Multiple Regression Analysis of Boston House Values

Project Final Report
(due May 03, 2016)

Group Members:
Changhui Xu (Computer Science Grad)
Haonan Guo (Actuarial Science Undergrad)
Jiayao Ji (Actuarial Science Undergrad)

Course Project for
Statistical Methods and Computing

Instructor: Professor Kate Cowles

****

University of Iowa

Contents

I: Introductions

As we know the house price is really important for our daily life. Everyone will take serious considerations before him/her purchase or sell a house. We want to have a formula to calculate the house values. Thus, we need different variables to determine the value of the house.

The purpose of this project is to analyze the Boston House values dataset. We use the dataset from StatLib library which is maintained at Carnegie Mellon University. This dataset provides us the information of housing in suburbs of Boston. We are going to find out what are the dominate factors for house values based on the 13 variables in the dataset and try to generalize a predictive model using multivariable linear regression. First, we use basic SAS univariate process to get the statistical summary and boxplot of each individual variable. So that we can obtain descriptive statistic that has information on the location, spread and range of each variable. Second, we use SAS multiple regression analysis to determine the model of predicting house value in Boston.

Data source is https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

In this dataset, there are 506 instances in total and no missing value. It has 1 dependent variable and 13 independent variables. The 13 variables focus on quality and quantity of many physical attributes of the property. Most of these variables are exactly what a typical home buyer would want to consider a lot about a candidate property, for example, the crime rate in the neighborhood, the rooms per dwelling, the distance to job place, pupil-teacher ratio and so on.

The data for all variables are numeric values. Therefore, it is not necessarily to convert data types. The SAS version used for this project is SAS 9.3, 64bit. The whole analysis requires no out of SAS preprocessing and post processing tools or procedures.

This report contains 5 sections. After introduction, we begin to show the procedures of data examinations, which show the distributions and other statistics of total 14 variables. Then in section 3, we show the multiple regression analysis of house value model. Section 4 is a brief discussion on our regression model and section 5 is the conclusions.

II: Dataset Examinations

Step 1. Input the dataset into SAS and Print.

```
/* title 'The Boston house-price data'; */

data boston;
   tract=_n_;
   input town $1-14
         crime biglots industry river nox rooms age distance
         highway tax ptratio black lowstat value;
   label
     crime   = 'Per capita crime rate'
     biglots = '% res. land zoned for lots>25,000 sq.ft.'
     industry= '% non-retail business acres per town'
     river   = 'Charles River dummy variable'
     nox     = 'Nitric oxides concentration'
     rooms   = 'Average number of rooms per dwelling'
     age     = '% owner-occupied units built prior to 1940'
     distance= 'Distance to 5 Boston employment centres'
     highway = 'Accessibility to radial highways'
     tax     = 'Property-tax rate per $10,000'
     ptratio = 'Pupil-teacher ratio by town'
     black   = 'Transformed proportion of blacks'
```

```
     lowstat = '% population of lower status'
     value   = 'Median value of owner-occupied homes'
     town    = 'Name of town' ;
datalines;
Nahant          .00632 18.0  2.31 0 .5380 6.575 65.2 4.0900  1 296 15.3 396.90  5.0 24.0
Swampscott      .02731  0.0  7.07 0 .4690 6.421 78.9 4.9671  2 242 17.8 396.90  9.1 21.6
Savin Hill      9.7242  0.0 18.10 0 .7400 6.406 97.2 2.0651 24 666 20.2 385.96 19.5 17.1
Savin Hill      5.6664  0.0 18.10 0 .7400 6.219  100 2.0048 24 666 20.2 395.69 16.6 18.4
Savin Hill      9.9665  0.0 18.10 0 .7400 6.485  100 1.9784 24 666 20.2 386.73 18.9 15.4
(------------------------------ more data lines here --------------------------------)
Dorchester      6.8012  0.0 18.10 0 .7130 6.081 84.4 2.7175 24 666 20.2 396.90 14.7 20.0
Dorchester      3.6931  0.0 18.10 0 .7130 6.376 88.4 2.5671 24 666 20.2 391.43 14.7 17.7
Hyde Park       5.6918  0.0 18.10 0 .5830 6.114 79.8 3.5459 24 666 20.2 392.68 15.0 19.1
Hyde Park       4.8357  0.0 18.10 0 .5830 5.905 53.2 3.1523 24 666 20.2 388.22 11.5 20.6
Chelsea         .15086  0.0 27.74 0 .6090 5.454 92.7 1.8209  4 711 20.1 395.09 18.1 15.2
Chelsea         .18337  0.0 27.74 0 .6090 5.414 98.3 1.7554  4 711 20.1 344.05 24.0  7.0
Revere          .26838  0.0  9.69 0 .5850 5.794 70.6 2.8927  6 391 19.2 396.90 14.1 18.3
Revere          .23912  0.0  9.69 0 .5850 6.019 65.3 2.4091  6 391 19.2 396.90 12.9 21.2
Revere          .17783  0.0  9.69 0 .5850 5.569 73.5 2.3999  6 391 19.2 395.77 15.1 17.5
(------------------------------ more data lines here --------------------------------)
Winthrop        .06076  0.0 11.93 0 .5730 6.976 91.0 2.1675  1 273 21.0 396.90  5.6 23.9
Winthrop        .10959  0.0 11.93 0 .5730 6.794 89.3 2.3889  1 273 21.0 393.45  6.5 22.0
Winthrop        .04741  0.0 11.93 0 .5730 6.030 80.8 2.5050  1 273 21.0 396.90  7.9 11.9
;

proc print data = boston ;
run;
```

The SAS System

Obs	tract	town	crime	biglots	industry	river	nox	rooms	age	distance	highway	tax	ptratio	black	lowstat	value
1	1	Nahant	0.0063	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	5.0	24.0
2	2	Swampscott	0.0273	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.1	21.6
3	3	Swampscott	0.0273	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.0	34.7
4	4	Marblehead	0.0324	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.9	33.4
5	5	Marblehead	0.0691	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.3	36.2
6	6	Marblehead	0.0299	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.2	28.7
7	7	Salem	0.0883	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.4	22.9
8	8	Salem	0.1446	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.2	27.1
9	9	Salem	0.2112	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.9	16.5
10	10	Salem	0.1700	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	11	Salem	0.2249	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.5	15.0

Step 2. Examine each variable in the dataset.

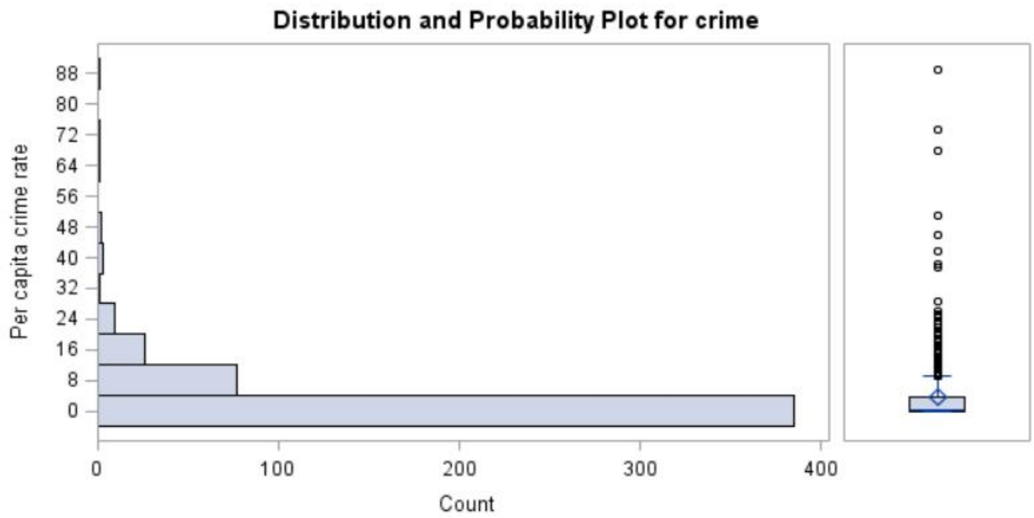
All variable use the proc univariate procedure to analyze

```
proc univariate plot data = boston;
var (variable);
run;
```

Variable#1: Crime (Per capita crime rate)

The “crime” variable shows per capita crime rate by town. Our sample gives 506 observations. As the distribution plots show, the overall shape of the “crime” data is skewed to the right, the range is max (88.976) minus the min (0.00632) which is 88.96968, and the spread of the crime variable is fairly large. Overall the mean of the dataset is 3.61352125, the median is 0.256510 and the 3rd quantile is 3.67820 which gives us a lot of higher outliers in the dataset. It means that some areas in Boston are dangerous because of high crime rates and people may try to avoid to living in those places.

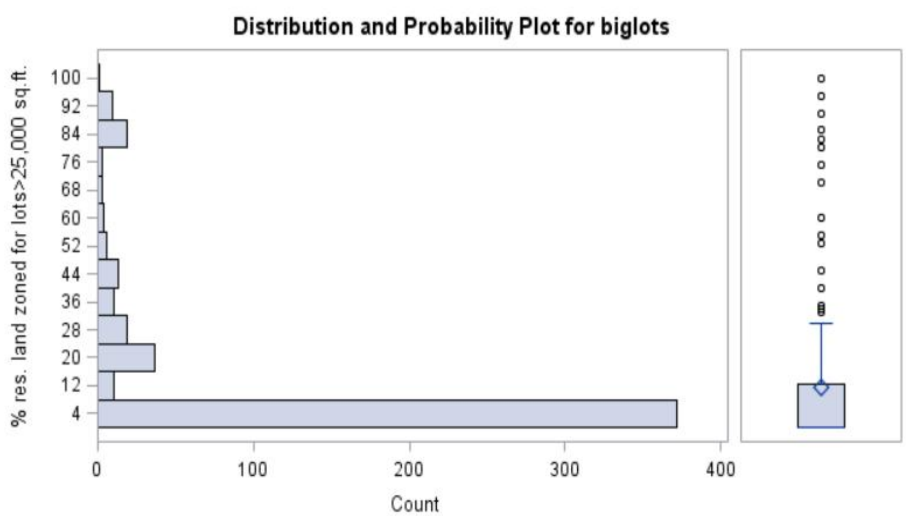
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	88.97600
99%	41.52900
95%	15.86000
90%	10.83400
75% Q3	3.67820
50% Median	0.25651
25% Q1	0.08199
10%	0.03768
5%	0.02763
1%	0.01360
0% Min	0.00632



Variable#2: biglots (% res. land zoned for lots>25,000 sq.ft.)

The "biglots" variable shows % residence land zoned for lots>25,000 square feet. With 506 observations, the shape of the dataset is skewed to the right. The range is max (100) minus the min (0) which is 100, the spread of the "biglots" data is fairly large. Overall the mean of the dataset is 11.3636364, the median is 0.000, and the 3rd quantile is 12.5. The "biglots" has lots of higher outliers, which may happened in rich areas.

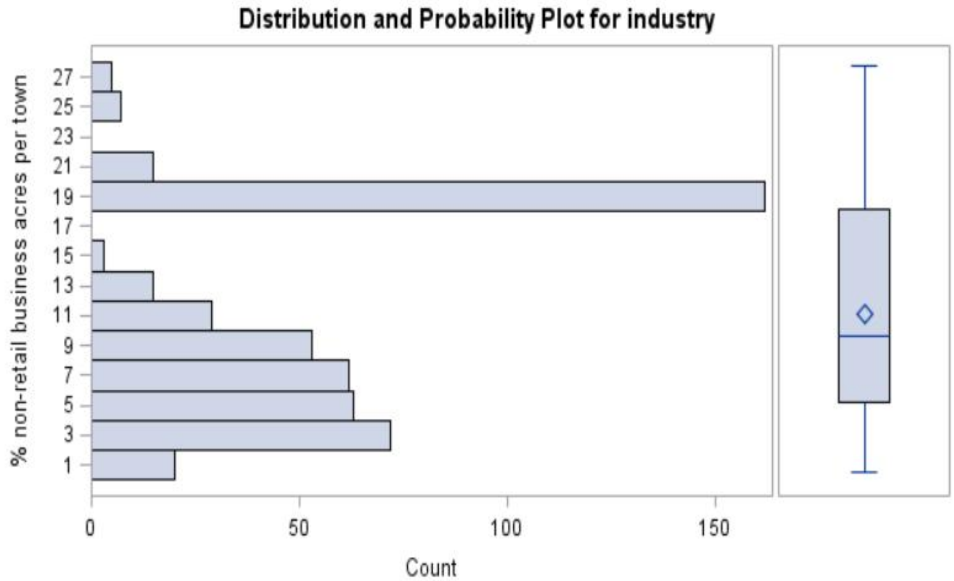
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	100.0
99%	90.0
95%	80.0
90%	45.0
75% Q3	12.5
50% Median	0.0
25% Q1	0.0
10%	0.0
5%	0.0
1%	0.0
0% Min	0.0



Variable#3: industry (% non-retail business acres per town)

The "industry" variable shows % non-retail business acres per town. With 506 observations, the shape of the data is skewed to the right. The range is max (27.74) minus the min (0.46) which is 27.28. Overall the mean of the data is 11.1367787, the median is 9.69, and the 3rd quantile is 18.10. From the boxplot, we observe that there is no outlier.

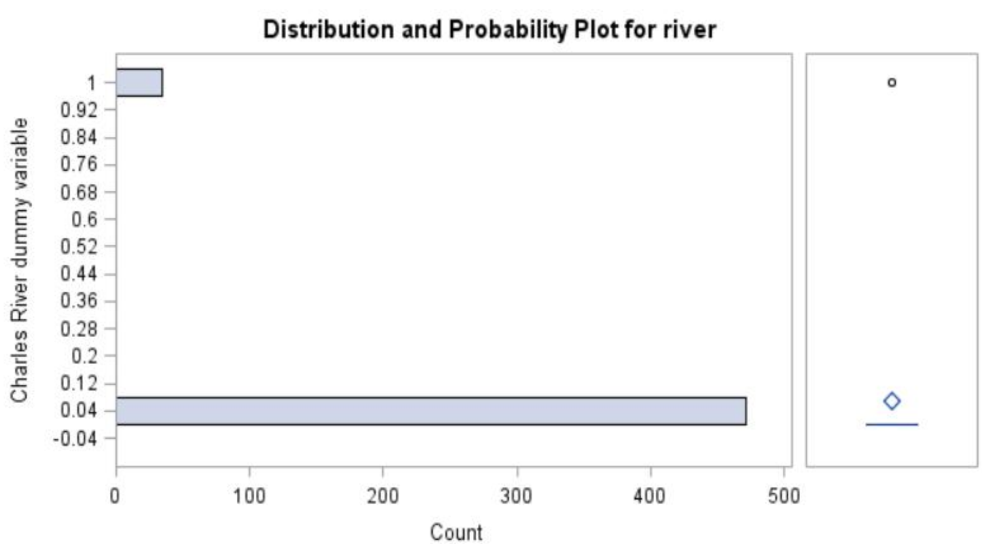
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	27.74
99%	25.65
95%	21.89
90%	19.58
75% Q3	18.10
50% Median	9.69
25% Q1	5.19
10%	2.89
5%	2.18
1%	1.25
0% Min	0.46



Variable#4: river (Charles River dummy variable)

The “river” variable is a Charles River dummy/ indicator variable. It has two values, “0” not nearby river and “1” close to Charles River. SAS shows that the distribution is 2 subgroups. We will not need to consider “river” variable when we do regression analysis.

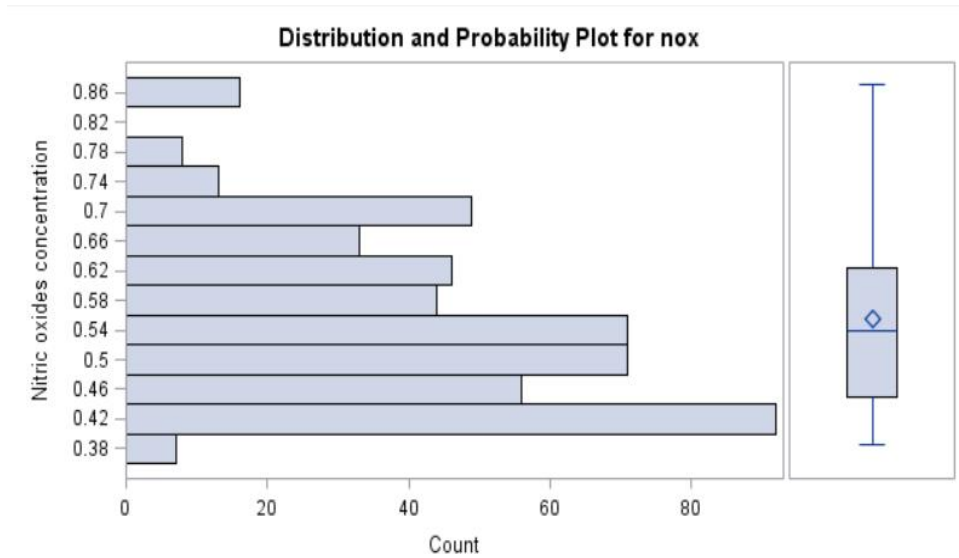
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1
99%	1
95%	1
90%	0
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0



Variable#5: nox (Nitric oxides concentration)

The “nox” variable shows Nitric oxides concentration by town. The shape of the data is slightly skewed to the right. The range is max (0.871) minus the min (0.385) which is 0.486. Overall the mean of the dataset is 0.55469506, the median is 0.538, and the 3rd quantile is 0.624. The boxplot shows that there are no outliers.

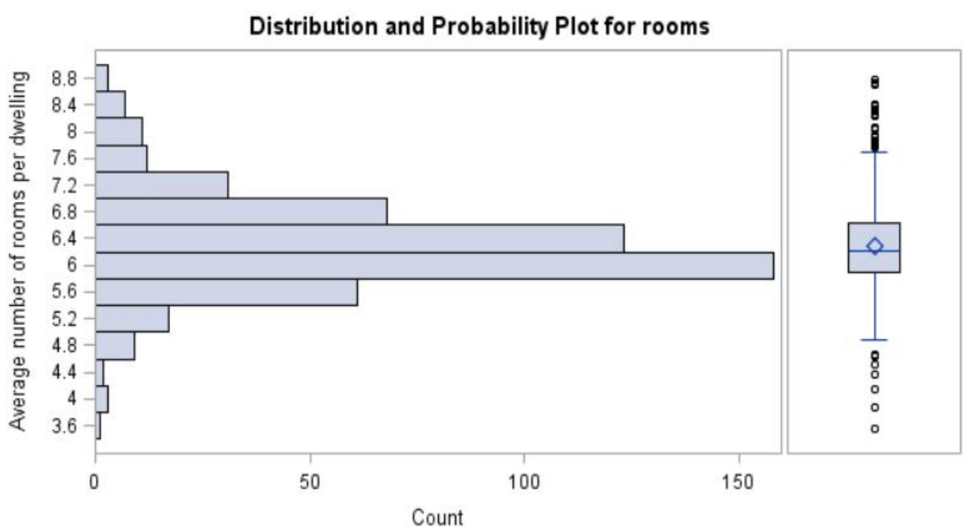
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.871
99%	0.871
95%	0.740
90%	0.713
75% Q3	0.624
50% Median	0.538
25% Q1	0.449
10%	0.426
5%	0.409
1%	0.398
0% Min	0.385



Variable#6: rooms (Average number of rooms per dwelling)

The “rooms” variable shows the average number of rooms per dwellings by town. The SAS output shows that the shape of the dataset is fairly symmetric. The range is max (8.78) minus the min (3.561) which is 5.219. Overall the mean of the dataset is 6.70261714, the median is 6.2085, and Q1 is 5.885, Q3 is 6.625. From the box plot, we can see that there are some outliers.

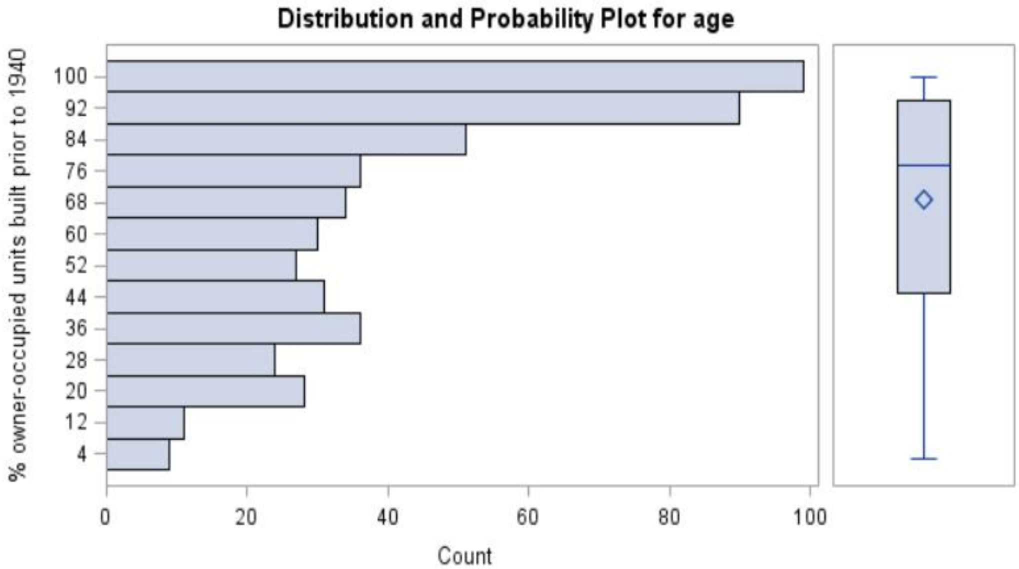
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	8.7800
99%	8.3370
95%	7.6100
90%	7.1550
75% Q3	6.6250
50% Median	6.2085
25% Q1	5.8850
10%	5.5930
5%	5.3040
1%	4.5190
0% Min	3.5610



Variable#7: age (% owner-occupied units built prior to 1940)

The “age” variable shows the % owner-occupied units built prior to 1940 by town. The shape of the dataset is skewed to the left, which means new houses are relatively less than older houses. The range is max (100) minus the min (2.9) which is 97.1; the spread of the “age” variable is large. Overall the mean of the dataset is 28.1488614, the median is 77.5, and the 1st quantile is 45, 3rd quantile is 94.1. There are no outliers.

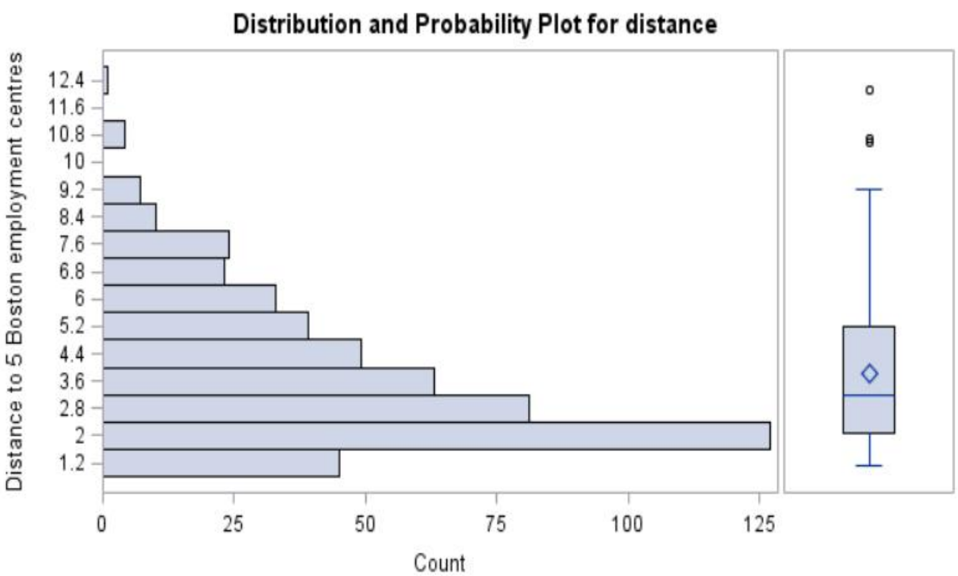
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	100.0
99%	100.0
95%	100.0
90%	98.8
75% Q3	94.1
50% Median	77.5
25% Q1	45.0
10%	26.3
5%	17.7
1%	6.6
0% Min	2.9



Variable#8: distance (Distance to 5 Boston employment centres)

The "distance" variable shows distance to 5 Boston employment centers by town. The overall shape of the data is 1.01179514 skew to the right, the range is max (12.127) minus the min (1.1296) which is 10.9974. Overall the mean of the data is 3.79504368, the median is 3.20745. There are small amount of outliers for the variable "distance", which are houses located in very remote areas.

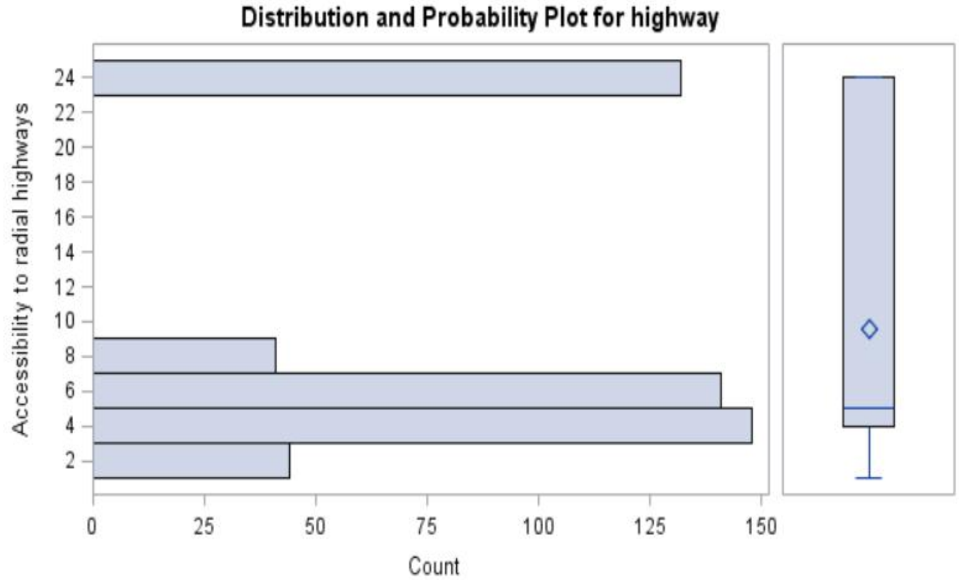
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	12.12700
99%	9.22290
95%	7.82780
90%	6.81850
75% Q3	5.21190
50% Median	3.20745
25% Q1	2.10000
10%	1.62320
5%	1.46080
1%	1.20240
0% Min	1.12960



Variable#9: highway (Accessibility to radial highways)

The "highway" variable shows accessibility to radial highways by town. The overall shape of the data is bimodal, the range is max (24) minus the min (1) which is 23. Overall the mean of the dataset is 9.54940711, the median is 5.

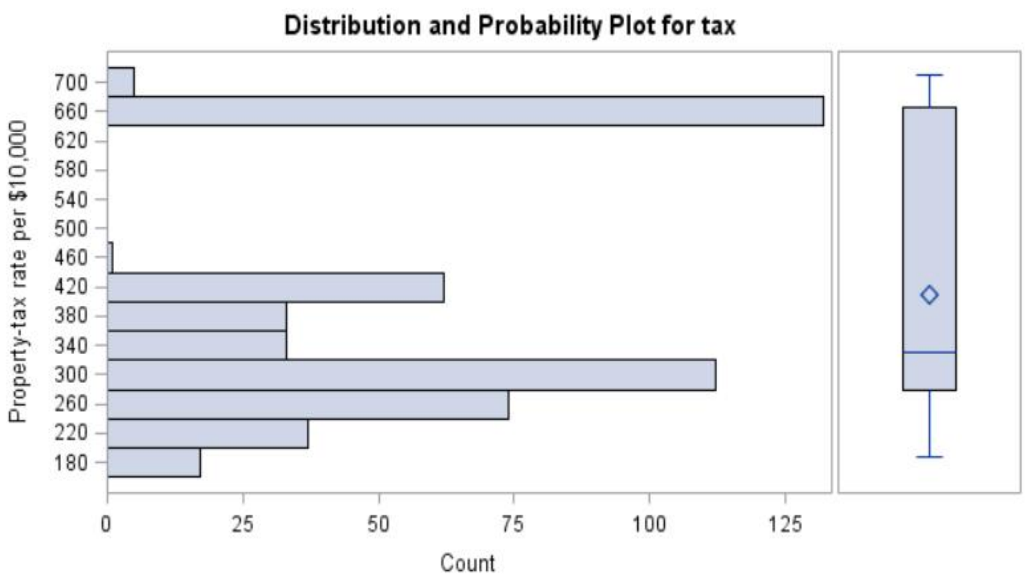
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	24
99%	24
95%	24
90%	24
75% Q3	24
50% Median	5
25% Q1	4
10%	3
5%	2
1%	1
0% Min	1



Variable#10: tax (Property-tax rate per $10,000)

The “tax” variable shows property-tax rate per $10,000 by town. The overall shape of the data is bimodal, the range is max (711) minus the min (187) which is 524, and the spread of the tax variable is quite large. Overall the mean of the “tax” data is 408.237154, the median is 330. There are no outlier for the variable tax.

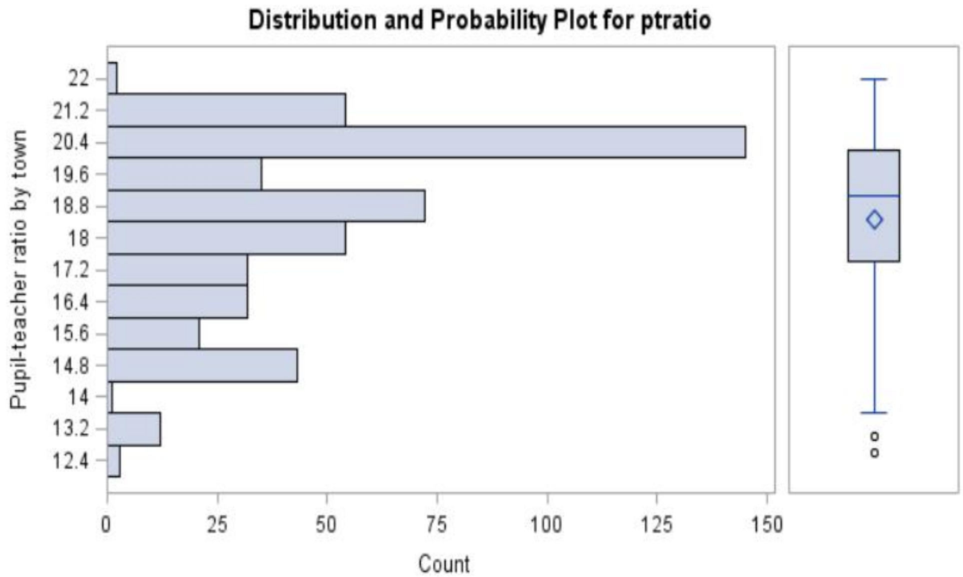
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	711
99%	666
95%	666
90%	666
75% Q3	666
50% Median	330
25% Q1	279
10%	233
5%	222
1%	188
0% Min	187



Variable#11: ptratio (Pupil-teacher ratio by town)

The “ptratio” variable shows pupil-teacher ratio by town. The overall shape of the data is skewed to the left, the range is max (22) minus the min (12.6) which is 9.4. Overall the mean of the “ptratio” data is 18.4555336, the median is 19.05. The box plot shows there are some lower outlier for the variable “ptratio”.

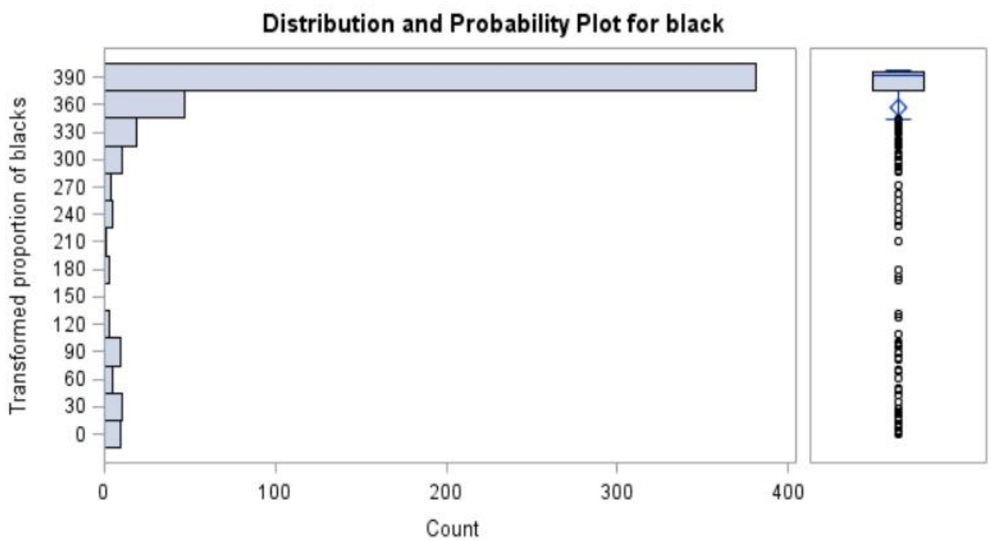
Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	22.00
99%	21.20
95%	21.00
90%	20.90
75% Q3	20.20
50% Median	19.05
25% Q1	17.40
10%	14.70
5%	14.70
1%	13.00
0% Min	12.60



Variable#12: black (Transformed proportion of blacks)

The “black” variable shows transformed proportion of blacks by town. The overall shape of the data is skewed to the left, the range is max (396.9) minus the min (0.32) which is 396.58, and the spread of the “black” variable is very large. Overall the mean is 356.674032, the median is 391.44 and the 1st quantile is 375.33 which indicate that there are lots of lower outlier for the variable “black”.

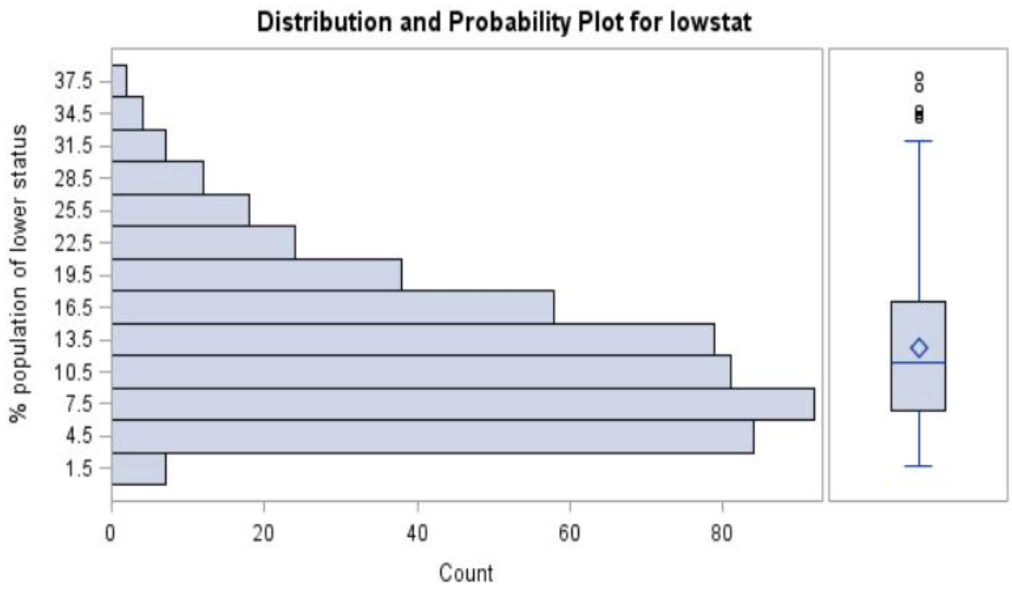
Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	396.90
99%	396.90
95%	396.90
90%	396.90
75% Q3	396.23
50% Median	391.44
25% Q1	375.33
10%	288.99
5%	83.45
1%	6.68
0% Min	0.32



Variable#13: lowstat (% population of lower status)

The “lowstat” variable shows the percent of population of lower status by town. The overall shape of the data is slightly skew to the right. There are some higher outliers for this variable.

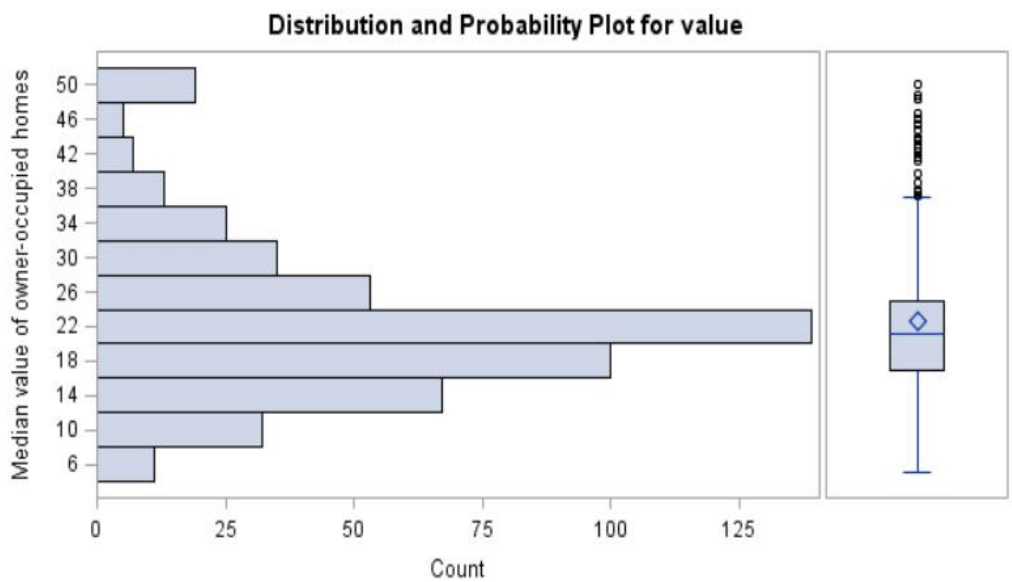
Quantiles (Definition 5)	
Quantile	Estimate
100% Max	38.00
99%	34.00
95%	26.80
90%	23.10
75% Q3	17.00
50% Median	11.35
25% Q1	6.90
10%	4.70
5%	3.70
1%	2.90
0% Min	1.70



Variable#14: value (Median value of owner-occupied homes)

The "value" variable shows the Median value of owner-occupied homes by town. The overall shape of the data is slightly skew to the right, the range is max (50) minus the min (5) which is 45. Overall the mean of the dataset is 22.5328063, the median is 21.2 and the 3rd quantile is 25 which indicate that there are quite a few higher outliers for the variable "value".

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	50.0
99%	50.0
95%	43.5
90%	34.9
75% Q3	25.0
50% Median	21.2
25% Q1	17.0
10%	12.7
5%	10.2
1%	7.0
0% Min	5.0



In conclusion, no variable distribution is absolutely normal. Most of the 14 variables are slightly skewed and have outliers. It is common to apply log or square root transformations to non-normal-distribution variables in regression analysis. Since the skewness effects may canceled out between dependent variable and independent variables, we will leave them as they are first.

Step 3. Examine the independence of variables in the dataset.

The regression analysis may give spurious results if the variables are not strongly independent. When variables are highly correlated in the regression model, we may get contradictive results from *t*-test and *F*-test and may get estimated model parameters which could have opposite signs from what are expected.

So, we have to examine the coefficient of correlation between each pair of numeric variables in the dataset. If one or more correlation coefficients are close to 1 or -1, then these variables are highly correlated, which would result in a severe multicollinearity problem. If that situation happens, then we need to remove one of the correlated variables in our prediction model.

```
proc corr data = boston ;
var crime biglots industry river nox rooms age distance highway tax ptratio
black lowstat value ;
run ;
```

Pearson Correlation Coefficients, N = 506
Prob > |r| under H0: Rho=0

	crime	biglots	industry	river	nox	rooms	age	distance	highway	tax	ptratio	black	lowstat	value
crime	1.00000	-0.20047	0.40658	-0.05589	0.42097	-0.21925	0.35273	-0.37967	0.62550	0.58276	0.28995	-0.38506	0.45545	-0.38830
Per capita crime rate		<.0001	<.0001	0.2094	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
biglots	-0.20047	1.00000	-0.53383	-0.04270	-0.51660	0.31199	-0.56954	0.66441	-0.31195	-0.31456	-0.39168	0.17552	-0.41259	0.36045
% res. land zoned for lots>25,000 sq. ft.	<.0001		<.0001	0.3378	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
industry	0.40658	-0.53383	1.00000	0.06294	0.76365	-0.39168	0.64478	-0.70803	0.59513	0.72076	0.38325	-0.35698	0.60354	-0.48373
% non-retail business acres per town	<.0001	<.0001		0.1575	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
river	-0.05589	-0.04270	0.06294	1.00000	0.09120	0.09125	0.08652	-0.09918	-0.00737	-0.03559	-0.12152	0.04879	-0.05392	0.17526
Charles River dummy variable	0.2094	0.3378	0.1575		0.0403	0.0402	0.0518	0.0257	0.8687	0.4244	0.0062	0.2733	0.2260	<.0001
nox	0.42097	-0.51660	0.76365	0.09120	1.00000	-0.30219	0.73147	-0.76923	0.61144	0.66802	0.18893	-0.38005	0.59064	-0.42732
Nitric oxides concentration	<.0001	<.0001	<.0001	0.0403		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
rooms	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1.00000	-0.24026	0.20525	-0.20985	-0.29205	-0.35550	0.12807	-0.61377	0.69536
Average number of rooms per dwelling	<.0001	<.0001	<.0001	0.0402	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0039	<.0001	<.0001
age	0.35273	-0.56954	0.64478	0.08652	0.73147	-0.24026	1.00000	-0.74788	0.45602	0.50646	0.26152	-0.27353	0.60210	-0.37695
% owner-occupied units built prior to 1940	<.0001	<.0001	<.0001	0.0518	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
distance	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1.00000	-0.49459	-0.53443	-0.23247	0.29151	-0.49668	0.24993
Distance to 5 Boston employment centres	<.0001	<.0001	<.0001	0.0257	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
highway	0.62550	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1.00000	0.91023	0.46474	-0.44441	0.48848	-0.38163
Accessibility to radial highways	<.0001	<.0001	<.0001	0.8687	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
tax	0.58276	-0.31456	0.72076	-0.03559	0.66802	-0.29205	0.50646	-0.53443	0.91023	1.00000	0.46085	-0.44181	0.54385	-0.46854
Property-tax rate per $10,000	<.0001	<.0001	<.0001	0.4244	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
ptratio	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.35550	0.26152	-0.23247	0.46474	0.46085	1.00000	-0.17738	0.37406	-0.50779
Pupil-teacher ratio by town	<.0001	<.0001	<.0001	0.0062	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
black	-0.38506	0.17552	-0.35698	0.04879	-0.38005	0.12807	-0.27353	0.29151	-0.44441	-0.44181	-0.17738	1.00000	-0.36579	0.33346
Transformed proportion of blacks	<.0001	<.0001	<.0001	0.2733	<.0001	0.0039	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
lowstat	0.45545	-0.41259	0.60354	-0.05392	0.59064	-0.61377	0.60210	-0.49668	0.48848	0.54385	0.37406	-0.36579	1.00000	-0.73750
% population of lower status	<.0001	<.0001	<.0001	0.2260	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
value	-0.38830	0.36045	-0.48373	0.17526	-0.42732	0.69536	-0.37695	0.24993	-0.38163	-0.46854	-0.50779	0.33346	-0.73750	1.00000
Median value of owner-occupied homes	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

As shown in the result, the highest correlation coefficient is 0.91 between “tax” and “highway”. Also, some other high correlation coefficients are -0.76 between “industry” and “nox”, 0.72 between “industry” and “tax”, -0.71 between “industry” and “distance” and -0.77 between “nox” and “distance”. It makes sense that “tax” and “highway” has high positive correlation relationship because houses which are closer to highway pay more tax, as well as high negative correlation relationship between “nox” and “industry”/

"distance". These are possible sources of multicollinearity. Each pair of variables explains the same thing as far as how they affect variation in "value".

As to the "value" itself, the "rooms" has the highest positive correlation (about 0.7), while "ptratio" and "lowstat" have the highest negative correlations. It is understandable that these three variables are dominant variables to the house value.

III: Multiple Regression Analysis

Using the dataset "boston.SAS", our objective is to build a multiple regression model to predict the value of a house. The "value" (y) is modeled as a function of "crime", "biglots", "industry", "river", "nox", "rooms", "age", "distance", "highway", "tax", "ptratio", "black" and "lowstat".

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Where,

Response variable (y) = "value"

Independent variables (x_i) = "crime", "biglots", "industry", "river", "nox", "rooms", "age", "distance", "highway", "tax", "ptratio", "black" and "lowstat".

R-Squared (R^2) value represents the fraction of the sample variation of the y values that is explained by the particular variable(s). However, there is one drawback of R^2 that the model will eventually have a R^2 close to 1 when more and more variables are added to the model.

On the other hand, Adjusted R^2 takes into account of the sample size and the number of β parameters in the model. As we know, R^2_{adj} is closely related to Mean Square Error (MSE). As R^2_{adj} increases, MSE decreases. The largest R^2_{adj} (or smallest MSE) indicates the best fit of the model.

In addition, Mallows's C_p value is another good criterion to check the goodness of the regression model. A small value of C_p indicates that the total MSE and the regression bias are minimized.

Yet another indicator of the goodness of the regression model is PRESS criterion. A small PRESS (small differences of $y_i - \hat{y}_i$) or Residual value indicates the model has a well predictive ability.

Step 1. Model Building by Stepwise Regression Analysis.

First, we run a Stepwise Regression analysis. Stepwise Regression determines the independent variables added to the model at each step using *t*-test. SAS will give out Partial R-Square and P-value for each variable added to the model. Because "river" is a dummy variable, as described in section 1, we remove it from out regression model.

```
proc reg data = boston ;
model value = crime biglots industry nox rooms age distance highway tax
ptratio black lowstat / selection = stepwise ;
```

```
run ;
```

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	lowstat		% population of lower status	1	0.5439	0.5439	347.755	601.04	<.0001
2	rooms		Average number of rooms per dwelling	2	0.0945	0.6384	173.679	131.47	<.0001
3	ptratio		Pupil-teacher ratio by town	3	0.0401	0.6785	101.016	62.57	<.0001
4	distance		Distance to 5 Boston employment centres	4	0.0116	0.6901	81.3684	18.78	<.0001
5	nox		Nitric oxides concentration	5	0.0178	0.7079	50.1771	30.50	<.0001
6	black		Transformed proportion of blacks	6	0.0073	0.7153	38.5195	12.85	0.0004
7	biglots		% res. land zoned for lots>25,000 sq.ft.	7	0.0042	0.7195	32.6918	7.46	0.0065
8	crime		Per capita crime rate	8	0.0026	0.7221	29.8420	4.65	0.0314
9	highway		Accessibility to radial highways	9	0.0056	0.7277	21.3741	10.23	0.0015
10	tax		Property-tax rate per $10,000	10	0.0075	0.7351	9.4652	13.95	0.0002

The above results show R^2 value increases as we add variables to the model. Note that the final R^2 is 0.7351. Also note that 2 variables ("industry" and "age") are removed from the selection list. SAS only keep variables in the model if they are significant at the 0.1500 level.

As mentioned above, we want to find a model that has high R^2, high R^2_{adj}, low MSE, low C_p and low PRESS. So we examined these criteria using SAS.

```
proc rsquare cp adjrsq mse jp data = boston ;
model value = crime biglots industry nox rooms age distance highway tax
ptratio black lowstat ;
run;
```

Number in Model	R-Square	Adjusted R-Square	C(p)	J(p)	MSE	Variables in Model
1	0.5439	0.5430	347.7546	38.8087	38.65590	lowstat
1	0.4835	0.4825	460.2539	43.9466	43.77357	rooms
1	0.2578	0.2564	880.7193	63.1494	62.90082	ptratio
1	0.2340	0.2325	925.1683	65.1794	64.92283	industry
1	0.2195	0.2180	952.1167	66.4102	66.14873	tax
9	0.7221	0.7170	31.8295	24.4094	23.93635	crime biglots nox rooms distance tax ptratio black lowstat
10	0.7351	0.7298	9.4652	23.3531	22.85621	crime biglots nox rooms distance highway tax ptratio black lowstat
10	0.7292	0.7237	20.6139	23.8807	23.37259	crime industry nox rooms distance highway tax ptratio black lowstat
10	0.7291	0.7237	20.6568	23.8827	23.37457	crime nox rooms age distance highway tax ptratio black lowstat
11	0.6768	0.6696	120.1034	28.6073	27.94454	crime biglots industry nox rooms age distance highway tax ptratio black
12	0.7354	0.7289	13.0000	23.5163	22.92729	crime biglots industry nox rooms age distance highway tax ptratio black lowstat

Based on the result, as shown in the above cropped screenshot, we find that 10 variables model with "crime", "biglots", "nox", "rooms", "distance", "highway", "tax", "ptratio", "black", and "lowstat". This model gives R^2 of 0.7351, R^2_{adj} of 0.7298, Cp of 9.4652 and MSE of 22.85621, which is the optimal model among the 133 possible models.

As we have seen in the section 1.2, the distribution of the dependent variable "value" is right skewed. So, we would have to examine the residuals carefully, which will be described in Step 3 of this section. For now, we use the model without log or square-root transformation of variables.

Based on the stepwise regression analysis, we keep 10 independent variables in our model.

Step 2. Model Adequacy.

We need to examine several other parameters of the regression.

(1) *F* test. We need to check the P-value with respect to the *F* test. We define out level of significance as 0.05. Thus, if Pr>F has value of "<0.05", then that variable is statistically significant in this model.

(2) Confidence interval and *t*-test. These tell us the inferences about the β parameters.

```
proc glm data = boston ;
model value = crime biglots nox rooms distance highway tax ptratio black
lowstat / solution clparm ;
run ;
```

Results are shown below.

The GLM Procedure

Dependent Variable: value Median value of owner-occupied homes

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	31402.47333	3140.24733	137.39	<.0001
Error	495	11313.82209	22.85621		
Corrected Total	505	42716.29542			

R-Square	Coeff Var	Root MSE	value Mean
0.735140	21.21714	4.780816	22.53281

Source	DF	Type I SS	Mean Square	F Value	Pr > F
crime	1	6440.77288	6440.77288	281.80	<.0001
biglots	1	3554.33840	3554.33840	155.51	<.0001
nox	1	1268.85577	1268.85577	55.51	<.0001
rooms	1	12949.68810	12949.68810	566.57	<.0001
distance	1	1208.34595	1208.34595	52.87	<.0001
highway	1	114.71893	114.71893	5.02	0.0255
tax	1	796.50266	796.50266	34.85	<.0001
ptratio	1	1647.81234	1647.81234	72.09	<.0001
black	1	641.26094	641.26094	28.06	<.0001
lowstat	1	2780.17735	2780.17735	121.64	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
crime	1	272.843944	272.843944	11.94	0.0006
biglots	1	257.519613	257.519613	11.27	0.0008
nox	1	490.396564	490.396564	21.46	<.0001
rooms	1	2014.205302	2014.205302	88.13	<.0001
distance	1	1518.678718	1518.678718	66.44	<.0001
highway	1	558.606646	558.606646	24.44	<.0001
tax	1	318.894296	318.894296	13.95	0.0002
ptratio	1	1296.131243	1296.131243	56.71	<.0001
black	1	298.835865	298.835865	13.07	0.0003
lowstat	1	2780.177347	2780.177347	121.64	<.0001

By inspecting the results, we find that all variables are statistically significant in the *F* tests. For now, all variables have passed the *F* tests and *t*-tests.

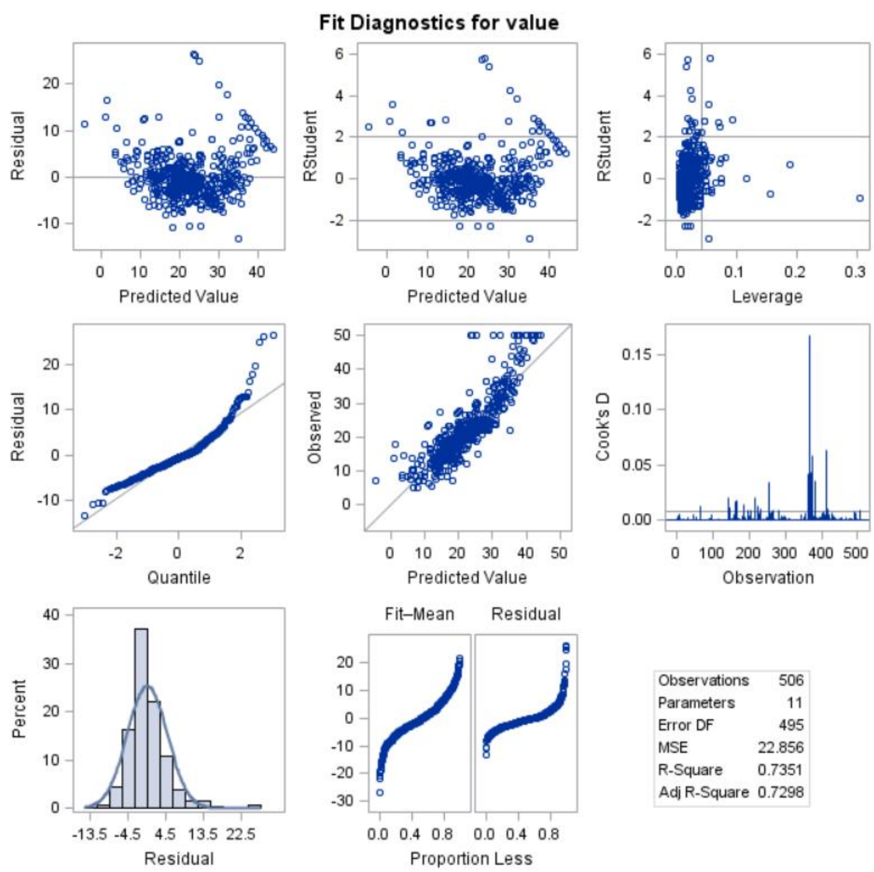
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	36.58437526	5.11388829	7.15	<.0001	26.53677123	46.63197929
crime	-0.11415148	0.03303897	-3.46	0.0006	-0.17906539	-0.04923758
biglots	0.04581788	0.01364998	3.36	0.0008	0.01899884	0.07263692
nox	-16.47576472	3.55691875	-4.63	<.0001	-23.46428484	-9.48724460
rooms	3.84768655	0.40987355	9.39	<.0001	3.04238012	4.65299297
distance	-1.52575958	0.18717818	-8.15	<.0001	-1.89352128	-1.15799788
highway	0.31540505	0.06379956	4.94	<.0001	0.19005372	0.44075638
tax	-0.01267077	0.00339220	-3.74	0.0002	-0.01933567	-0.00600588
ptratio	-0.97813520	0.12989023	-7.53	<.0001	-1.23333936	-0.72293104
black	0.00974596	0.00269532	3.62	0.0003	0.00445028	0.01504164
lowstat	-0.52738483	0.04781823	-11.03	<.0001	-0.62133655	-0.43343311

Step 3. Model Assumptions.

We need to have a regression model that has (1) random error $\varepsilon \sim N(0,\ \sigma^2)$ and (2) all pairs of random errors are independent.

SAS can do residual tests to detect violations in regression modeling assumptions. We mainly check the Residuals and Partial Residuals Plots. If there are any trends or patterns in these plots, we can conclude that the model is lack of fit and has potential problems.

```
proc reg data = boston ;
model value = crime biglots nox rooms distance highway tax ptratio black
lowstat ;
run ;
```



As can be seen from above plots, no obvious trends or patterns can be found in the Residual plots, neither any significant outliers can be found. So our model is safe for our assumptions.

Step 4. Potential Modeling Problems and Solutions.

1. Check the multicollinearity.

For this project, we mainly check the multicollinearity problem. SAS can check the Variance Inflation Factors (VIF) for the β's. As a rule of thumb, if VIF is greater than 10, then a severe multicollinearity problem exists in the model.

```
proc reg data = boston ;
model value = crime biglots nox rooms distance highway tax ptratio black lowstat /
VIF ;
run ;
```

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	36.58438	5.11389	7.15	<.0001	0
crime	Per capita crime rate	1	-0.11415	0.03304	-3.46	0.0006	1.78440
biglots	% res. land zoned for lots>25,000 sq.ft.	1	0.04582	0.01365	3.36	0.0008	2.23923
nox	Nitric oxides concentration	1	-16.47576	3.55692	-4.63	<.0001	3.75348
rooms	Average number of rooms per dwelling	1	3.84769	0.40987	9.39	<.0001	1.83242
distance	Distance to 5 Boston employment centres	1	-1.52576	0.18718	-8.15	<.0001	3.43239
highway	Accessibility to radial highways	1	0.31541	0.06380	4.94	<.0001	6.81845
tax	Property-tax rate per $10,000	1	-0.01267	0.00339	-3.74	0.0002	7.22174
ptratio	Pupil-teacher ratio by town	1	-0.97814	0.12989	-7.53	<.0001	1.74717
black	Transformed proportion of blacks	1	0.00975	0.00270	3.62	0.0003	1.33783
lowstat	% population of lower status	1	-0.52738	0.04782	-11.03	<.0001	2.57636

In the above result table, we find that all the VIFs are less than 10. So, the model is safe for this criterion.

2. Model Extrapolation.

As shown above in Part II Step 2, we have the result of 95% confidence interval of all the β's for all independent variables, as well as the 95% confidence interval for the interception.
We need to be very cautious that we can only predict house values using data within the min to max range of each variable.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	36.58437526	5.11388829	7.15	<.0001	26.53677123	46.63197929
crime	-0.11415148	0.03303897	-3.46	0.0006	-0.17906539	-0.04923758
biglots	0.04581788	0.01364998	3.36	0.0008	0.01899884	0.07263692
nox	-16.47576472	3.55691875	-4.63	<.0001	-23.46428484	-9.48724460
rooms	3.84768655	0.40987355	9.39	<.0001	3.04238012	4.65299297
distance	-1.52575958	0.18717818	-8.15	<.0001	-1.89352128	-1.15799788
highway	0.31540505	0.06379956	4.94	<.0001	0.19005372	0.44075638
tax	-0.01267077	0.00339220	-3.74	0.0002	-0.01933567	-0.00600588
ptratio	-0.97813520	0.12989023	-7.53	<.0001	-1.23333936	-0.72293104
black	0.00974596	0.00269532	3.62	0.0003	0.00445028	0.01504164
lowstat	-0.52738483	0.04781823	-11.03	<.0001	-0.62133655	-0.43343311

Step 5. Model Representation.

For this project, we have examined all variables in the dataset and checked several regression criteria. We finally come to our result model.

$$\begin{aligned} value = & -0.114 * crime + 0.046 * biglots - 16.476 * nox + 3.848 * rooms \\ & -1.526 * distance + 0.315 * highway - 0.013 * tax \\ & -0.0370 * ptratio + 0.0004 * black - 0.0289 * lowstat + 36.58 \end{aligned}$$

The model is valid when each independent variable is in its data range.

IV: Discussions

The regression model has a R^2 of 0.7351, which means that 73.5% of "house value" variance can be explained by this model. Based on the model, we know that house value is higher in low crime rate, low NO level, short distance to CBD, low tax level, low pupil-teacher ratio and less lower-status population area. The house value is higher for property has big lots and more rooms, for property that are away from highway. The β for "black" variable is 0.0004 and "black" has mean value of 356, so the weight of "black" is very small.

Due to different metrics for different variables, we cannot determine which variables are more important to the house value based on the regression coefficients. However, we can infer some relationship from the results of the correlation analysis as described in section 2.3.

The correlation analysis shows that there are several pairs of variables are highly correlated, which are potential sources of multicollinearity. After check the VIF's, we have confirmed that there is no multicollinearity in our model.

The variable "rooms" and "value" has a positive relation with a correlation coefficient of 0.7. It means that the more rooms in a property, the higher the house value, which is totally reasonable. Another variable "lowstat" has a negative relation with "value" with a correlation coefficient of -0.74. It can be interpreted as that the greater lower status population in a neighborhood, the lower the house value in that area, which makes sense because low income people tend to buy inexpensive properties. The other variables have less correlation coefficient with house value. Therefore, the most two dominant factors of house value are "rooms" and "lowstat".

The regression model has many limitations, such as we assume that house value is linearly dependent on other variables. The real model may be very complicated and need more samples to fit more sophisticated equations. Also, the reality may need more variables to represent house value, for example, interior/exterior materials and decorations, electrical/heating/central AC systems, and so on.

V: Conclusions

The goal of this report was to determine the neighborhood and property attributes that can best explain the variations of house pricing. We have used SAS univariate techniques to examine the sample observations and we have carried out multiple regression analysis of the boston dataset. In examining the final model, one finds – quite reasonably – that house prices are higher in areas with lower crime and lower pupil-teacher ratios. House prices also tend to be higher closer to the business districts, and houses with more rooms are pricier. The number of rooms in the property and the lower-status population in the neighborhood are more closely related to house value.

The most interesting factors to consider are nitrogen oxide levels and distance to the main employment centers. People are aware of the pollutions in Boston. Talking of pollution, it is not just nitrogen oxide levels that are higher in industry districts, but also noise levels. Our regression model shows that the house value is less in polluted areas, which indicates that those areas have to use lower housing to attract people.

Last but not least, we want audience to note that the data for this report was collected several decades ago. Nowadays, the inflation rate and other society/business changes need to take into account when we study house values. We expect a more current dataset would appear for people to study housing price model.