Boston Housing Data Analysis

Josh Janda

Introduction

This project, the Boston Housing Data Analysis, utilizes SAS to analyze the Boston Housing dataset and ultimately work with the data to get it to a state that is clean, valid, and ready for modeling and analyzing. The dataset, *Boston_damaged.dat*, contains the classic Boston Housing dataset that was published by Harrison and Rubinfield in 1978. This original dataset was utilized to analyze housing prices in the Boston area and how they are related to the demand for clean air. The dataset we are using is an updated version of this, created by Gilley and Pace, which provides corrections and examined censoring as well as adding georeferencing and spatial estimation in the data. The data file uses an input style of delimited data, with each line representing one observation. The data is delimited by a tab rather than a space, which allows our variables to contain spaces without sacrificing the ease of reading it in. The variables in the dataset that will be analyzed in this project are:

- TOWN a factor with levels given by town names
- TOWN# a numeric vector corresponding to TOWN
- TRACT a numeric vector of tract ID numbers
- LON a numeric vector of tract point longitudes in decimal degrees
- LAT a numeric vector of tract point latitudes in decimal degrees
- MEDV a numeric vector of median values of owner-occupied housing in USD 1000
- CMEDV a numeric vector of corrected median values of owner-occupied housing in
- USD 1000
- CRIM a numeric vector of per capita crime
- ZN a numeric vector of proportions of residential land zoned for lots over 25000 sq. ft
- per town (constant for all Boston tracts)
- INDUS a numeric vector of proportions of non-retail business acres per town (constant
- for all Boston tracts)
- CHAS a factor with levels 1 if tract borders Charles River; 0 otherwise
- NOX a numeric vector of nitric oxides concentration (parts per 10 million) per town
- RM a numeric vector of average numbers of rooms per dwelling
- AGE a numeric vector of proportions of owner-occupied units built prior to 1940
- DIS a numeric vector of weighted distances to five Boston employment centers
- RAD a numeric vector of an index of accessibility to radial highways per town (constant
- for all Boston tracts)
- TAX a numeric vector full-value property-tax rate per USD 10,000 per town (constant for
- all Boston tracts)
- PTRATIO a numeric vector of pupil-teacher ratios per town (constant for all Boston
- tracts)
- B a numeric vector of 1000*(Bk - 0.63)^2 where Bk is the proportion of blacks
- LSTAT a numeric vector of percentage values of lower status population

The variable descriptions above are referenced from the original and updated dataset. The objectives of my project are as follows:

- Read in the raw data file using SAS
- Appropriately format and label all variables to be in a readable and understandable state
- Subset the data for variable analyzing and validation
- Utilizing different SAS procedures and tools to check the data for errors
- Validate the data and clean it using data editing techniques

Ultimately, the goal of this project is to read in the data for validation, analyzing, and finally cleaning the data for further use such as modeling and visualization.

Methods

The guidelines I am following for this data is to check each variable utilizing the frequency procedure for character variables, and the univariate procedure for numeric variables. If I notice any obscurity, I do a deeper analysis on the variable to find the obscurity and remedy it. Any variable not mentioned below can be assumed to be validated, and any variable mentioned below has a discrepancy in it. All figures are included below in this section that are referenced.

I first checked all variables simultaneously using the frequency procedure, which can be seen in **Figure 1**, which gave me the number of levels for each variable and any associated missing values in each variable. Right away, I noticed that the levels of "Town" and "Town_Number" were not equal. This tells me that that either some towns were misspelled, or town numbers were used twice. I first explored the misspelling option using a frequency table of the town variable, seen in **Figure 2**. I noticed that there were two towns misspelled, "Somervile" meaning to be "Somerville" and "Welesley" meaning to be "Wellesley". This was remedied using an if statement in a data step.

The next noticeable data errors were the missing values in "Indus" and "Tax". Both variables are constant across town, so should not be missing. I used the print procedure to print rows where "Indus" was missing, along with the associated town number (**Figure 3**). I used the same procedure for the "Tax" variable (**Figure 4**). After finding the towns with missing values, I found the correct value for each variable in the associated town(s) using the print procedure (**Figure 5 & 6**). I then remedied this issue using an if statement in a data step.

For the next error, I moved onto using the univariate procedure to check each numeric variable. The first issue I found was in the longitude variable, "lon". Since all observations are in the same general area, their longitude should be relatively the same. Looking at **Figure 7**, this is not the case. The positive values were obviously meant to be negative. This was remedied using an if statement in a data step.

I then analyzed the latitude variable, "lat", using the univariate procedure. Similarly, to the longitude variable, all latitudes should be extremely similar. Looking at **Figure 8**, we can see there are some extreme observations that are incorrect and obviously a 1 was mis keyed at the beginning of the entry. This was remedied using an if statement in a data step and removing the 1 from the beginning of the variable value for the given observations.

The next error I found was in the "nox" variable, seen in **Figure 9**. The observations of "9999.00" are very extreme, and obviously incorrect. Knowing this variable is constant between towns, we can see in **Figure 10** the towns these extreme values belong to. **Figure 11** shows us the correct values for these extreme values, and then an if statement in a data step was used to remedy these errors.

The next variable with an error was "ptratio". This variable should be constant across all towns, so the standard deviation of this variable between each town should be zero or null. Looking at **Figure 12**, we can see there is a town with a standard deviation greater than zero. Using **Figure 13**, we print out all observations where town is "Boston Hyde Park" and its associated observation number and ptratio. One observation is 28.2 rather than 20.2, which is a mistake. This was remedied using an if statement in a data step.

After all variables were analyzed, I utilized a data step to remedy all issues. I used multiple "if then" and "else if then" statements to look for the observation numbers with incorrect values for variables and set them to the correct values. I also created a format for the "chas" variable in this step and labeled each variable for better understanding. We can see in the next section that these methods ultimately resulted in a clearer, more valid dataset.

Figure 1

Number of Variable Levels			
Variable	Levels	Missing Levels	Nonmissing Levels
obs	506	0	506
town	94	0	94
town_number	92	0	92
tract	506	0	506
lon	375	0	375
lat	377	0	377
medv	229	0	229
cmedv	228	0	228
crim	504	0	504
zn	26	0	26
indus	77	1	76
chas	2	0	2
nox	82	0	82
rm	446	0	446
age	356	0	356
dis	412	0	412
rad	9	0	9
tax	67	1	66
ptratio	47	0	47
b	357	0	357
lstat	455	0	455

Figure 2

The FREQ Procedure

town	Frequency
Arlington	7
Ashland	2
Bedford	2
Belmont	8
Beverly	6
Boston Allston-Brighton	8
Boston Back Bay	6
Boston Beacon Hill	3
Boston Charlestown	6
Boston Dorchester	11
Boston Downtown	8
Boston East Boston	12
Boston Forest Hills	7
Boston Hyde Park	4
Boston Mattapan	6
Boston North End	2
Boston Roxbury	19
Boston Savin Hill	23
Boston South Boston	13
Boston West Roxbury	4
Braintree	8
Brookline	12
Burlington	4
Cambridge	30
Canton	3
Chelsea	5
Cohasset	1
Concord	3
Danvers	4
Dedham	5
Dover	1
Duxbury	1
Everett	7
Framingham	10
Hamilton	1
Hanover	1
Hingham	2

Figure 2

The FREQ Procedure

town	Frequency
Holbrook	2
Hull	1
Lexington	6
Lincoln	1
Lynn	22
Lynnfield	2
Malden	9
Manchester	1
Marblehead	3
Marshfield	2
Medfield	1
Medford	11
Melrose	4
Middleton	1
Millis	1
Milton	4
Nahant	1
Natick	6
Needham	5
Newton	18
Norfolk	1
North Reading	2
Norwell	1
Norwood	5
Peabody	9
Pembroke	2
Quincy	12
Randolph	3
Reading	4
Revere	8
Rockland	2
Salem	7
Sargus	4
Scituate	2
Sharon	3
Sherborn	1
Somervile	1

Figure 2

The FREQ Procedure

town	Frequency
Somerville	14
Stoneham	3
Sudbury	2
Swampscott	2
Topsfield	1
Wakefield	4
Walpole	3
Waltham	11
Watertown	4
Wayland	2
Welesley	1
Wellesley	3
Wenham	1
Weston	2
Westwood	3
Weymouth	8
Wilmington	3
Winchester	5
Winthrop	5
Woburn	6

Figure 2

Obs	obs	town	town_number
128	128	Somerville	27
129	129	Somerville	27
130	130	Somerville	27
131	131	Somerville	27
132	132	Somerville	27
133	133	Somerville	27
134	134	Somerville	27
135	135	Somerville	27
136	136	Somerville	27
137	137	Somerville	27
138	138	Somerville	27
139	139	Somerville	27
140	140	Somerville	27
141	141	Somerville	27
142	142	*Somervile*	27
280	280	Wellesley	48
281	281	*Welesley*	48
282	282	Wellesley	48
283	283	Wellesley	48

Figure 3

Obs	obs	town_number	indus
136	136	27	.

Figure 4

Obs	obs	town_number	tax
213	213	38	.
315	315	59	.

Figure 5

Obs	obs	town_number	tax
206	206	38	277
207	207	38	277
208	208	38	277
209	209	38	277
210	210	38	277
211	211	38	277
212	212	38	277
213	213	38	.
214	214	38	277
215	215	38	277
216	216	38	277
309	309	59	304
310	310	59	304
311	311	59	304
312	312	59	304
313	313	59	304
314	314	59	304
315	315	59	.
316	316	59	304
317	317	59	304
318	318	59	304
319	319	59	304
320	320	59	304

Figure 6

Obs	obs	town_number	indus
128	128	27	21.89
129	129	27	21.89
130	130	27	21.89
131	131	27	21.89
132	132	27	21.89
133	133	27	21.89
134	134	27	21.89
135	135	27	21.89
136	136	27	.
137	137	27	21.89
138	138	27	21.89
139	139	27	21.89
140	140	27	21.89
141	141	27	21.89
142	142	27	21.89

Figure 7

The UNIVARIATE Procedure
Variable: lon

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-71.2895	255	-70.8300	354
-71.2807	254	-70.8100	353
-71.2690	256	71.0243	121
-71.2685	253	71.0312	122
-71.2630	201	71.0377	123

Figure 8

The UNIVARIATE Procedure
Variable: lat

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
42.0300	356	42.3715	58
42.0485	354	42.3740	57
42.0520	355	42.3810	56
42.0590	353	142.1150	336
42.0590	300	142.1374	339

Figure 9

The UNIVARIATE Procedure
Variable: nox

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.385	287	0.871	157
0.389	286	0.871	160
0.392	256	9999.000	82
0.392	255	9999.000	116
0.394	257	9999.000	214

Figure 10

Obs	obs	town_number	nox
82	82	19	9999
116	116	25	9999
214	214	38	9999

Obs	obs	town_number	nox
81	81	19	0.43
82	82	19	9999.00
83	83	19	0.43
84	84	19	0.43
112	112	25	0.55
113	113	25	0.55
114	114	25	0.55
115	115	25	0.55
116	116	25	9999.00
117	117	25	0.55
118	118	25	0.55
119	119	25	0.55
120	120	25	0.55
206	206	38	0.49
207	207	38	0.49
208	208	38	0.49
209	209	38	0.49
210	210	38	0.49
211	211	38	0.49
212	212	38	0.49
213	213	38	0.49
214	214	38	9999.00
215	215	38	0.49
216	216	38	0.49

Figure 12

Obs	town	_TYPE_	_FREQ_	stdPtratio
1		0	506	2.20652
15	Boston Hyde Park	1	4	4.00000

Figure 13

Obs	obs	Town	Ptratio
485	485	Boston Hyde Park	20.2
486	486	Boston Hyde Park	20.2
487	487	Boston Hyde Park	20.2
488	488	Boston Hyde Park	28.2

Results

After performing all methods mentioned above in the methods section, we resulted in a valid and clean dataset. We now have all variables with invalid values set to what I believe to be their correct values. Using **Results Figure 1**, we can see that now town and town_number have an equal amount of levels, and that there are no missing values for any variables. Using **Results Figure 2**, we can see the transformation of all observations with invalid values to the clean data with fixed values (I have only included variables that had errors in them). This figure shows that we have utilized cleaning techniques and have resulted in a dataset that is valid and much more useable. I have bolded and italicized any noticeable results that are mentioned above in the figures. Any figures mentioned above are included directly below.

Results Figure 1

Number of Variable Levels		
Variable	**Label**	**Levels**
obs		506
town	factor with levels given by town names	***92***
town_number	Unique Town Identifier	***92***
tract	Unique ID for each observation	506
lon	Longitude of Observation	375
lat	Latitude of Observation	376
medv	Median values of owner-occupied housing in USD 1000	229
cmedv	Corrected median values of owner-occupied housing in USD 1000	228
crim	Per Capita Crime	504
zn	Proportions of residential land zoned for lots over 25000 sq. ft per town	26
indus	Proportions of non-retail business acres per town	76
chas	factor with levels 1 if tract borders Charles River; 0 otherwise	2
nox	Nitric oxides concentration (parts per 10 million) per town	83
rm	Average numbers of rooms per dwelling	446
age	Proportions of owner-occupied units built prior to 1940	356
dis	Weighted distances to five Boston employment centers	412
rad	Index of accessibility to radial highways per town	9
tax	Full-value property-tax rate per USD 10,000 per town	66
ptratio	Pupil to teacher ratios per town	46
b	1000*(Bk - 0.63)^2 where Bk is the proportion of blacks	357
lstat	Percentage values of lower status population	455

Results Figure 2

Obs	obs	town	town_number	lon	lat	indus	nox	tax	ptratio
82	82	Reading	19	-71.0690	42.315	4.86	***9999.00***	281	19.0
116	116	Malden	25	-71.0355	42.255	10.01	***9999.00***	432	17.8
121	121	Everett	26	***71.0243***	42.248	25.65	0.58	188	19.1
122	122	Everett	26	***71.0312***	42.251	25.65	0.58	188	19.1
123	123	Everett	26	***71.0377***	42.247	25.65	0.58	188	19.1
136	136	Somerville	27	-71.0750	42.236	.	0.62	437	21.2
142	142	***Somervile***	27	-71.0543	42.227	21.89	0.62	437	21.2
213	213	Waltham	38	-71.1335	42.225	10.59	0.49	.	18.6
214	214	Waltham	38	-71.1375	42.236	10.59	***9999.00***	277	18.6
281	281	***Welesley***	48	-71.1660	42.187	3.33	0.44	216	14.9
315	315	Quincy	59	-71.0000	42.153	9.90	0.54	.	18.4
336	336	Weymouth	63	-70.9700	***142.115***	5.19	0.52	224	20.2
339	339	Weymouth	63	-70.9633	***142.137***	5.19	0.52	224	20.2
488	488	Boston Hyde Park	88	-71.0650	42.161	18.10	0.58	666	***28.2***

Obs	obs	town	town_number	lon	lat	indus	nox	tax	ptratio
82	82	Reading	19	-71.0690	42.3150	4.86	***0.4300***	281	19.0
116	116	Malden	25	-71.0355	42.2545	10.01	***0.5500***	432	17.8
121	121	Everett	26	***-71.0243***	42.2483	25.65	0.5810	188	19.1
122	122	Everett	26	***-71.0312***	42.2505	25.65	0.5810	188	19.1
123	123	Everett	26	***-71.0377***	42.2470	25.65	0.5810	188	19.1
136	136	Somerville	27	-71.0750	42.2362	***21.89***	0.6240	437	21.2
142	142	***Somerville***	27	-71.0543	42.2265	21.89	0.6240	437	21.2
213	213	Waltham	38	-71.1335	42.2250	10.59	0.4890	***277***	18.6
214	214	Waltham	38	-71.1375	42.2355	10.59	***0.4900***	277	18.6
281	281	***Wellesley***	48	-71.1660	42.1870	3.33	0.4429	216	14.9
315	315	Quincy	59	-71.0000	42.1530	9.90	0.5440	***304***	18.4
336	336	Weymouth	63	-70.9700	***42.1150***	5.19	0.5150	224	20.2
339	339	Weymouth	63	-70.9633	***42.1374***	5.19	0.5150	224	20.2
488	488	Boston Hyde Park	88	-71.0650	42.1610	18.10	0.5830	666	***20.2***

Appendix

With our data cleaned and validated using the above figures and techniques, we will analyze some variables to verify cleaning was completed thoroughly. These variables will be analyzed in terms of a question.

The first question to answer is “Which Town is represented by the most tracts (town that appears the most)”. As we can see in **Appendix Figure 1**, the town that is most frequent in our data is Cambridge.

The second question to answer is “Which Towns have the highest and lowest average per capita crime rate?” The towns that have the highest and lowest average per capita crime rate are shown in **Appendix Figure 2**. The top 5 are “Boston Charlestown:”, “Boston South Boston”, “Boston Downtown”, “Boston Roxbury”, and “Boston North End”. The bottom 5 are “Nahant”, “Medfield”, “Millis”, “Cohasset”, and “Topsfield”.

The third, and final, question to answer is “What is the distribution of the variable MEDV?”. Using **Appendix Figure 3**, we can answer that question. We can see right away from the histogram that the variable is right skewed, with a mean of 22.53 and standard deviation of 9.20. So, our data has a center of 22.53 and a spread of 9.20. For deciding on the distribution of this variable, we must also look at the skewness and kurtosis. A normal distribution has a skewness of 0, and kurtosis of 3. Our variable has a skewness of 1.11, confirming that it is right-skewed, and kurtosis of 1.50. Our data seems to be better fit by a more right-skewed distribution, such as the Chi-Square, Beta, or F-Distributions. One interesting feature of this variable is that it is scaled by 1/1000. So, the true values are really 1000 times greater than shown. Another interesting feature is that the range is 5.0-50.0. I believe that values extended beyond 50.0 but were capped there for unknown reasons. We can see this in the histogram, as the percent of homes at 50.0 is higher than those at 46. If there was not a cap at 50.0, I believe this variable would have been even more right-skewed.

Appendix Figure 1

town	Frequency
Cambridge	30
Boston Savin Hill	23
Lynn	22
Boston Roxbury	19
Newton	18
Somerville	15
Boston South Boston	13
Boston East Boston	12
Brookline	12
Quincy	12
Boston Dorchester	11
Medford	11
Waltham	11
Framingham	10
Malden	9
Peabody	9
Belmont	8
Boston Allston-Brighton	8
Boston Downtown	8
Braintree	8
Revere	8
Weymouth	8
Arlington	7
Boston Forest Hills	7
Everett	7
Salem	7
Beverly	6
Boston Back Bay	6
Boston Charlestown	6
Boston Mattapan	6
Lexington	6
Natick	6
Woburn	6
Chelsea	5
Dedham	5
Needham	5
Norwood	5
Winchester	5
Winthrop	5
Boston Hyde Park	4
Boston West Roxbury	4

Appendix Figure 1

town	Frequency
Burlington	4
Danvers	4
Melrose	4
Milton	4
Reading	4
Sargus	4
Wakefield	4
Watertown	4
Wellesley	4
Boston Beacon Hill	3
Canton	3
Concord	3
Marblehead	3
Randolph	3
Sharon	3
Stoneham	3
Walpole	3
Westwood	3
Wilmington	3
Ashland	2
Bedford	2
Boston North End	2
Hingham	2
Holbrook	2
Lynnfield	2
Marshfield	2
North Reading	2
Pembroke	2
Rockland	2
Scituate	2
Sudbury	2
Swampscott	2
Wayland	2
Weston	2
Cohasset	1
Dover	1
Duxbury	1
Hamilton	1
Hanover	1
Hull	1
Lincoln	1

Appendix Figure 1

town	Frequency
Manchester	1
Medfield	1
Middleton	1
Millis	1
Nahant	1
Norfolk	1
Norwell	1
Sherborn	1
Topsfield	1
Wenham	1

Appendix Figure 2

Obs	town	meanCrim
1	Boston Charlestown	29.2019
2	Boston South Boston	21.2049
3	Boston Downtown	20.8953
4	Boston Roxbury	17.8646
5	Boston North End	14.8032

Obs	town	meanCrim
1	Nahant	0.00632
2	Medfield	0.00906
3	Millis	0.01096
4	Cohasset	0.01301
5	Topsfield	0.01311

Appendix Figure 3

The UNIVARIATE Procedure

Variable: medv (Median values of owner-occupied housing in USD 1000)

Moments			
N	506	**Sum Weights**	506
Mean	22.5328063	**Sum Observations**	11401.6
Std Deviation	9.19710409	**Variance**	84.5867236
Skewness	1.10809841	**Kurtosis**	1.49519694
Uncorrected SS	299626.34	**Corrected SS**	42716.2954
Coeff Variation	40.8165053	**Std Error Mean**	0.40886115

Basic Statistical Measures			
Location		**Variability**	
Mean	22.53281	**Std Deviation**	9.19710
Median	21.20000	**Variance**	84.58672
Mode	50.00000	**Range**	45.00000
		Interquartile Range	8.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	55.11115	**Pr > |t|**	<.0001
Sign	**M**	253	**Pr >= |M|**	<.0001
Signed Rank	**S**	64135.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
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

Appendix Figure 3

The UNIVARIATE Procedure
Variable: medv (Median values of owner-occupied housing in USD 1000)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
5.0	406	50	369
5.0	399	50	370
5.6	401	50	371
6.3	400	50	372
7.0	490	50	373