. *"The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The correlation coefficient takes on values ranging between +1 and -1."* (Friedman, Tibshirani, Hastie, 2001) The heat correlation visualization is illustrated in *Figure 11* and numerical correlation values are presented in the Appendix D.

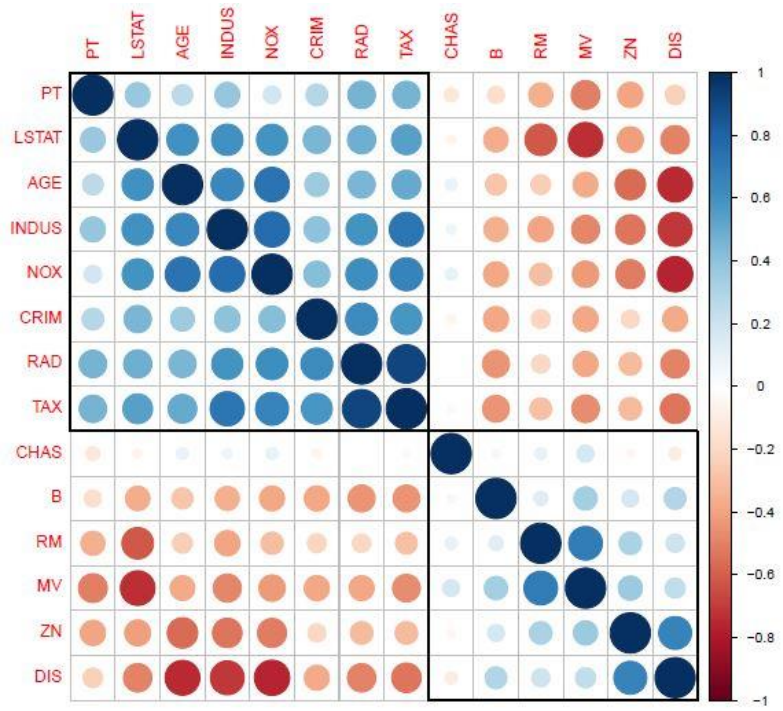


Figure 11 The heat correlation illustration for Boston data set

Furthermore, the statistical hypothesis will be tested. *"In statistical hypothesis testing, the p-value or probability value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or more extreme than the actual observed results."*(Wasserstein, Lazar, 2016). All computed p-values are presented in the Appendix G.

The relationship between the CRIM and DIS variable is illustrated in the scatterplot (*Figure 12).* According to the scatter plot we can see that negative linear relationship between variables CRIM and DIS exist, however, two of the variables don't have strong relationship.

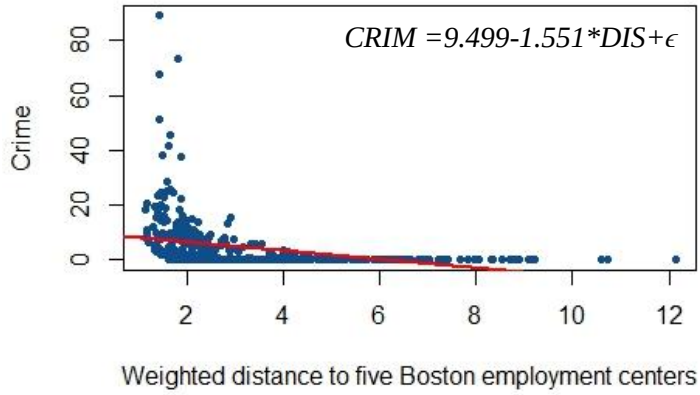


Figure 12 The regression between CRIM and DIS variables

The R studio software has indicated that correlation coefficient between two variables is equal to -0.38 which represent medium correlation and amount of variability (R^2) is 0.1425. The intercept for CRIM and DIS was computed and represents β_0= 9.499 while in the meantime slope is β_1= -1.551. The equation is illustrated in the *Figure 12*. The equation shows that variable CRIM is a dependent variable and variable DIS is predictor variable, for this reason we can assume, that the if weighted distance to five Boston employment centres increases per capita crime rate by town decrease by 1.551. For this reason the assumption which has been made is that areas further from five Boston employments centres have a higher per capita crime rate in the area is not correct.

The DIS variable in this particular case determines the mean value of the CRIM variable, which is a specific point on the line of the means. Furthermore, null hypothesis was computed H_0 *:* $\beta_1 = 0$ and computed *p – value* is equal to $2.2*10^{-16}$ and calculation shows that the H_0 can be rejected and research hypothesis can be accepted. The graphical analysis of regression residual is illustrated in the *Figure 13*.

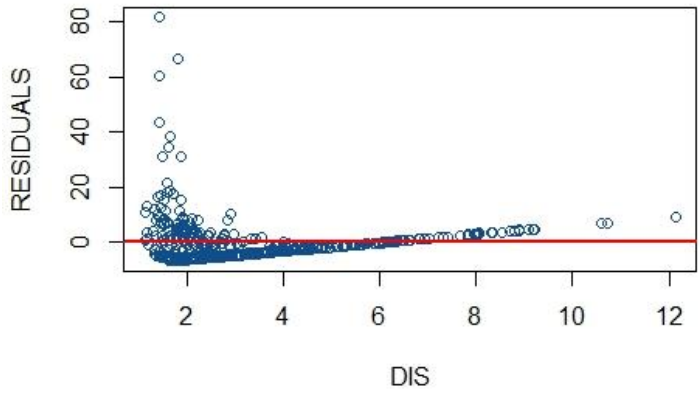


Figure 13 The residual regression model for DIS variable

This model for CRIM variable represents the change in per capita crime rate by town dependent on weighted distances to five Boston employment centres, however, this model illustrates that residuals are not randomly distributed. The *Figure 13* shows that the residuals

between 0 and 4 on the DIS axis (the x axis) has a large distribution and that there is a large error between observed values and fitted values, however, the distance value increase from 4 and higher the error becomes smaller.

The second single linear regression model is illustrated in the *Figure 14*, between dependent variable CRIM and independent variable PT. According to the scatter plot we can interpret that there is a positive correlation existing between the two variables. The correlation coefficient between two variables is indicated to be equal to 0.290 and amount off variability (R^2) is 0.08225. This indicates a very weak positive correlation.

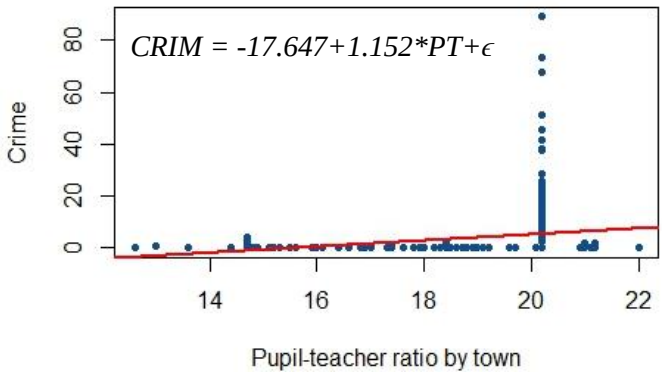


Figure 14 The regression between CRIM and PT variables

The intercept for CRIM and PT was calculated and represents β_0= -17.647 while the slope is β_1= 1.152. The equation is illustrated in the *Figure 14*. The equation shows that variable CRIM is a dependent variable and the variable PT is a predictor variable, for this reason we can assume, that the pupil-teacher ratio by town increases per capita crime rate by town increase accordingly by 1.152. This regression model proves that if an area has a higher ratio of students to teacher per capita crime rate the area is vaster. Also, null hypothesis was computed then $H_0 : \beta_1 = 0$ and computed *p – value* is equal to $2.94*10^{-11}$. Due to the small p – value H_0 fails to be accepted.

The third plot illustrates that there is a decreasing linear relationship between the dependent variable CRIM and the independent variable B (1000(Bk - 0.63)^2 where Bk is the proportion of coloured people by town). The *Figure 15* illustrates negative correlation between two variables.

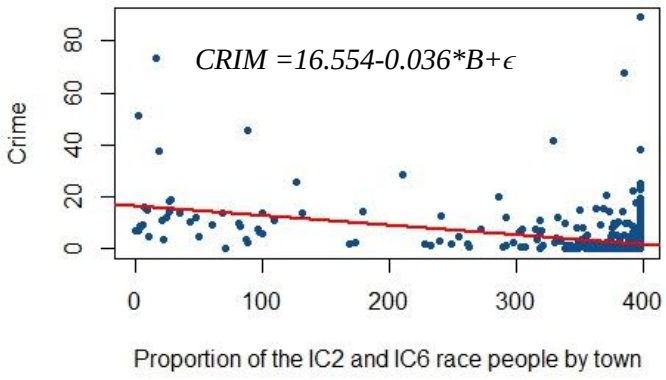


Figure 15 The regression between CRIM and B variables

The computed correlation coefficient between two variables is equal to -0.385 this indicates a negative medium correlation and amount of variability (R^2) is 0.1466. The intercept for CRIM and B variable was calculated and represents β_0= 16.554 while the mean slope is β_1= -0.036. The equation is illustrated in *Figure 15*. The equation shows that variable CRIM is a dependent variable and variable B is a predictor variable. Interesting to see, that the proportion of the IC2 and IC6 race people by town increase by B, per capita crime rate by town decrease accordingly by 0.0336. For this reason we can assume that if the town has a larger proportion of the IC2 and IC6 race people, per capita crime rate in the town is lower. The null hypothesis was computed $H_0 : \beta_1 = 0$ and computed $p - value$ is equal to $2.2*10^{-16}$ for this reason H_0 can't be accepted.

The fourth plot illustrates that positive linear relationship between the dependent variable CRIM and the independent variable LSTAT which represents the percentage of lower status population in the area. The scatter plot illustrates positive correlation between two variables (*Figure 16).*

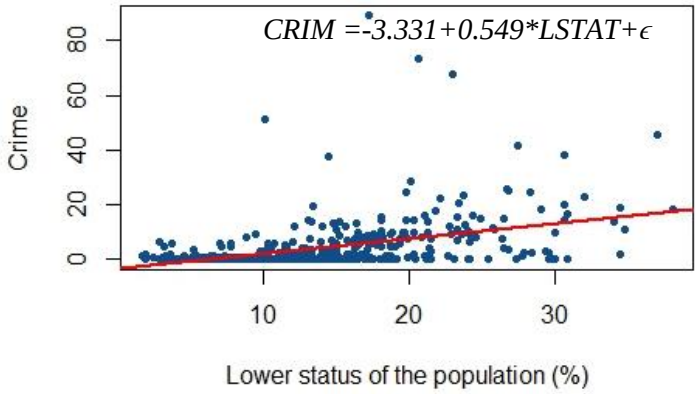


Figure 16 The regression between CRIM and LSTAT variables

The computed correlation coefficient between two variables is equal to 0.455 that indicates the medium correlation and amount of variability (R^2) is 0.206. The LSTAT variable has the highest correlation coefficient with the CRIM variable compared with the previously analyzed variables. The intercept for regression model was calculated and represents β_0= -3.335 while in the mean slope is β_1= 0.549. The equation is illustrated in the *Figure 16*. The equation shows that variable CRIM is dependent variable and variable LSTAT is predictor variable. According, to the regression model the percentage of the lower status of the population increases, per capita crime rate by town increase by 0.549, for this reason it can be assumed that the larger percentage of lower status of the population in the area, the bigger per capita crime in town can be observed. The null hypothesis was computed then $H_0 : \beta_1 = 0$ and computed $p - value$ is equal to $2.2*10^{-16}$ for this reason H_0 can be rejected. The graphical analysis of regression residual is illustrated in the *Figure 17.*

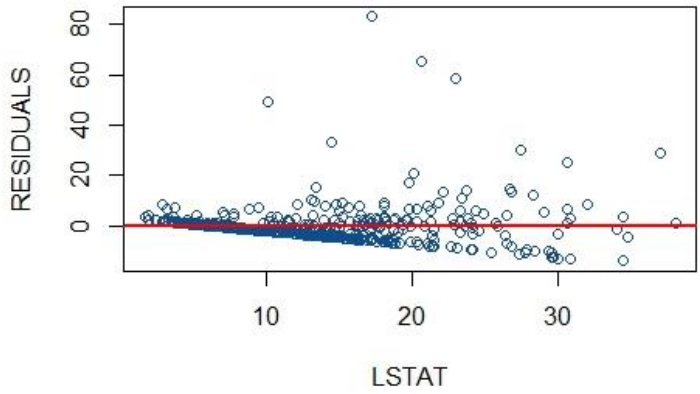


Figure 17 The residual regression model for LSTAT variable

The *Figure 17* shows that the linear model has a good fit for relatively small LSTAT ratio values, but it is not a good predictor for a LSTAT values which are larger than 15.

The numerical correlation values between variables are illustrated in the Appendix D. The analysis shows that the RAD and TAX variables have the most significant correlation with the CRIM variable which is equal to RAD=0.626 and TAX=0.583 and represent a high correlation. Furthermore, regressions models, Q-Q plots, residual plots, calculated p-values, amount of variability (R^2), intercepts and slopes between the variable CRIM and variables ZN, INDUS, CHAS, NOX, RM, AGE, RAD, TAX, PT and MV are presented in the Appendix E and Appendix G.

The table in the Appendix G helps to estimate the significant predictors. To estimate predictors the null hypothesis ($H_0 : \beta_1 = 0$) has been tested. From the results in the Appendix G we can conclude that all predictors have a p-value close to zero (less than 0.05) except the variable CHAS. For this reason we can assume that there is a statistically significant association between each predictor and CRIM variable excluding the CHAS predictor.

A multiple regression model was fitted to predict the CRIM variable response using all the predictors which are listed in the *Table 1*. Summary is illustrated in the *Figure 18*.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.924  -2.120  -0.353   1.019  75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033225   7.234903   2.354 0.018949 *
ZN            0.044855   0.018734   2.394 0.017025 *
INDUS        -0.063855   0.083407  -0.766 0.444294
CHAS         -0.749134   1.180147  -0.635 0.525867
NOX         -10.313532   5.275537  -1.955 0.051152 .
RM            0.430130   0.612830   0.702 0.483089
AGE           0.001452   0.017925   0.081 0.935488
DIS          -0.987176   0.281817  -3.503 0.000502 ***
RAD           0.588209   0.088049   6.680 6.46e-11 ***
TAX          -0.003780   0.005156  -0.733 0.463793
PT           -0.271081   0.186450  -1.454 0.146611
B            -0.007538   0.003673  -2.052 0.040702 *
LSTAT         0.126211   0.075725   1.667 0.096208 .
MV           -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Figure 18 The multiple regression model

According to the *Figure 20* the null hypothesis $H_0 : \beta_1 = 0$ can be rejected for the ZN, DIS, RAD, B and LSTAT variables, because p-value are less than 0.05.

The plot which illustrated in the *Figure 19* displays coefficients from the simple linear regression for each variable with the CRIM on the x axis. The coefficients estimate from the multiple linear regression for each variable with the CRIM is on the y axis. Each predictor is displayed as a single point in the *Figure 21*.

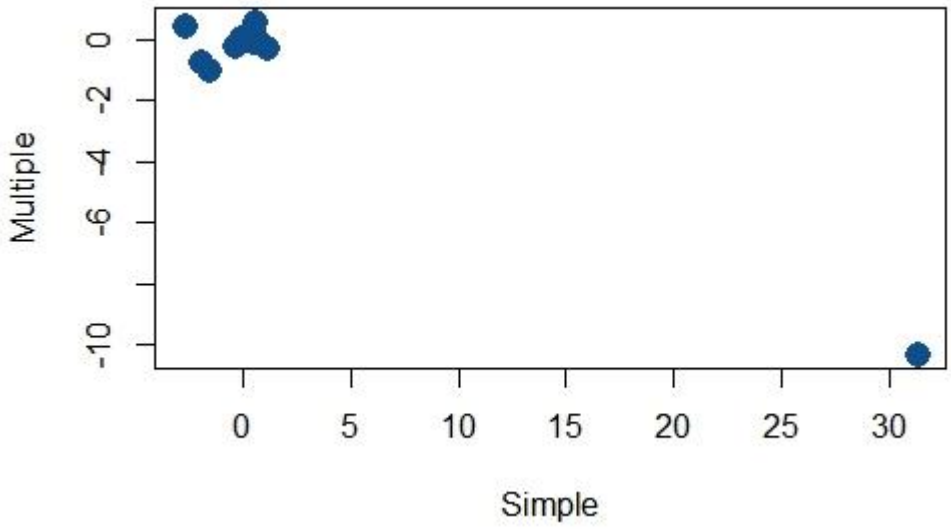


Figure 19 The plot of simple and multiple linear regression coefficients

The difference between the simple and multiple regression coefficients is that in the simple regression model the slopes represents the change in the predictor variable ignoring other predictors and shows how this change affects dependent variable (in this particular case the

variable CRIM). In the multiple regression model the slopes represents the change in the predictor variable, with no change on other predictors (all other slopes maintain fixed values) and represents how this change affects the dependent variable CRIM.

To investigate a non-linear relationship between the dependent variable CRIM and all 13 predictors the model $\boldsymbol{CRIM = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon}$ was fitted. The p-values for linear, quadratic and cubic models are computed and illustrated in the *Table 7*. The main goal of this analysis was to investigate relationship between the CRIM and the DIS, PT, B and LSTAT variables. According to the *Table 7* the DIS and PT variables are predictor, the p-values suggest the satisfaction of the cubic model fit, while in the meantime the LSTAT p-value shows that the cubic co-efficiency is not statically meaningful. The variable B is the only one variable for which the linear relationship model is the best model, due to high p-values in the quadratic and cubic models.

Table 7 The p-values for the $\mathbf{CRIM = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3}$ model

Dependent variable	**Independent variable**	**x (P-value)**	**x^2 (P-value)**	**x^3 (P-value)**
CRIM	*ZN*	$4.7*10^{-6}$***	0.00442**	0.22954
	INDUS	$2*10^{-16}$***	0.00109**	$1.2*10^{-12}$***
	CHAS	0.209	N/A	N/A
	NOX	$2*10^{-16}$***	$7.74*10^{-5}$***	$6.96*10^{-16}$***
	RM	5.13**10^{-7}***	0.00151**	0.50857
	AGE	$2*10^{-16}$***	$2.29*10^{-6}$***	0..0068**
	DIS	$2*10^{-16}$***	$7.87*10^{-14}$***	$1.09*10^{-8}$***
	RAD	$2*10^{-16}$***	0.00912**	0.48231
	TAX	$2*10^{-16}$***	$3.67*10^{-6}$***	0.244
	PT	$1.57*10^{-11}$***	0.00241**	0.00630**
	B	$2*10^{-16}$***	0.457	0.544
	LSTAT	$2*10^{-16}$***	0.0378*	0.1299
	MV	$2*10^{-16}$***	$2*10^{-16}$***	$1.05*10^{-12}$***

The p-values for the ZN, RM, RAD, TAX variables show the same results as the LSTAT variable. This means the cubic coefficient is not statically meaningful. While the p-values for the INDUS, NOX, AGE, and MV variables illustrate the same results as the DIS and PT variables the cubic model is a reasonable and a satisfying model.

BIBLIOGRAPHY

1. C. D. Steele, D. S. (November 2014). *Worldwide Fst Estimates Relative to Five Continental-Scale Populations*.
2. D. Haririson JR, D. L. (December 22, 1976). Hedonic Housing Prices and Demand for Clean Air. *Journal of environmental economics and management 5, 81-108*.
3. Harvey, G. (2015). *Excel 2016 All-in-One For Dummies (For Dummies (Computer/Tech).* Hoboken, New Jersey: John Wiley & Sons.
4. J. T. McClave, T. S. (2012). *Statistics, 12th edition.* New Jersey: Pearson.
5. Jerome H. Friedman, Robert Tibshirani, Trevor Hastie. (2003). *The Elements of Statistical Learning.* Springer.
6. Seltman, H. J. (2015). *Experimental Design and Analysis.* Retrieved from http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf
7. Wassertein, R. L., & Lazar, N. A. (2016 March 7). *The ASA's Statement on p-Values: Context, Process and Purpose.* The American Statistician.

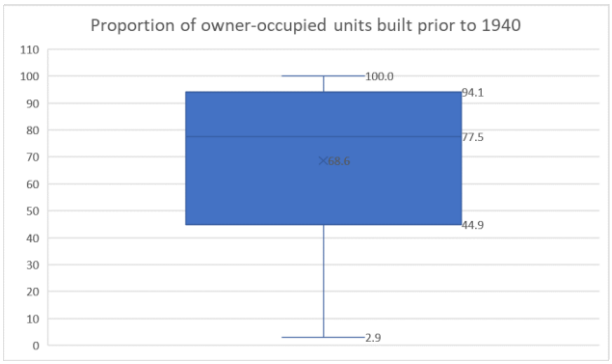
APPENDIX A

	CRIM	***ZN***	***INDUS***	***CHAS***	***NOX***	***RM***	***AGE***
Mean	3.614	11.364	11.137	0.069	0.555	6.285	68.575
Standard Error	0.382	1.037	0.305	0.011	0.005	0.031	1.251
Median	0.257	0.000	9.690	0.000	0.538	6.209	77.500
Mode	0.069	0.000	18.100	0.000	0.538	5.713	100.000
Standard Deviation	8.602	23.322	6.860	0.254	0.116	0.703	28.149
Sample Variance	73.987	543.937	47.064	0.065	0.013	0.494	792.358
Kurtosis	37.130	4.032	-1.234	9.638	-0.065	**1.892**	-0.968
Skewness	5.223	2.226	0.295	3.406	0.729	**0.404**	-0.599
Range	88.970	100.000	27.280	1.000	0.486	5.219	97.100
Minimum	0.006	0.000	0.460	0.000	0.385	3.561	2.900
Maximum	88.976	100.000	27.740	1.000	0.871	8.780	100.000
Sum	1828.440	5750.000	5635.210	35.000	280.676	3180.025	34698.900
Count	506.000	506.000	506.000	506.000	506.000	506.000	506.000
Confidence Level(95.0%)	0.751	2.037	0.599	0.022	0.010	0.061	2.459

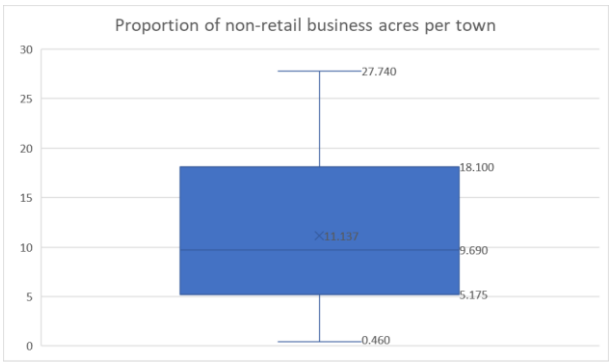
	DIS	*RAD*	*TAX*	*PT*	*B*	*LSTAT*	*MV*
Mean	3.795	9.549	408.237	18.456	356.674	12.653	22.533
Standard Error	0.094	0.387	7.492	0.096	4.059	0.317	0.409
Median	3.208	5.000	330.000	19.050	391.440	11.360	21.200
Mode	5.401	24.000	666.000	20.200	396.900	8.050	50.000
Standard Deviation	2.106	8.707	168.537	2.165	91.295	7.141	9.197
Sample Variance	4.434	75.816	28404.759	4.687	8334.752	50.995	84.587
Kurtosis	0.488	-0.867	-1.142	-0.285	7.227	0.493	1.495
Skewness	1.012	1.005	0.670	-0.802	-2.890	0.906	1.108
Range	10.997	23.000	524.000	9.400	396.580	36.240	45.000
Minimum	1.130	1.000	187.000	12.600	0.320	1.730	5.000
Maximum	12.127	24.000	711.000	22.000	396.900	37.970	50.000
Sum	1920.299	4832.000	206568.000	9338.500	180477.060	6402.450	11401.600
Count	506.000	506.000	506.000	506.000	506.000	506.000	506.000
Confidence Level(95.0%)	0.184	0.760	14.720	0.189	7.974	0.624	0.803

APPENDIX B

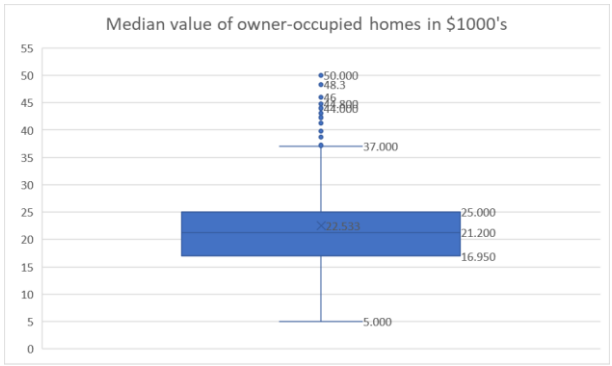
The variable **AGE**



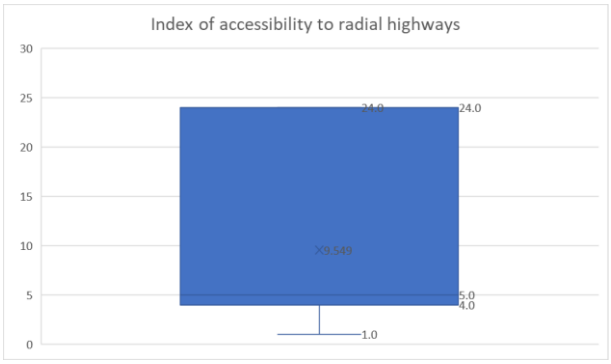
The variable **INDUS**



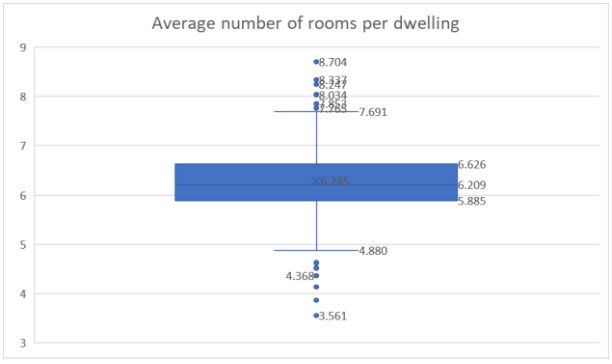
The variable **MV**



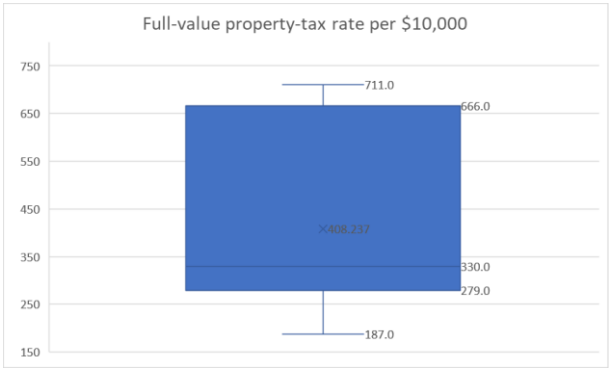
The variable **RAD**



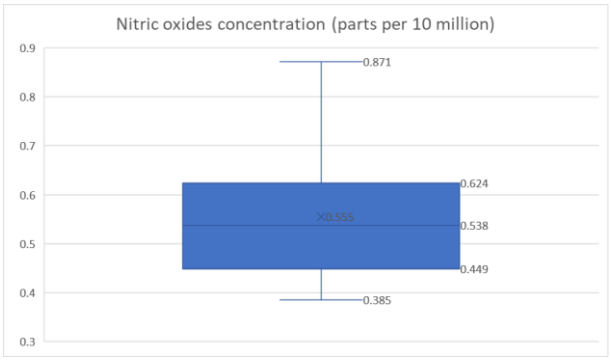
The variable **RM**



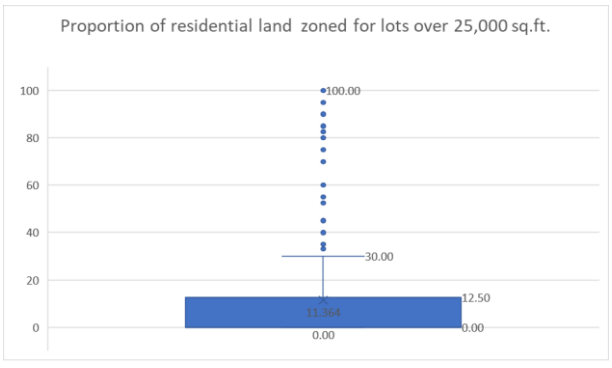
The variable **TAX**



The variable **NOX**



The variable **ZN**

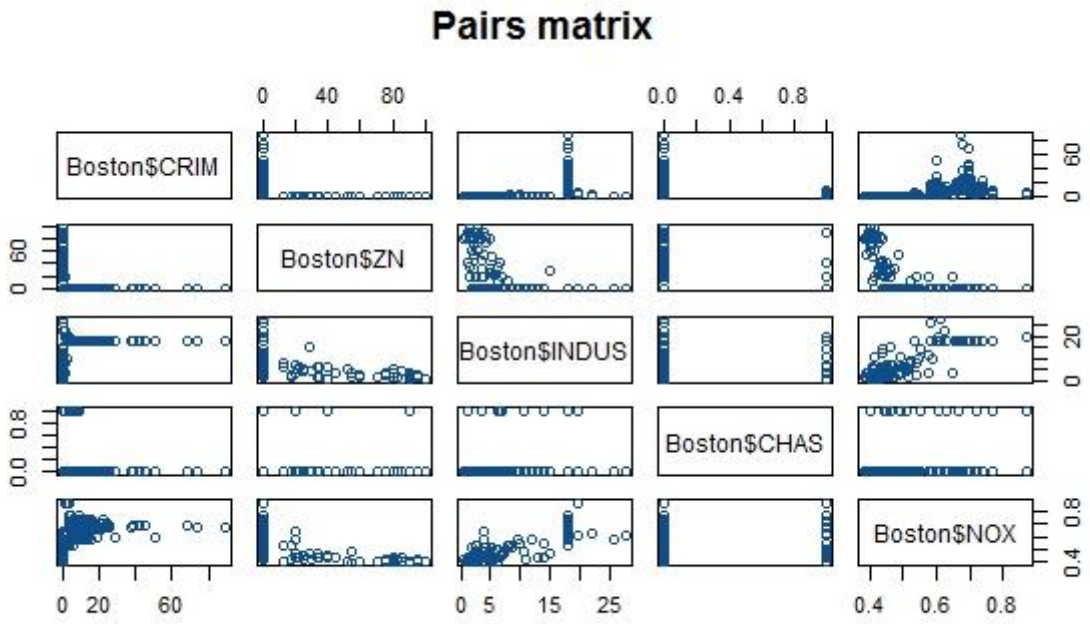


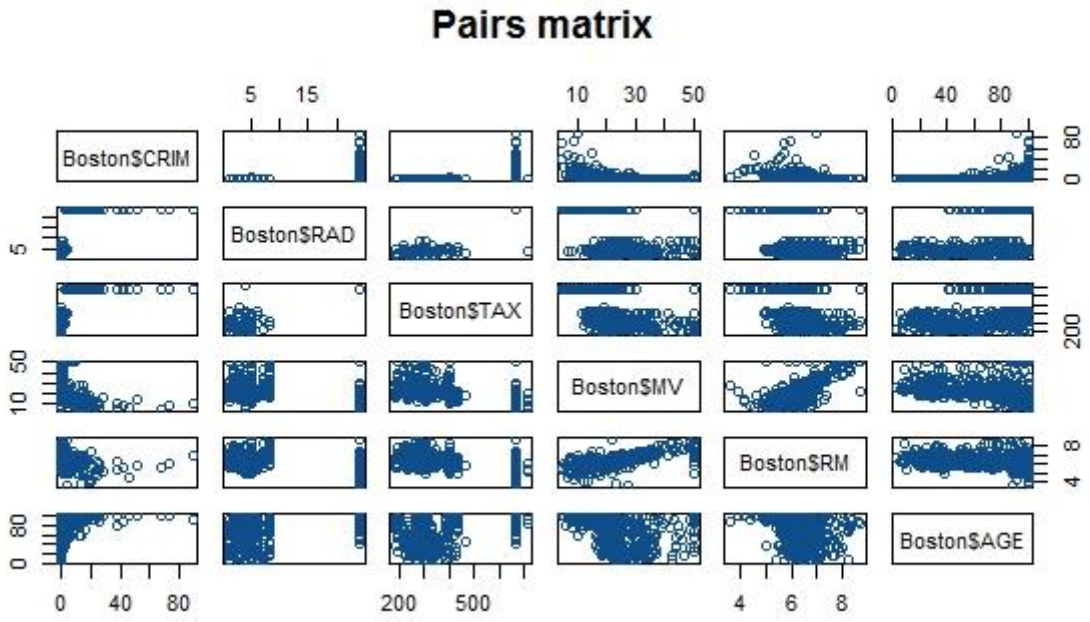
	CRIM	*ZN*	*INDUS*	*CHAS*	*NOX*	*RM*	*AGE*
Minimum	0.006	0.000	0.460	0.000	0.385	3.561	2.900
QL(25%)	0.082	0.000	5.175	0.000	0.449	5.885	44.850
Median	0.257	0.000	9.690	0.000	0.538	6.209	77.500
QL(75%)	3.682	12.500	18.100	0.000	0.624	6.626	94.100
Maximum	88.976	100.000	27.740	1.000	0.871	8.780	100.000
IQR	3.600	12.500	12.925	0.000	0.175	0.741	49.250
Lower inner fence	-5.318	-18.750	-14.213	0.000	0.187	4.773	-29.025
Upper inner fence	9.081	31.250	37.488	0.000	0.887	7.738	167.975
Lower outer fence	-10.717	-37.500	-33.600	0.000	-0.076	3.661	-102.900
Upper outer fence	14.481	50.000	56.875	0.000	1.149	8.850	241.850
Between inner fences	440 (87%)	438 (86.6%)	506 (100%)	N/A	506 (100%)	476 (94.07%)	506 (100%)
Between upper inner fence and upper outer fence	37 (7.3%)	23 (4.5%)	0	N/A	0	22 (4.35%)	0
Between lower inner fence and lower outer fence	0	0	0	N/A	0	7 (1.38%)	0
Outside lower outer fence	0	0	0	N/A	0	1 (0.2%)	0
Outside upper outer fence	30 (5.9%)	45 (8.9%)	0	N/A	0	0	0

	DIS	*RAD*	*TAX*	*PT*	*B*	*LSTAT*	*MV*
Minimum	1.130	1.000	187.000	12.600	0.320	1.730	5.000
QL(25%)	2.100	4.000	279.000	17.375	375.300	6.928	16.950
Median	3.208	5.000	330.000	19.050	391.440	11.360	21.200
QL(75%)	5.213	24.000	666.000	20.200	396.233	16.993	25.000
Maximum	12.127	24.000	711.000	22.000	396.900	37.970	50.000
IQR	3.113	20.000	387.000	2.825	20.933	10.065	8.050
Lower inner fence	-2.569	-26.000	-301.500	13.138	343.901	-8.170	4.875
Upper inner fence	9.882	54.000	1246.500	24.438	427.631	32.090	37.075

Lower outer fence	-7.237	-56.000	-882.000	8.900	312.503	-23.268	-7.200
Upper outer fence	14.550	84.000	1827.000	28.675	459.030	47.188	49.150
Between inner fences	500 (98.81%)	506 (100%)	506 (100%)	491 (97.04%)	430 (84.98%)	500 (98,81%)	469 (92.69%)
Between upper inner fence and upper outer fence	6 (1.19%)	0	0	0	0	6 (1.19%)	21 (4.15%)
Between lower inner fence and lower outer fence	0	0	0	15 (2.96%)	18 (3.56%)	0	0
Outside lower outer fence	0	0	0	0	58 (11,46%)	0	0
Outside upper outer fence	0	0	0	0	0	0	16 (3.16%)

APPENDIX C


Pairs matrix
Boston$CRIM
Boston$ZN
Boston$INDUS
Boston$CHAS
Boston$NOX


Pairs matrix
Boston$CRIM
Boston$RAD
Boston$TAX
Boston$MV
Boston$RM
Boston$AGE

APPENDIX D

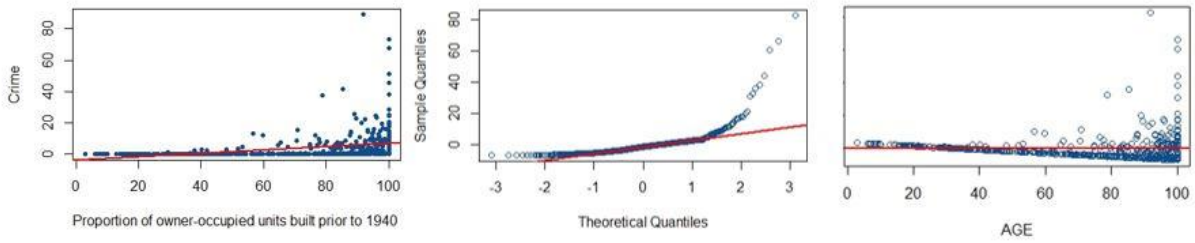
	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD
CRIM	1.0000000	-0.2004692	0.4065834	0.4209717	-0.2192467	0.3527343	-0.3796701	0.6255051
ZN	-0.2004692	1.0000000	-0.5338282	-0.5166037	0.3119906	-0.5695373	0.6644082	-0.3119478
INDUS	0.4065834	-0.5338282	1.0000000	0.7636515	-0.3916759	0.6447785	-0.7080270	0.5951293
NOX	0.4209717	-0.5166037	0.7636515	1.0000000	-0.3021882	0.7314701	-0.7692301	0.6114406
RM	-0.2192467	0.3119906	-0.3916759	-0.3021882	1.0000000	-0.2402649	0.2052462	-0.2098467
AGE	0.3527343	-0.5695373	0.6447785	0.7314701	-0.2402649	1.0000000	-0.7478805	0.4560225
DIS	-0.3796701	0.6644082	-0.7080270	-0.7692301	0.2052462	-0.7478805	1.0000000	-0.4945879
RAD	0.6255051	-0.3119478	0.5951293	0.6114406	-0.2098467	0.4560225	-0.4945879	1.0000000
TAX	0.5827643	-0.3145633	0.7207602	0.6680232	-0.2920478	0.5064556	-0.5344316	0.9102282
PT	0.2899456	-0.3916785	0.3832476	0.1889327	-0.3555015	0.2615150	-0.2324706	0.4647413
B	-0.3850640	0.1755203	-0.3569765	-0.3800506	0.1280686	-0.2735340	0.2915117	-0.4444128
LSTAT	0.4556215	-0.4129946	0.6037997	0.5908789	-0.6138083	0.6023385	-0.4969958	0.4886763
MV	-0.3883046	0.3604453	-0.4837252	-0.4273208	0.6953599	-0.3769546	0.2499287	-0.3816262

	TAX	PT	B	LSTAT	MV
CRIM	0.5827643	0.2899456	-0.3850640	0.4556215	-0.3883046
ZN	-0.3145633	-0.3916785	0.1755203	-0.4129946	0.3604453
INDUS	0.7207602	0.3832476	-0.3569765	0.6037997	-0.4837252
NOX	0.6680232	0.1889327	-0.3800506	0.5908789	-0.4273208
RM	-0.2920478	-0.3555015	0.1280686	-0.6138083	0.6953599
AGE	0.5064556	0.2615150	-0.2735340	0.6023385	-0.3769546
DIS	-0.5344316	-0.2324706	0.2915117	-0.4969958	0.2499287
RAD	0.9102282	0.4647413	-0.4444128	0.4886763	-0.3816262
TAX	1.0000000	0.4608531	-0.4418080	0.5439934	-0.4685359
PT	0.4608531	1.0000000	-0.1773833	0.3740444	-0.5077867
B	-0.4418080	-0.1773833	1.0000000	-0.3660869	0.3334608
LSTAT	0.5439934	0.3740444	-0.3660869	1.0000000	-0.7376627
MV	-0.4685359	-0.5077867	0.3334608	-0.7376627	1.0000000

APPENDIX E

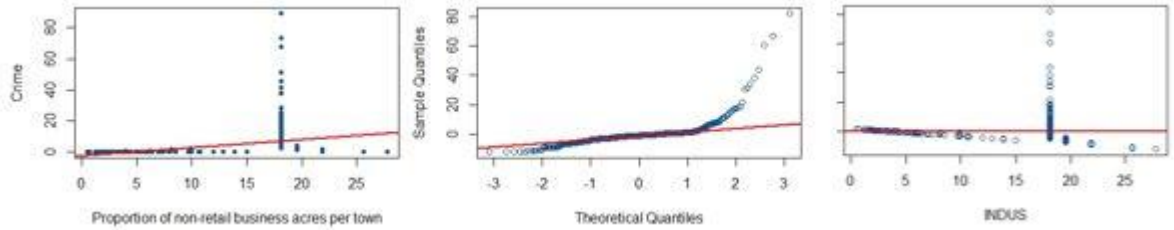
The variable **AGE**

1) Regression Normal Q-Q Plot Residuals



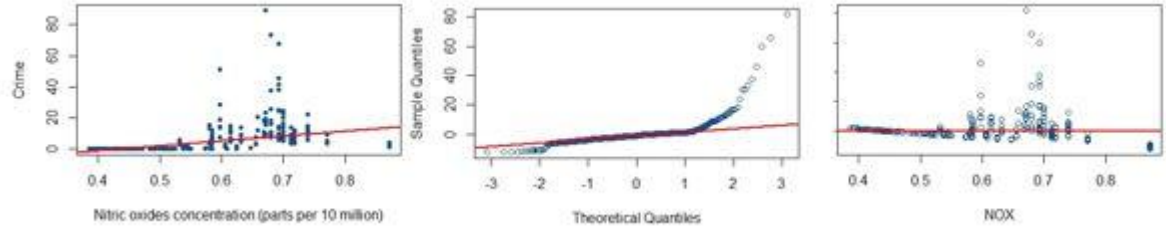
The variable **INDUS**

2) Regression Normal Q-Q Plot Residuals



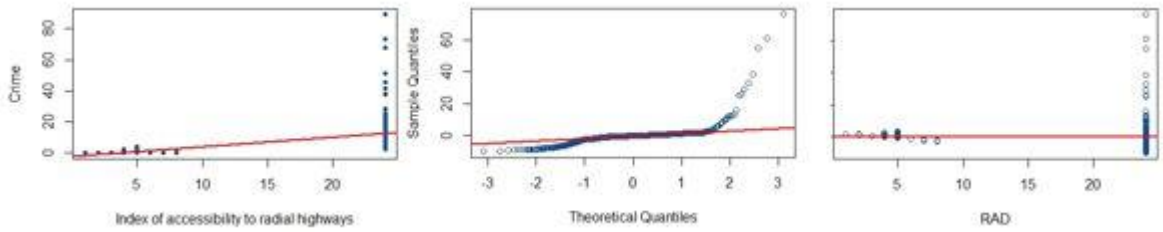
The variable **NOX**

3) Regression Normal Q-Q Plot Residuals



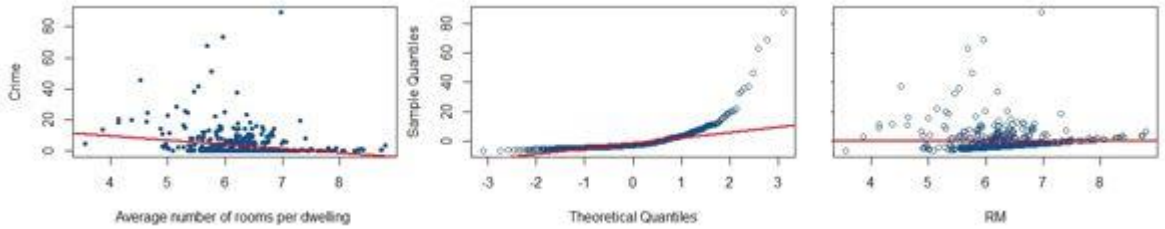
The variable **RAD**

4) Regression Normal Q-Q Plot Residuals



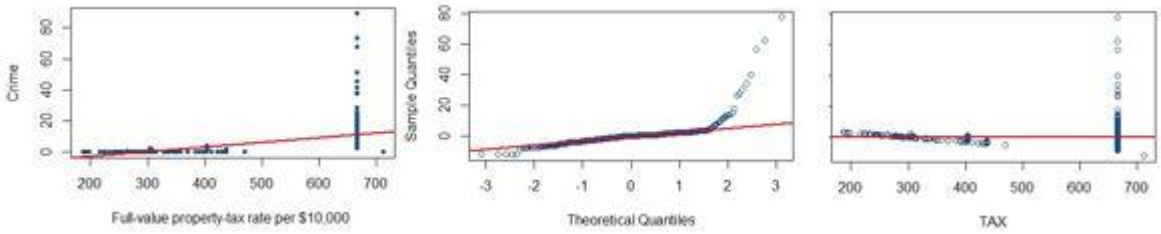
The variable **RM**

5) Regression Normal Q-Q Plot Residuals



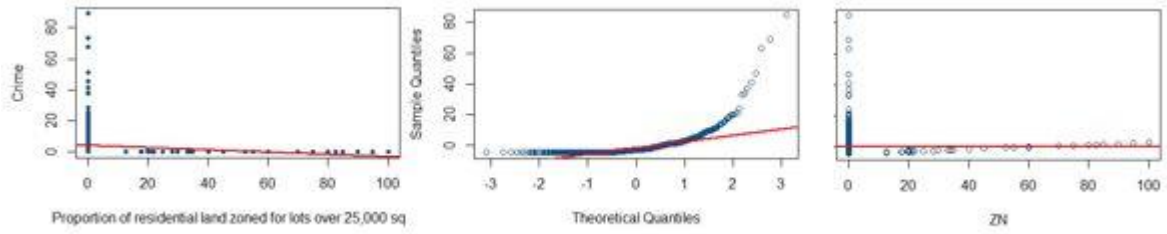
The variable **TAX**

6) Regression Normal Q-Q Plot Residuals



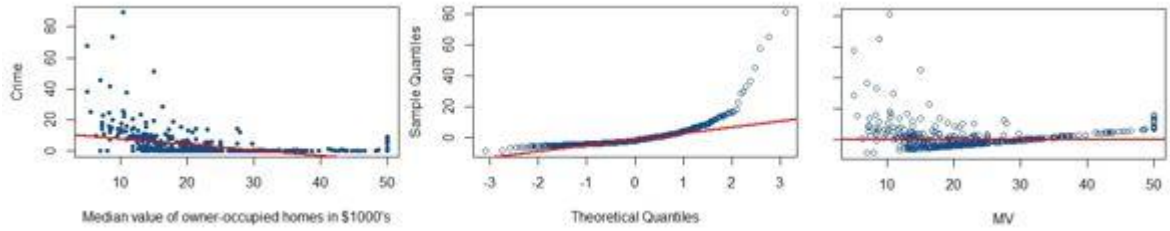
The variable **ZN**

7) Regression Normal Q-Q Plot Residuals



The variable **MV**

8) Regression Normal Q-Q Plot Residuals



APPENDIX G

Dependent variable	Independent variable	P-value	Adjusted R^2	Intercept (β_0)	Slope (β_1)
CRIM	*ZN*	$5.51*10^{-6}$	0.03828	4.45369	-0.07393
	INDUS	$2*10^{-16}$	0.1637	-2.0637	0.5098
	CHAS	0.001146	0.209	3.744	-1.893
	NOX	$2*10^{-16}$	0.1756	-13.72	31.25
	RM	$6.35*10^{-7}$	0.04618	20.482	-2.684
	AGE	$2.855*10^{-16}$	0.1227	-3.7779	0.1078
	DIS	$2*10^{-16}$	0.1425	9.499	-1.551
	RAD	$2*10^{-16}$	0.39	-2.2872	0.6179
	TAX	$2*10^{-16}$	0.3383	-8.52837	0.02974
	PT	$2.94*10^{-11}$	0.08225	-17.647	1.152
	B	$2*10^{-16}$	0.1466	16.55353	-0.03628
	LSTAT	$2*10^{-16}$	0.206	-3.3305	0.5488
	MV	$2*10^{-16}$	0.1491	11.7965	-0.3632