Kennesaw State University
IS 8935 Business Intelligence: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 3
January 28, 2019

Boston Housing Data Correlation Analysis

By
Karis Kim

Executive Summary

This analysis examines the correlation between the 14 selected attributes in the Boston Housing data set to provide insight regarding factors that affect housing values in the suburbs of Boston. The analysis yielded 3 very strong positive correlations: 1) crime rate and accessibility to radial highways, 2) accessibility to radial highways and property tax rate, and 3) crime rate and property tax rate. Attributes with the strongest correlation to median value of homes were: 1) average number of rooms per dwelling, and 2) percentage of lower status population. Attributes with no correlation to median value of homes were: 1) distance to employment centres, 2) proximity to Charles River, 3) accessibility to radial highways, and 4) proportion of African-Americans in town, among others. While crime rate and accessibility to radial highways are very strongly correlated to each other, they are not strongly correlated to the median value of homes.

Business Understanding

The goal of this correlation analysis is to examine the statistical measure of how strong the relationships are between the variables (hereafter attributes) in the Boston Housing data set. Strong positive or negative correlations between certain attributes should provide insight regarding factors that may affect housing values in the suburbs of Boston.

Assumptions

- The attributes listed in the data set are significant and relevant factors affecting housing values and are valid considerations in the determination of housing values.
- While housing values and variables that affect housing values fluctuate constantly over time and require ongoing investigation, this analysis is premised on the assumption that a one-time snapshot correlation analysis will still yield useful insight.
- Identifying certain attributes that affect housing values could be used as objective and standard factors in housing appraisal.
- Boston housing data set pertains to owner-occupied single family residential dwellings, such as detached homes, attached homes/townhomes, or individually appraised condominium units.

Business Questions

- What attributes have a strong relationship to Boston housing values?
- Which attributes seek consideration in Boston housing appraisals/assessments because of their strong correlations?

Data Understanding

The Boston Housing Data consists of 14 attributes and 506 instances (a.k.a. examples, observations) that relate to housing values in suburbs of Boston circa 1978.

Data Storage

The dataset is from StatLib library of Carnegie Mellon University

Creator

D. Harrison and D. L. Rubinfeld

Past Usage

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management,5*(1), 81-102.

Belsley, D. A., Kuh, E., & Welsch, R. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity* (pp. 244-261). New York: J. Wiley.

Quinlan, R. (1993). Combining Instance-Based and Model-Based Learning. In *Tenth International Conference of Machine Learning*(pp. 236-243). University of Massachusetts, Amherst, MA: Morgan Kaufmann.

Attribute Information

No.	Attribute	Description
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	PTRATIO	pupil-teacher ratio by town
12	B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13	LSTAT	% lower status of the population
14	MEDV	Median value of owner-occupied homes in $1000's

Data Quantity

Boston Housing data set consists of 14 attributes (columns) and 506 instances (rows).

Data Assumptions

No.	Attribute	Assumptions / Implications
1	CRIM	Lower crime rate has a negative correlation to housing value. 0 = No crime reported and does not indicate a missing value.
2	ZN	0 = There is no land zoned for residential, and is zoned for 1) commercial/retail, 2) rural/agricultural, 3) historic/aesthetic, or 4) industrial. That is, 0 is not a missing value. Zoning type and proportion affects housing value.
3	INDUS	If value > 0, then land is zoned for non-retail industrial. Zoning type and proportion affects housing value.
4	CHAS	Proximity to Charles River affects housing value.
5	NOX	Higher NOX value is more detrimental/undesirable because NOx contributes to air pollution.
6	RM	Higher number of rooms will have positive correlation to MEDV. An average owner-occupied single family dwelling in the data set should not have more than 10 rooms per dwelling.
7	AGE	Age of owner-occupied homes will have a correlation to housing value.
8	DIS	Distance to Boston employment centres will affect housing value.
9	RAD	Higher value indicates easier/better accessibility to radial highways, so greater accessibility indicates closer distances to urban/inner city areas.
10	TAX	Higher value indicates higher property-taxes.

11	PTRATIO	Lower pupil-teacher ratio is favorable to housing value.
12	B	African-Americans "in town" refers to residents of said town. Higher value indicates higher population of African-Americans. Proportion of African-American residents has a correlation to housing value.
13	LSTAT	Higher percentage of lower status population has a negative correlation to housing value.
14	MEDV	Appraisal of home values that yielded these median values of owner-occupied homes was accurate, appropriate, and consistent. "Owner-occupied homes" indicates single family dwelling, whether it be categorized as a detached home, attached unit, or condominium.

Missing Values

MEDV (median value of owner-occupied homes) is the only attribute with missing values. 54 of 506 examples are missing MEDV values.

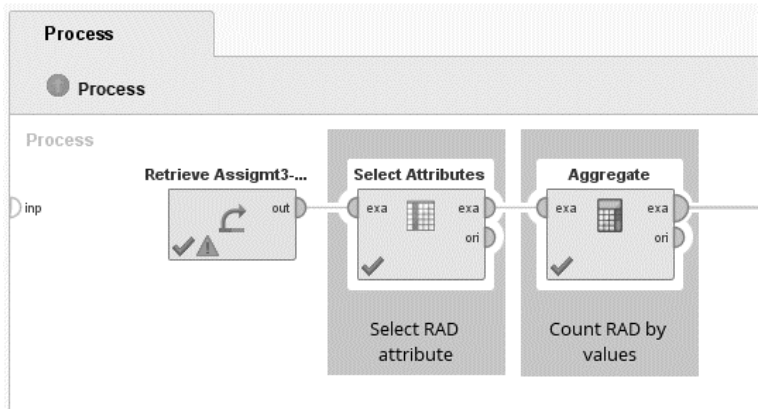
Inconsistent Values

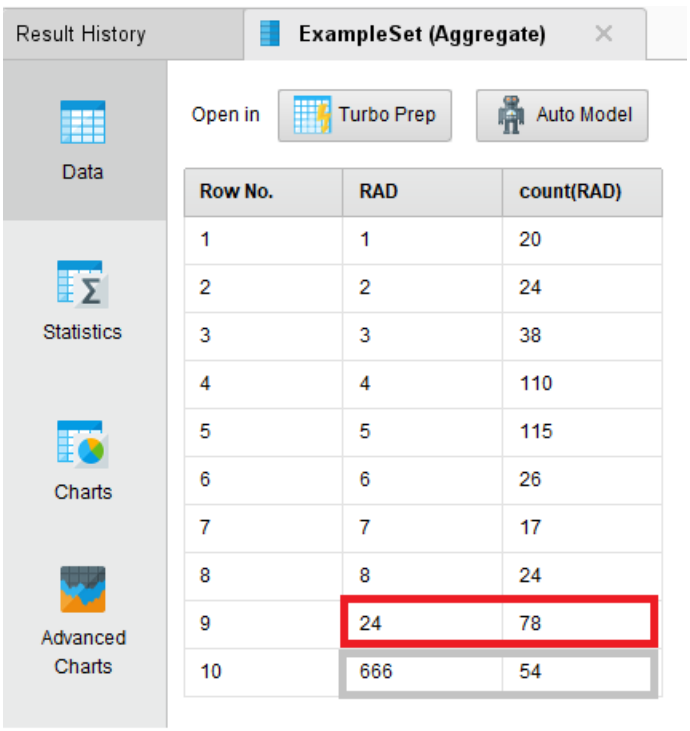
1. RM (average number of rooms per dwelling) has high range values from 56.7 to 100. These values are deemed inconsistent because a single dwelling typically does not have 50 plus rooms, even if it is a castle. Residential properties with high numbers of rooms like 50 – 100 are typically classified as apartments and condominiums.

   There are 54 of 506 examples with RM values that exceed the normal number of rooms per dwelling. These rows are also the same 54 examples with missing MEDV values. This indicates the possibility/probability that MEDV could not be reported because the dwelling type was not an owner-occupied home (such as an apartment complex for rent) or because the number of rooms were reported from an aggregated total number of rooms for the entire multi-resident condominium, thereby preventing MEDV values per owner-occupied homes to be determined.

2. RAD (index of accessibility to radial highways) mostly shows single digit values except 54 examples with RAD value of 666, and 78 examples with RAD value of 24. Those 54 rows also coincide with the 54 rows with missing MEDV and inconsistent RM values.

However, the 78 examples with RAD values outside of the normal range are an anomaly that need reckoning.





Row No.	RAD	count(RAD)
1	1	20
2	2	24
3	3	38
4	4	110
5	5	115
6	6	26
7	7	17
8	8	24
9	24	78
10	666	54

Figure 1. RapidMiner Count of RAD values

3. PTRATIO (pupil-teacher ratio by town) shows inconsistent values with ratios as high as 332.09 pupil: 1 teacher. We can safely presume that no legitimate school in the U.S. would permit a pupil to teacher ratio exceeding a few dozen to one, even in the worst of conditions. The top 54 highest PTRATIO examples are also the same 54 examples with missing MEDV, inconsistent RM, and RAD. This indicates the possibility that the high PTRATIO values may have been derived by some type of aggregation for the multi-tenant dwellings these examples potentially represent.

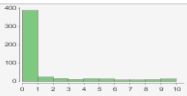
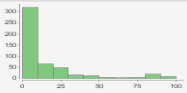
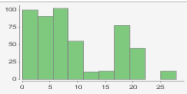
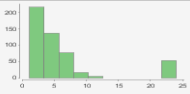
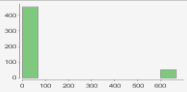
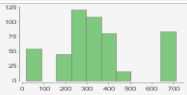
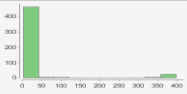
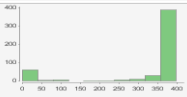
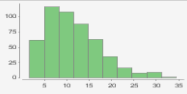
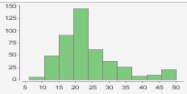
Data
Statistics
Charts
Advanced Charts
Annotations

Open in Turbo Prep Auto Model

Row No.	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
374	0	18.100	0	0	4.906	100	1.174	24	666	20	396.900	34.770	13.800	?
375	0	18.100	0	0	4.138	100	1.137	24	666	20	396.900	37.970	13.800	?
376	0	18.100	0	0	7.313	97.900	1.316	24	666	20	396.900	13.440	15	?
379	0	18.100	0	0	6.380	96.200	1.386	24	666	20	396.900	23.690	13.100	?
381	0	18.100	0	0	6.968	91.900	1.417	24	666	20	396.900	17.210	10.400	?
382	0	18.100	0	0	6.545	99.100	1.519	24	666	20	396.900	21.080	10.900	?
386	0	18.100	0	0	5.277	98.100	1.426	24	666	20	396.900	30.810	7.200	?
387	0	18.100	0	0	4.652	100	1.467	24	666	20	396.900	28.280	10.500	?
388	0	18.100	0	0	5	89.500	1.518	24	666	20	396.900	31.990	7.400	?
393	0	18.100	0	0	5.036	97	1.770	24	666	20	396.900	25.680	9.700	?
395	0	18.100	0	0	5.887	94.700	1.782	24	666	20	396.900	16.350	12.700	?
399	0	18.100	0	0	5.453	100	1.490	24	666	20	396.900	30.590	5	?
401	0	18.100	0	0	5.987	100	1.589	24	666	20	396.900	26.770	5.600	?
402	0	18.100	0	0	6.343	100	1.574	24	666	20	396.900	20.320	7.200	?
404	0	18.100	0	0	5.349	96	1.703	24	666	20	396.900	19.770	8.300	?
470	0	18.100	0	0	5.713	56.700	2.824	24	666	20	396.900	14.760	20.100	?
380	0	18.100	0	0	6.223	100	1.386	24	666	20	393.740	21.780	10.200	?
441	0	18.100	0	0	5.818	92.400	1.866	24	666	20	391.450	22.110	10.500	?
406	0	18.100	0	0	5.683	100	1.425	24	666	20	384.970	22.980	5	?
480	0	18.100	0	0	6.229	88	1.951	24	666	20	383.320	13.110	21.400	?
479	0	18.100	0	0	6.185	96.700	2.171	24	666	20	379.700	18.030	14.600	?
389	0	18.100	0	0	4.880	100	1.589	24	666	20	372.920	30.620	10.200	?
407	0	18.100	0	0	4.138	100	1.178	24	666	20	370.220	23.340	11.900	?
469	0	18.100	0	0	5.926	71	2.908	24	666	20	368.740	18.130	19.100	?
377	0	18.100	0	0	6.649	93.300	1.345	24	666	20	363.020	23.240	13.900	?
478	0	18.100	0	0	5.304	97.300	2.101	24	666	20	349.480	24.910	12	?
408	0	18.100	0	0	5.608	100	1.285	24	666	20	332.090	12.130	27.900	?
405	0	18.100	0	0	5.531	85.400	1.607	24	666	20	329.460	27.380	8.500	?
421	0	18.100	0	0	6.411	100	1.859	24	666	20	318.750	15.020	16.700	?
423	0	18.100	0	0	5.648	87.600	1.951	24	666	20	291.550	14.100	20.800	?
385	0	18.100	0	0	4.368	91.200	1.440	24	666	20	285.830	30.630	8.800	?
445	0	18.100	0	0	5.854	96.600	1.896	24	666	20	240.520	23.790	10.800	?
414	0	18.100	0	0	5.155	100	1.589	24	666	20	210.970	20.080	16.300	?
410	0	18.100	0	0	6.852	100	1.466	24	666	20	179.360	19.780	27.500	?
368	0	18.100	0	0	3.863	100	1.511	24	666	20	131.420	13.330	23.100	?
418	0	18.100	0	0	5.304	89.100	1.647	24	666	20	127.360	26.640	10.400	?
436	0	18.100	0	0	6.629	94.600	2.125	24	666	20	109.850	23.270	13.400	?
435	0	18.100	0	0	6.208	95	2.222	24	666	20	100.630	15.170	11.700	?
415	0	18.100	0	0	4.519	100	1.658	24	666	20	88.270	36.980	7	?
432	0	18.100	0	0	6.833	94.300	2.088	24	666	20	81.330	19.690	14.100	?
439	0	18.100	0	0	5.935	87.900	1.821	24	666	20	68.950	34.020	8.400	?
420	0	18.100	0	0	6.824	76.500	1.794	24	666	20	48.450	22.740	8.400	?
446	0	18.100	0	0	6.459	94.800	1.988	24	666	20	43.060	23.980	11.800	?
412	0	18.100	0	0	6.657	100	1.528	24	666	20	35.050	21.220	17.200	?
413	0	18.100	0	0	4.628	100	1.554	24	666	20	28.790	34.370	17.900	?
437	0	18.100	0	0	6.461	93.300	2.003	24	666	20	27.490	18.050	9.600	?
416	0	18.100	0	0	6.434	100	1.835	24	666	20	27.250	29.050	7.200	?
427	0	18.100	0	0	5.837	59.700	1.998	24	666	20	24.650	15.690	10.200	?
355	0.043	80	1.910	0	0.413	5.663	21.900	10.586	4	334	22	382.800	8.050	18.200

Figure 2. RapidMiner Missing MEDV with Other Inconsistent Attributes

Descriptive Statistics Overview

Name	Type	Missing	Statistics				
CRIM	Real	0	 Open chart	Min 0	Max 9.967	Average 1.269	Deviation 2.399
ZN	Real	0	 Open chart	Min 0	Max 100	Average 13.295	Deviation 23.049
INDUS	Real	0	 Open chart	Min 0	Max 27.740	Average 9.205	Deviation 7.170
CHAS	Integer	0	 Open chart	Min 0	Max 1	Average 0.069	Deviation 0.254
NOX	Real	0	 Open chart	Min 0.385	Max 7.313	Average 1.101	Deviation 1.647
RM	Real	0	 Open chart	Min 3.561	Max 100	Average 15.680	Deviation 27.220
AGE	Real	0	 Open chart	Min 1.137	Max 100	Average 58.745	Deviation 33.104
DIS	Real	0	 Open chart	Min 1.130	Max 24	Average 6.173	Deviation 6.476
RAD	Integer	0	 Open chart	Min 1	Max 666	Average 78.063	Deviation 203.542
TAX	Integer	0	 Open chart	Min 20	Max 711	Average 339.296	Deviation 180.708
PTRATIO	Real	0	 Open chart	Min 2.600	Max 396.900	Average 42.615	Deviation 87.585
B	Real	0	 Open chart	Min 0.320	Max 396.900	Average 332.791	Deviation 125.322
LSTAT	Real	0	 Open chart	Min 1.730	Max 34.410	Average 11.538	Deviation 6.065
MEDV	Real	54	 Open chart	Min 6.300	Max 50	Average 23.750	Deviation 8.809

Data Preparation

Missing Values

54 examples with missing MEDV values have been eliminated from the correlation analysis, because these examples also exhibited multiple attributes with inconsistent values and therefore deemed invaluable to the analysis.

Inconsistent Values

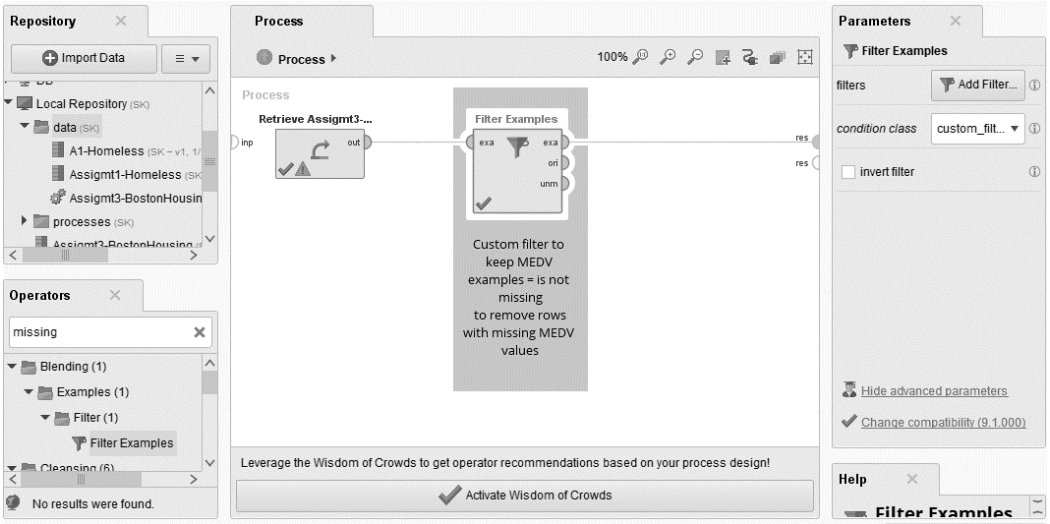
The 54 examples with missing MEDV values that also had inconsistent RM values and inconsistent PTRATIO values have been eliminated from the correlation analysis for reasons aforementioned in data preparation of missing values. These 54 eliminated examples also had the outlier RAD value of 666.

The 78 examples with the next highest RAD value of 24 will remain in the data set for analysis as is, because 78 instances is deemed too many to be considered outliers or entry errors, though a rationale for the wide variance between this value and the majority of other RAD values cannot be offered without access to further information on the data.

Data Transformation

With the exception of the removal of the 54 examples mentioned above, no other data transformations have been performed in the preparation phase.

Using Filter Examples operator in RapidMiner, 54 examples with missing MEDV values were filtered out of/removed from result data set.



Removing those 54 examples also resolved the inconsistent RM values, outlier RAD values, and inconsistent PTRATIO values.

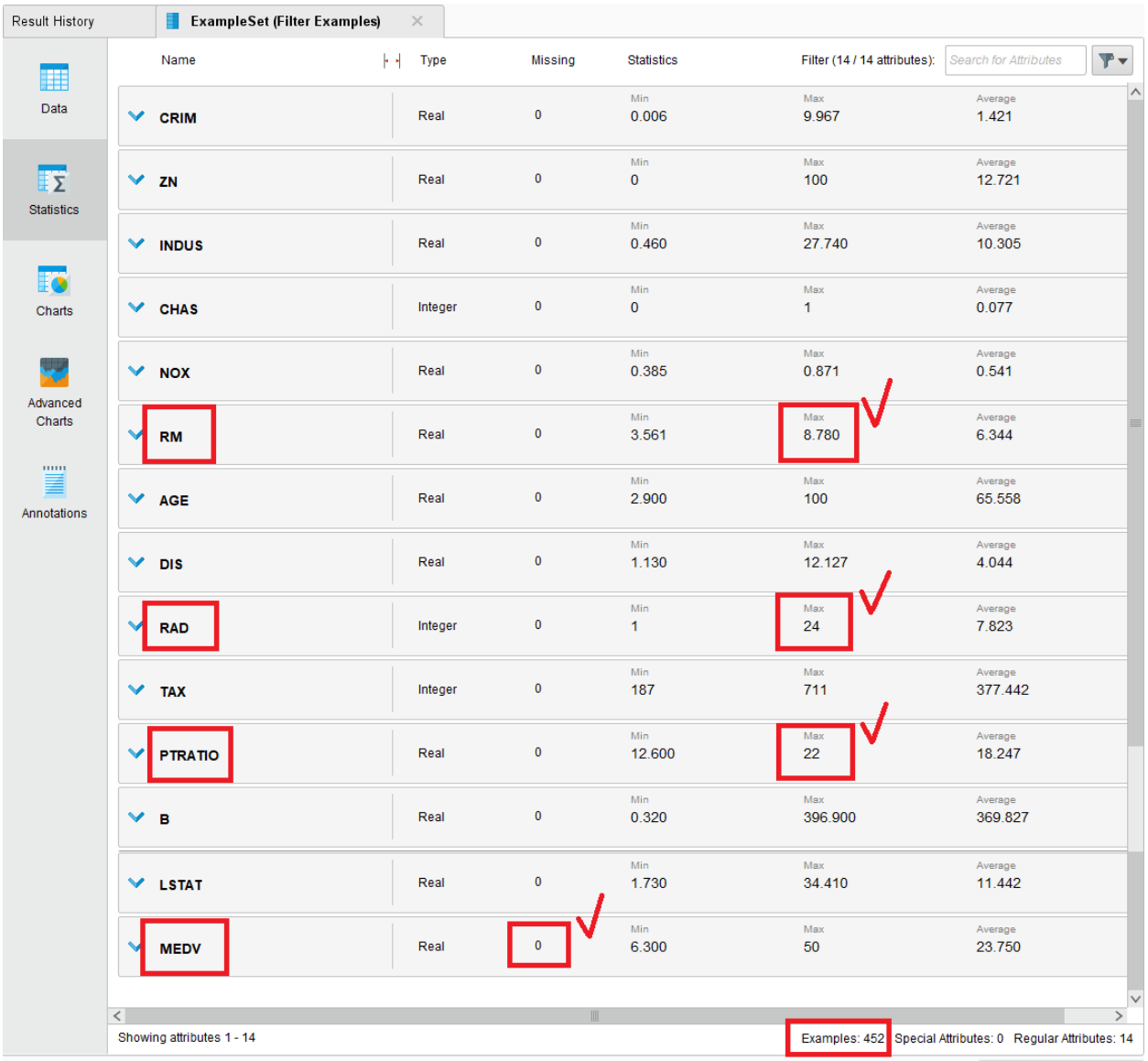
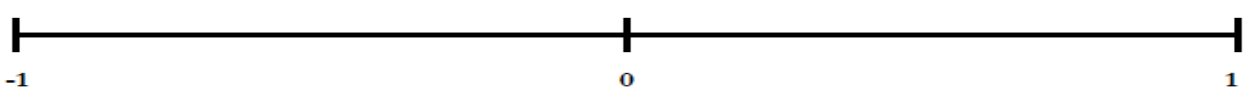


Figure 3. RapidMiner Descriptive Statistics after removal of 54 examples with missing MEDV

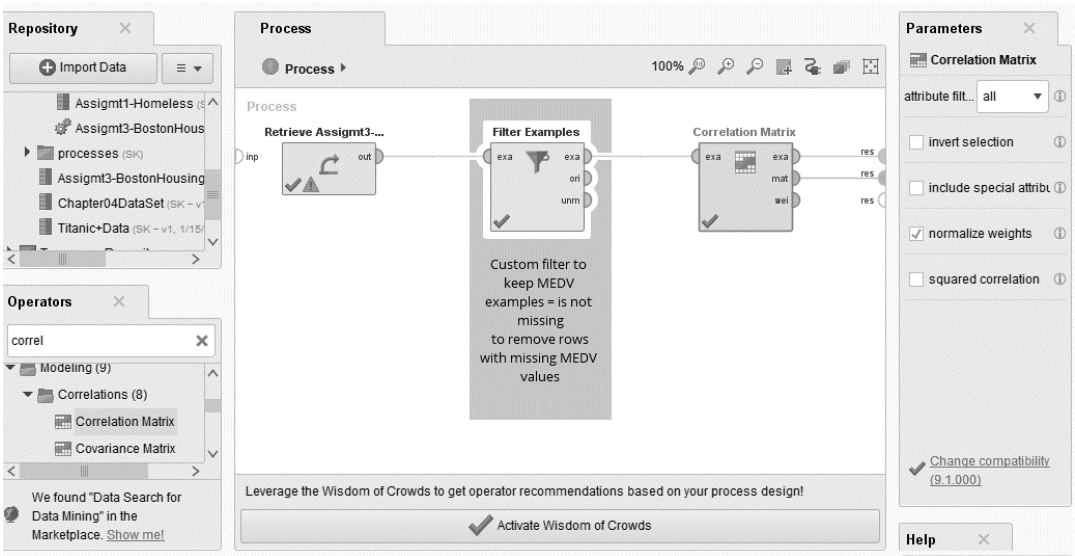
Modeling Correlations

The following scale will be used to determine correlation strength between -1 to 1.



-1 ← -0.8 Very Strong Correlation	-0.8 ← -0.6 Strong Correlation	-0.6 ← -0.4 Some Correlation	-0.4 ← 0 No Correlation	0 → 0.4 No Correlation	0.4 → 0.6 Some Correlation	0.6 → 0.8 Strong Correlation	0.8 → 1 Very Strong Correlation

In RapidMiner, a Correlation Matrix operator was executed on the Boston Housing data set.



The Correlation Matrix yielded the following results, with the strongest correlation coefficients highlighted in darker colors and gradually highlighted in lighter colors with corresponding coefficients.

Correlation Matrix (Correlation Matrix) | ExampleSet (Filter Examples)

Attribut...	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1	-0.281	0.574	0.050	0.637	-0.142	0.448	-0.462	0.898	0.826	0.319	-0.413	0.425	-0.286
ZN	-0.281	1	-0.514	-0.060	-0.501	0.307	-0.556	0.656	-0.267	-0.269	-0.364	0.150	-0.411	0.332
INDUS	0.574	-0.514	1	0.103	0.739	-0.365	0.606	-0.669	0.513	0.673	0.317	-0.317	0.565	-0.412
CHAS	0.050	-0.060	0.103	1	0.134	0.077	0.123	-0.141	0.057	0.017	-0.100	0.013	-0.009	0.154
NOX	0.637	-0.501	0.739	0.134	1	-0.265	0.707	-0.746	0.542	0.615	0.103	-0.358	0.537	-0.333
RM	-0.142	0.307	-0.365	0.077	-0.265	1	-0.188	0.139	-0.096	-0.215	-0.334	0.108	-0.607	0.740
AGE	0.448	-0.556	0.606	0.123	0.707	-0.188	1	-0.720	0.359	0.427	0.193	-0.224	0.573	-0.300
DIS	-0.462	0.656	-0.669	-0.141	-0.746	0.139	-0.720	1	-0.388	-0.444	-0.152	0.234	-0.424	0.139
RAD	0.898	-0.267	0.513	0.057	0.542	-0.096	0.359	-0.388	1	0.873	0.387	-0.353	0.310	-0.218
TAX	0.826	-0.269	0.673	0.017	0.615	-0.215	0.427	-0.444	0.873	1	0.385	-0.367	0.411	-0.346
PTRATIO	0.319	-0.364	0.317	-0.100	0.103	-0.334	0.193	-0.152	0.387	0.385	1	-0.090	0.303	-0.461
B	-0.413	0.150	-0.317	0.013	-0.358	0.108	-0.224	0.234	-0.353	-0.367	-0.090	1	-0.291	0.265
LSTAT	0.425	-0.411	0.565	-0.009	0.537	-0.607	0.573	-0.424	0.310	0.411	0.303	-0.291	1	-0.706
MEDV	-0.286	0.332	-0.412	0.154	-0.333	0.740	-0.300	0.139	-0.218	-0.346	-0.461	0.265	-0.706	1

The following Pairwise Correlation Table in RapidMiner is marked with strong correlation in the red box and some correlation in the yellow box from positive correlation coefficients down to the negative coefficients; middle coefficients that fall into the no correlation range have been omitted from the following view.

Result History | Correlation Matrix (Correlation Matrix)

Data | Pairwise Table | Charts | Annotations

First Attribute	Second Attri...	Correlation
CRIM	RAD	0.898
RAD	TAX	0.873
CRIM	TAX	0.826
RM	MEDV	0.740
INDUS	NOX	0.739
NOX	AGE	0.707
INDUS	TAX	0.673
ZN	DIS	0.656
CRIM	NOX	0.637
NOX	TAX	0.615
INDUS	AGE	0.606
CRIM	INDUS	0.574
AGE	LSTAT	0.573
INDUS	LSTAT	0.565
NOX	RAD	0.542
NOX	LSTAT	0.537
INDUS	RAD	0.513
CRIM	AGE	0.448
AGE	TAX	0.427
CRIM	LSTAT	0.425
TAX	LSTAT	0.411
RAD	PTRATIO	0.387
TAX	PTRATIO	0.385
AGE	RAD	0.359

Pairwise Table | Charts | Annotations

NOX	MEDV	-0.333
RM	PTRATIO	-0.334
TAX	MEDV	-0.346
RAD	B	-0.353
NOX	B	-0.358
ZN	PTRATIO	-0.364
INDUS	RM	-0.365
TAX	B	-0.367
DIS	RAD	-0.388
ZN	LSTAT	-0.411
INDUS	MEDV	-0.412
CRIM	B	-0.413
DIS	LSTAT	-0.424
DIS	TAX	-0.444
PTRATIO	MEDV	-0.461
CRIM	DIS	-0.462
ZN	NOX	-0.501
ZN	INDUS	-0.514
ZN	AGE	-0.556
RM	LSTAT	-0.607
INDUS	DIS	-0.669
LSTAT	MEDV	-0.706
AGE	DIS	-0.720
NOX	DIS	-0.746

Correlation Result Summary

Very Strong Positive Correlation:

There is a very strong positive correlation is between per capita crime rate and accessibility to radial highways, followed closely by accessibility to radial highways and property tax rate, then followed by crime rate and property tax rate. So, when accessibility to radial highways rises, the crime rate also rises. When accessibility to radial highways rises, the property tax rates also rise. When crime rates rise, tax rates also rise.

Strong Positive Correlation:
There is a strong positive correlation between average number of rooms per dwelling and median home value, industrial zoned land and nitric oxide levels, nitric oxide levels and age of homes built before 1940, industrial zoned land and property tax rate, residential zoned land and distance to employment centres, crime rate and nitric oxide levels, and nitric oxide levels and property tax rate. When one attribute rises, the other rises also; when one attribute falls, the other falls also.

		Correlation Strength		
↑ **CRIM**	↑ **RAD**	**0.898**	↓ **CRIM**	↓ **RAD**
↑ **RAD**	↑ **TAX**	**0.873**	↓ **RAD**	↓ **TAX**
↑ **CRIM**	↑ **TAX**	**0.826**	↓ **CRIM**	↓ **TAX**
↑ RM	↑ MEDV	0.740	↓ RM	↓ MEDV
↑ INDUS	↑ NOX	0.739	↓ INDUS	↓ NOX
↑ NOX	↑ AGE	0.707	↓ NOX	↓ AGE
↑ INDUS	↑ TAX	0.673	↓ INDUS	↓ TAX
↑ ZN	↑ DIS	0.656	↓ ZN	↓ DIS
↑ CRIM	↑ NOX	0.637	↓ CRIM	↓ NOX
↑ NOX	↑ TAX	0.615	↓ NOX	↓ TAX
↑ INDUS	↑ AGE	0.606	↓ INDUS	↓ AGE

Some positive correlation was observed between 10 other attribute combinations, but for this report, focus will remain on the strong and very strong correlations.

Strong Negative Correlation:
There is a strong negative correlation between nitric oxide levels and distance to employment centres, age of homes built before 1940 and distance to employment centres, lower status population percentage and median home value, industrial zoned land and distance to employment centres, average number of rooms per dwelling and lower status population

percentage. So, when nitric oxide levels rise, distance to employment centres fall and visa versa. When the percentage of lower status population rises, the median value of homes fall and visa versa. Table below shows the relationship between attributes with a strong negative correlation.

		Correlation Strength		
↑ NOX	↓ DIS	-0.746	↓ NOX	↑ DIS
↑ AGE	↓ DIS	-0.720	↓ AGE	↑ DIS
↑ LSTAT	↓ MEDV	-0.706	↓ LSTAT	↑ MEDV
↑ INDUS	↓ DIS	-0.669	↓ INDUS	↑ DIS
↑ RM	↓ LSTAT	-0.607	↓ RM	↑ LSTAT

Some negative correlation was observed between 10 other attribute combinations, but for this report, focus will remain on the strong correlations. There were no attribute combinations that showed very strong negative correlations.

Evaluation

By Attribute

Explanation of the most significant correlation for each attribute and possible implications:

1) Per capita crime rate has the strongest positive correlation with accessibility to radial highways. The higher the crime rate, the greater the accessibility to radial highways. Greater accessibility to radial highways indicates closer proximity to urban/inner city areas, so this correlation implies that crime rate is higher in areas closer to urban city. Crime rate also has a very strong positive correlation to tax rate, indicating that the higher the tax rate, the higher the crime. This second correlation is rather unexpected, since general preconception is that crime rates would be higher in lower income areas where property values and the corresponding property tax rates are lower.
2) Proportion of residential land zoned has the strongest positive correlation with distance to employment centres. The greater the proportion of residential land, the greater/farther the distance to employment centers. This implies that employment centres are located away from residential areas, in non-residential/commercial or industrial, possibly urban areas.

3) Proportion of non-retail business acres per town has the strongest positive correlation to nitric oxide levels. The greater the proportion of industrial land, the higher the levels of nitric oxide levels.
4) Charles River proximity has the highest correlation coefficient to median value of owner-occupied homes, but the correlation strength was a mere 0.154, so this attribute has basically no correlation to any other attribute.
5) Nitric Oxide level has the highest negative correlation with the distance to employment centres. So, the greater the distance to employment centres, the lower the nitric oxide levels. Since nitric oxides are emitted into the air in areas of high motor vehicle traffic such as in large cities, this indicates that employment centres are likely located in urban/inner city areas. This correlation is as expected, since we would not expect to find employment centers located far outside urban cities.
6) Average number of rooms per dwelling has the strongest positive correlation to median value of owner-occupied homes. The greater the number of rooms, the higher the median value of homes, which is as expected since larger homes tend to be more expensive.
7) Proportion of owner-occupied units built prior to 1940 has the highest negative correlation with the distance to employment centres. This indicates that the lower the proportion of homes built before 1940, the farther the distance to employment centres. So, a higher proportion of owner-occupied homes built before 1940 appear to be located closer in distance to the employment centres.
8) Distance to Boston employment centres has the strongest negative correlation with nitric oxide levels. This correlation is the same observation made in number 5) Nitric Oxide Levels.
9) Index of accessibility to radial highways has the strongest positive correlation to per capita crime rate. This correlation is the same observation made in number 1) Per Capita Crime Rate.
10) Full-value property-tax rate per $10,000 has the strongest positive correlation with accessibility to radial highways. This correlation indicates that the higher the property tax rate, the greater the accessibility to radial highways. In other words, urban city areas with greater accessibility to radial highways have higher tax rates. This could be explained by the fact that the cost of building and maintenance of highway infrastructure must be paid by those who benefit most from the accessibility, so the properties with greater accessibility are taxed proportionately higher.
11) Pupil-teacher ratio has the strongest positive correlation with accessibility to radial highways, but the correlation was a weak 0.387. This shows that there is a very slight correlation between pupil-teacher ratio rising as accessibility to radial highways rise, but with the correlation coefficient under 0.4, this attribute has basically no correlation to

another. This result is rather surprising, since preconceptions would dictate that better pupil-teacher ratio (i.e.- lower pupil-teacher ratio) exists at better schools and better schools lead to higher home values, but this report shows that there is basically no correlation between pupil-teacher ratio and other expected attributes like median value of homes.

12) Proportion of African-Americans in town has the strongest negative correlation with per capita crime rate. This indicates that the lower the proportion of African-Americans, the higher the crime rate and visa versa. However, the correlation coefficient is -0.413, which barely meets the threshold to be considered some correlation. With such a weak correlation coefficient, there is not enough to draw any meaningful implications. The creator of this data set appears to have held preconceptions that the proportion of African-Americans would have meaningful correlations to Boston housing values, but the data has shown otherwise.
13) Percentage of the lower status of the population has the strongest negative correlation with median value of owner-occupied homes. This indicates that the smaller the percentage of lower status population, the higher the median value of homes. This correlation is expected, since "lower status" could be referring to economic class and lower economic class would be expected to own homes with lower median values.
14) Median value of owner-occupied homes has the strongest positive correlation to the average number of rooms per dwelling. This correlation is the same observation made in number 6) average number of rooms per dwelling, and falls under the expected correlation direction.

In General

The correlation model has proven effective in showing expected correlations to be true and also in showing baseless correlation assumptions with certain variables to be unsubstantiated and irrelevant. The correlation model has also been effective in showing a couple of unexpected correlations to hold very strong correlations.

This correlation model is derived from a housing data snapshot, and is only accurate and generalizable insofaras the reliability of that data and the timeframe (circa 1978) of the snapshot. In order for more accurate and generalizable correlations, multiple data sets from sequential timeframes with greater number of examples/observations need to be incorporated.

Business Recommendations

Add Time Attribute

In the investigation of attributes that may affect Boston housing values, we need to focus on which attributes have a strong relationship to the Median Value of Owner-Occupied Homes

(MEDV). This correlation analysis has revealed that the average number of rooms per dwelling has the strongest correlation to MEDV, but this is a natural, expected, and obvious correlation.

Add Average Household Income

The next strongest relationship is MEDV to percentage of lower status population, with a negative correlation. However, since the correlation does not measure or indicate causation, this relationship bears minimal business implications. If the data set could be expanded to include a time attribute and bring in additional timeframes and examples, perhaps more information could be obtained with a historical angle to examine fluctuations over time. Then MEDV to LSTAT correlation may show a deeper relationship. Additional attributes like average household income per owner-occupied homes may be a better indicator for potential correlations than a generic attribute such as percentage of "lower status" population.

Add Attributes to Investigate School Correlation

There is some negative correlation between MEDV and pupil-teacher ratio. The PTRATIO to MEDV may have been expected to have a higher correlation, since housing prices tend to be affected by quality of schools. However, pupil-teacher ratio alone may not be enough of a measurement to infer any valid assumptions about quality of schools or how that affects home values, so adding other attributes such as the number of schools in proximity, average age of owners in owner-occupied homes, average number of school children per home, along others may be required to investigate the home value and school/education aspect.

Correlations of Every Attribute to Median Value of Homes			
MEDV	RM	0.740	Strong positive correlation
MEDV	ZN	0.332	No correlation
MEDV	B	0.265	No correlation
MEDV	CHAS	0.154	No correlation
MEDV	DIS	0.139	No correlation
MEDV	RAD	-0.218	No correlation
MEDV	CRIM	-0.286	No correlation
MEDV	AGE	-0.300	No correlation
MEDV	NOX	-0.333	No correlation
MEDV	TAX	-0.346	No correlation
MEDV	INDUS	-0.412	Some negative correlation
MEDV	PTRATIO	-0.461	Some negative correlation
MEDV	LSTAT	-0.706	Strong negative correlation

Disengage INDUS Attribute

Some relationship does exist between MEDV and proportion of non-retail business acres. This slight negative correlation could imply that higher the median value of homes, the lower the

proportion of non-retail or industrial zoned land. One may infer that higher value homes may be located in more residential areas, but the correlation between MEDV to ZN (proportion of residential zoned land) is very weak and refutes that assumption. This report recommends disengaging the INDUS to MEDV correlation from consideration of housing values.

Drop CHAS and B Attributes

Of the 9 attributes that show no correlation to median value of homes, this report recommends that the Charles River proximity attribute be dropped, because the correlation is close to non-existent. The report also recommends that the proportion of African-Americans attribute be dropped, because the correlation is next to non-existent and also because the mere inclusion of such an attribute is offensive and introduces racial bias into housing value examinations. Even with strong correlations, causation cannot be inferred, and in the case of such low correlation, the only potentially viable reason to maintain the attribute in consideration would be to prove that whatever preconceived prejudicial notions drove this attribute to be included is baseless and futile.